ANALYZING CANADIAN DEMOGRAPHIC AND HOUSING DATA

Building skills and community to analyze Canadian demographic and housing data

Jens von Bergmann

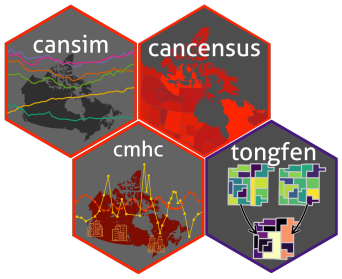


Table of contents

Preface

This book is intended for people interested in learning how to access, process, analyze, and visualize Canadian demographic, economic, and housing data using R. The target audience ranges from individuals wanting to understand their environment through data, to community activists and community groups seeking to introduce or solidify data-based approached into their work, to journalists wanting to enrich their reporting with data or aim to incorporate their own descriptive data analysis, to non-profits or people involved in policy who are looking for data-based answers to their questions.

The key prerequisite for this book is a keen interest in using data to help understand how demographics, economic indicators, housing, and transportation, are reflected in and shape cities and rural areas in Canada. Prior knowledge of R is not necessary, but may be beneficial.

Canada has high quality demographic, economic and housing data. While significant data gaps exist, the available data often remains under-utilized in policy and planning analyses. This under-utilization is accentuated by analysis frequently getting done in silos, relying out data that's already outdated at the time of release, and lack of transparency in the analysis.

In this book aims to help close the gap in high-quality data analysis by

- expanding the group of people doing analysis, increasing the perspectives and interests people bring to data analysis,
- providing guidance on data analysis workflows to increase quality and clarity of data analysis, and
- putting a high emphasis on reproducible and adaptable work flows to ensure the analysis is transparent, can easily be updated as new data becomes available, and can be tweaked or adapted to address related questions.

Under construction

This book provides a basic introduction into how to analyze Canadian data in R and can be used as a standalone resource to cover basic data analysis and visualization workflows in R, as well as be a comprehensive introduction into Canadian data sources. At the same time, we view this as a resources that requires continuous updating as discover new gaps in data analysis and the needs of the Canadian data community changes.

We are planning to add to this book as we find time and come across good examples that are simple enough to slowly build skills, as well as interesting enough to be engaging and motivating to the reader. In this process the order of sections will change as we add new material, and we will come back and revise existing sections as we receive feedback from readers, which we encourage and is ideally submitted as a GitHub issue. When referring to this book we recommend to refer by section name rather than section number. Links in the online version of the book are based on section names and will be unaffected by re-numbering of sections.

Project based approach

We are taking a project based approach to teach through examples, with one project per section. Each project will be broken up into distinct steps. This standardized workflow acts as scaffolding for data analysis and helps ensure key components of analysis are adequately addressed.

1. **Formulating the question.** What is the question we are interested in? Asking a clear question will help focus our efforts and ensure that we don't aimlessly trawl through data. This also involves being clear about the quantities of interest, as well as the target population that we seek to understand.
2. **Identifying possible data sources.** Here we try to identify data sources that can speak to our question. We will also take the time to read up on definitions and background concepts to better understand the data and prepare us for data analysis,

and understand how well the concepts in the data match our original question from step 1.

3. **Data acquisition.** In this step we will import the data into our current working session. This could be as simple as an API call, or more complicated like scraping a table from the web, or involve even more complex techniques to acquire the data.
4. **Data preparation.** In this step we will reshape and filter the data to prepare it for analysis.
5. **Analysis.** This step could be as simple as computing percentages or even doing nothing, if the quantities we are interested in already come with the dataset, if our question can be answered by a simple descriptive analysis. In other cases, when our question is more complex, this step may be much more involved. The book will try to slowly build up analysis skills along the way, with increasing complexity of questions and required analysis.
6. **Visualization.** The final step in the analysis process is to visualize and communicate the results. In some cases this can be done via a table or a couple of paragraphs of text explaining the results, but in most cases it is useful to produce graphs or maps or even interactive visualizations to effectively communicate the results.
7. **Interpretation.** What's left to wrap this up is to interpret the results. How does this answer our question, where does it fall short. What does this mean in the real-world context? What new questions emerge from this?

While we won't always follow this step by step process to the letter, it will be our guiding principle throughout the book. Sometimes things won't go so clean, where after the visualization step we notice that something looks off or is unexpected, and we may jump back up a couple of steps and add more data and redo parts of the analysis to better understand our data and how it speaks to our initial questions. We might even come to understand that our initial question was not helpful or was ill-posed, and we will come back to refine it.

This approach to data analysis leans in parts on (Lundberg, Johnson, and Stewart 2021), who describe a more rigorous framework how data analysis questions should get approached and is a great resource for expert users.

Goals

By taking this approach we have several goals in mind:

- Provide guidance and guardrails for basic data analysis tasks.
- Teach basic data literacy, appreciate definitions and quirks in the data.

- Expose the world of Canadian data and increase accessibility.
- Learn how data can be interpreted in different ways, and data and analysis is not necessarily "neutral".
- Learn how to effectively communicate results.
- Learn how to adapt and leverage off of previous work to answer new questions.
- Learn how to reproduce and critique data analysis.
- Build a community around Canadian data, where people interested in similar questions, or people using the same data, can learn from each other.
- Raise the level of understanding of Canadian data and data analysis so we are better equipped to tackle the problems Canada faces.
- Stay motivated by using real world Canada-focused and (hopefully) interesting examples.

This is setting a very high goal for this book, and we are not sure we can achieve all of this. But we will try our best to be accessible and interesting as possible.

Why use R?

Most people reading this book will not have used R before, or only used it peripherally, maybe during a college course many years in the past. Instead, readers may be familiar with working through housing and demographic data in Excel or similar tools. Or making maps in QGIS or similar tools when dealing with spatial data. And the type of analysis outlined above that this book will teach can in general terms be accomplished using these tools.

But where tools like spreadsheets and desktop GIS fall short is in another important focus of this book: **transparency**, **reproducibility**, and **adaptability**.

An analysis in a spreadsheet or desktop GIS typically involves a lot of manual steps, the work is not **reproducible** without repeating these steps. We can't easily inspect how the result was derived, the analysis lacks **transparency**. When we just compute a ratio or percentage this may not be problematic, but trying to understand how a more complex analysis was done in a spreadsheet easily turns into a nightmare. Analysis that involves a lot of manual steps is not auditable without putting in the work to repeat those manual steps.

But why does this matter? It's always been this way, some experts produce analysis and produce a glossy paper to present the results. One can argue if this was an adequate modus

operandi in the past, but we feel strongly that it isn't in today's world. The lines between experts and non-experts has become blurred, and the value we place on lived experience has increased relative to more formal expertise. We argue this places different demands on policy-relevant analysis, it needs to be open and transparent, in principle anyone should be able to understand how the analysis was done and the conclusions were reached. That's where reproducibility and transparency come in. Additionally, it requires bringing up data analysis skills in the broader population, so that the ability to reproduce and critique an analysis in principle can be realized in practice.

The remaining reason for using R, **adaptability**, has likewise become increasingly important. The amount of data available to us has increased tremendously, but our collective ability to analyse data and extract information has not kept up. Doing analysis in R allows us to efficiently reuse previous analysis to perform a similar one. Or to build on previous analysis to deepen it. Which turbocharges our ability to do analysis, covering more ground and going deeper.

R is not the only framework to do this in, there are other options like python or julia. But we believe that R is best suited for people transitioning into this space, and we can rely on an existing ecosystem of packages to access and process Canadian data. People already proficient in python, julia, STATA, SAS or SPSS will have little difficulty translating what we do into their preferred framework, or dynamically switch back and forth between R, python, julia, or whatever other tools they prefer as needed and convenient.

Building a Canadian data community

Which brings us to our most ambitious goal, to help create a community around Canadian data analysis. When analysis is transparent, reproducible and adaptable people can piggyback of each others work, reusing parts of analysis others have done and building and improving upon it. Or critiquing and correcting analysis, or taking it toward a different direction. A community that grows in their understanding of data, and a community using a shared set of tools to access and process Canadian data, enabling discussions to move forward instead of in circles. A community that builds up expertise from the bottom up.

The book tries to address both of these requirements for building a Canadian data community, a principled approach to data and data analysis, while introducing R as a common framework to work in hoping that the reader will come away with

- better data literacy skills to understand and critique data analysis,
- technical skills to reproduce and perform their own data analysis, and

- a common tool set for acquiring, processing and analyzing Canadian data that facilitates collaborative practices.

1. Introduction

In this section we give a taste of what's to come. Some of the concepts introduced in the preface may be too abstract to picture for people just starting out in this space. People probably grasp the importance of having a principled approach to data analysis, from formulating a question all the way to sharing results. But why so much emphasis on reproducibility and adaptability? And do we really need to learn a new framework like R for this?

This is best understood by walking through a simple example of what analysis of Canadian data in R, and a Canadian data community might look like. We won't explain all steps in full detail here, this is to serve to illustrate the concepts talked above in the preface and give the reader a taste of what's to come.

If you don't understand all the code now, don't worry, that's part of the point of this book. We will work out and explain these examples in detail in the first chapter of the book. What's important right now is to illustrate the principle of reproducible and adaptable code, and how this can function to foster a community of Canadian data analysis. And to note how little code is needed to make this work.

1.1. A hypothetical example

Imagine Amy, a Toronto-based social services worker looking to pilot a community intervention targeted at children in low income. She is in the process of putting together a proposal describing her intervention and is trying to locate a good neighbourhood for her pilot and make a compelling case to possible funders.

Amy knows that census data has a good geographic breakdown of children in poverty, but the latest available data is from 2016, using 2015 income data. CRA tax data is available up to 2019, but also has information on families in low income, but nothing directly on children in the standard release tables at fine geographies. As a first step she settles on census data, with the goal to re-run the analysis once the 2021 data comes out later in the year.

She refers to the Census Dictionary to understand the various low income measures, and uses CensusMapper's interactive map that allows to explore these concepts. She would

have liked to use the Market Based Measure, but due to data availability she settles for LICO-AT.

She sets up a new Notebook and loads in the R libraries that she will need for this, ggplot2 for graphing and **cancensus** for ingesting the data.

```
library(cancensus)
library(ggplot2)
```

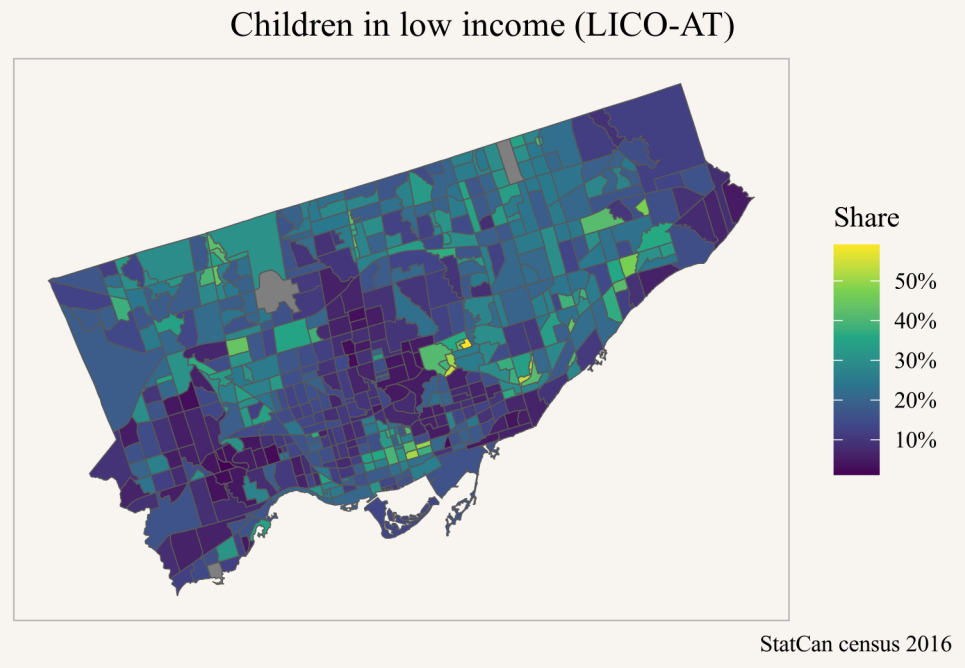
Next she pull in the data. the CensusMapper API GUI tool helps her locate the StatCan geographic identifier for Toronto, (3520005), and the internal CensusMapper vector for the percentage of children in LICO-AT (v_CA16_2573).

```
lico_yyz <- get_census("CA16",regions=list(CSD="3520005"), vectors=c(lico="v_CA16_2573"),
                       level="CT",geo_format="sf")
```

Here Amy specified that she wants data for the 2016 Canadian census ("CA16"), the region and vectors, at the census tract ("CT") level, with geographies as well as the low income data.

Now that she has the data at her finger tips her first step is to make a map. For that she needs to tell **ggplot** is what variable to use as fill colour, and maybe give it a nicer colour scale and some labels to explain what the map is about.

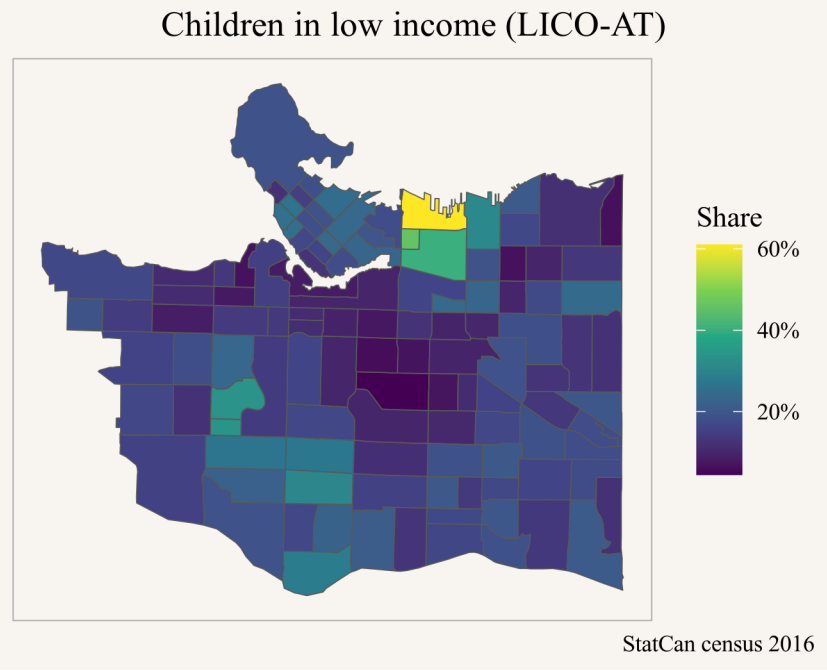
```
ggplot(lico_yyz, aes(fill=(lico/100))) +
  geom_sf() +
  scale_fill_viridis_c(labels=scales::percent) +
  coord_sf(datum=NA) +
  labs(title="Children in low income (LICO-AT)",
       fill="Share",
       caption="StatCan census 2016")
```



Based on this she locates a couple of good candidate neighbourhoods for her pilot and sends the map in a email to her colleague Peter to get input on which neighbourhood might be best suited.

Peter has some good feedback for Amy, but also gets an idea to try and set up something similar in Vancouver. Peter asks Amy if she can share the code, and Amy sends along the above code snippets. Peter looks up the geographic identifier for Vancouver and subs that in instead of Toronto's.

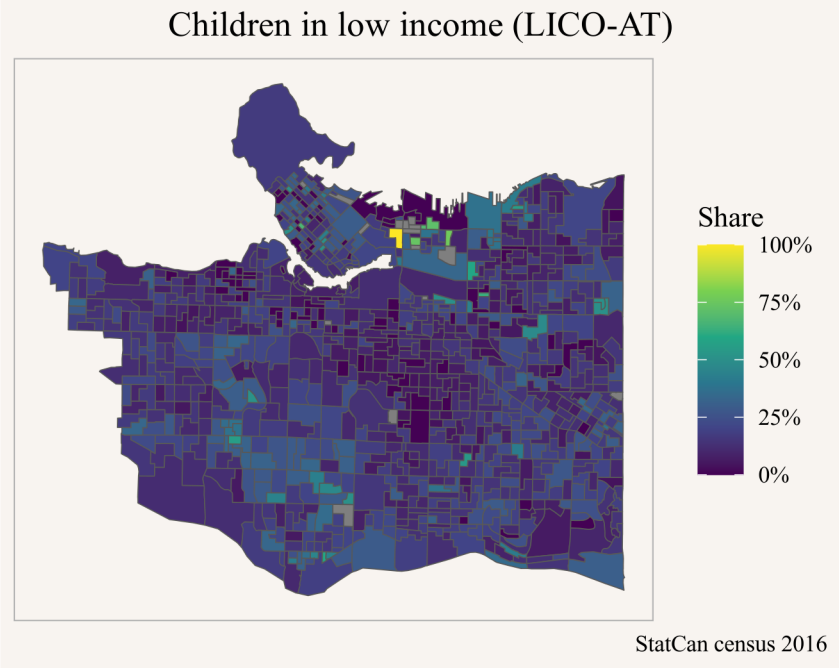
```
lico_yvr <- get_census("CA16",regions=list(CSD="5915022"), vectors=c(lico="v_CA16_2573"),
                       level="CT",geo_format="sf")

ggplot(lico_yvr, aes(fill=(lico/100))) +
  geom_sf() +
  scale_fill_viridis_c(labels=scales::percent) +
  coord_sf(datum=NA) +
  labs(title="Children in low income (LICO-AT)",
       fill="Share",
       caption="StatCan census 2016")
```



Easy peasy, thanks to Amy's previous work. Peter takes the map to his friend Yuko and asks her for advice where a community-based intervention for low-income children might make sense in Calgary. Yuko asks for the code from Peter to take a closer look herself.

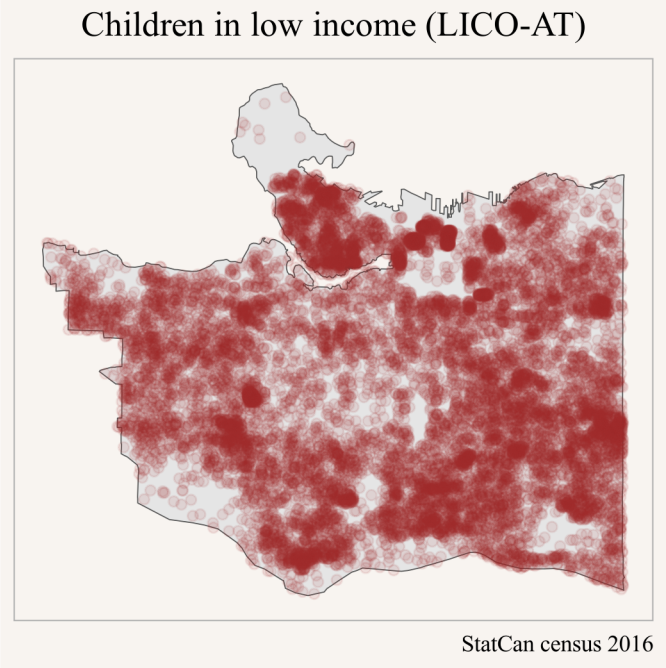
Yuko is interested in a finer geographic breakdown, so she swaps our the geographic level from census tracts to dissemination areas.

```
lico_yvr_da <- get_census("CA16",regions=list(CSD="5915022"), vectors=c(lico="v_CA16_2573")
                          level="DA",geo_format="sf")

ggplot(lico_yvr_da, aes(fill=(lico/100))) +
  geom_sf() +
  scale_fill_viridis_c(labels=scales::percent) +
  coord_sf(datum=NA) +
  labs(title="Children in low income (LICO-AT)",
       fill="Share",
       caption="StatCan census 2016")
```



But then Yuko pauses to think that maybe looking at share of the low income population is not the right metric. She decides to query the number of children in low income (vector "v_CA16_2558") and prepare the data for a dot-density map.

```
# remotes::install_github("mountainmath/dotdensity")
library(dotdensity)

lico_dots_yvr <- get_census("CA16",regions=list(CSD="5915022"),geo_format="sf",
                            vectors=c(lico="v_CA16_2558"), level="DA") |>
  compute_dots("lico")
yvr_city <- get_census("CA16",regions=list(CSD="5915022"),geo_format="sf")

ggplot(lico_dots_yvr) +
  geom_sf(data = yvr_city) +
  geom_sf(colour="brown",alpha=0.1) +
  coord_sf(datum=NA) +
  labs(title="Children in low income (LICO-AT)",
       fill="Share",
       caption="StatCan census 2016")
```



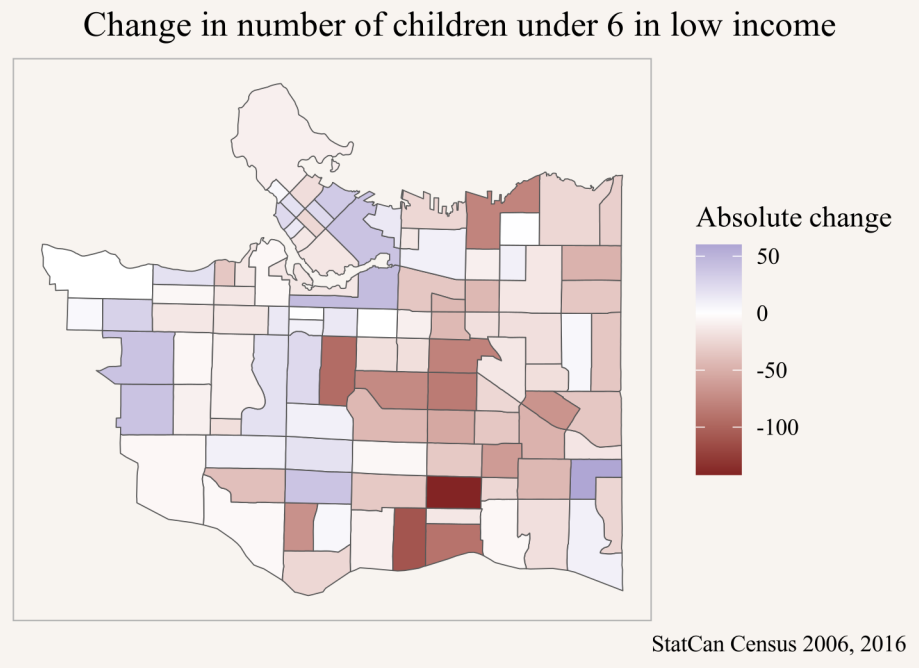
That paints a somewhat different picture, and Yuko feels this is much better suited to pinpoint where to best stage a community intervention. She lets Peter and Amy know and emails them her modifications to the code.

Meanwhile, Yuko's Vancouver friend Stephanie is looking specifically at children below the age of 6 in low income, and wants to understand how the geographic distribution of low income children has changed over time. Comparing census data through time can be tricky because census geographies change, but this problem has been completely solved via the **tongfen** R package. Looking at Yuko's work she thinks it might be best to look at both, the change in share of children in low income as well as the change in absolute number.

