Introduction to cancensus

Jens von Bergmann

2022-04-01

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 we have in mind ranges from interested individuals interested in understanding their environment through data, community activists and community groups interested in introducing data-based approached into their work, journalists who want to report on data in their stories or aim to incorporate their own descriptive data analysis, non-profits or people involved in policy who are looking for data-based answers to their questions.

The most important prerequisite is a keen interest in using data to help understand how housing and demographics shape cities and rural areas in Canada, and a willingness to learn. 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. Moreover, many analyses that do come out go quickly out of date and can't easily be updated because they rely on non-reproducible and non-adaptable workflows.

In this book we will maintain a strong 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.

Project based approach

This book will take a project based approach to to teach through examples, with one project per section. Each project will be loosely broken up into four parts.

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.
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.
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 we 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, is 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 analyusis 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.

Goals

By taking this approach we have several goals in mind:

- Stay motivated by using real world Canada-focused and (hopefully) interesting examples.
- Teach basic data literacy, appreciate definitions and quirks in the data.
- Expose the world of Canadian data and make it more accessible.
- 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.

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 so bad, 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's not 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. And it also 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 also 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 will have no problem translating what we do into their preferred framework, or dynamically switch back and forth between R, python 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 piggy-back of each other's 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, ggplot2for graphing and **cancensus** for ingesting the data.

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

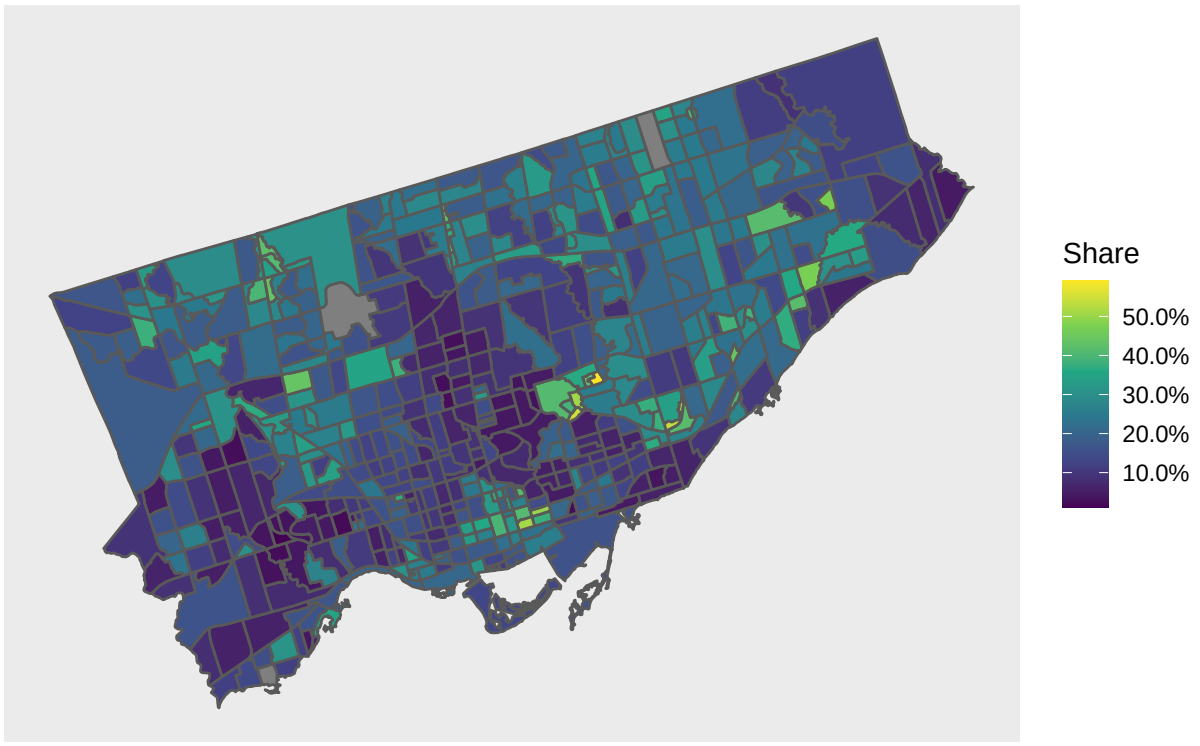
Next she pull in the data. the CensusMapper API GUI tool helps her locate the StatCan geographic identifier for Toronto, (3520005), and the internal CensusMapper vector for the percentage of children in LICO-AT (v_CA16_2573).

```
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 times 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")
```