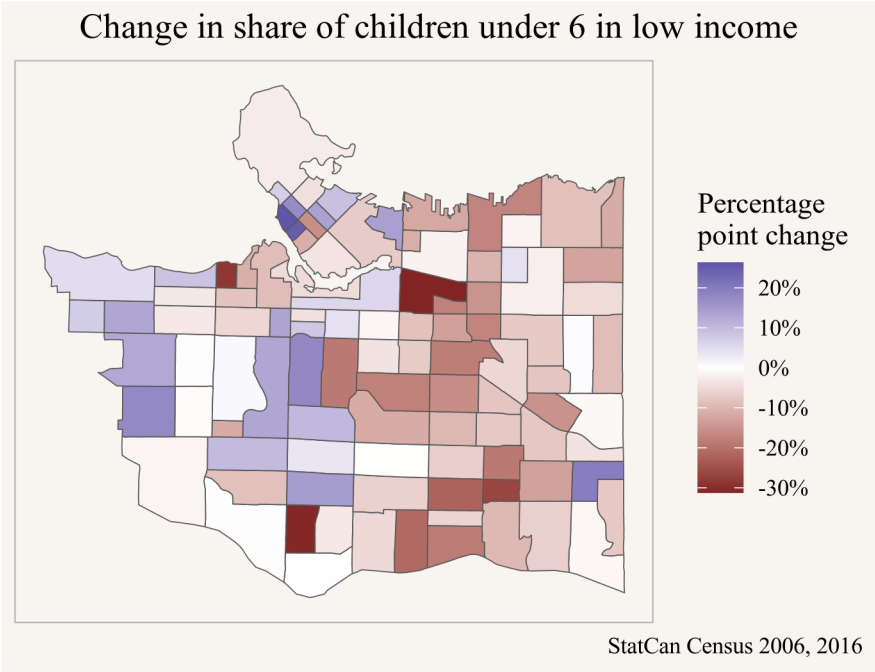
```
library(tongfen)
meta <- meta_for_ca_census_vectors(c(total_2006="v_CA06_1982",lico_share_2006="v_CA06_1984"
                                     lico_2016="v_CA16_2561",lico_share_2016="v_CA16_2576")

lico_data <- get_tongfen_ca_census(regions=list(CSD="5915022"),meta,level="CT") |>
  mutate(lico_2006=total_2006*lico_share_2006/100) |>
  mutate(`Absolute change`=lico_2016-lico_2006,
         `Percentage point change`=lico_share_2016-lico_share_2006)
```

Armed with this data Stephanie can plot the absolute and percentage point change in children below 6 in low income.

```
ggplot(lico_data,aes(fill=`Absolute change`)) +
  geom_sf() +
  scale_fill_gradient2() +
  coord_sf(datum=NA) +
  labs(title="Change in number of children under 6 in low income",
       caption="StatCan Census 2006, 2016")
```



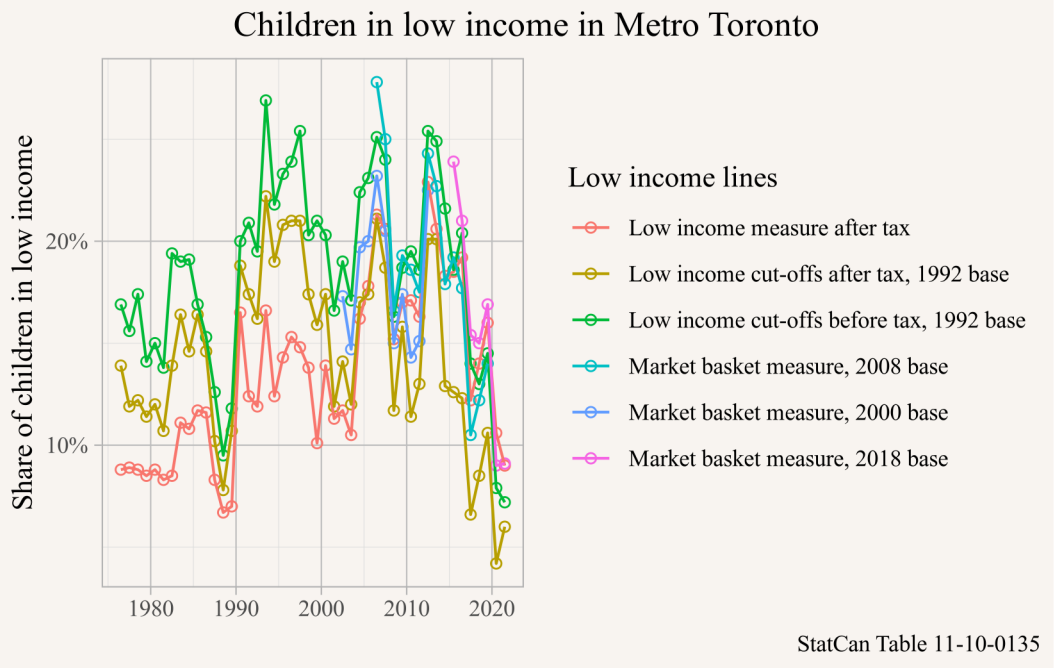

```
ggplot(lico_data,aes(fill=`Percentage point change`/100)) +
  geom_sf() +
  scale_fill_gradient2(labels=scales::percent) +
  coord_sf(datum=NA) +
  labs(title="Change in share of children under 6 in low income",
       fill="Percentage\npoint change",
       caption="StatCan Census 2006, 2016")
```



Stephanie shares her results with Amy in Toronto in case there are components of Amy's pilot specifically targeting children below 6 in low income.

Meanwhile Amy has been trying to understand more broadly how the share of low income children has evolved since the 2016 census (using 2015 income data) at the metropolitan level over longer time spans, so she looks through the StatCan socioeconomic tables and settles on table 11-10-0135, which also allows her to compare various low income concepts.

```
library(cansim)
mbm_timeline <- get_cansim("11-10-0135") |>
  filter(`Persons in low income`=="Persons under 18 years",
         GEO=="Toronto, Ontario",
         Statistics=="Percentage of persons in low income")

ggplot(mbm_timeline,aes(x=Date,y=val_norm,colour=`Low income lines`)) +
  geom_point(shape=21) +
  geom_line() +
  scale_y_continuous(labels=scales::percent) +
  labs(title="Children in low income in Metro Toronto",
       y="Share of children in low income",
       x=NULL,
       caption="StatCan Table 11-10-0135")
```



She notes that there has been a substantial overall drop in children in low income since 2015 across all measures, which is excellent news. She considers pushing off her pilot project until after the 2021 census data comes out to first understand if the geographic patterns have changed.

1.2. What you will learn in this book

Looking at R code for the first time can be intimidating. If the code looks opaque right now, there is no need to worry. It will be explained in detail in the first chapter and is very much part of the rationale for writing this book. If decisions around what low income metric to pick, or why **tongfen** is needed to compare census data through time are not clear, again, that will be explained in this book in detail and expanding understanding of data and data analysis is the other big rationale for this book.

Readers will learn how to reproduce analysis, how to critique analysis, and adapt it for their own purposes. And readers will learn how to conduct their own analysis in the Canadian context, based on questions and use cases relevant to them.

Hopefully the above hypothetical scenario have explained how the adaptability of the R code has made life much easier for several of the subsequent analysis steps, and how little code was needed to gain some insights and communicate results.

Part I.

Getting started

Before we dive into data projections we want to give a quick overview over R, RStudio, tidyverse and the main data acquisition and processing packages for Canadian data that we will use.

Statistics Canada produces a lot of high quality demographic and economic data for Canada. CMHC complements this with housing data, and municipalities across Canada often provide relevant data through their Open Data portals.

2. R, RStudio, and the tidyverse

Statistics Canada produces a lot of high quality demographic and economic data for Canada. CMHC complements this with housing data, and municipalities across Canada often provide relevant data through their Open Data portals.

2.1. R and RStudio

We will be working in R and the RStudio IDE, although using a different editor like Visual Studio Code works just as well, especially if you are already familiar with it. Within R we will be operating within the tidyverse framework, a group of R packages that work well together and allow for intuitive operations on data via pipes.

While an introduction to R is part of the goal of this book, an we will slowly build up skills as we go, we not give a systematic introduction but rather build up skills slowly as we work on concrete examples. It may be beneficial to supplement this with a more principled introduction to **R** and the **tidyverse.**

2.2. Packages

Packages are bundled sets of functionality that expand base R. We install or upgrade packages with the `install.packages`. For example, to install the tidyverse framework we type

```
install.packages("tidyverse")
```

into the R console. This will install or upgrade the package and required dependencies. To make the functionality, for example the `tibble` function from the **tibble** package that is part of **tidyverse**, available to use we can then either access functions from the package using the `::` namespace selector `tibble::tibble()` or first load the **tibble** or **tidyverse** package via `library(tidyverse)` that makes the `tibble()` function available without having to use the namespace selector.

Additionally, we will need a number of packages that handle data acquisition and processing for Canadian data.

```
install.packages(c("cancensus","cansim","cmhc","tongfen"))
```

2.3. Basic data manipulation patterns

There are several basic data manipulation patterns that we will use throughout, and we want to give a quick overview using the Palmer Penguins dataset from the **palmerpenguins** package.

```
install.packages("palmerpenguins") # install the package if needed
```

We will at times require additional packages like this to accomplish specialized tasks, installing packages in R is generally a simple and pain-free procedure.

```
# install.packages("palmerpenguins") # install the package if needed
library(palmerpenguins)
```

Now we have all the functionality of the **palmerpenguins** package available.

2.3.1. Exploring the data

With the **palmerpenguins** package comes the `penguins` dataset, we can expect the first few rows using the `head()` function which displays the first few rows.

```
head(penguins)
```

```
# A tibble: 6 x 8
  species island    bill_length_mm bill_depth_mm flipper_l~1 body_~2 sex    year
  <fct>   <fct>              <dbl>         <dbl>       <int>   <int> <fct> <int>
1 Adelie  Torgersen           39.1          18.7         181    3750 male   2007
2 Adelie  Torgersen           39.5          17.4         186    3800 fema~  2007
3 Adelie  Torgersen           40.3          18           195    3250 fema~  2007
4 Adelie  Torgersen           NA            NA            NA      NA <NA>   2007
5 Adelie  Torgersen           36.7          19.3         193    3450 fema~  2007
6 Adelie  Torgersen           39.3          20.6         190    3650 male   2007
# ... with abbreviated variable names 1: flipper_length_mm, 2: body_mass_g
```

The `str()` function offers another convenient way to get an overview over the data.

```
str(penguins)
```

```
tibble [344 x 8] (S3: tbl_df/tbl/data.frame)
 $ species          : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ island           : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ bill_length_mm   : num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...
 $ bill_depth_mm    : num [1:344] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...
 $ flipper_length_mm: int [1:344] 181 186 195 NA 193 190 181 195 193 190 ...
 $ body_mass_g      : int [1:344] 3750 3800 3250 NA 3450 3650 3625 4675 3475 4250 ...
 $ sex              : Factor w/ 2 levels "female","male": 2 1 1 NA 1 2 1 2 NA NA ...
 $ year             : int [1:344] 2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...
```

We can also type `View(penguins)` into the console to view the dataset in a spreadsheet form.

2.3.2. Basic data manipulation

To manipulate and visualize the data we load the tidyverse package.

```
library(tidyverse)
```

We will explore some common data manipulation and visualization workflows.

2.3.2.1. Count groups

To see how many rows there are for each species we 'pipe' the `penguins` dataset into the `count()` verb. Pipes are how we can stepwise transform data, the pipe operator is given by `%>%` within the **tidyverse** framework and now also available natively in base R via `|>`. These two function (almost) the same way, and we will use both in this book.

```
penguins |> count(species)
```

```
# A tibble: 3 x 2
  species       n
  <fct>     <int>
```

```
1 Adelie      152
2 Chinstrap    68
3 Gentoo      124
```

This gives us the count of each species in the dataset, the pipe `|>` inserts the left hand side as the first argument in the `count()` function. We could have equivalently written this without the pipe operator as `count(penguins,species)`.

2.3.2.2. Group and summarize

The usefulness of the pipe operator becomes clear when we chain several data transformations. If we want to know the mean bill length by species, we group by species and summarize the data.

```
penguins |>
  group_by(species) |>
  summarize(bill_length_mm=mean(bill_length_mm, na.rm=TRUE))
```

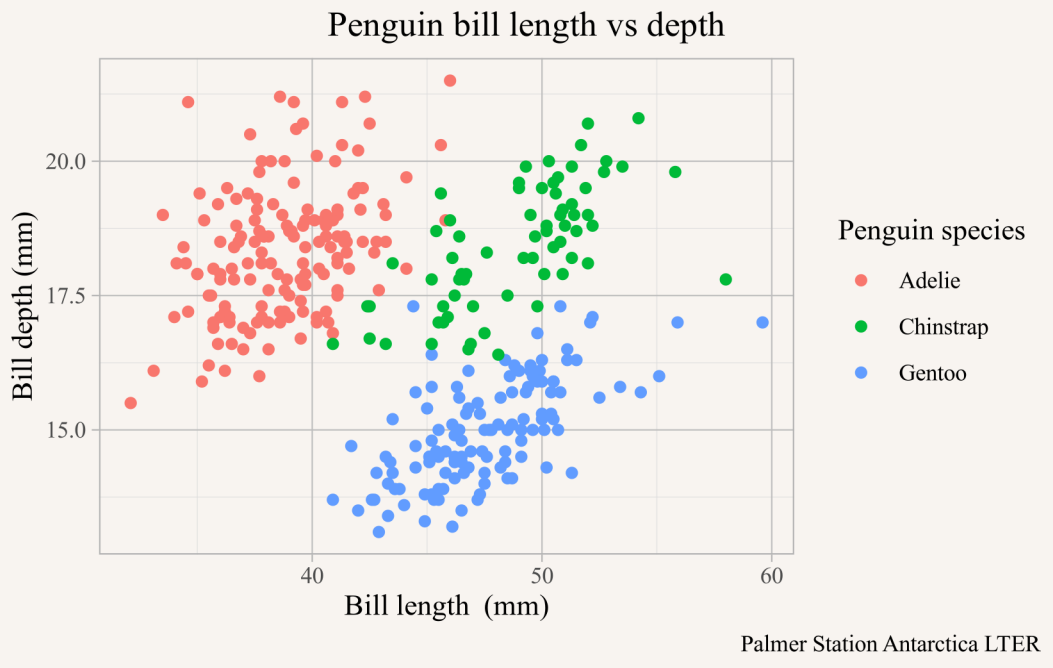
```
# A tibble: 3 x 2
  species   bill_length_mm
  <fct>              <dbl>
1 Adelie              38.8
2 Chinstrap           48.8
3 Gentoo              47.5
```

Here we explicitly specify how missing values should be treated when summarizing, `na.rm=TRUE` says that NA values should be ignored when computing the mean.

2.3.3. Visualizing data

We can visualize the data using `ggplot`. For this we have to specify the mapping aesthetics, we plot the bill length on the x-axis, the depth on the y-axis, colour by species and plot the data as points. The `labs()` function allows us to customize the graph labels.

```
ggplot(penguins,aes(x=bill_length_mm,y=bill_depth_mm,colour=species)) +
  geom_point() +
  labs(title="Penguin bill length vs depth",
       x="Bill length  (mm)",y="Bill depth (mm)",
       colour="Penguin species",
       caption="Palmer Station Antarctica LTER")
```

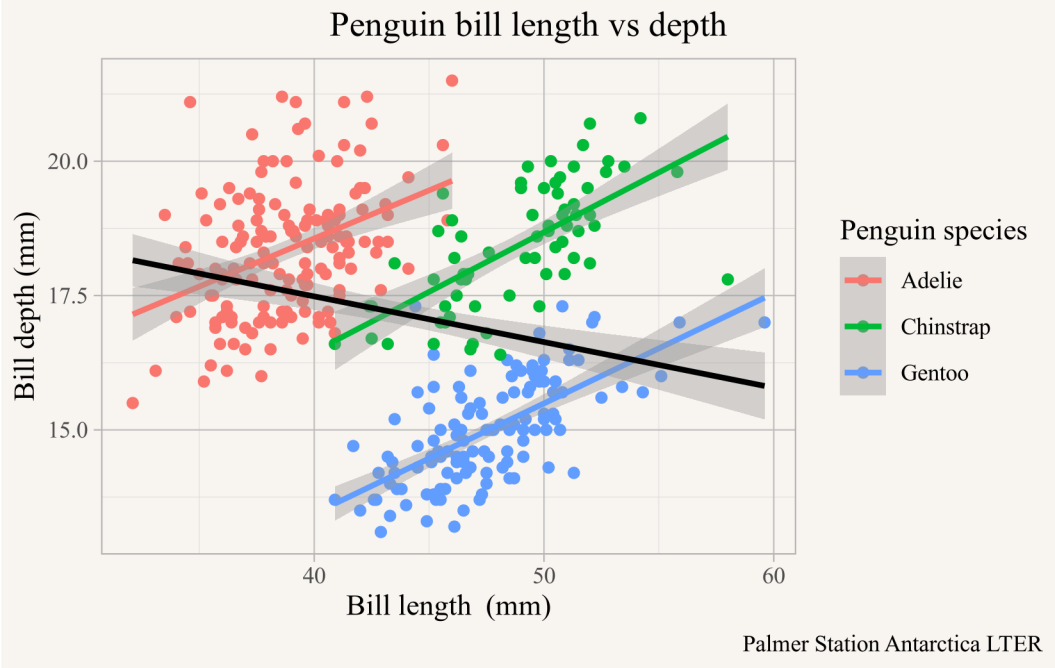


2.3.3.1. Add regression lines

As an aside we note the Simpson's paradox, in the overall dataset the bill depth declines with length, but if we look separately within each species the bill depth increases with bill length. To make that explicit we can add regression lines using the `geom_smooth` function using `lm` (linear model) as the smoothing method.

```
ggplot(penguins,aes(x=bill_length_mm,y=bill_depth_mm,colour=species)) +
  geom_point() +
  geom_smooth(method="lm") +
  geom_smooth(method="lm", colour="black") +
  labs(title="Penguin bill length vs depth",
```

```
       x="Bill length  (mm)",y="Bill depth (mm)",
       colour="Penguin species",
       caption="Palmer Station Antarctica LTER")
```



The first `geom_smooth()` function will add a regression line for each species, distinguished by colour in the plot aesthetics. Overriding the `colour` argument in the second `geom_smooth()` function will forget that the data was coloured by species and add the black regression line run on the entire dataset.

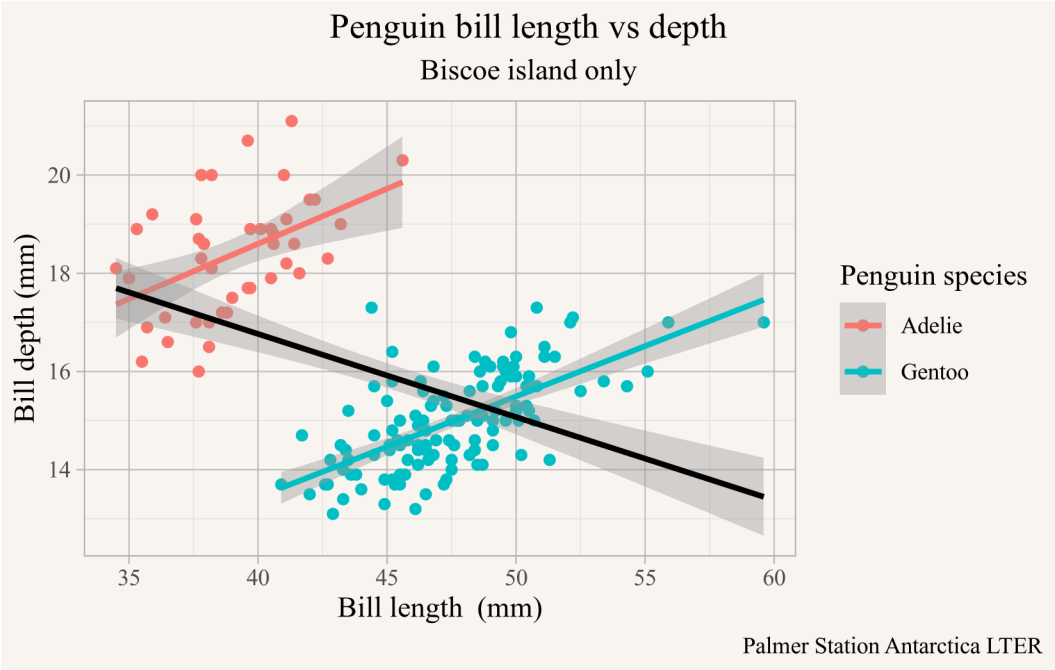
2.3.4. More data manipulations

There are several common data manipulation steps that we will employ frequently.

2.3.4.1. Filtering rows

Often we are only interested in subsets of the data, we can filter the rows in the dataset by using the `filter` verb from the **dplyr** package that is part of **tidyverse**. For example, if we want to take the previous plot but only show it for penguins on the island of Biscoe we can filter the data accordingly before plotting it.

```
penguins |>
  filter(island=="Biscoe") |>
ggplot(aes(x=bill_length_mm,y=bill_depth_mm,colour=species)) +
  geom_point() +
  geom_smooth(method="lm") +
  geom_smooth(method="lm", colour="black") +
  labs(title="Penguin bill length vs depth",
       subtitle="Biscoe island only",
       x="Bill length  (mm)",y="Bill depth (mm)",
       colour="Penguin species",
       caption="Palmer Station Antarctica LTER")
```



2.3.4.2. Selecting columns

Instead of filtering rows it can be useful to select a subset of the columns to remove columns we don't need and de-clutter the dataset. This is especially useful when producing tables. If we want a table of the numeric data fields of all female Adelie penguins on the island of Biscoe observed in 2007 we can filter by sex and island and select the columns we want.

```
penguins |>
  filter(island=="Biscoe", sex=="female", species=="Adelie", year==2007) |>
  select(where(is.numeric),-year)
```

```
# A tibble: 5 x 4
  bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
           <dbl>         <dbl>             <int>       <int>
1           37.8          18.3               174        3400
2           35.9          19.2               189        3800
3           35.3          18.9               187        3800
4           40.5          17.9               187        3200
5           37.9          18.6               172        3150
```

2.3.4.3. Mutating data

We often want to change data fields, or compute new columns from existing ones. For example, if we want to convert the body mass from *g* to *kg* we can add a new column using `mutate` for that.

```
penguin_selection <- penguins |>
  filter(island=="Biscoe", sex=="female", species=="Adelie", year==2007) |>
  mutate(body_mass_kg=body_mass_g/1000) |>
  select(where(is.numeric),-year,-body_mass_g)

penguin_selection
```

```
# A tibble: 5 x 4
  bill_length_mm bill_depth_mm flipper_length_mm body_mass_kg
           <dbl>         <dbl>             <int>        <dbl>
1           37.8          18.3               174         3.4
2           35.9          19.2               189         3.8
3           35.3          18.9               187         3.8
4           40.5          17.9               187         3.2
5           37.9          18.6               172         3.15
```

2.3.4.4. Pivoting data

The data in our `penguin_selection` dataset above is in **wide form**, all the different variables are in their own column. Often it is useful to convert it to **long form**, where we only have one value column with the numeric values and another column specifying the type of measurement. In this case it is useful to add an identification column so that we know which measurements belong to the same penguin. We can just label the penguins by row number.

```
penguin_selection_long <- penguin_selection |>
  mutate(ID=row_number()) |>
  pivot_longer(-ID,names_to="Metric",values_to="Value")

penguin_selection_long |> head()
```

```
# A tibble: 6 x 3
     ID Metric            Value
  <int> <chr>             <dbl>
1     1 bill_length_mm     37.8
2     1 bill_depth_mm      18.3
3     1 flipper_length_mm 174
4     1 body_mass_kg        3.4
5     2 bill_length_mm     35.9
6     2 bill_depth_mm      19.2
```

We can do the reverse transformation, going **from long** form to **wide form**, using `pivot_wider`.

```
penguin_selection_long |>
  pivot_wider(names_from = Metric,values_from = Value)
```

```
# A tibble: 5 x 5
     ID bill_length_mm bill_depth_mm flipper_length_mm body_mass_kg
  <int>          <dbl>         <dbl>              <dbl>        <dbl>
1     1           37.8          18.3                174         3.4
2     2           35.9          19.2                189         3.8
3     3           35.3          18.9                187         3.8
4     4           40.5          17.9                187         3.2
5     5           37.9          18.6                172         3.15
```

This recovers the previous form of the data, with the added `ID` column.

2.4. Canadian data packages

During the course of this book we will make heavy use of several R packages to facilitate data access to Canadian data, we will introduce them in this chapter.

3. Introduction to the cansim package

The **cansim** R package (von Bergmann and Shkolnik 2021) interfaces with the StatCan NDM that replaces the former CANSIM tables. It can be queried for

- whole tables
- specific vectors
- data discovery searching through tables

It encodes the metadata and allows to work with the internal hierarchical structure of the fields.

Larger tables can also be imported into a local SQLite database for reuse across sessions without the need to re-download the data, and better performance when subsetting the data or performing other basic data operations at the database level before loading the data into memory.

Data discovery can be cumbersome, the `list_cansim_cubes` function from the **cansim** package fetches the newest list of all available tables and can be filtered by survey, release

date or dates of data coverage. The table list is cached for the duration of the R session. The `search_cansim_cubes` function provides a convenient shortcut to narrow down this list.

In some cases searching the web for "StatCan Table xxxx", where "xxxx" contains search phrases for the data of interest, is sometimes a useful way to discover data. In reverse, we can bring up the StatCan webpage for a specific table number using the `view_cansim_webpage` function and explore the data via the web interface. Especially for large datasets this can be a faster way to determine if a specific table contains the information we are interested in without first having to download the data.

To get overview information for a table we have already downloaded the `get_cansim_table_overview` function provides a high-level overview over the variables contained in the table. The `get_cansim_column_list` function returns a list of the available columns or dimensions in the table, and `get_cansim_column_categories` returns the list of levels in a specific dimension. The `get_cansim_table_nots` provides the data notes that can hold important information to guide interpretation of some of the dimensions or levels.

The data is accessed via `get_cansim` function, or alternatively the `get_cansim_sqlite` function that stores the data permanently for use across R sessions in a local SQLite database. By default the English language tables are accessed, setting the `language="fr"` parameter changes that to the French version. The SQLite option is especially useful for larger tables. The **cansim** package will emit a warning if an SQLite table is outdated and newer data is available, if the `auto_refresh=TRUE` option is passed to the function call it will automatically download any new data if available. When accessing data from the SQLite version we can use normal `dplyr` verbs to filter the data or perform basic `select`, `group_by` or `summarize` operations before calling `collect_and_normalize` to fetch the result from the database and enrich it with metadata.

Metadata added by the **cansim** package includes converting the dimension values to factors and adding information on the hierarchical structure of the levels. Moreover, the package creates a native `Date` field and a `val_norm` field with normalized values. The values shipped by StatCan are sometimes expressed in "thousands of units", the `val_norm` converts this to base units for easier interpretation and uniformity across tables.

More information can be found in the package documentation and the package vignettes.

To install the package from cran use

```
install.packages("cansim")
```

4. Introduction to the cancensus package

The **cancensus** R package (von Bergmann, Shkolnik, and Jacobs 2022) interfaces with the CensusMapper API server. It can be queried for

- census geographies for census years 1996 through 2021
- census profile data for census years 1996 through 2021
- some census custom tabulations
- hierarchical metadata of census variables
- some non-census data that comes on census geographies, e.g. T1FF taxfiler data

A slight complication, the **cancensus** package needs an API key. You can sign up for one for free on CensusMapper.

Once you have your API key it's useful to store it as an environment variable in your .Renviron configuration file so it's available in all your R sessions.

```
install.packages("cancensus")

cancensus::set_api_key(key = "CensusMapper_XXXX...XXXX", install=TRUE)
```

By default {cancensus} caches downloaded census data, which makes it easier and faster to re-run analysis and protects from overusing the CensusMapper API quota. To use caching a local path needs to be designated for data caching. The cache is shared across R sessions.

```
cancensus::set_cache_path(cache_path = "<local path of cache data>", install=TRUE)
```

{cancensus} provides convenient access to census data. Calls to {cancensus} require to spectify

- The `dataset`, for example "CA21" for the 2021 Canadian census. A list of available datasets can be accessed via `cancensus::list_census_datasets()`.
- The `regions` to access the data for, this is a list keyed by geographic levels. For example, to access data for the Vancouver census metropolitan area it would be `list(CMA="59933")`, for the City of Toronto it would be `list(CSD="3520005")`. Region parameters can contain several regions of the same type or mix regions of different type. For example, to access data for the region covered by the Vancouver School Board, we need to assemble two CSDs and three CTs `list(CSD=c("5915022","5915803"),CT=c("93300`
  This allows pinpointing what geographic region we are interested in.
- The geographic `level` to query the data for. This simply are the regions specified in the `regions` parameter, but it could also be any geographic level equal to or lower than the lowest level geographic region specified in the `regions` parameter. Valid level identifiers are `DB` for dissemination blocks, `DA` for dissemination areas, `EA` for enumeration areas for the 1996 census, `CT` for census tracts, `CSD` for census subdivisions, `CMA` for census metropolitan areas or census agglormerations, `CD` for census districts, `PR` for provinces or territories and `C` for country level data. Geographic regions can also be assembled using the CensusMapper API GUI tool, CSD and higher level geographies can be explored or searched programmatically via the `list_census_regions()` or `search_census_regions()` functions.
- The `vectors` parameter allows to specify which census variables to query. By default the data comes with population, dwelling and household counts, other census variables can be explored and selected via the CensusMapper API GUI tool or explored or searched programmatically via the `list_census_vectors()` or `find_census_vectors()` functions. There are also helper functions to select variables using the

internal CensusMapper metadata and hierarchy of census variables via the `child_census_vectors()` function. For finer control over the names of the returned census variable the `vectors` parameter can be a named vector.

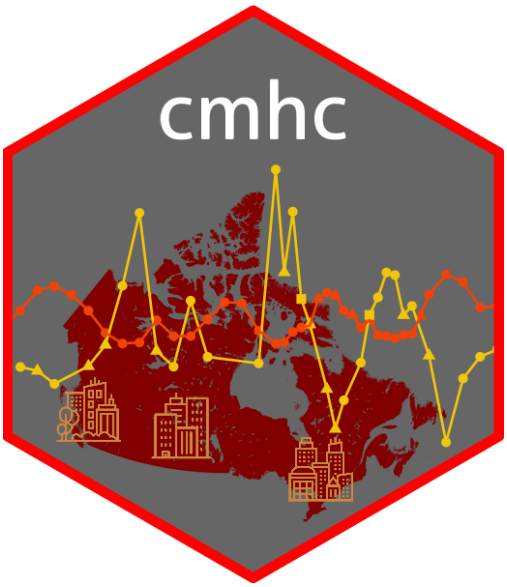
- The `geo_format` parameters allows to select if geographic data should also be downloaded, and if yes, in what format. In this post we will only access data via the modern "sf" format, but data can also be accessed in the legacy "sp" spatial data format.

As an example we will retrieve the share of the population in Toronto, Mississauga, and Brampton spending 30% or more of income on shelter costs in 2016.

```
library(cancensus)
library(dplyr)
regions <- list(CSD=c("3520005","3521005","3521010"))
vectors <- c(shelter_cost_burdened="v_CA16_4889", shelter_base = "v_CA16_4886")

data <- get_census(dataset = "CA16", regions=regions, vectors=vectors)
data %>%
  mutate(`Share shelter cost burdened`=shelter_cost_burdened/shelter_base) |>
  select(GeoUID,`Region Name`,`Share shelter cost burdened`)
```

```
# A tibble: 3 x 3
  GeoUID  `Region Name`    `Share shelter cost burdened`
  <chr>   <fct>                                   <dbl>
1 3520005 Toronto (C)                             0.296
2 3521005 Mississauga (CY)                        0.264
3 3521010 Brampton (CY)                           0.305
```

5. Introduction to the cmhc package



The cmhc R package (von Bergmann 2022) interfaces with the CMHC Housing Market Information Portal (HMIP) and allows programmatic and reproducible access to CMHC data. This gives access to data from four major CMHC surveys

- Starts and Completions Survey (Scss), which has data on housing construction covering starts, completions, units under construction, length of construction, absorbed and unabsorbed units and their prices.
- Rental Market Survey (Rms), which surveys the purpose-built rental market on an annual (and for some time twice-annual) basis. It has data on vacancy rates, availability rates, rents, fixed sample rent change and the overall rental universe by bedroom type, structure size, and year of construction.
- Secondary Rental Market Survey (Srms), which covers parts of the secondary market rental market with data on condominium apartment vacancy rates, rents, and number and share of rented units, as well as some information on other secondary rentals.

- Senior's housing (Seniors), which gives data on seniors housing of various levels of assistance.
- Census data (Census), which holds several housing related cross-tabulations.
- Core Housing Need (Core Housing Need) related cross-tabulations.

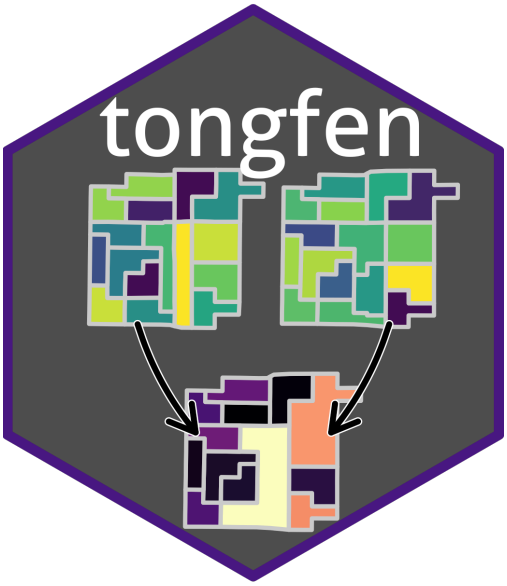
The package is designed to work in conjunction with the **cancensus** package and census geographic identifiers.

To install the package from CRAN use.

```
install.packages("cmhc")
```

The nature of the CMHC backend makes it at times challenging to find data, the **cmhc** package has several convenience functions to facilitate data discovery. The `list_…` functions, for example `list_cmhc_surveys()` list options. The `select_cmhc_table()` allows the interactive selection of data tables in the console, and returns the syntax for the desired function call to acquire the data.

6. Introduction to the tongfen package



The tongfen R package (von Bergmann 2021) facilitates making data on different geometries comparable.

TongFen ( ) means to convert two fractions to the least common denominator, typically in preparation for further manipulation like addition or subtraction. In English, that's a mouthful and sounds complicated. But in Chinese there is a word for this, TongFen, which makes this process appear very simple.

When working with geospatial datasets we often want to compare data that is given on different regions. For example census data and election data. Or data from two different censuses. To properly compare this data we first need to convert it to a common geography. The process to do this is quite analogous to the process of TongFen for fractions, so we appropriate this term to give it a simple name. Using the **tongfen** package, preparing data on disparate geographies for comparison by converting them to a common geography is as easy as typing `tongfen`.

In particular, the package has a number of convenience functions to facilitate making Canadian census data comparable through time, making it easy to perform longitudinal

analysis on fine geographies based on the Canadian Census. Essentially, the **tongfen** package creates a semi-custom tabulation based on Dissemination Block, Dissemination Area, or Census Tract geographies.

These semi-custom tabulations are created in three steps:

1. Create a correspondence table for geographies from different censuses. By default the official StatCan correspondence files are used for that, but these only exist back to 2001 when the current geographic system based on *dissemination blocks* and *dissemination areas* started. To go back further, when *enumeration areas* were the basic building block, we need to rely on geospatial matching of the areas to create a harmonized common geography.
2. Create metadata that contains information about how the census variables of interest can be aggregated in the case where geographies get joined. For example, if we are interested in the share of households in low income, we need to know what this share is based on in order to correctly aggregate it up. CensusMapper holds detailed metadata, so this process is automated.
3. Join geographies and aggregate census data as described in the correspondence table from Step 1 and the metadata in step 2.

The result of this process is a semi-custom tabulation of the data we want that is created on the fly, at the price of coming on a slightly coarser geography than the original input geographies in cases where geographies had to be joined to create the harmonized geography.

To install the package from CRAN use

```
install.packages("tongfen")
```

Part II.

Basic descriptive analysis

In this section we will look at how to do basic descriptive analysis. The questions we ask here will be quite simple, for example: How has income changed over time? Or: Which areas of Toronto have the highest incomes?

The accompanying analysis won't be very involved, sometimes we will compute percentages or make other simple data manipulations, but generally the analysis will be quite straightforward. We will focus on how to find data sources that can inform on our question, how to get the data, and how to present and interpret it.

7. Geography of CERB

In 2020 Canada introduced the COVID-19 Emergency Recovery Benefit (CERB), a program to support people ring the pandemic.

7.1. Question

Where did CERB benefits go?

7.2. Data sources

Standard T1FF taxfiler data has this for large geographies, to understand fine geographic distribution we turn to Census data from the 2021 census, which reports on 2020 income. The census dictionary explains

> Canada Emergency Response Benefit (CERB) payments received during the reference period. This benefit was intended to provide financial support to employees and self-employed Canadians who had lost their job or were working fewer hours due to the COVID-19 pandemic and the public health measures implemented to minimize the spread of the virus.

Census income data is taken directly from T1 tax returns and linked at the individual person level.

7.3. Data acquisition

We can use the CensusMapper API tool and search for "COVID-19" in the *Variable Selection* tab to locate available census variables. Since we are interested in where people lived that received the benefit we select **v_CA21_593**, the number of recipients, as well as **v_CA21_554**, the baseline of people 15 years or older who are in principle eligible for this benefit.

We also need to decide which region we want to investigate, let's take a look at the City of Ottawa. We can select the city in the *Region Selection* tab and read off the geographic identifier **3506008** for the City of Ottawa in the *Overview* tab.

Now we have all we need to pull in the data, we just need to decide on the geographic granularity. Let's use **census tracts**, a standard geographic region aiming to capture between 2,500 and 7,500 people in metropolitan areas. We also specify that we want the geographies, not just the tabular data.

```
library(cancensus)
ottawa_cerb <- get_census("CA21",regions=list(CSD="3506008"),
                          vectors=c(cerb="v_CA21_593", base="v_CA21_554"),
                          level="CT", geo_format="sf")
```

7.4. Data preparation

To understand the geographic distribution we compute the percentage of people 15 years and over receiving CERB. Generally in this book we work in the **tidyverse** to help with data manipulation and visualization, so we load that library too.

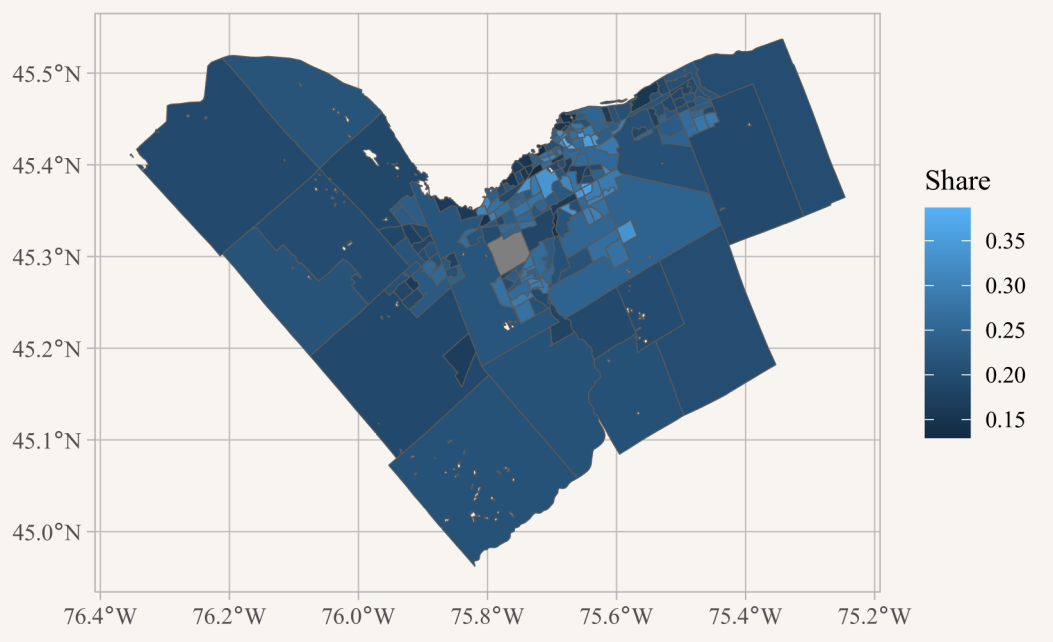
```
library(tidyverse)
plot_data <- ottawa_cerb |>
  mutate(Share=cerb/base)
```

There is not much to do, computing a percentage is a simple division. The `mutate` verb creates a new column called `Share` holding the computed ratios.

7.5. Analysis and visualization

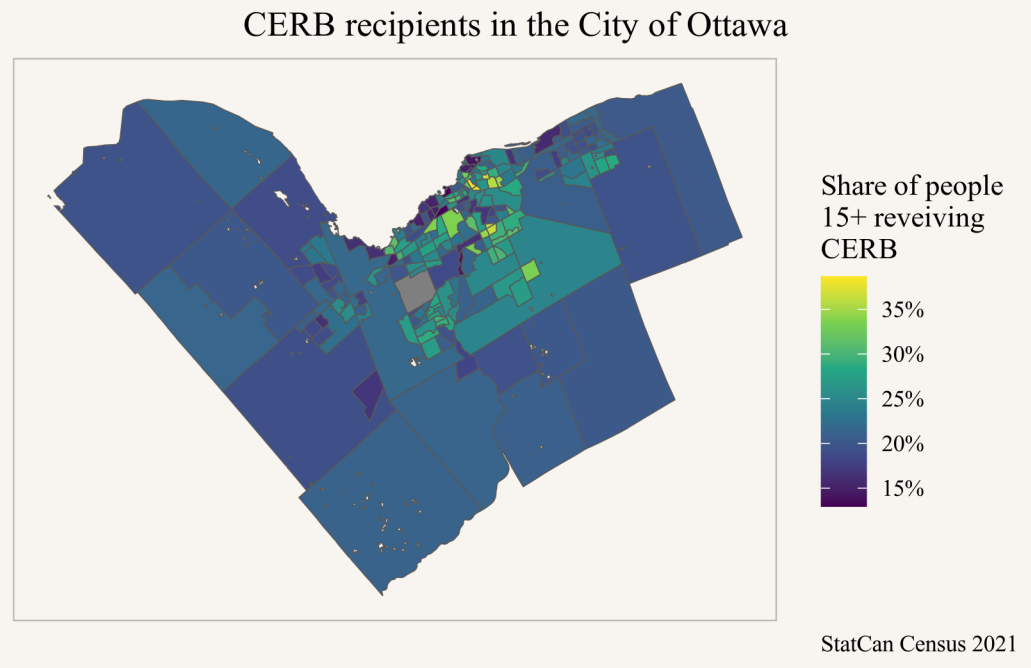
All that's left is to visualize the data. To plot geographic data we use `ggplot` and the `geom_sf` geometry. We need to tell it how to colour the map, the *aesthetic*, and we specify to fill each area by the share of CERB recipients.

```
ggplot(plot_data) +
  geom_sf(aes(fill=Share))
```



To make this a little nicer we add labels, remove the coordinate grid and choose nicer colours and reduce the boundary line size.

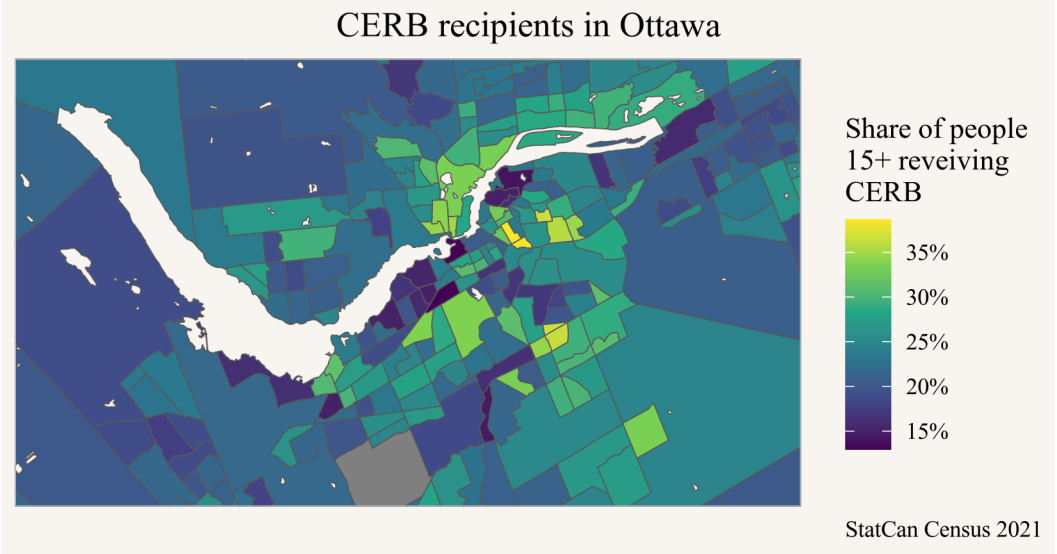
```
ggplot(plot_data) +
  geom_sf(aes(fill=Share)) +
  scale_fill_viridis_c(labels=scales::percent) +
  coord_sf(datum=NA) +
  labs(title="CERB recipients in the City of Ottawa",
       fill="Share of people\n15+ reveiving\nCERB",
       caption="StatCan Census 2021")
```



It is difficult to see the central parts, we might want to zoom in a little. At the same time, it might be useful to add in Gatineau and surrounding municipalities, so maybe we want the data for the entire metro area.

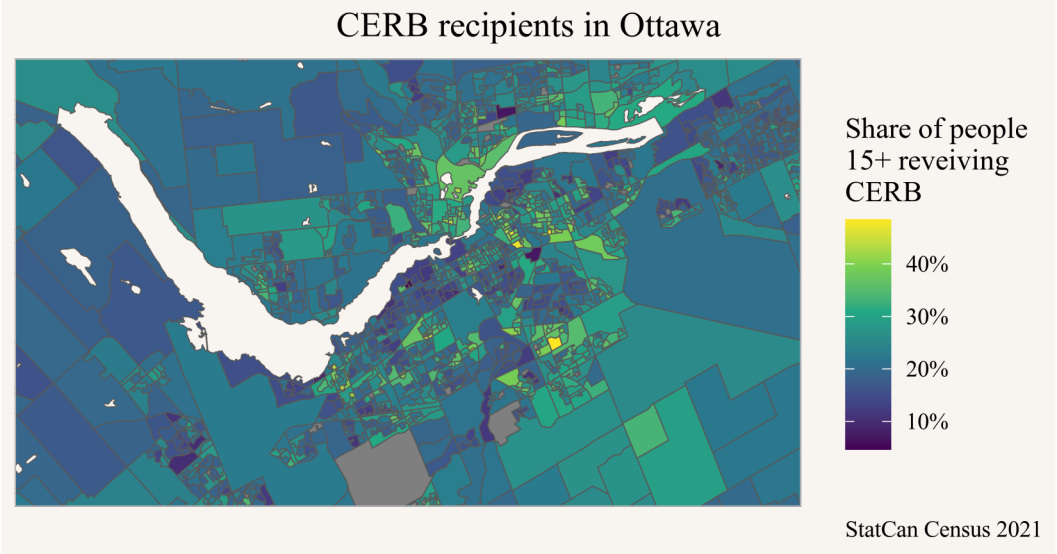
To do this we copy and paste the code from above and chain it into a single pipe, from data acquisition (using the CMA 505 for Ottawa CMA), computing the share, to plotting and cutting the region to the central parts by looking at the grid from the first map.

```
get_census("CA21",regions=list(CMA="505"),
                          vectors=c(cerb="v_CA21_593", base="v_CA21_554"),
                          level="CT", geo_format="sf") |>
  mutate(Share=cerb/base) |>
  ggplot() +
  geom_sf(aes(fill=Share)) +
  scale_fill_viridis_c(labels=scales::percent) +
  coord_sf(datum=NA, xlim=c(-76,-75.5), ylim=c(45.3,45.5)) +
  labs(title="CERB recipients in Ottawa",
       fill="Share of people\n15+ reveiving\nCERB",
       caption="StatCan Census 2021")
```



This brings out the central regions much better. We could also try this with finer geographies, setting the level to dissemination areas instead of census tracts. The same code as before works, except changing the `level="CT"` to `level="DA"`.

```
get_census("CA21",regions=list(CMA="505"),
                          vectors=c(cerb="v_CA21_593", base="v_CA21_554"),
                          level="DA", geo_format="sf") |>
  mutate(Share=cerb/base) |>
  ggplot() +
  geom_sf(aes(fill=Share)) +
  scale_fill_viridis_c(labels=scales::percent) +
  coord_sf(datum=NA, xlim=c(-76,-75.5), ylim=c(45.3,45.5)) +
  labs(title="CERB recipients in Ottawa",
       fill="Share of people\n15+ reveiving\nCERB",
       caption="StatCan Census 2021")
```



7.6. Interpretation

We notice substantial differences in the share of people receiving CERB benefits, with rural areas generally having lower shares, and central areas being more mixed, varying between under 10% to well over 40% of people 15 years and over receiving CERB. Generally areas with lower incomes have benefited more from CERB.