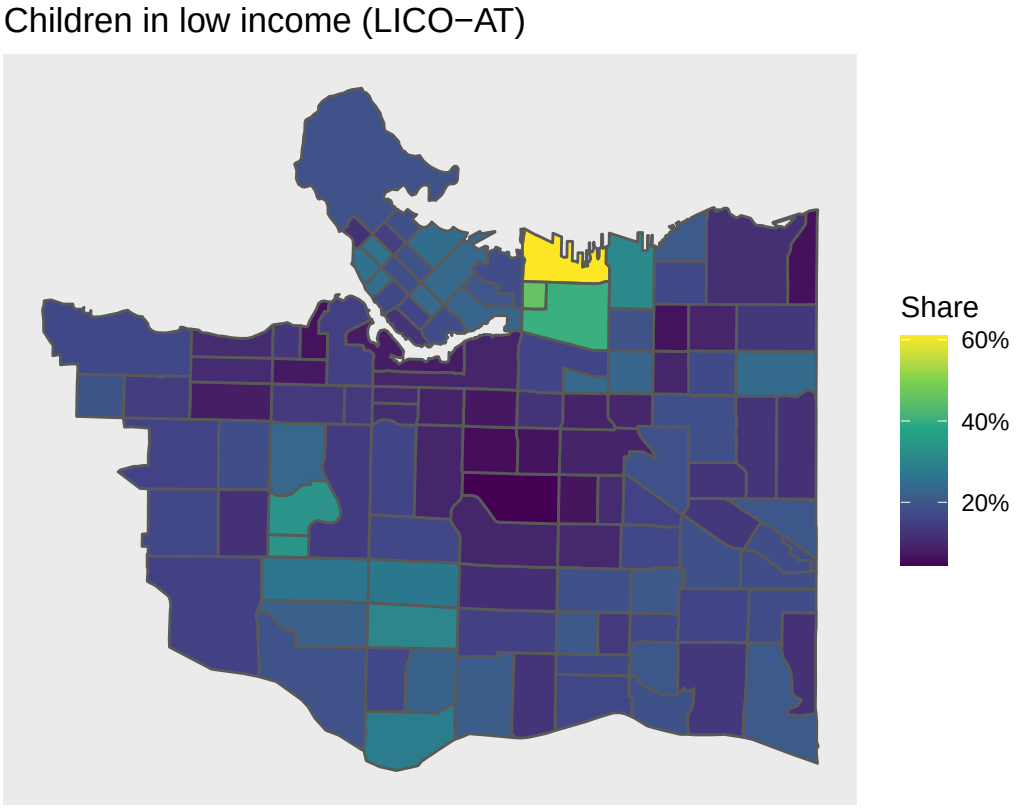


Based on this she locates a couple of good candidate neighbourhoods for her pilot and sends the map in a email to her colleague Peter to get input on which neighbourhood might be best suited.