8. Cars vs SUVs in Canada

Private motor vehicles in Canada seem to be getting larger and it feels like SUVs and light trucks are taking over. This subjective feeling prompts us to ask the following question to check if this is just our imagination or a real phenominon.

8.1. Question

Are SUVs and light trucks taking over in Canada?

This question is somewhat vague, it's not clear what *taking over* means. But the question is clear enough to get us started on some descriptive analysis.

8.2. Data sources

Data discovery can be challenging, but just typing "statcan motor vehicle sales" into a search engine is a good start and gets us to the StatCan table enumerating data on new motor vehicle sales. We can also use the built-in search functionality in the {cansim} package.

```
library(dplyr)
library(ggplot2)
library(cansim)

search_cansim_cubes("motor vehicle sales") |>
  select(cansim_table_number,cubeTitleEn,cubeStartDate,cubeEndDate)
```

```
# A tibble: 2 x 4
  cansim_table_number cubeTitleEn                      cubeStartDate cubeEndDate
  <chr>               <chr>                            <date>        <date>
1 20-10-0001          New motor vehicle sales          1946-01-01    2024-02-01
2 20-10-0002          New motor vehicle sales, by typ~ 2010-01-01    2023-01-01
```

There are two tables with motor vehicle sales, we can inspect them on the web or via the {cansim} package. The second table covers a much shorter time period, and is also less recent. We will check out the first table to see if it fits our needs.

To access the web we can simply type `view_cansim_webpage("20-10-0001")` into the console, which will open the StatCan webpage for Table 20-10-0001. Getting table overview data via the {cansim} package requires to load the table first, which can be slow for larger tables.

```
get_cansim_table_overview("20-10-0001")
```

This tells us that Table 20-10-0001 might have the information we need, we check the table notes to better understand what "Trucks" entails, selecting the two columns we are interested in.

```
get_cansim_table_notes("20-10-0001") %>%
  select(`Member Name`,Note) %>%
  knitr::kable()
```

Member Name	Note
NA	Prior to 1953, data for Canadian and United States manufactured vehicles and overseas manufactured vehicles are not segregated.
British Columbia and the Territories	Includes Yukon, Northwest Territories and Nunavut.
Trucks	Trucks include minivans, sport-utility vehicles, light and heavy trucks, vans and buses.
Total, overseas	Includes Japan and other countries.
NA	Seasonally adjusted data for the New Motor Vehicle Sales survey are available for the period between January 1991 to February 2012.

It looks like "Trucks" does includes SUVs, but next to light trucks it also includes heavy trucks and buses. It also includes minivans, and thinking back at our original question, we might want to refine it to include minivans.

This allows us to separate passenger cars from basically everything else. Thinking that heavy truck and bus sales probably only make up a small portion, we could use that as a stand-in for our "SUVs and light trucks" in our question. But the match is not ideal and this leaves questions open.

Maybe Table 20-10-0002 works better for our purposes, time to look at the table overview.

```
get_cansim_table_overview("20-10-0002")
```

The frequency is only annual as opposed to the monthly data from the previous table, but the breakdown of vehicle types looks much better for our purposes, it allows us to distinguish light trucks from heavy trucks and buses. Time to check the table notes for more details on the definitions.

```
get_cansim_table_notes("20-10-0002") %>%
  select(`Member Name`,Note) %>%
  knitr::kable()
```

Member Name	Note
British Columbia and the Territories	Includes Yukon, Northwest Territories and Nunavut.
Trucks	Trucks include minivans, sport-utility vehicles, light and heavy trucks, vans and buses.
Light trucks	Light trucks: include minivans, sport-utility vehicles, light trucks and vans.
Heavy trucks	Heavy trucks: include class 4, 5, 6, 7 and 8 trucks.

This looks like it fits what we need, we want to compare unit sales of Passenger cars to Light trucks.

8.3. Data acquisition

Getting the data is easy now. The {cansim} package will automatically add a native `Date` column, to convert annual data to dates it defaults to July 1st of that year. While it is a sensible default to assign a mid-year date to annual data, later on for plotting it will be more convenient for us to set the date at January 1st, so we override the default using the optional `default_month` argument.

```
data <- get_cansim("20-10-0002", default_month = 1)

data |> select(Date,GEO,`Vehicle type`,Sales,val_norm) %>%
  head()
```

```
# A tibble: 6 x 5
  Date       GEO    `Vehicle type`            Sales      val_norm
  <date>     <chr>  <fct>                     <fct>         <dbl>
1 2010-01-01 Canada Total, new motor vehicles Units       1584499
2 2010-01-01 Canada Total, new motor vehicles Dollars 52315609000
3 2010-01-01 Canada Passenger cars            Units        710214
4 2010-01-01 Canada Passenger cars            Dollars 18982437000
5 2010-01-01 Canada Trucks                    Units        874285
6 2010-01-01 Canada Trucks                    Dollars 33333173000
```

Quick inspection of the data, using the columns we identified in the overview, helps identify the basic structure of the data.

8.4. Data preparation

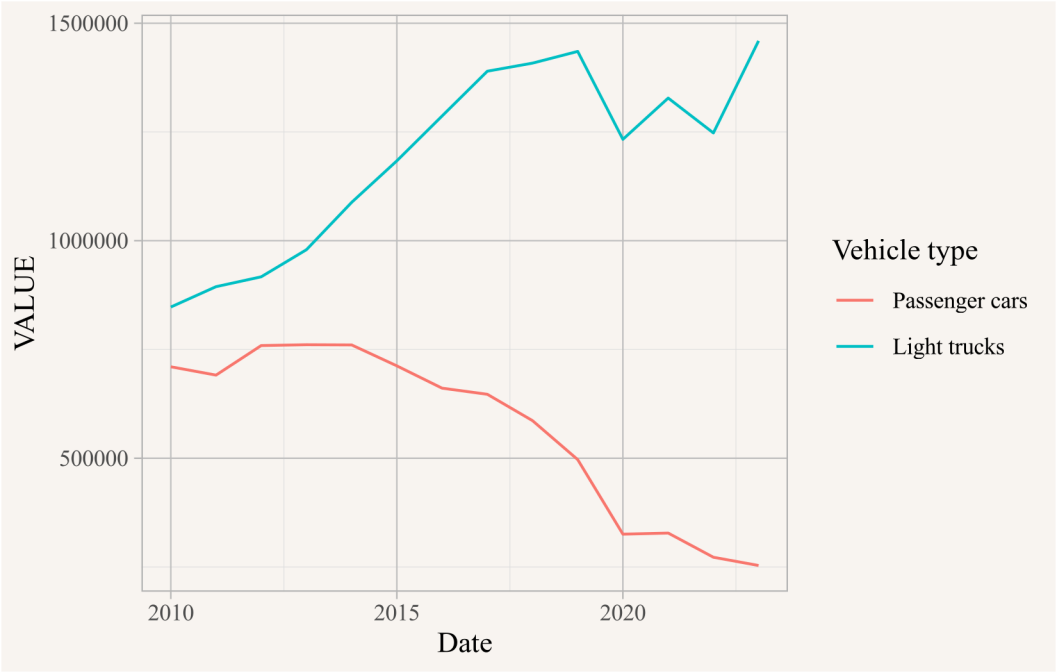
There is not much data preparation needed, we just filter down to the data we are interested in.

```
plot_data <- data |>
  filter(`Vehicle type` %in% c("Passenger cars","Light trucks"),
         Sales=="Units",
         GEO=="Canada")
```

8.5. Analysis and visualization

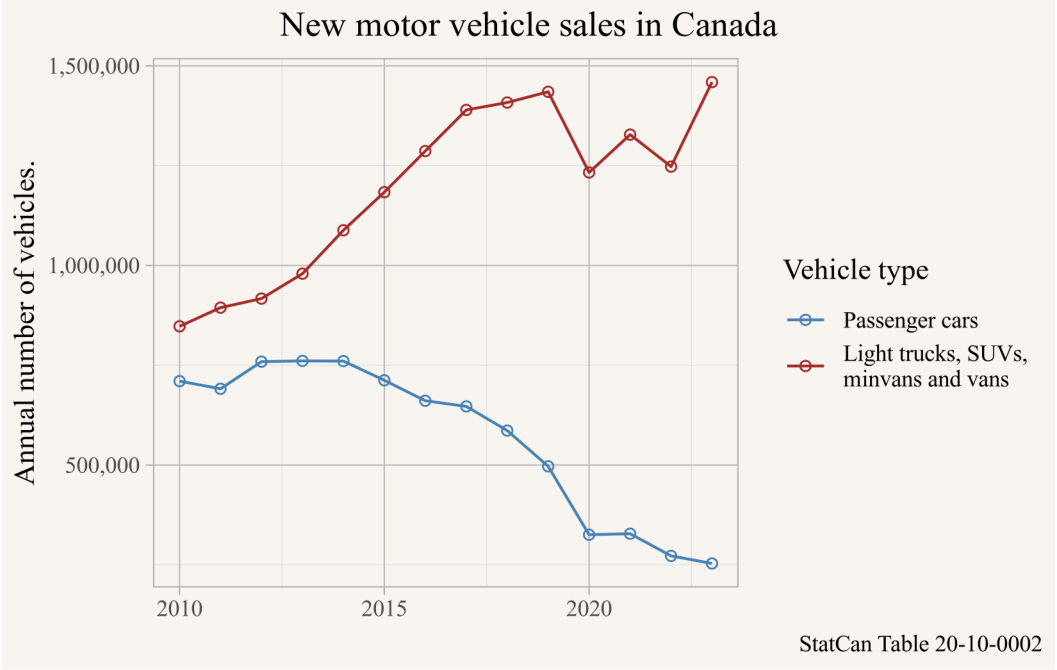
Time to lake a look what this looks like, to plot we filter for the overall Canadian data series, and tell `ggplot` to map *Date* on the x-axis, the values colum '`VALUE`' on the y-axis, and colour by vehicle type.

```
ggplot(plot_data,aes(x=Date,y=VALUE,colour=`Vehicle type`)) +
  geom_line()
```



This is looking good, time to clean up the graph a bit. We add markers for the data points, nicer axis labels, as well as a title and labels.

```
ggplot(plot_data,aes(x=Date,y=VALUE,colour=`Vehicle type`)) +
  geom_point(shape=21) +
  geom_line() +
  scale_y_continuous(labels=scales::comma) +
  scale_color_manual(labels=c("Light trucks"="Light trucks, SUVs,\nminvans and vans"),
                     values=c("Light trucks"="brown","Passenger cars"="steelblue")) +
  labs(title="New motor vehicle sales in Canada",
       x=NULL,y="Annual number of vehicles.",
       caption="StatCan Table 20-10-0002")
```



This answers our question in that more light trucks, SUVs, minivans and vans are sold than cars, and the gap has been growing. But the data only starts in 2010, and we suspect that things weren't always this way. At what point did SUVs and light trucks overtake new car sales?

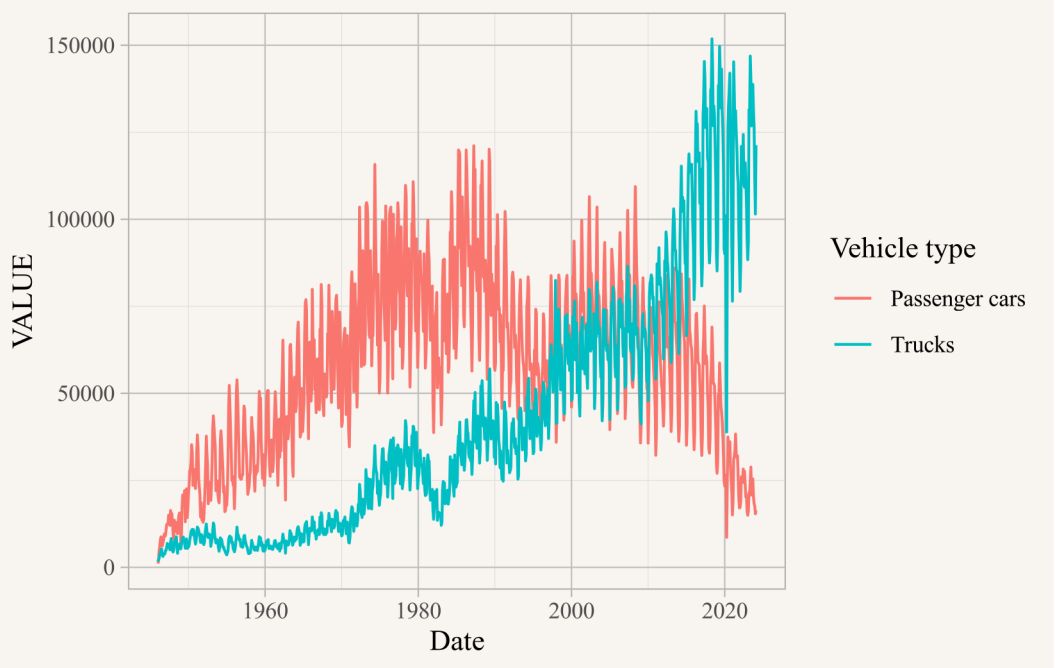
To answer than we need to jump back and load the other time series. It won't let us separate out heavy trucks and buses, but we can estimate how bad the difference is by comparing it to this data.

We load the data and filter it down to the parts that we are interested in.

```
data2 <- get_cansim("20-10-0001")

plot_data2 <- data2 |>
  filter(`Vehicle type` %in% c("Passenger cars","Trucks"),
         Sales=="Units",
         `Origin of manufacture`=="Total, country of manufacture",
         `Seasonal adjustment`=="Unadjusted",
         GEO=="Canada")
```

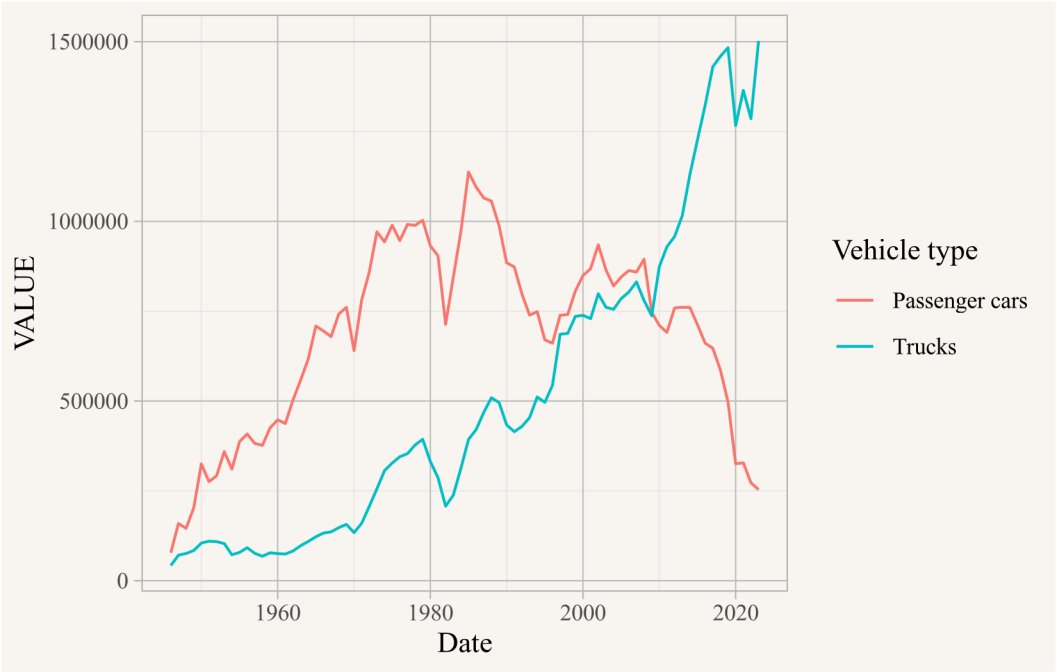
A quick plot gives us a general idea what this looks like.

```
plot_data2 %>%
  ggplot(aes(x=Date,y=VALUE,colour=`Vehicle type`)) +
  geom_line()
```



There is a strong seasonal pattern in vehicle sales, for now we will just aggregate it to annual sales so we can compare it with the previous data. For this we extract the `Year` from the `Date` column, group by `Year` and `Vehicle type` and summarize by adding up the 'VALUE' column. We added a count column to keep track how many months we added up so we can later ensure we are only showing years for which we have complete data.

```
plot_data2_annual <- plot_data2 |>
  mutate(Year=strftime(Date,"%Y")) |>
  group_by(Year,`Vehicle type`) |>
  summarise(VALUE=sum(VALUE), n=n(),.groups="drop") |>
  mutate(Date=as.Date(paste0(Year,"-01-01"))) |>
  filter(n==12)  # only use years with full 12 months of data

ggplot(plot_data2_annual,aes(x=Date,y=VALUE,colour=`Vehicle type`)) +
  geom_line()
```



Time to combine this with our previous data. This tells us that the most interesting change happened 1985 and onward, so we will discard earlier years. One quick sanity check is to see if the annual passenger car sales derived from the two series agree for the years where they are in common. Here we join the two data tables by `Date` and `Vehicle type` in order to compare the two estimate. We rename the `VALUE` column on the first one in order to avoid name conflicts.

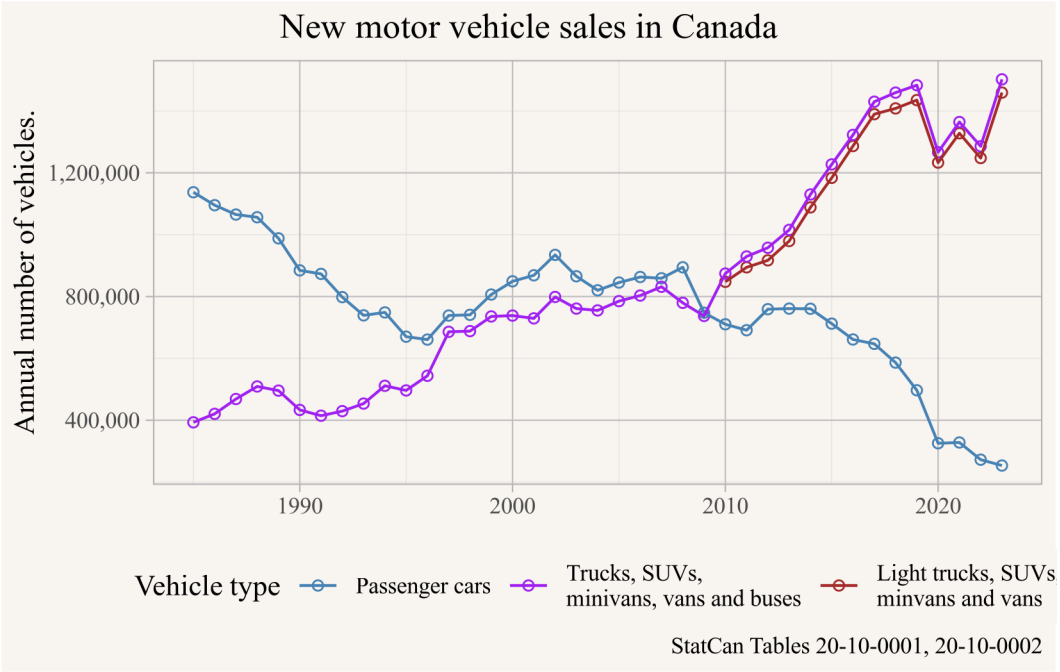
```
plot_data %>%
  filter(`Vehicle type`=="Passenger cars") %>%
  select(Date,`Vehicle type`,VALUE1=VALUE) %>%
  left_join(plot_data2_annual,by=c("Date","Vehicle type")) %>%
  mutate(diff=VALUE1-VALUE) %>%
  select(Year,VALUE1,VALUE,diff)
```

```
# A tibble: 14 x 4
   Year  VALUE1  VALUE  diff
   <chr>  <dbl>  <dbl> <dbl>
 1 2010  710214 710214     0
 2 2011  691079 691079     0
 3 2012  759024 759024     0
 4 2013  760924 760920     4
```

```
 5 2014  760449 760449     0
 6 2015  712322 712322     0
 7 2016  661088 661088     0
 8 2017  646960 646960     0
 9 2018  586357 586357     0
10 2019  496851 496851     0
11 2020  325494 325494     0
12 2021  327994 327994     0
13 2022  272408 272408     0
14 2023  253453 253453     0
```

The data for all years agrees, except for 2013 where one series counts 4 more passenger cars.

```
bind_rows(plot_data %>% filter(`Vehicle type`!="Passenger cars"),
          plot_data2_annual %>% filter(Date>=as.Date("1985-01-01"))) %>%
  ggplot(aes(x=Date,y=VALUE,colour=`Vehicle type`)) +
  geom_point(shape=21) +
  geom_line() +
  scale_y_continuous(labels=scales::comma) +
  scale_color_manual(labels=c("Light trucks"="Light trucks, SUVs,\nminvans and vans",
                              "Trucks"="Trucks, SUVs,\nminivans, vans and buses"),
                     values=c("Light trucks"="brown","Passenger cars"="steelblue",
                              "Trucks"="purple")) +
  theme(legend.position = "bottom") +
  labs(title="New motor vehicle sales in Canada",
       x=NULL,y="Annual number of vehicles.",
       caption="StatCan Tables 20-10-0001, 20-10-0002")
```



8.6. Interpretation

This confirms our initial suspicion that the "Trucks" category is dominated by light trucks, SUVs, minivans and vans, at least for the years 2010 onwards where we have data for both. Which gives us confidence to say that truck and SUV sales caught up to passenger cars sales by around 1997, and the two evolved fairly parallel until 2009, after which SUVs and light trucks increased dramatically and passenger car sales fell.

9. Under construction

Units under construction give some indication of construction activity beyond starts and completions.

9.1. Question

How many homes are currently under construction in Toronto?

9.2. Data sources

CMHC tracks information on housing starts and completions. And the number of homes under construction, that is dwelling units that have started but aren't yet completed.

CMHC defines a housing "start" as the time when the foundation is finished, so digging a parking crater and building below ground happens before what CMHC calls a building "start". This might differ from how one might colloquially think about units under construction, as there can be significant construction activity before a "start". But this probably comes reasonably close to our question of interest.

9.3. Data acquisition

The **cmhc** package facilitates importing data from CMHC. This pries data out the Housing Market Information Portal where data is organized across a variety of tables. The easiest way to locate a table of interest is to use the `select_cmhc_table()` function from the **cmhc** package in the console to interactively step through the process. In our case, we are interested in data from the Starts and Completions Survey (**Scss**), look at the **Under Construction** series, after which we can select to have data broken down by B**edroom Type** or I**ntended Market**, where we select the former. Lastly we need to decide the breakdown type, either a level of geography or **Historical Time Periods** for a fixed geography, which is what we are interested in.

Going through this process gives us the code we need to access the data, all we need to do is fill in the geographic identifier. The **cmhc** package is designed to work in conjunction with other census data, so it uses the same geographic identifiers and translates them to CMHC's own internal geographic identifiers under the hood. For Toronto, we need to decide if we are interested in the City of the metro area and grab the geographic identifier from the CensusMapper API tool. We will query data for the City of Toronto with standard StatCan geographic identifier "3520005" .

```
library(tidyverse)
library(cmhc)
under_construction <- get_cmhc(survey = "Scss",
                               series = "Under Construction",
                               dimension = "Dwelling Type",
                               breakdown = "Historical Time Periods",
                               geo_uid = "3520005")
```

9.4. Data preparation

There is really not much to do here, let's just inspect what the data looks like

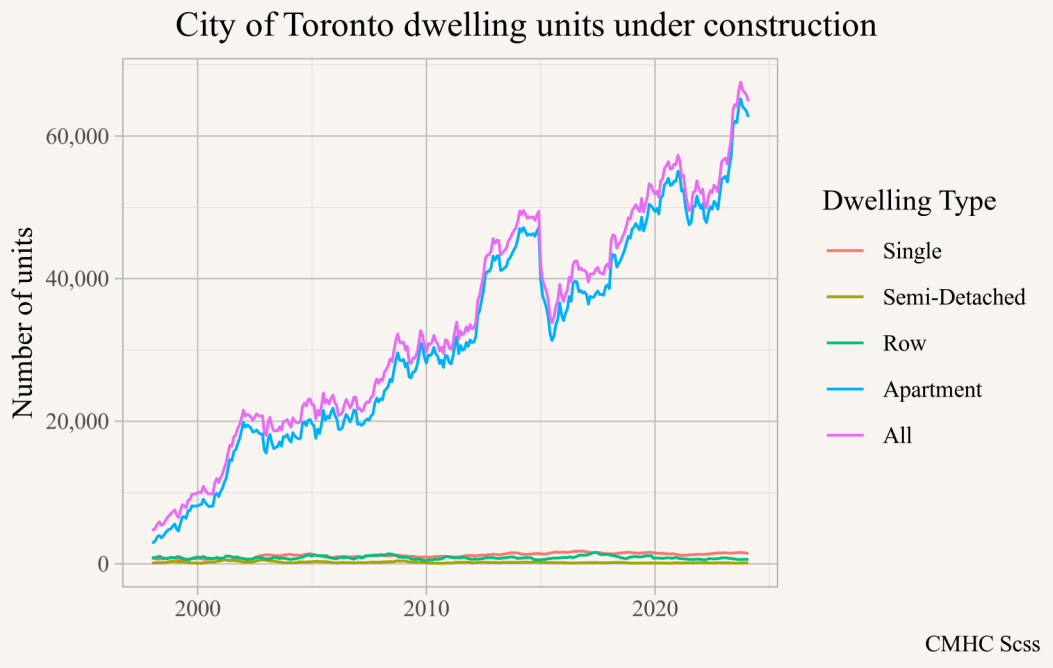
```
under_construction |> head()
```

```
# A tibble: 6 x 7
  GeoUID  Date       DateString `Dwelling Type` Value Survey Series
  <chr>   <date>     <chr>      <fct>           <dbl> <chr>  <chr>
1 3520005 1998-01-01 Jan 1998   Single            838 Scss   Under Construction
2 3520005 1998-01-01 Jan 1998   Semi-Detached     132 Scss   Under Construction
3 3520005 1998-01-01 Jan 1998   Row               838 Scss   Under Construction
4 3520005 1998-01-01 Jan 1998   Apartment        3008 Scss   Under Construction
5 3520005 1998-01-01 Jan 1998   All              4816 Scss   Under Construction
6 3520005 1998-02-01 Feb 1998   Single            738 Scss   Under Construction
```

9.5. Analysis and visualization

What's left is to plot the data, broken out by dwelling type.

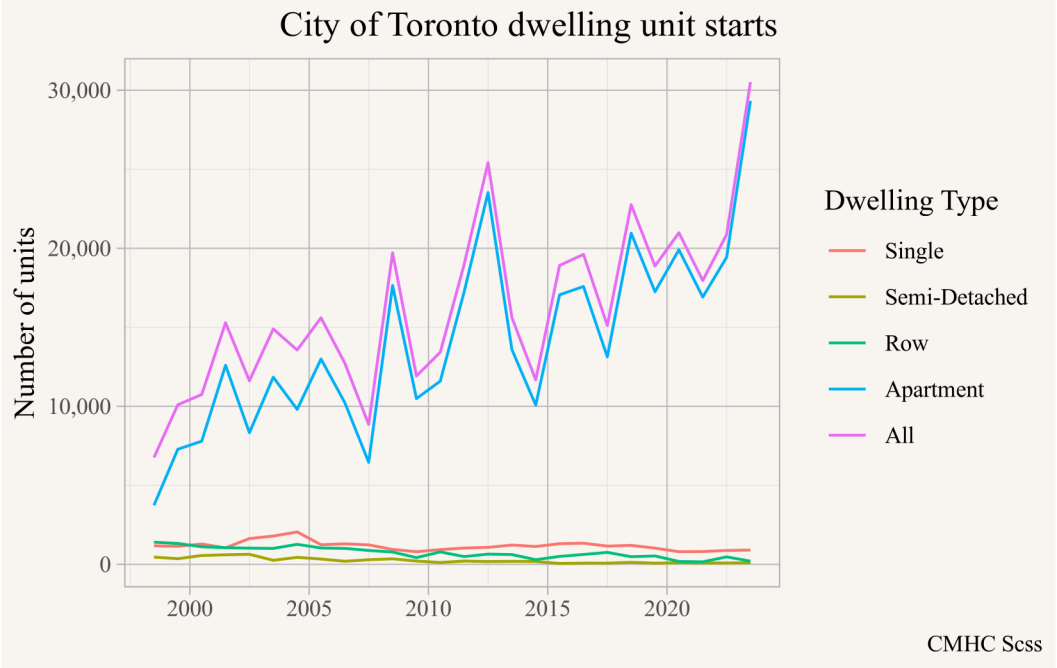
```
ggplot(under_construction, aes(x=Date,y=Value,colour=`Dwelling Type`)) +
  geom_line() +
  scale_y_continuous(labels=scales::comma) +
  labs(title="City of Toronto dwelling units under construction",
       x=NULL,y="Number of units", caption="CMHC Scss")
```



It looks like the number of units under construction, especially apartment units, has increased considerably over time. Let's cross-check that against housing starts. These tend to be quite noisy, so we go to annual frequency instead of monthly. We can adapt the code above for data acquisition and graphing into one chunk.

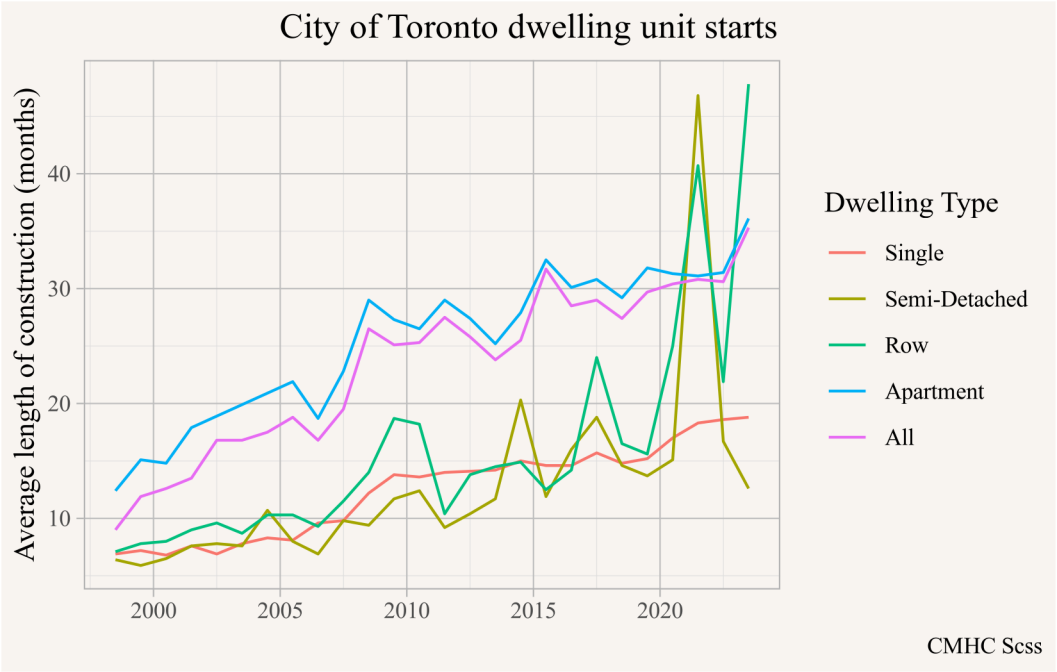
```
get_cmhc(survey = "Scss",
         series = "Starts",
         dimension = "Dwelling Type",
         breakdown = "Historical Time Periods",
         frequency = "Annual",
         geo_uid = "3520005") |>
  ggplot(aes(x=Date,y=Value,colour=`Dwelling Type`)) +
  geom_line() +
  scale_y_continuous(labels=scales::comma) +
```

```
  labs(title="City of Toronto dwelling unit starts",
       x=NULL,y="Number of units", caption="CMHC Scss")
```



Starts have increased, but not that much. Something else must be at play too, let's look at how length of construction has changed over this timeframe, again using annual data to cut down on noise.

```
get_cmhc(survey = "Scss",
         series = "Length of Construction",
         dimension = "Dwelling Type",
         breakdown = "Historical Time Periods",
         frequency = "Annual",
         geo_uid = "3520005") |>
  ggplot(aes(x=Date,y=Value,colour=`Dwelling Type`)) +
  geom_line() +
  scale_y_continuous(labels=scales::comma) +
  labs(title="City of Toronto dwelling unit starts",
       x=NULL,y="Average length of construction (months)",
       caption="CMHC Scss")
```



And indeed, the length of construction shot up a lot, for apartments from around 13 months in the late 90s to about 30 months around 2020. That means we now have over twice as many construction sites for the same number of units coming to market compared to the late 90s.

The sharp increase in construction time for Semi-detached and row houses might well be a data anomaly, where low and dropping number of starts of such units can be disproportionally impacted by a couple of stalled projects.

9.6. Interpretation

The units under construction has increased a lot in the City of Toronto, due to the combined effects of increasing building starts as well as a more than doubling of average time to complete these units.

10. Geography of income change

Incomes in a region change by people getting higher (or lower) incomes as well as people moving in and out of a region. We can observe the aggregate effects by looking at change in income statistics.

10.1. Question

Where and how did incomes change in the City of Vancouver?

10.2. Data sources

The main data sources for fine-geography income data is the census, although custom tabulations of T1FF taxfiler data can offer insight of this on an annual basis at the census tract geography. For our question we are interested in broad temporal ranges, so the 5-year census data will work well.

We need to decide which income concept is best suited for our question, it is worthwhile to spend some time with the Census Income Reference Guide to understand how the data was collected and what income concept to use. Prior to 2011 the income data was part of the long form census. In 2011 the mandatory long form was replaced with the voluntary NHS, given people the option to link directly to T1FF taxfiler data or to detail the income data manually. Starting 2016 income data was linked for all people to the T1FF taxfiler data.

The question what income concept to use, e.g. individual income, household income, family income, employment income, etc, depends on the particular question we are interested in. For now we will go with family income, trying to understand how the income situation of families varies across Vancouver and across time. Family income is less affected by demographic factors like the distribution of single vs multiple person households, but is still impacted by e.g. differences in shares of seniors vs young families vs families at the peak of their earnings.

10.3. Data acquisition

We again use the CensusMapper API tool to locate the internal CensusMapper identifiers for Median Total Income of Economic Families for the years 2006 through 2021. For 2001 the standard census products reported income for census families instead of economic families, so they aren't directly comparable. As geographic breakdown we choose census tracts.

```
library(tidyverse)
regions <- list(CSD="5915022")
income_vectors <- c("2021"="v_CA21_965",
                    "2016"="v_CA16_2447",
                    "2011"="v_CA11N_2456",
                    "2006"="v_CA06_1741")
```

To facilitate the data import we write a wrapper function to acquire the census data for each of our four years. For a given census year we create the corresponding dataset identifier and select the appropriate income variable. To reduce clutter we select just the income variable and also keep the geographic identifier, and add the census year to the table.

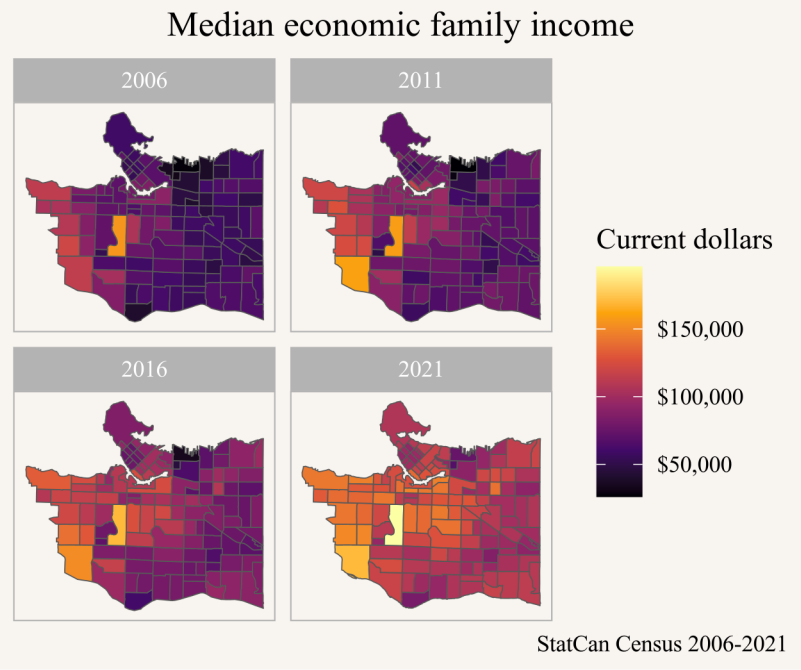
```
get_census_data <- function(year){
  year <- as.character(year)
  get_census(year,regions=regions,
             vectors=c("ef_income"=as.character(income_vectors[year])),
             geo_format="sf",level="CT") |>
    select(GeoUID,ef_income) |>
    mutate(Year=year)
}
```

Importing the data is easy now, we just call our function for each census year and collect it into a data frame.

```
library(cancensus)
income_data <- seq(2006,2021,5) |>
  map_df(get_census_data)
```

Let's take a quick look.

```
ggplot(income_data) +
  geom_sf(aes(fill=ef_income)) +
  scale_fill_viridis_c(option="inferno", labels=scales::dollar) +
  facet_wrap(~Year) +
  coord_sf(datum=NA) +
  labs(title="Median economic family income",
       fill="Current dollars",
       caption="StatCan Census 2006-2021")
```



10.4. Data preparation

Looking at the above graph we can see the geographic variation in each year, but it is difficult to discern geographic trends over time as incomes have gone up a lot during this timeframe. It makes sense to look at inflation-adjusted incomes instead. For this we use annual consumer price index data from StatCan Table 18-10-0005. To simplify things we locate the specific vector **v41693271** for the all-time CPI.

```
library(cansim)
inflation <- get_cansim_vector("v41693271") |>
  mutate(Year=strftime(Date,"%Y")) |>
  select(Year,CPI=val_norm) |>
  filter(Year %in% names(income_vectors))

inflation
```

```
# A tibble: 4 x 2
  Year    CPI
  <chr> <dbl>
1 2006   109.
2 2011   120.
3 2016   128.
4 2021   142.
```

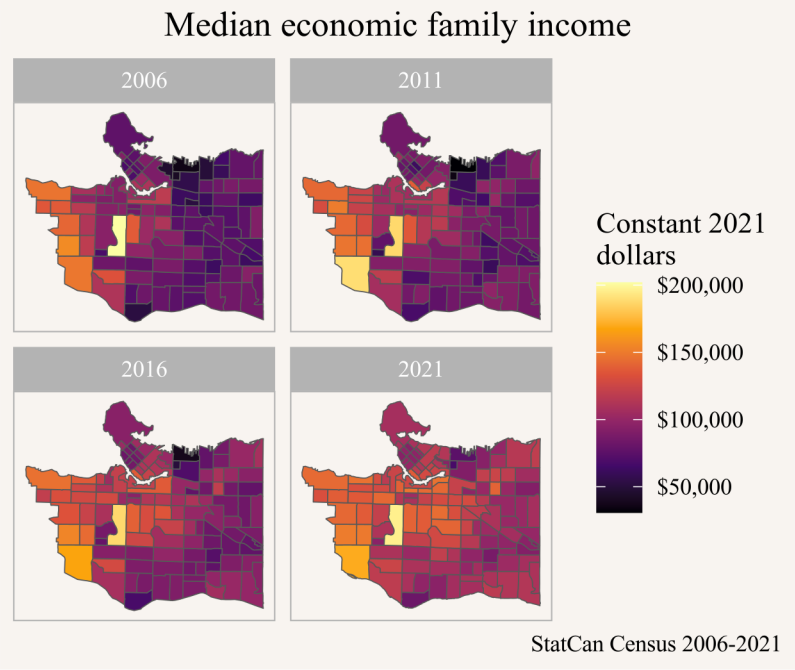
10.5. Analysis and visualization

With this, we can adjust the census data by inflation. We choose to base everything on 2021 dollars.

```
inflation <- inflation |>
  mutate(CPI=CPI/last(CPI,order_by = Year))
```

Now we just join the inflation data onto our income data by year, this adds the `CPI` column from the inflation data frame to our income with the CPI value corresponding to the value in the `Year` column in each of the two data frames. We then colour by inflation-adjusted income using the same code for graphing as above.

```
income_data |>
  left_join(inflation,by="Year") |>
  ggplot() +
  geom_sf(aes(fill=ef_income/CPI)) +
  scale_fill_viridis_c(option="inferno", labels=scales::dollar) +
  facet_wrap(~Year) +
  coord_sf(datum=NA) +
  labs(title="Median economic family income",
       fill="Constant 2021\ndollars",
       caption="StatCan Census 2006-2021")
```



This shows more clearly how incomes have increased over time, but it would be nice to compute the change in income 2006 to 2021 for each individual census tract. But keen observers will notice that some census tracts have changed over the years, making it very difficult to compare data directly.

10.6. Data acquisition (part 2)

Fortunately the problem of making census data comparable across time has been solved with the tongfen package. This allows us to create a semi-custom tabulation on the fly on a harmonized geography based on census tracts by aggregating census data appropriately. One problem is that medians can't be aggregated, so we need to either use average income instead or be content that medians can only be approximated. By default the **tongfen** package aggregates medians as if they were averages and emits a warning. This is the route we will take for this.

To start out, we need to create metadata for the **tongfen** procedure. This is automated for Canadian census data, leveraging the metadata built into CensusMapper.

```
library(tongfen)
meta <- meta_for_ca_census_vectors(income_vectors)
```

```
meta
```

```
# A tibble: 8 x 10
  variable  label dataset type  aggregation units rule  parent geo_dataset  year
  <chr>     <chr> <chr>   <chr> <chr>       <chr> <chr> <chr>  <chr>       <int>
1 v_CA21_9~ 2021  CA21    Orig~ Median of ~ Curr~ Medi~ v_CA2~ CA21         2021
2 v_CA16_2~ 2016  CA16    Orig~ Median of ~ Curr~ Medi~ v_CA1~ CA16         2016
3 v_CA11N_~ 2011  CA11N   Orig~ Median of ~ Curr~ Medi~ v_CA1~ CA11         2011
4 v_CA06_1~ 2006  CA06    Orig~ Median of ~ Curr~ Medi~ v_CA0~ CA06         2006
5 v_CA21_9~ v_CA~ CA21    Extra Additive    <NA>  Addi~ <NA>   CA21         2021
6 v_CA16_2~ v_CA~ CA16    Extra Additive    <NA>  Addi~ <NA>   CA16         2016
7 v_CA11N_~ v_CA~ CA11N   Extra Additive    <NA>  Addi~ <NA>   CA11         2011
8 v_CA06_1~ v_CA~ CA06    Extra Additive    <NA>  Addi~ <NA>   CA06         2006
```

The metadata contains our original income data, as well as extra variables needed to properly aggregate the data. Getting the income data on a common geography is easy now.

```
unified_income_data <- get_tongfen_ca_census(regions,meta)
```

10.7. Analysis and visualization

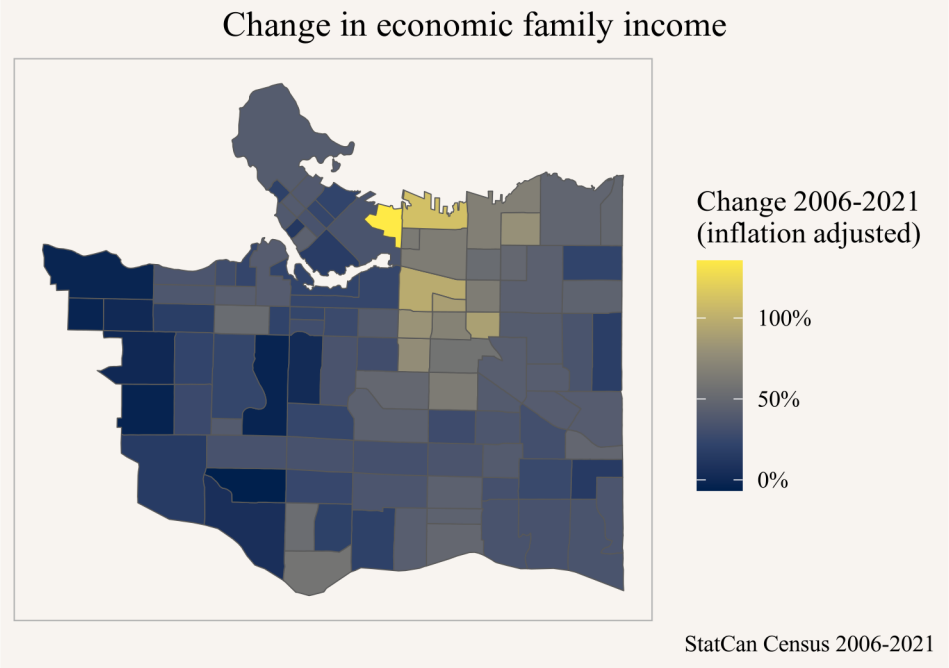
In line with what we did before we look at inflation-adjusted income change. To this end we extract the adjustment factor for the 2006-2021 timeframe.

```
inflation_2006_2021 <- inflation |>
  filter(Year=="2006") |>
  pull(CPI)
```

With that we can simply plot the data, mapping the inflation-adjusted percent change 2006 to 2021.

```
unified_income_data |>
  ggplot() +
  geom_sf(aes(fill=`2021`/`2006`*inflation_2006_2021-1)) +
  scale_fill_viridis_c(option="cividis", labels=scales::percent) +
  coord_sf(datum=NA) +
```

```
labs(title="Change in economic family income",
     fill="Change 2006-2021\n(inflation adjusted)",
     caption="StatCan Census 2006-2021")
```



10.8. Interpretation

In summary we see that income of economic families changed fastest in the Downtown Eastside, Grandview-Woodlands and Strathcona neighbourhoods, effectively doubling. Incomes increased least on the West Side, where they were already quite high to start with, and increased by about 50% throughout much of the East Side.