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")
```

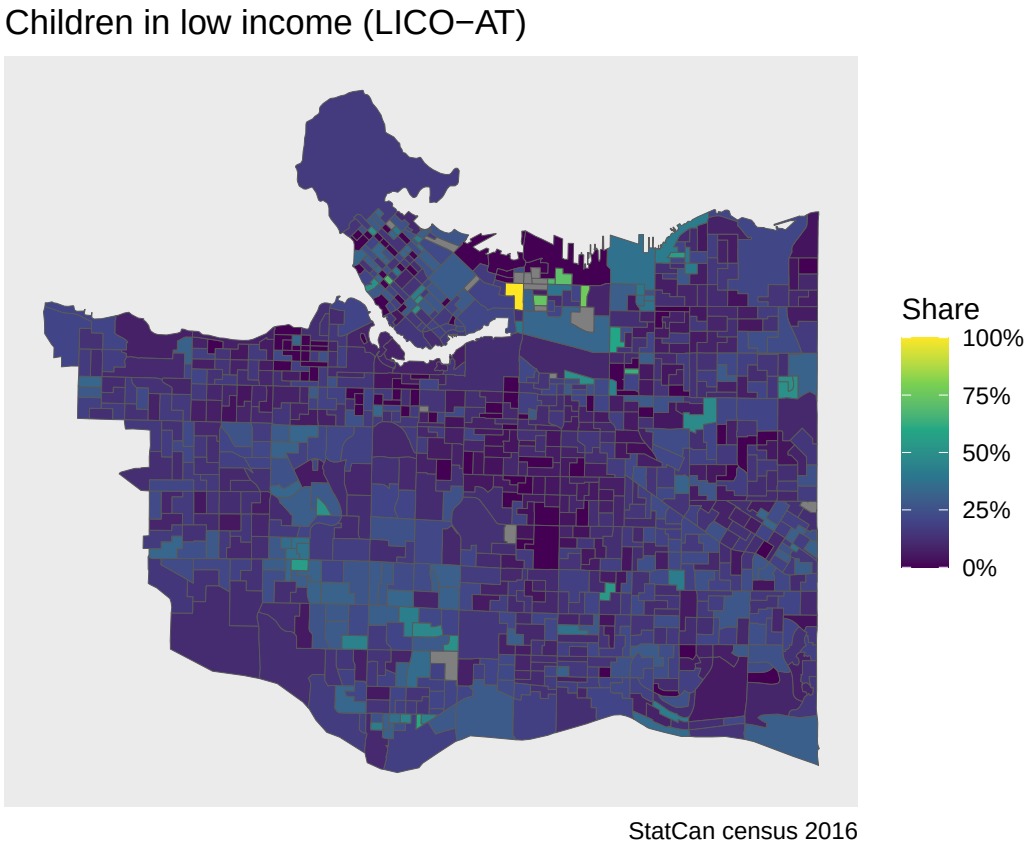


Easy peasy, thanks to Amy's previous work. Peter takes the map to his friend Yuko and asks her for advice where a community-based intervention for low-income children might make sense in Calgary. Yuko asks for the code from Peter to take a closer look herself.

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(size=0.1) +
  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")
```