11. Toronto children

We keep reading in the news that the number (and share) of children under the age of 15 declined in the City of Toronto, a fate that's shared by the City of Vancouver. But is this happening uniformly across the city or are there geographic differences?

11.1. Question

Where in Toronto is the number (and share) of children under 15 decreasing, and where is it increasing?

11.2. Data sources

In Canada we have data on the number of children from two main data sources and their derived products. The census and T1FF taxfiler data. Both are available at sub-city level, although that's a custom tabulation for T1FF taxfiler data. For this post we will go with census data since the T1FF custom tabulations we have on CensusMapper only go up to 2018.

11.3. Data acquisition

Since we are looking at comparing census data over time, and census geographies change over time, we will rely on tongfen to harmonize the census geographies. As the base we will use census tracts. We can use the CensusMapper API GUI to select the vectors we need, the children under 15 in 2021 and the children under 15 in 2001, assembled from 5 year age groups for males and females.

```
library(tidyverse)
library(tongfen)

meta <- meta_for_ca_census_vectors(c(children_2021="v_CA21_11",
                                     children1m_2001="v_CA01_7",
                                     children2m_2001="v_CA01_8",
                                     children3m_2001="v_CA01_9",
                                     children1f_2001="v_CA01_26",
                                     children2f_2001="v_CA01_27",
                                     children3f_2001="v_CA01_28"))
```

We still need to get the region identifier for Toronto, either by using the CensusMapper API GUI or searching for it using the cancensus package.

```
library(cancensus)
search_census_regions("Toronto","CA21")
```

```
# A tibble: 3 x 8
  region  name    level     pop municipal_status CMA_UID CD_UID PR_UID
  <chr>   <chr>   <chr>   <int> <chr>            <chr>   <chr>  <chr>
1 35535   Toronto CMA   6202225 B                <NA>    <NA>   35
2 3520    Toronto CD    2794356 CDR              <NA>    <NA>   35
3 3520005 Toronto CSD   2794356 C                35535   3520   35
```

Now we have everything in place to get the data on a harmonized geography based on census tracts.

```
toronto_children <- get_tongfen_ca_census(regions=list(CSD="3520005"), meta=meta,
                                          level = "CT", na.rm = TRUE)
```

11.4. Data preparation

This time around we have a little bit of data preparation to do, we need to add up all the children from the age and gender groups for 2001.

```
plot_data <- toronto_children |>
  mutate(children_2001=children1m_2001+children2m_2001+children3m_2001+
           children1f_2001+children2f_2001+children3f_2001) |>
  select(matches("children_\\d{4}|Population"))
```

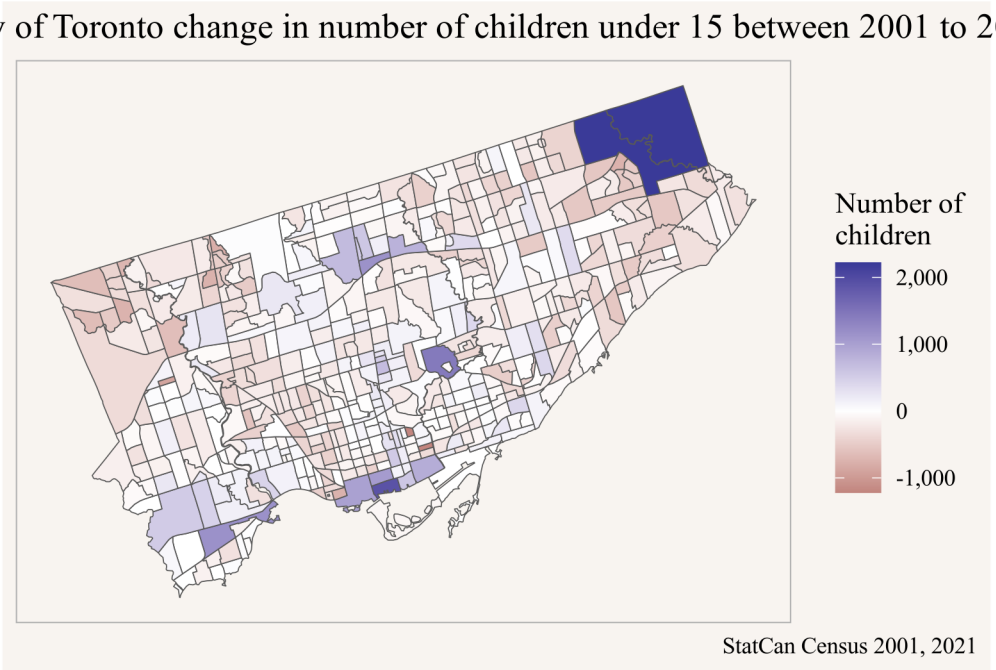
Sometimes it can be tedious to spell out all the variables we want to add up, and we can use more abstract selections to specify what and how to add. In this case that's complicated by also having geographic data attached, that we need to drop in order to perform row wise summation on our variables of interest. For reference, here is an alternative way to perform this summation that generalizes to more complex scenarios.

```
plot_data <- toronto_children %>%
  mutate(children_2001=select(.,matches("children.+_2001")) |>
           sf::st_drop_geometry() |>
           rowSums(na.rm=TRUE)) |>
  select(matches("children_\\d{4}|Population"))
```

11.5. Analysis and visualization

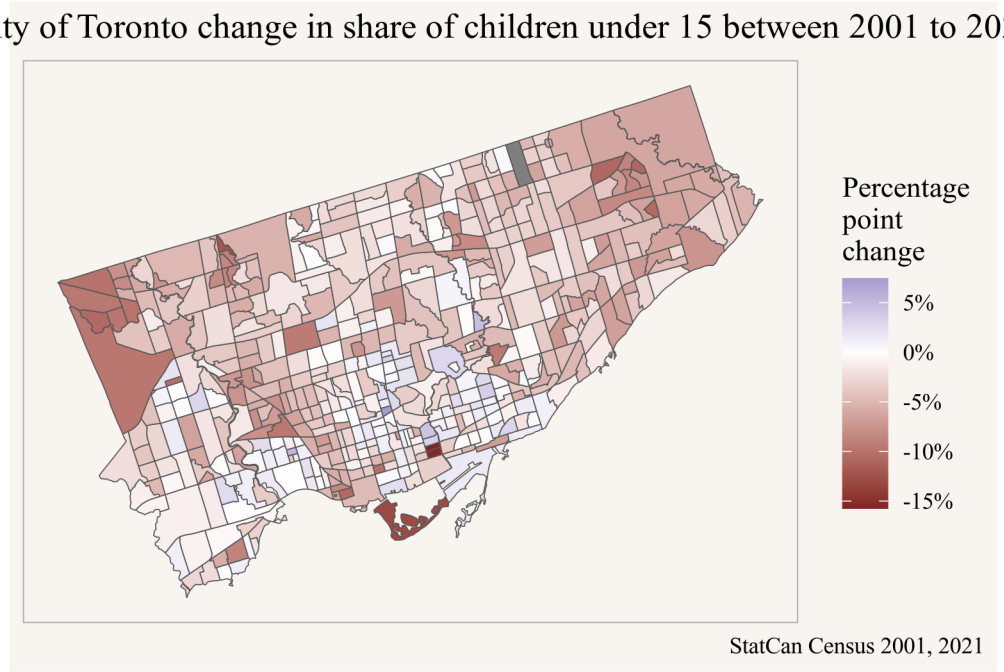
Here we need to simply map the difference in children.

```
ggplot(plot_data, aes(fill=children_2021-children_2001)) +
  geom_sf() +
  scale_fill_gradient2(labels=scales::comma) +
  coord_sf(datum=NA) +
  labs(title="City of Toronto change in number of children under 15 between 2001 to 2021",
       fill="Number of\nchildren",
       caption="StatCan Census 2001, 2021")
```



Another view into this is to look at the change in the share of children in each region between these years.

```
ggplot(plot_data, aes(fill=children_2021/Population_CA21-children_2001/Population_CA01)) +
  geom_sf() +
  scale_fill_gradient2(labels=scales::percent) +
  coord_sf(datum=NA) +
  labs(title="City of Toronto change in share of children under 15 between 2001 to 2021",
       fill="Percentage\npoint\nchange",
       caption="StatCan Census 2001, 2021")
```



11.6. Interpretation

The share of children has decreased in most areas, which is to be expected as Canada's overall age distribution shifts with people living longer and baby boomers aging into the retirement age. This means that if we want to keep the other age groups, we need to make more space for them.

Looking at the map with the absolute change in children we see that there are several areas where we did manage to make space for a shifting age distribution, and the total number of children increased even as their share decreased.

Additionally we see areas where not just the number but also the share of children increased. These are typically areas dominated by denser housing that traditionally weren't attractive to families with children. But with increasing constraints on housing availability the only alternative is to commute into the city from increasingly longer distances, and living in family-sized apartments in the central parts is increasingly becoming an attractive alternative.

12. Number of Household Maintainers

In a recent newspaper article it was reported that

> New Statistics Canada data show that households with three or more people contributing to shelter costs and other expenses grew 61 per cent compared with the overall household growth in the City of Vancouver in the past five years.

Reading this we might be interested in more context.

12.1. Question

How has the number of household maintainers changed in other municipalities, and what does this mean?

12.2. Data sources

The number of household maintainers is reported in the census:

> Refers to whether or not a person residing in the household is responsible for paying the rent, or the mortgage, or the taxes, or the electricity or other services or utilities. Where a number of people may contribute to the payments, more than one person in the household may be identified as a household maintainer. If no person in the household is identified as making such payments, the reference person is identified by default.

The census dictionary indicates that the ability to identify more than one household maintainer started with the 1996 census moving forward.

12.3. Data acquisition

We need the **tidyverse** and **cancensus** packages,

```
library(tidyverse)
library(cancensus)
```

and choose the appropriate census variables and decide what regions we are interested in. Moreover, we need to decide how long a timeframe we are interested in, the news article only referred to the 2016-2021 timeframe, but it might be worthwhile to also consider longer timeframes. The standard census profile data does not report on this before 2011 though, to keep things simple we collect data for the censuses 2011 and onward. We start by collecting the relevant census variables, labelled by **base** for the base number of households, and One, Two, and Three+ for 1, 2 or 3+ household maintainers.

```
vectors <- list(
  "2011"=c(base="v_CA11N_2259",One="v_CA11N_2260",Two="v_CA11N_2261","Three+"="v_CA11N_2262
  "2016"=c(base="v_CA16_4873",One="v_CA16_4874",Two="v_CA16_4875","Three+"="v_CA16_4876"),
  "2021"=c(base="v_CA21_4275",One="v_CA21_4276",Two="v_CA21_4277","Three+"="v_CA21_4278")
)
```

Selecting regions to compare Vancouver to is somewhat subjective. We go with a mix of larger cities in Metro Vancouver, as well as some from other provinces. One complication is that census geographies can change over time, so we need to be mindful of this. It is good practice to check this against the list of cities that changed 2011 to 2016 and 2016 to 2021, or explicitly inspect the geographies for changes.

We select from the list of all cities of at least 205k people (a number chosen to separate Richmond, BC from Richmond Hill, ON), and within that list narrow it down by matching by name.

```
regions <- list_census_regions("CA21") |>
  filter(level=="CSD",
         pop>205000,
         grepl("Vancouver|Surrey|Burnaby|Richmond|Toronto|Calgary|Edmonton|Halifax|Winnipeg

regions
```

```
# A tibble: 13 x 8
```

	region	name	level	pop	municipal_status	CMA_UID	CD_UID	PR_UID
	<chr>	<chr>	<chr>	<int>	<chr>	<chr>	<chr>	<chr>
1	3520005	Toronto	CSD	2794356	C	35535	3520	35
2	2466023	Montréal	CSD	1762949	V	24462	2466	24
3	4806016	Calgary	CSD	1306784	CY	48825	4806	48
4	3506008	Ottawa	CSD	1017449	CV	505	3506	35
5	4811061	Edmonton	CSD	1010899	CY	48835	4811	48
6	4611040	Winnipeg	CSD	749607	CY	46602	4611	46
7	5915022	Vancouver	CSD	662248	CY	59933	5915	59
8	5915004	Surrey	CSD	568322	CY	59933	5915	59
9	1209034	Halifax	CSD	439819	RGM	12205	1209	12
10	2465005	Laval	CSD	438366	V	24462	2465	24
11	4711066	Saskatoon	CSD	266141	CY	47725	4711	47
12	5915025	Burnaby	CSD	249125	CY	59933	5915	59
13	5915015	Richmond	CSD	209937	CY	59933	5915	59

This leaves us with 13 cities, and manual inspection shows that only Edmonton had a boundary change affecting population, resulting in a gain of 542 people 2016-2021. This should not make a noticeable difference for our analysis, so we will ignore this.

```
maintainer_data <- bind_rows(
  get_census("2011",regions=as_census_region_list(regions), vectors=vectors[["2011"]]) |>
    mutate(Year="2011"),
  get_census("2016",regions=as_census_region_list(regions), vectors=vectors[["2016"]]) |>
    mutate(Year="2016"),
  get_census("2021",regions=as_census_region_list(regions), vectors=vectors[["2021"]]) |>
    mutate(Year="2021")
) |>
  select(GeoUID,Year,base,One,Two,`Three+`) |>
  left_join(regions |> select(GeoUID=region,Name=name),by="GeoUID")
```

Getting the data for the three censuses and our selection of regions is easy, we join on the region names to make sure we have a uniform way to format and spell the names.

12.4. Data preparation

We are interested in shares rather than absolute numbers, for this we change the three different maintainer categories to **long form** and compute shares.

```
maintainer_levels <- c("One","Two","Three+")
plot_data <- maintainer_data |>
  pivot_longer(maintainer_levels, names_to="Maintainers",values_to = "Value") |>
  mutate(Share=Value/base)
```

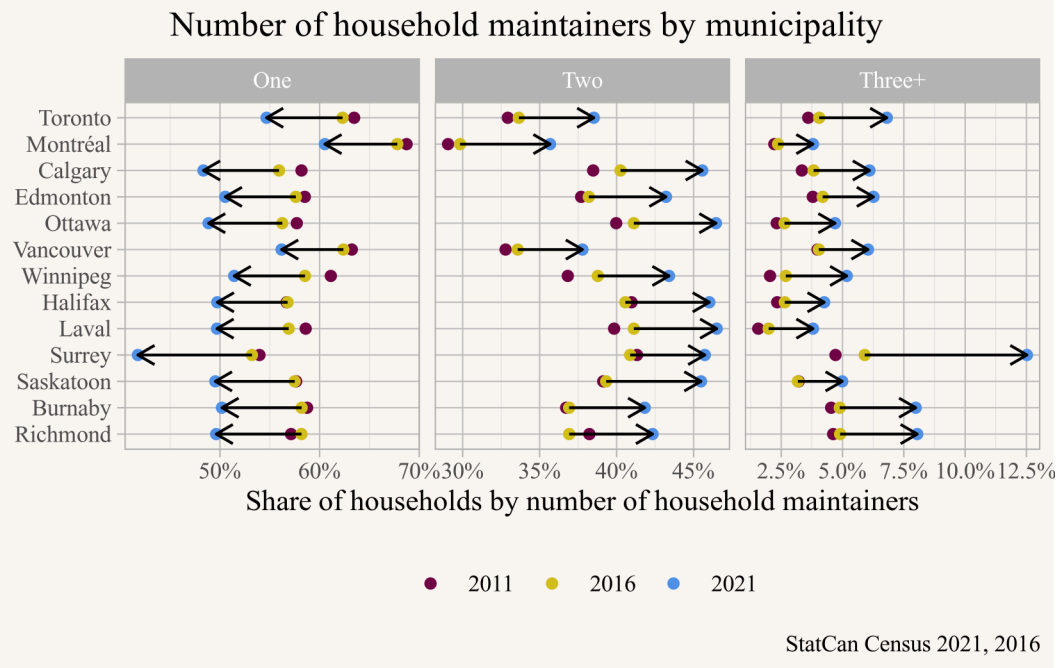
12.5. Analysis and visualization

To visualize the data it can be useful to be deliberate about the order of variables. In **R** we preferably use **factors** to deal with categorical data, and the factor levels can be set to fix their order in visualizations.

```
name_order <- plot_data |>
  filter(Year=="2021",Maintainers=="One") |>
  arrange(Value) |>
  pull(Name)

plot_data <- plot_data |>
  mutate(Name=factor(Name,name_order)) |>
  mutate(Maintainers=factor(Maintainers, levels = maintainer_levels))
```

With that in place we can plot the data. We are in particular interested in the change from 2016 to 2021, so we add in an arrow to emphasize this.

```
plot_data |>
  select(Name,Year,Maintainers,Share) |>
ggplot(aes(y=Name, x=Share)) +
  geom_point(aes(colour=Year)) +
  scale_colour_manual(values=sanzo::trios$c157)  +
  theme(legend.position = "bottom") +
  facet_wrap(~Maintainers,scales="free_x") +
  geom_segment(data=~pivot_wider(.,names_from = Year,values_from = Share),
               aes(x=`2016`,xend=`2021`,yend=Name),
               arrow = arrow(length = unit(0.1,"inches"))) +
  scale_x_continuous(labels=scales::percent) +
  labs(title="Number of household maintainers by municipality",
       y=NULL,x="Share of households by number of household maintainers",
       colour=NULL,
       caption="StatCan Census 2021, 2016")
```



To add the segments we took the plot data and pivoted it wider by **Year**, allowing us to select the start and endpoints for our arrows indicating the movement 2016 to 2021.

Looking at the graph for 2011 to 2016 we notice a general decrease in the share of one-person maintainer households, coupled with an increase in both two and three-person maintainer households. But this trend is not uniform and is at least partially reversed for some cities, for example Saskatoon or Halifax. But for 2016 to 2021 the trends are very large and uniform. To the extend that one gets suspicious, demographic changes usually happen gradually and don't show strong across the board changes like this. Either something big has happened, or we have some data issues.

At this point we should go back a step and try to understand what is going on here.

12.6. Analysis (revisited)

Let's try and understand what processes could be driving differences or changes in the number of household maintainers. Household size is a large factor, one person households can have at most one household maintainer. Similarly, a two-person household can have at most two household maintainers. We can try to filter some of this effect out by looking only at two or more person households, that might give a clearer picture of what is going on.

12.7. Data acquisition (revisited)

For this we need data on the number of one-person households for the three years. The process is easily adapted from the code above.

```
vectors2 <- list(
  "2011"=c(`One person households`="v_CA11F_210"),
  "2016"=c(`One person households`="v_CA16_419"),
  "2021"=c(`One person households`="v_CA21_444")
)

household_size_data <- bind_rows(
  get_census("CA11",regions=as_census_region_list(regions), vectors=vectors2[["2011"]]) |>
    mutate(Year="2011"),
  get_census("CA16",regions=as_census_region_list(regions), vectors=vectors2[["2016"]]) |>
    mutate(Year="2016"),
  get_census("CA11",regions=as_census_region_list(regions), vectors=vectors2[["2021"]]) |>
    mutate(Year="2021")
) |>
  select(GeoUID,Year,`One person households`)
```

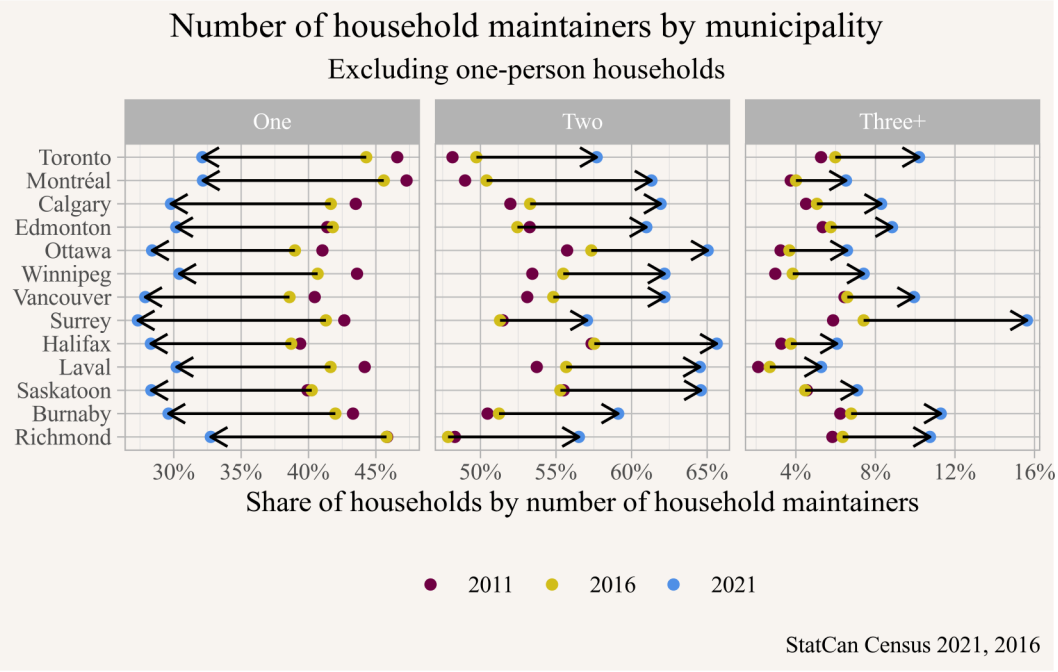
Now we need to combine the data on household maintainers with the data on one-person households, which we do by joining the data frames by geographic identifier and year. Then we subtract out one-person households from the denominator and the numerator of single maintainer households, and proceed as before.

```
plot_data2 <- maintainer_data |>
  full_join(household_size_data,by=c("GeoUID","Year")) |>
  mutate(base=base-`One person households`,
         `One`=`One`-`One person households`) |>
  pivot_longer(maintainer_levels, names_to="Maintainers",values_to = "Value") |>
  mutate(Share=Value/base) %>%
  mutate(Name=factor(Name,filter(.,Year=="2021",Maintainers=="One") |>
                       arrange(Value) |>
                       pull(Name))) |>
  mutate(Maintainers=factor(Maintainers, levels = maintainer_levels))
```

12.8. Visualization (revisited)

For visualization we copy the code from before and add in a subtitle to indicate that we excluded one-person households.

```
plot_data2 |>
  select(Name,Year,Maintainers,Share) |>
ggplot(aes(y=Name, x=Share)) +
  geom_point(aes(colour=Year)) +
  scale_colour_manual(values=sanzo::trios$c157)  +
  theme(legend.position = "bottom") +
  facet_wrap(~Maintainers,scales="free_x") +
  geom_segment(data=~pivot_wider(.,names_from = Year,values_from = Share),
               aes(x=`2016`,xend=`2021`,yend=Name),
               arrow = arrow(length = unit(0.1,"inches"))) +
  scale_x_continuous(labels=scales::percent) +
  labs(title="Number of household maintainers by municipality",
       subtitle="Excluding one-person households",
       y=NULL,x="Share of households by number of household maintainers",
       colour=NULL,
       caption="StatCan Census 2021, 2016")
```



Dropping one-person households does remove some of the variation between cities, but it amplifies the effect of the drop in single maintainer households 2016-2021, while the change for 2011-2016 is still ambiguous, although in most cases also dropping. It's hard to imagine what processes could cause such large change across all these cities. Shifting demographics, like Millennials aging into family formation years and coupling up could have some effect, but not of this magnitude. Changes in housing affordability, and people coupling or tripling up to pay for housing could also cause some shift, but these kind of shifts happen more gradually and won't affect such a large share of households.

This is a bit of a puzzle, time to dig a little deeper at the data sources.

12.9. Data sources (revisited)

We have looked through the census dictionary and found no indication that the concept of the number of household maintainers changed over our time period, with a minor caveat that 2011 data is from the voluntary National Household Survey instead of the mandatory long form census.

But the Housing Characteristics Reference Guide shows that StatCan flagged this issue of lack of ability to compare this concept over time. They write:

> In 2021, the household maintainer question was asked for every member of the household aged 15 years or older. Previously, in 2016, the question had a mark-all-that-apply format for the first five persons listed on the paper questionnaire. This alteration to the paper questionnaire brings better visual resemblance between the electronic and paper questionnaires. The wording was also modified to include a qualifying statement of "partly or entirely" when referring to the payments. The result of these changes is the capture of more household maintainers who are not the primary household maintainer.
>
> The growth rate of the number of primary household maintainers since the 2016 Census of Population was 6.4%, the same as the growth rate of private occupied dwellings, because every private occupied dwelling has one primary maintainer. At the same time, the growth rate of other household maintainers over the same period was 34.1%, indicating more comprehensive coverage of all household maintainers. This has not affected the characteristics of the primary household maintainer, which are often used to derive statistics such as the homeownership rates of different generations. However, the more complete coverage of other maintainers will allow for future analysis of the characteristics of these other household members contributing to housing payments.

We can track this further by looking at the 2011 NHS and 2016 census questionnaires, where the information on the number of household maintainers comes from the first question on the dwelling section, Question E1 and F1, respectively.

In the 2011 NHS the question reads:

> **E1** Who pays the rent or mortgage, taxes, electricity, etc., for this dwelling?
>
> - 1: Person 1
> - 2: Person 2
> - 3: Person 3
> - 4: Person 4
> - 5: Person 5
> - 6: A person who is listed on another questionnaire for this dwelling
> - 7: A person who does not live here

In the 2016 census the question reads identical, except it received an additional instruction on how to answer if more than one person contributes.

> F1. Who pays the rent or mortgage, taxes, electricity, etc., for this dwelling?
>
> If more than one person contributes to such payments, mark as many circles as apply.
>
> - 1: Person 1
> - 2: Person 2
> - 3: Person 3
> - 4: Person 4
> - 5: Person 5
> - 6: A person who is listed on another questionnaire for this dwelling
> - 7: A person who does not live here

This change in instruction could impact how people answer this question, plausibly increasing the number of people listing multiple people as household maintainers. This taints the overall drop in single household maintainers that we observed in the data 2011-2016.

For the 2021 census the question on household maintainers has been removed from the dwelling section and added to the section that is to be separately filled out for every person in the household. It is not question 58 of part D, which reads.

> **58.** Does this person pay, partly or entirely, the rent or mortgage, taxes, electricity, etc. for this dwelling?
>
> Mark "Yes" if this person pays the rent or mortgage, taxes, electricity, etc. for this dwelling, even if more than one person contributes to such payments.
>
> A **dwelling** is a separate set of living quarters with a **private entrance** from the outside or from a common hallway or stairway inside the building. This entrance should not be through someone else's living quarters.
>
> **Do not consider** payments for **other dwellings** such as the school residence of a child, the residence of a former spouse, or another dwelling that you may own or rent.
>
> - Yes
> - No

12.10. Interpretation

The number of household maintainer variable is not comparable across the 2016 to 2021 censuses, and may also be somewhat tainted for comparisons 2011 to 2016. We observe large changes in the shares of single and multiple household maintainer households 2016 to 2021 that are very likely dominated by changes to the census questionnaire.

Comparisons across regions for fixed years can still be informative, with data prior to 2021 likely being tainted by people misreading the question and selecting only one household maintainer where they should have selected more than one. Composition of households by household size also matter, removing one-person household when computing shares can remove some of the bias introduced by some municipalities having a significantly higher share of one-person households than others. This effect is particularly strong when comparing Vancouver and Surrey.

It would be worthwhile to look deeper into what drives three or more household maintainer households, either using cross tabulations or PUMF data. Surrey's high share of

multigenerational households is a likely contributor to Surrey's high share of three or more household maintainer households.

This exercise serves as a good reminder to be suspicious of implausibly large effects in data. Most probably these arise as a result of errors in the data analysis process, but may also come about due to changes in definitions or in the data generation process, in this case the questionnaire.

13. Land values

The Globe and Mail reported that

> "between 2006 and 2022, Vancouver building values stayed the same while the land value increased by more than 500 per cent".

To anyone familiar with Vancouver during this time frame the claim that "building values stayed the same" seems questionable. This brings us to our question.

13.1. Question

How have land and building values in Vancouver changed since 2005 (2006 tax assessment year)?

13.2. Data sources

Land and building values are assessed separately by BC Assessment, and we will piggy-back of their estimates instead of trying to estimate them ourselves. The City of Vancouver makes assessment data for the City available on their Open Data Portal.

13.3. Data acquisition

The VanouvR package ("VancouvR: Access the 'City of Vancouver' Open Data API" 2019) makes it easy to access this data in R and one of the package vignettes has code that does pretty much what we need.

```
library(tidyverse)
library(VancouvR)
```

The datasets in question are the `property-tax-report`, due to size the data is split over several datasets.

```
search_cov_datasets("property-tax-report") |>
  select(dataset_id,title)
```

```
# A tibble: 4 x 2
  dataset_id                    title
  <chr>                         <chr>
1 property-tax-report-2016-2019 Property tax report 2016-2019
2 property-tax-report-2011-2015 Property tax report 2011-2015
3 property-tax-report           Property tax report
4 property-tax-report-2006-2010 Property tax report 2006-2010
```

The first tax assessment year in the dataset is for 2006, the last one is for the current year, 2024 as of the writing of this. Assessments are pegged to July 1st of the previous year, so we have data for all years from July 2005 through 2023. This is likely the same data source that news article used, except that the article did not adjust to the date the assessments are pegged to.

For our purposes all we need is aggregates for each year, the Open Data Portal allows server side aggregation of data and the R package supports that. This cuts down on time and the amount of data we need to transfer. We simply group by tax assessment year and aggregate up the assessed land and building values for each year.

```
land_building_data_raw <-search_cov_datasets("property-tax-report") |>
    pull(dataset_id) |>
    map_df(function(ds) aggregate_cov_data(
      ds,
      group_by="tax_assessment_year as Year",
      select="sum(current_land_value) as Land, sum(current_improvement_value) as Building")
  arrange(Year)
```

This gives us a simple data frame with land and building values for each year. We check on the tax years in question, as well as the most recent one.

```
land_building_data_raw |>
  filter(Year %in% c(2006,2022,max(Year))) |>
  tinytable::tt()
```

Year	Land	Building
2006	88649668277	35086836003
2022	394201892132	102591491465
2024	415679362031	105858810250

13.4. Data preparation

There is not much to do here, we remember that assessments are pegged to July 1st in the previous year and reshape the data into long form.

```
land_building_data <- land_building_data_raw |>
    mutate(Date=as.Date(paste0(as.integer(Year)-1,"-07-01"))) |>
    pivot_longer(c("Land","Building"),names_to = "Component")
```

13.5. Analysis and visualization

Let's take a quick look what that data looks like.

```
ggplot(land_building_data,aes(x=Date,y=value,colour=Component)) +
  geom_line() +
  scale_y_continuous(labels=\(x)scales::dollar(x,scale=10^-9,suffix="bn")) +
  expand_limits(y=0) +
  labs(title="City of Vancouver assessed land and building values",
       y="Assessed value",
       x=NULL,
       caption="CoV Open Data")
```

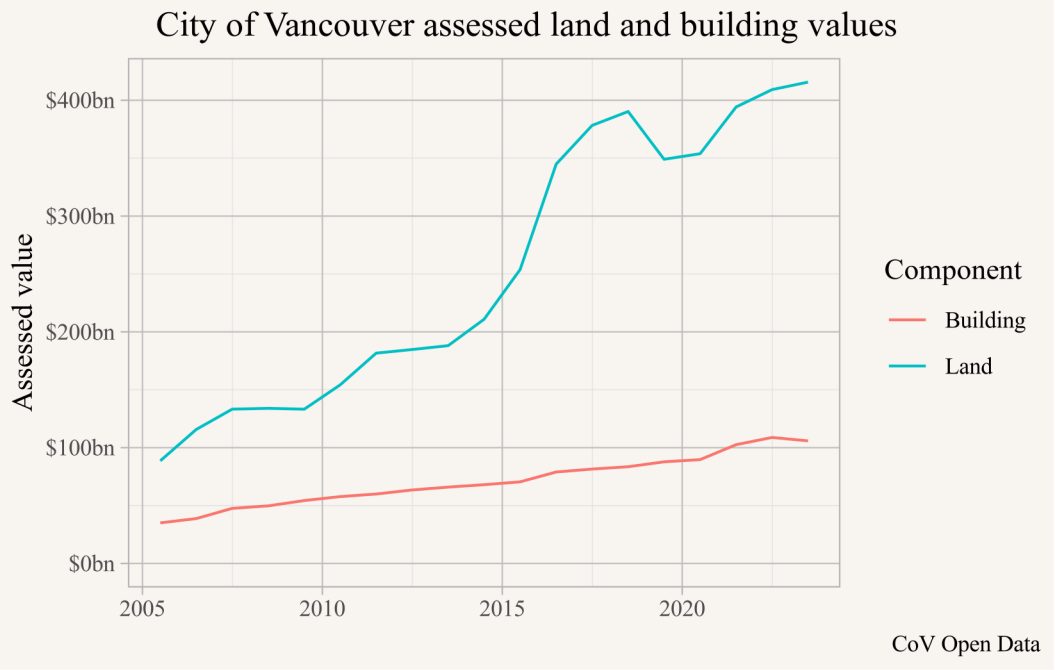


Figure 13.1.

So far so good, but we should probably account for inflation. We borrow code from the section on income change to pull CPI data and fold it in.

```
library(cansim)
inflation <- get_cansim_vector("v41693271") |>
  mutate(Date=Date %m+% months(6)) |>
  select(Date,CPI=val_norm) |>
  filter(Date %in% land_building_data$Date) |>
  mutate(CPI=CPI/last(CPI,order_by = Date))

land_building_data |>
  left_join(inflation,by="Date") |>
ggplot(aes(x=Date,y=value/CPI,colour=Component)) +
  geom_line() +
  scale_y_continuous(labels=\(x)scales::dollar(x,scale=10^-9,suffix="bn")) +
  expand_limits(y=0) +
  labs(title="City of Vancouver assessed land and building values",
       y="Assessed value (July 2023 dollars)",
```

```
       x=NULL,
       caption="CoV Open Data")
```

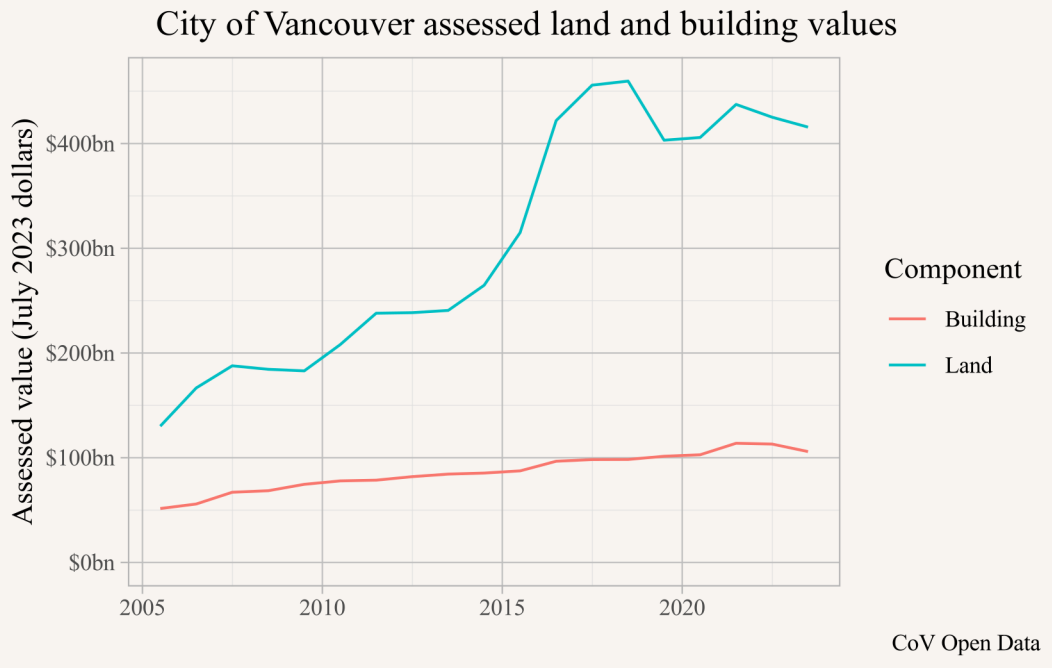


Figure 13.2.

As we might have expected, values rose faster than inflation, but they did so for buildings as well as for land. The land value change is impressive, but it's hard to judge that against the building value change, which started at a much lower value. The article looked at percentage change, so let's do the same.

```
plot_data <- land_building_data |>
  left_join(inflation,by="Date") |>
  mutate(real_value=value/CPI) |>
  mutate(real_ratio = real_value/first(real_value,order_by=Date),
         ratio = value/first(value,order_by=Date),
         .by=Component)

ggplot(plot_data,aes(x=Date,y=real_ratio,colour=Component)) +
  geom_line() +
  scale_y_continuous(labels=scales::percent,trans="log",breaks=seq(1,5)) +
```

```
labs(title="City of Vancouver assessed land and building values",
     y="Real change since July 1, 2005",
     x=NULL,
     caption="CoV Open Data")
```

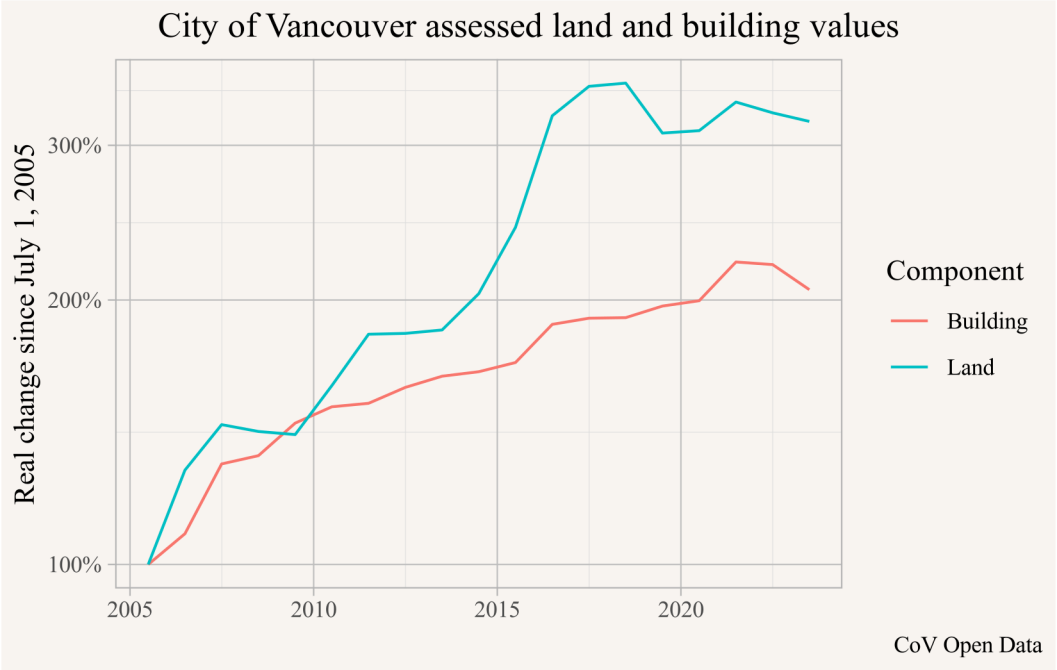


Figure 13.3.

Since this is ratio data we chose a logarithmic scale on the y-axis. This shows that between 2005 and 2021 (so using assessment years 2006 and 2022) real land values increased by 336% and building values by 221% . Maybe the article was using nominal value increases, in nominal terms land increased by 445% and building values by 292%.

13.6. Interpretation

The increase in land values is lower than the "more than 500 per cent" claimed in the article, and the claim that "building values stayed the same" is clearly false.

What is clear is that land values have risen faster than building values, likely in large part because restrictive zoning has prevented buildings from making adequate use of the land they are on.

It could be that the article mis-quoted its sources and the claim was about a sub-set of Vancouver properties, maybe just residential properties, or just single-family properties. We make a rather crude estimate by filtering the data on RS-1 and R1-1 zoning districts. This will under-estimate the growth a bit as properties that got rezoned within this timeframe will be included in the earlier years but not in the later ones.

```
search_cov_datasets("property-tax-report") |>
    pull(dataset_id) |>
    map_df(function(ds) aggregate_cov_data(
      ds,
      group_by="tax_assessment_year as Year",
      where="zoning_district like 'RS-' or zoning_district like 'R1-1'",
      select="sum(current_land_value) as Land, sum(current_improvement_value) as Building")
    mutate(Date=as.Date(paste0(as.integer(Year)-1,"-07-01"))) |>
    pivot_longer(c("Land","Building"),names_to = "Component") |>
  left_join(inflation,by="Date") |>
  mutate(real_value=value/CPI) |>
  mutate(real_ratio = real_value/first(real_value,order_by=Date),
         ratio = value/first(value,order_by=Date),
         .by=Component) |>
ggplot(aes(x=Date,y=real_ratio,colour=Component)) +
  geom_line() +
  scale_y_continuous(labels=scales::percent,trans="log",breaks=seq(1,5)) +
  expand_limits(y=0) +
  labs(title="City of Vancouver assessed land and building values in RS/R1-1 zones",
       y="Real change since July 1, 2005",
       x=NULL,
       caption="CoV Open Data")
```

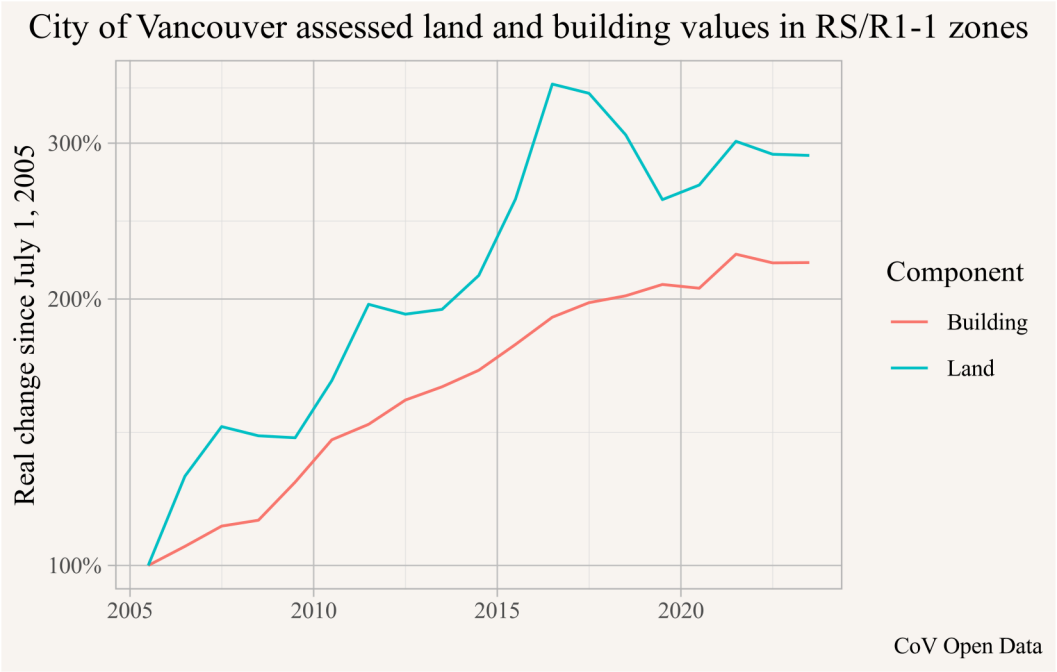


Figure 13.4.

Again, the claim that building values stayed the same has no basis in reality. Readers interested in more detail are encouraged to use individual property data and match individual lots over time to further refine these estimates.

Part III.

Advanced descriptive analysis

Building on the section of basic descriptive analysis we will move into more advanced data processing and descriptive analysis. This will involve mixing of different datasets to tease out finer aspects. We will learn how to group and summarize data, and how to use joins.

14. BC migration

This example is motivated by a BC government press release titled "**B.C. welcomes more than 100,000 people – the most in 60 years**". This is the type of attention-grabbing headline where our gut reaction usually is to question if this is true.

Let's first try and understand what the headline really means. B.C. "welcoming" people refers to people moving to the province from elsewhere, either from other provinces or internationally. So this is referring to gross in-migration. But reading the text of the press release it immediately pivots to a different concept, saying that "B.C.'s net migration reached 100,797 people in 2021". It helpfully explains that net migration is the difference between people moving here and people moving away. Which is quite different from the number of people B.C. "welcomed" that year, or the number of people "moving to the province in 2021" as implied by the title and the first sentence of the press release.

So here comes the first difficulty, the press release is contradicting itself by mixing two concepts. That leads us to formulate a fairly broad question that should help clear this up.

14.1. Question

How many people has B.C. welcomed, net and gross, how has that changed over the last 6 decades, and how should this be interpreted?

14.2. Data sources

To start, let's figure out where that data point comes from.

The press release references StatCan as the source, let's search through the StatCan tables. Google usually works reasonably well, but we can also search programmatically. We are looking for migration estimates from the quarterly demographic estimates to get the most up-to-data population estimates from StatCan. For results we just need the first two columns, that table number and the title.

```
library(tidyverse)
library(cansim)

search_cansim_cubes("migration") |>
  filter(grepl("quarterly",cubeTitleEn)) |>
  arrange(desc(cubeEndDate)) |>
  select(1:2)
```

```
# A tibble: 2 x 2
  cansim_table_number cubeTitleEn
  <chr>               <chr>
1 17-10-0020          Estimates of the components of interprovincial migration,~
2 17-10-0040          Estimates of the components of international migration, q~
```

It looks like Table 17-10-0020 and 17-10-0040 are what we are looking for. Let's load in the data and inspect the first couple of rows for BC.

14.3. Data acquisition

```
interprovincial <- get_cansim("17-10-0020")
international <- get_cansim("17-10-0040")

interprovincial |>
  filter(GEO=="British Columbia") |>
  select(GEO,Date,`Interprovincial migration`,val_norm) |>
  tail()
```

```
# A tibble: 6 x 4
  GEO              Date       `Interprovincial migration` val_norm
  <chr>            <date>     <fct>                         <dbl>
1 British Columbia 2023-04-01 In-migrants                   22371
2 British Columbia 2023-04-01 Out-migrants                  22671
3 British Columbia 2023-07-01 In-migrants                   12552
4 British Columbia 2023-07-01 Out-migrants                  17186
5 British Columbia 2023-10-01 In-migrants                    7892
6 British Columbia 2023-10-01 Out-migrants                  10620
```

For inter-provincial migration we get in and out migration counts for every quarter. Let's also inspect the international migration data.

```
international |>
  filter(GEO=="British Columbia") |>
  select(GEO,Date,`Components of population growth`,val_norm) |>
  tail()
```

```
# A tibble: 6 x 4
  GEO              Date       `Components of population growth` val_norm
  <chr>            <date>     <fct>                                <dbl>
1 British Columbia 2023-10-01 Net emigration                        2890
2 British Columbia 2023-10-01 Emigrants                             4853
3 British Columbia 2023-10-01 Returning emigrants                   1963
4 British Columbia 2023-10-01 Net non-permanent residents          20516
5 British Columbia 2023-10-01 Non-permanent residents, inflows     54008
6 British Columbia 2023-10-01 Non-permanent residents, outflows    33492
```

Here we get immigrants, emigrants, returning emigrants, but for temporary emigrants and non-permanent residents we only get net change. That puts a bit of a damper on our ambition to look at gross migration, for those last two categories net is all we have.