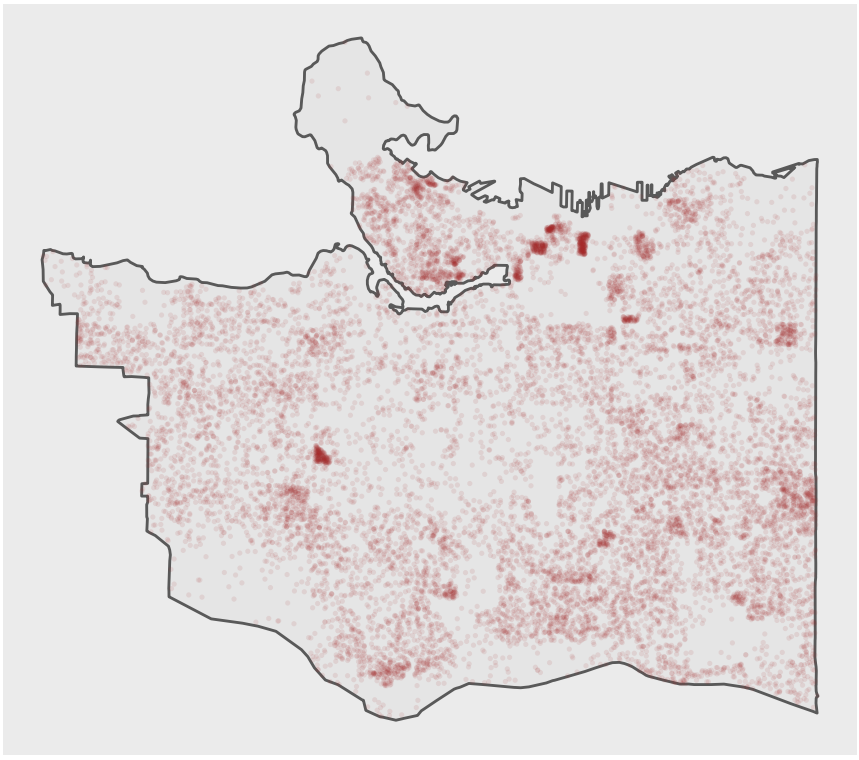


But then Yuko pauses to think that maybe looking at share of the low income population is not the right metric. She decides to query the number of children in low income (vector "v_CA16_2558") and prepare the data for a dot-density map.

```
library(dotdensity)
library(dplyr)

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(size=0.25,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.

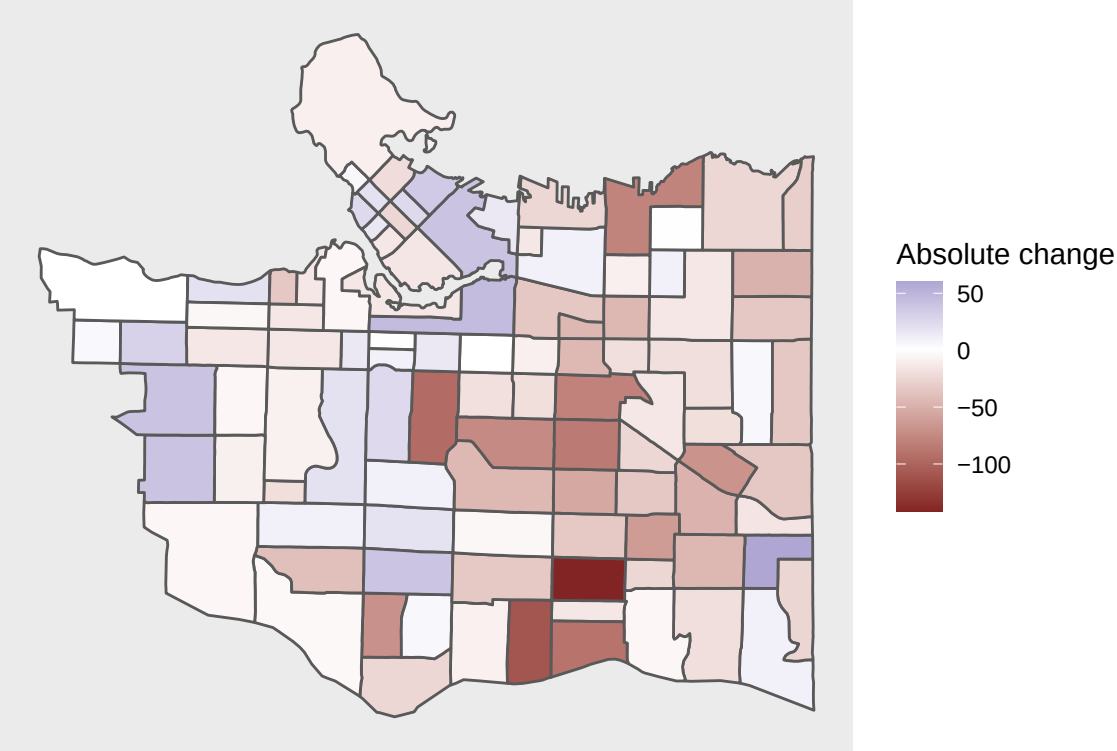
```
library(tongfen)
meta <- meta_for_ca_census_vectors(c(total_2006="v_CA06_1982",lico_share_2006="v_CA06_1984
                                     lico_2016="v_CA16_2561",lico_share_2016="v_CA16_2576"

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.

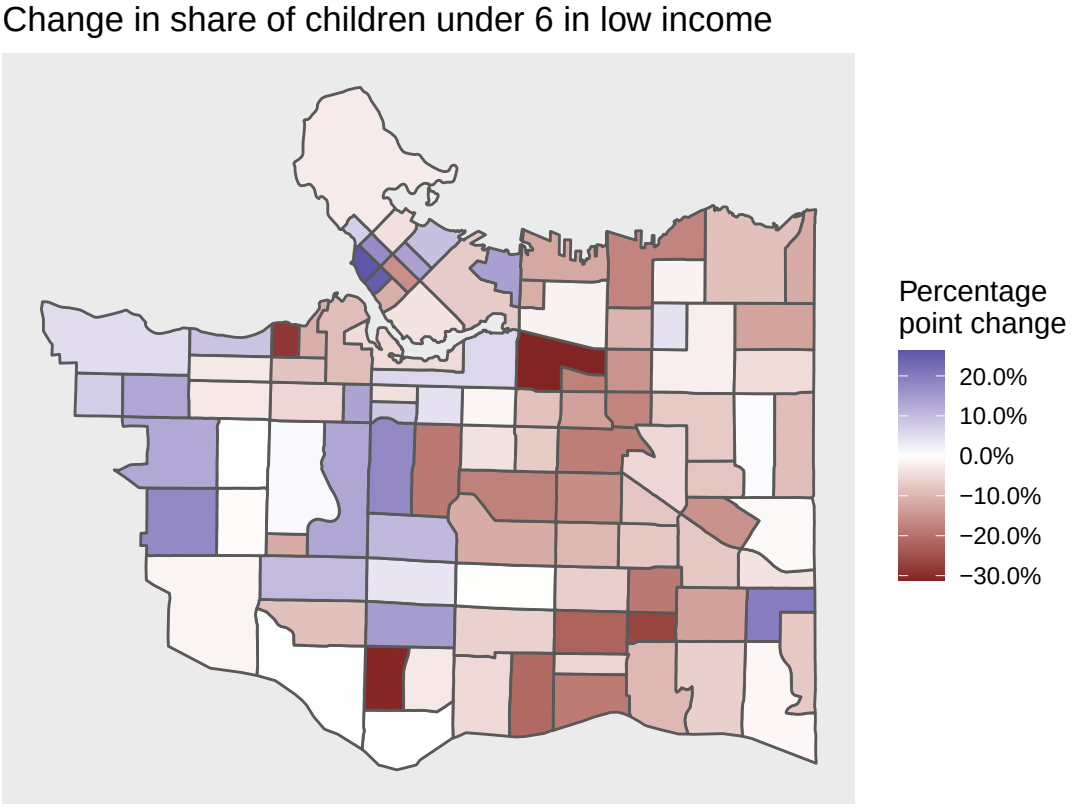
```
ggplot(lico_data,aes(fill=`Absolute change`)) +
  geom_sf() +
  scale_fill_gradient2() +
  coord_sf(datum=NA) +
  labs(title="Change in number of children under 6 in low income",
       caption="StatCan Census 2006, 2016")
```