14.4. Data preparation

Next we got to wrangle this data into a useful format. We are interested in all of these components, so we need to join these two data series together. We will retain the `GeoUID`, `GEO`, `Components of population growth`, `Date` and `val_norm` columns, which requires some renaming and then defining factor levels so that they stack nicely later in our plots. We also flip the sign on out-migrants and emigrants, as these are out-flows. To make sure those two time series start at the same time we cut it off appropriately.

The press release talked about annual change, so we do a rolling sum over 4 quarters, right-aligning the data so it's for the period of the preceding year.

```
migration_data <- bind_rows(
  interprovincial |>
    select(GeoUID,GEO,Date,
           `Components of population growth`=`Interprovincial migration`,val_norm) |>
    mutate(`Components of population growth`=
```

```
            paste0("Interprovincial ",tolower(`Components of population growth`))),
  international |>
    select(GeoUID,GEO,Date,`Components of population growth`,val_norm)
) |>
  mutate(`Components of population growth`=
           factor(`Components of population growth`,
                  levels=c("Interprovincial out-migrants",
                           "Emigrants",
                           "Interprovincial in-migrants",
                           "Immigrants",
                           "Returning emigrants",
                           "Net temporary emigrants",
                           "Net non-permanent residents"))) |>
  mutate(value=ifelse(`Components of population growth` %in%
                        c("Interprovincial out-migrants","Emigrants"),
                      -val_norm,val_norm)) |>
  filter(Date>=pmax(min(interprovincial$Date),min(international$Date))) |>
  group_by(GeoUID,`Components of population growth`) |>
  arrange(Date) |>
  mutate(annual=zoo::rollsum(value,k=4,na.pad = TRUE,align = "right")) |>
  filter(!is.na(annual)) |>
  ungroup()
```

We will also need net migration stats, so let's compute these by summing of the components,

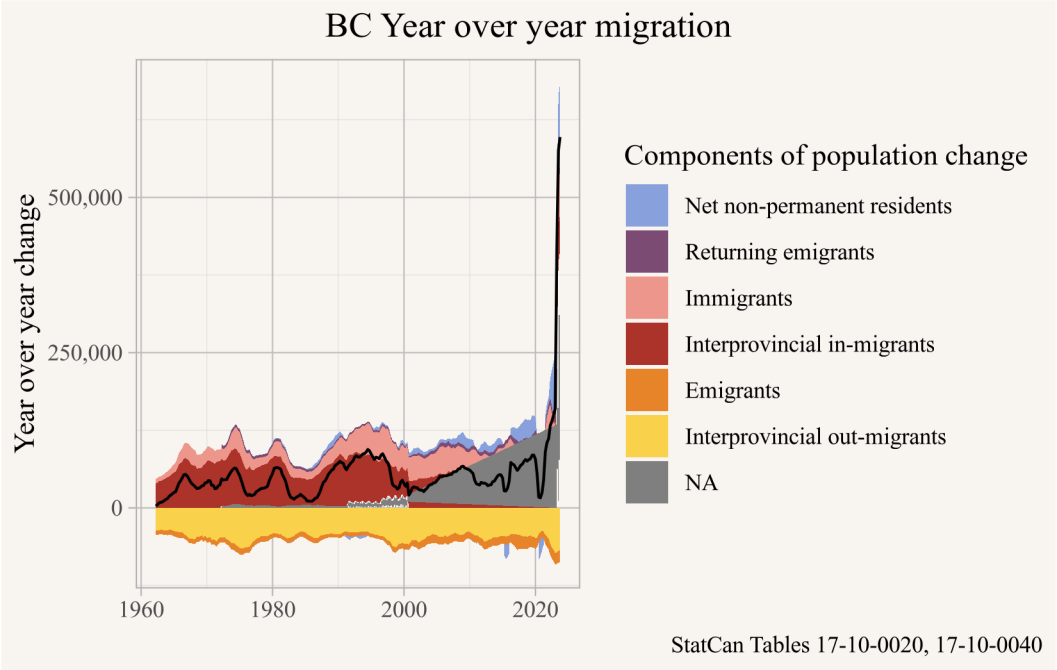
```
net_migration <- migration_data |>
  group_by(Date,GEO,GeoUID) |>
  summarize(value=sum(value),annual=sum(annual),.groups="drop") |>
  mutate(`Components of population growth`="Net migration")
```

14.5. Analysis and visualization

Time to make a graph.

```
migration_colours <- setNames(MetBrewer::met.brewer("Archambault",7),
                              migration_data$`Components of population growth` %>%
                                levels %>% rev)
```

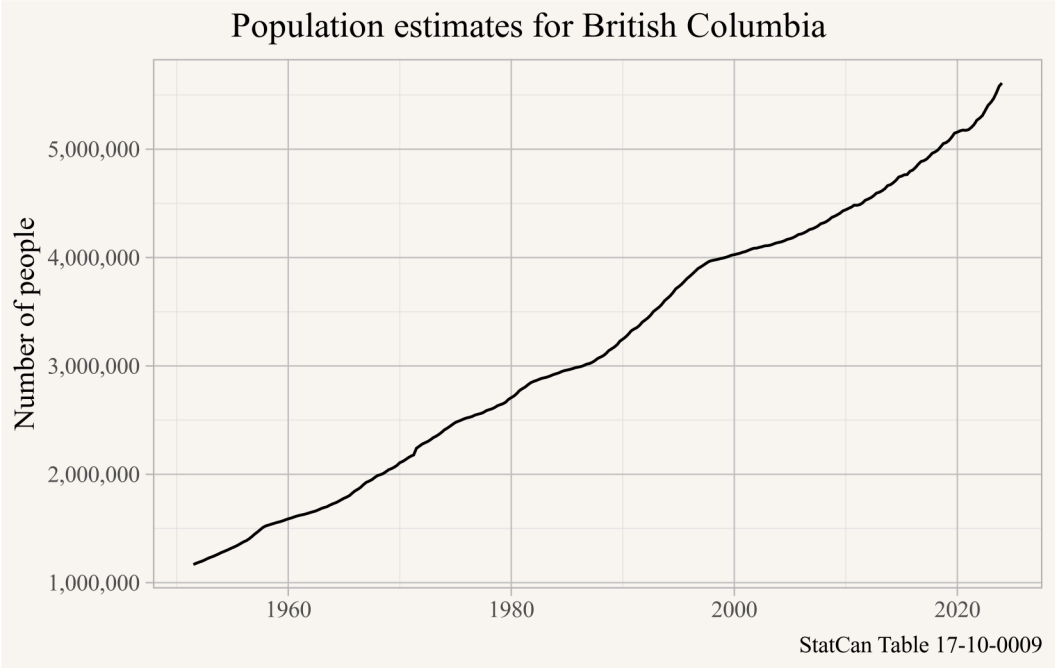
```
ggplot(migration_data |> filter(GEO=="British Columbia")) +
  geom_area(aes(x=Date,y=annual,fill=fct_rev(`Components of population growth`)),
            stat="identity") +
  scale_y_continuous(labels=scales::comma) +
  geom_line(data=net_migration |> filter(GEO=="British Columbia"),
            aes(x=Date,y=annual)) +
  scale_fill_manual(values=migration_colours) +
  labs(title="BC Year over year migration",
       y="Year over year change",x=NULL,fill="Components of population change",
       caption="StatCan Tables 17-10-0020, 17-10-0040")
```



This shows us that the press report did not mean to talk about number of people B.C. has "welcomed" or that "moved to the province" but instead the difference between the number of people it welcomed and the number of people it bid farewell.

And the net migration is indeed at record levels. At least in absolute terms. But B.C. now is very different from B.C. in the 60s at the start of this time series. How can we compare net migration over time in a more meaningful way? Normalizing by population is a good option here. Let's grab the data and take a look how B.C. population has changed.

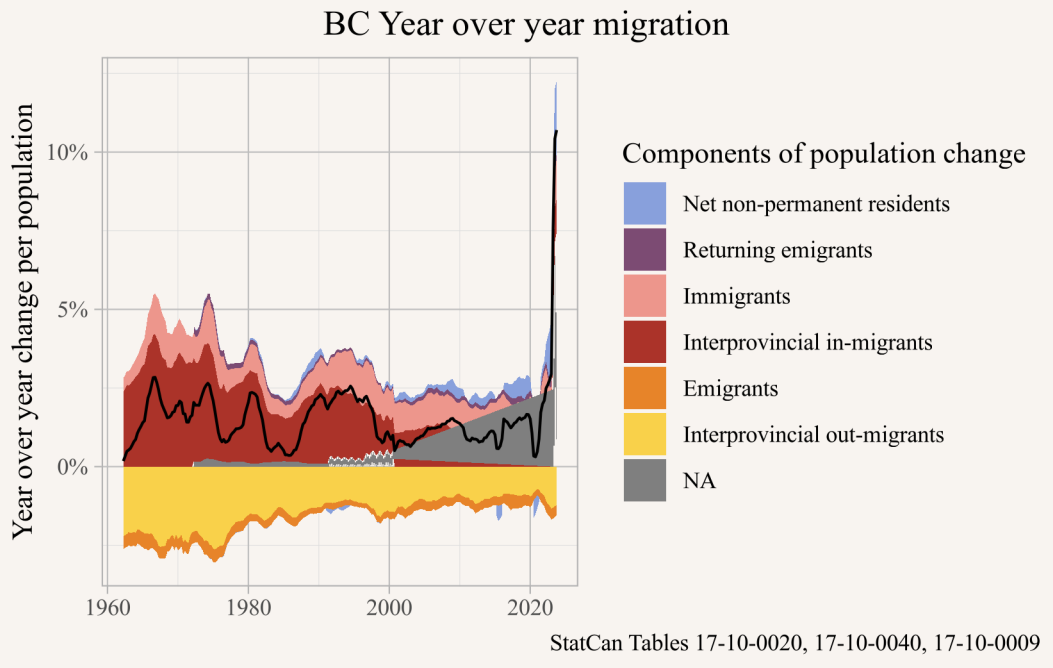
```
pop_data <- get_cansim("17-10-0009") |>
  select(GEO,Date,Population=val_norm)

pop_data |>
  filter(GEO=="British Columbia") |>
  ggplot(aes(x=Date,y=Population)) +
  geom_line() +
  scale_y_continuous(labels=scales::comma) +
  labs(title="Population estimates for British Columbia",
       y="Number of people",
       x=NULL,
       caption="StatCan Table 17-10-0009")
```



Indeed, the trend is quite strong. Let's fold that in and normalize by population.

```
migration_data |>
  left_join(pop_data, by=c("GEO","Date")) |>
  filter(GEO=="British Columbia") |>
  ggplot() +
  geom_area(aes(x=Date,y=annual/Population,fill=fct_rev(`Components of population growth`)),
            stat="identity") +
```

```
  scale_y_continuous(labels=scales::percent) +
  geom_line(data=net_migration |>
              left_join(pop_data, by=c("GEO","Date")) |>
              filter(GEO=="British Columbia"),
            aes(x=Date,y=annual/Population)) +
  scale_fill_manual(values=migration_colours) +
  labs(title="BC Year over year migration",
       y="Year over year change per population",x=NULL,
       fill="Components of population change",
       caption="StatCan Tables 17-10-0020, 17-10-0040, 17-10-0009")
```

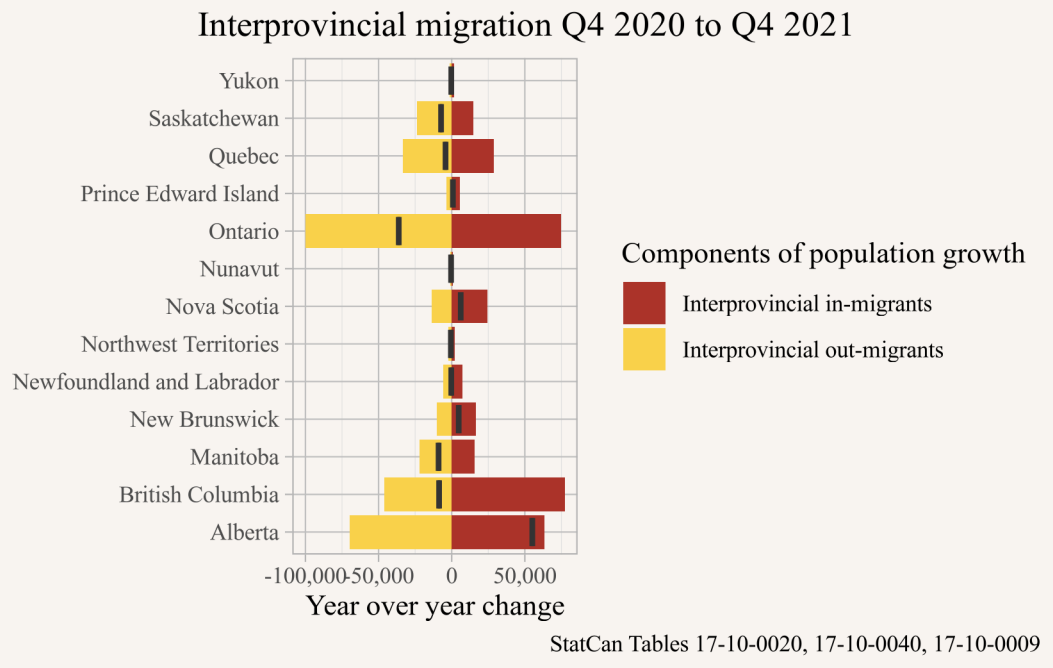


Here the picture looks a little different. Net migration per capita is at its highest since the 90s, but the past 60 years there were several periods where it was larger.

The press report also mentioned that B.C.'s interprovincial migration numbers are higher than any other province. This is easy to check now. In line with the press release that prompted our question we are pinning the data at the forth quarter of 2021.

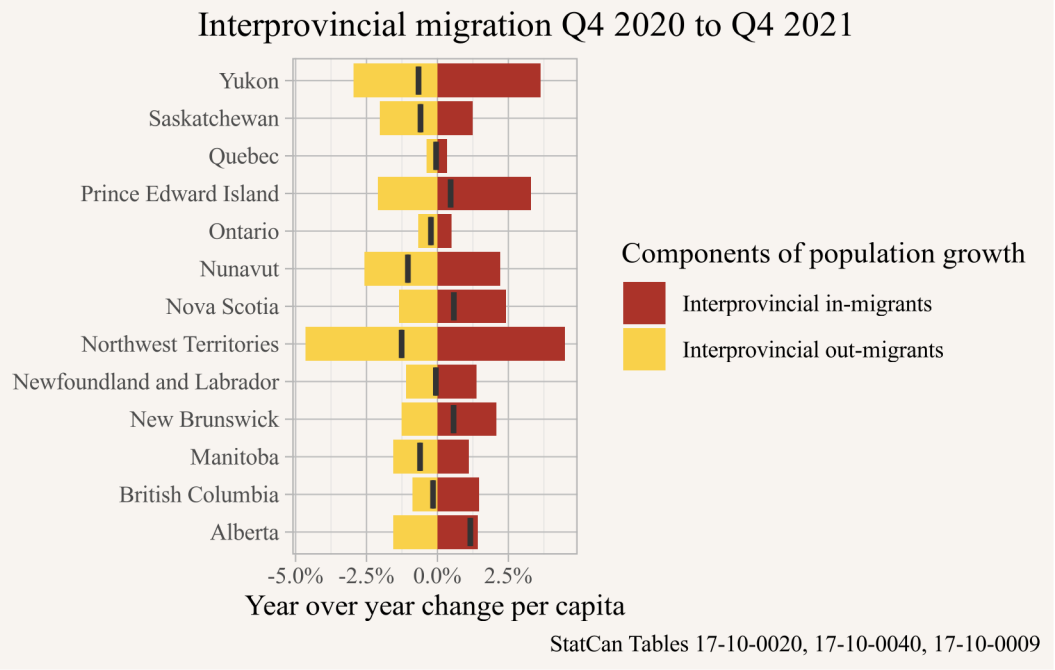
```
pinned_date <- as.Date("2021-10-01")
```

```
migration_data_interprovinicial <- migration_data |>
  left_join(pop_data, by=c("GEO","Date")) |>
  filter(grepl("Interprovincial",`Components of population growth`))

net_interprovincial <- migration_data_interprovinicial |>
  group_by(GEO,Date) |>
  summarize(value=sum(value),
            annual=sum(annual),
            Population=first(Population),
            .groups="drop")

migration_data_interprovinicial |>
  filter(GEO!="Canada") |>
  filter(Date==pinned_date) |>
  ggplot(aes(y=GEO,x=annual)) +
  geom_bar(stat="identity",
           aes(fill=fct_rev(`Components of population growth`))) +
  geom_boxplot(data=net_interprovincial |>
                 filter(GEO!="Canada") |>
                 filter(Date==max(Date))) +
  scale_fill_manual(values=migration_colours[grepl("Interprov",names(migration_colours))])
  scale_x_continuous(labels=scales::comma) +
  labs(title="Interprovincial migration Q4 2020 to Q4 2021",
       fill="Components of population growth",
       y=NULL,x="Year over year change",
       caption="StatCan Tables 17-10-0020, 17-10-0040, 17-10-0009")
```



In absolute number B.C. indeed has both the highest interprovincial in-migration and interprovincial net-migration among all provinces. But the provinces have vastly different sizes, so that's not really a fair comparison. Again, we can normalize by population.

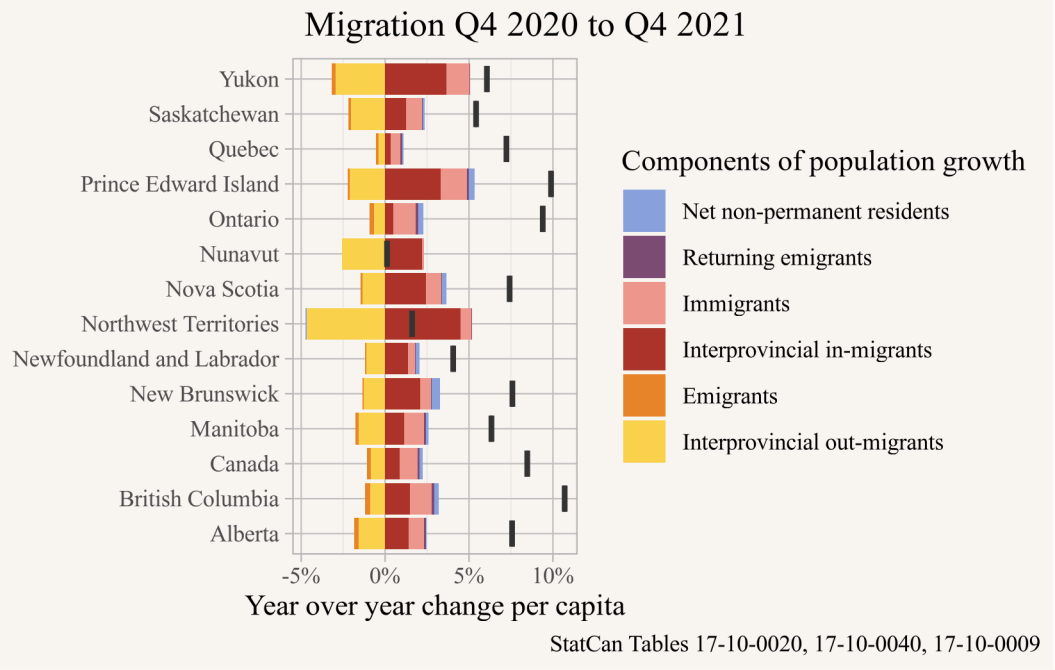
```
migration_data_interprovinicial |>
  filter(GEO!="Canada") |>
  filter(Date==pinned_date) |>
  ggplot(aes(y=GEO,x=annual/Population)) +
  geom_bar(stat="identity",
           aes(fill=fct_rev(`Components of population growth`))) +
  geom_boxplot(data=net_interprovincial |>
                 filter(GEO!="Canada") |>
                 filter(Date==max(Date))) +
  scale_fill_manual(values=migration_colours[grepl("Interprov",names(migration_colours))])
  scale_x_continuous(labels=scales::percent) +
  labs(title="Interprovincial migration Q4 2020 to Q4 2021",
       fill="Components of population growth",
       y=NULL,x="Year over year change per capita",
       caption="StatCan Tables 17-10-0020, 17-10-0040, 17-10-0009")
```



Viewed this way B.C.'s interprovincial in-migration and net migration still looks good, but many of the other provinces beat out that growth rate.

For completeness we can also just show the full graph that includes the international migration components.

```
migration_data |>
  left_join(pop_data, by=c("GEO","Date")) |>
  filter(Date==pinned_date) |>
  ggplot(aes(y=GEO,x=annual/Population)) +
  geom_bar(stat="identity",aes(fill=fct_rev(`Components of population growth`))) +
  geom_boxplot(data=net_migration |>
                 left_join(pop_data, by=c("GEO","Date")) |>
                 filter(Date==max(Date))) +
  scale_fill_manual(values=migration_colours) +
  scale_x_continuous(labels=scales::percent) +
  labs(title="Migration Q4 2020 to Q4 2021",
       fill="Components of population growth",
       y=NULL,x="Year over year change per capita",
       caption="StatCan Tables 17-10-0020, 17-10-0040, 17-10-0009")
```



14.6. Interpretation

This answers our question, the Q4 2021 year over year net migration edges over the 100,000 people mark, and in absolute terms this is the highest it's been over at least 60 years. (And it climbed even higher in the following quarters.) And B.C.'s interprovincial (gross) in-migration was the highest in Canada in absolute terms. But what can we learn from that?

B.C. 60 years ago is very different from B.C. today. To account for that we can normalize by population, and the relative net migration has been higher at several times during the past 60 years, most recently in the 90s.

Viewed relative to population size we note that other provinces, in particular some of the Atlantic Provinces, vastly outperform BC's interprovincial net as well as gross in-migration in that timeframe, mostly fuelled by the 50k people that left Ontario during that time period.

We also note the big dip in net-migration during COVID-19. It is not clear if the current heights are a bounce-back to make up for the comparatively low net in-migration during the pandemic, or if it is simply reverting back to the increasing trend we have seen over the past 10 years.

References

Lundberg, Ian, Rebecca Johnson, and Brandon M. Stewart. 2021. "What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory." *American Sociological Review* 86 (3): 532–65. https://doi.org/10.1177/00031224211004187.

"VancouvR: Access the 'City of Vancouver' Open Data API." 2019. The R Foundation. https://doi.org/10.32614/cran.package.vancouvr.

von Bergmann, Jens. 2021. *Tongfen: R Package to Make Data Based on Different Geographies Comparable*. https://mountainmath.github.io/tongfen/.

———. 2022. *Cmhc: R Package to Access, Retrieve, and Work with CMHC Data*. https://mountainmath.github.io/cmhc/.

von Bergmann, Jens, and Dmitry Shkolnik. 2021. *Cansim: Functions and Convenience Tools for Accessing Statistics Canada Data Tables*. https://mountainmath.github.io/cansim/.

von Bergmann, Jens, Dmitry Shkolnik, and Aaron Jacobs. 2022. *Cancensus: R Package to Access, Retrieve, and Work with Canadian Census Data and Geography*. https://mountainmath.github.io/cancensus/.