Change in number of children under 6 in low income



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")
```

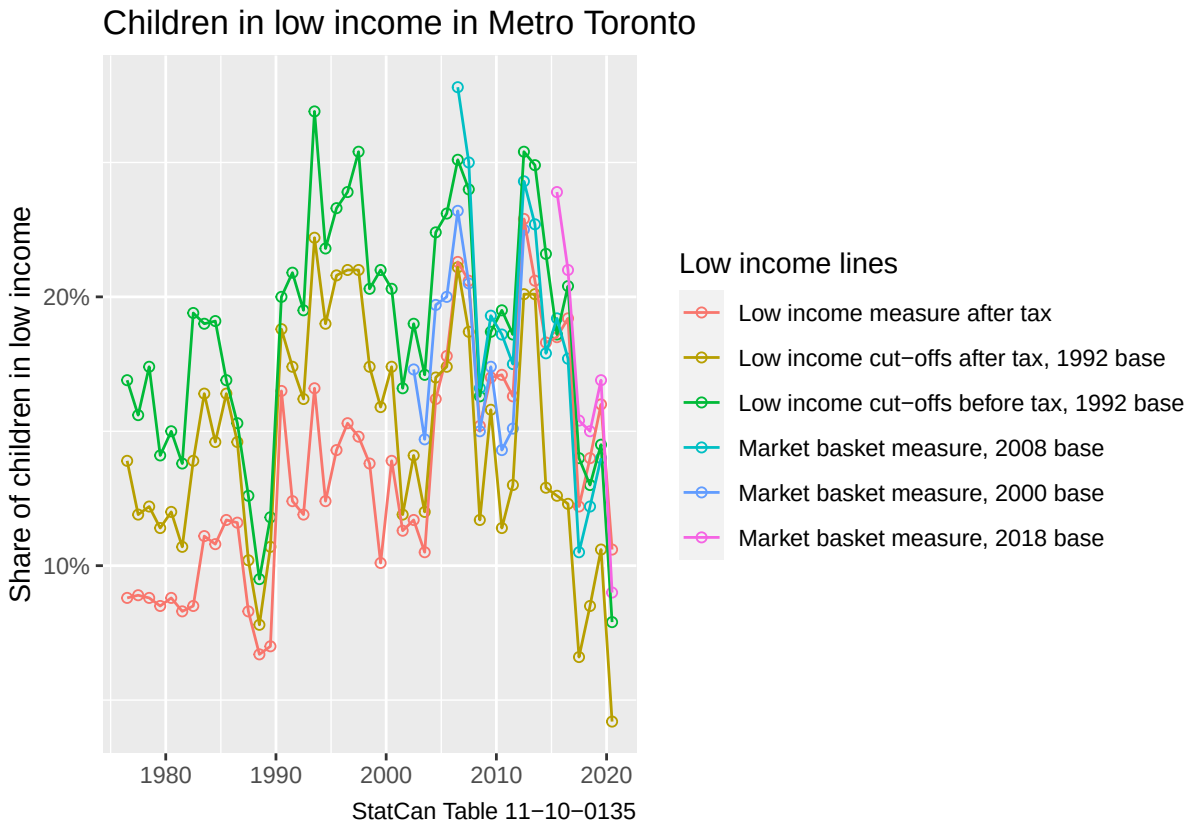


Stephanie shares her results with Amy in Toronto in case there are components of Amy's pilot specifically targeting children below 6 in low income.

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 with R and RStudio

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.

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**.

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

2 Introduction to cansim

2.1 The cansim R package

The **cansim** R package 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.

3 Introduction to cancensus

3.1 The cancensus R package

The **cancensus** R package interfaces with the CensusMapper API server. It can be queried for

- census geographies
- census data
- hierarchical metadata of census variables
- some non-census data that comes on census geographies, e.g. T1FF taxfiler data

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

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.

3.2 BC population growth

This example is motivated by a BC government press release titeled "**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 take that as our question:

3.2.1 Question

Has BC's population really grown by over 100,000, is this the most in 60 years, and how should I interpret this number?

3.2.2 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 population 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("population estimates") %>%
  filter(surveyEn=="Quarterly Demographic Estimates") %>%
  select(1:2)
```

```
# A tibble: 1 x 2
  cansim_table_number cubeTitleEn
  <chr>               <chr>
1 17-10-0009          Population estimates, quarterly
```

It looks like Table 17-10-0009 is exactly what we are looking for. Let's load in the data and inspect the first couple of rows for BC.

3.2.3 Data aquisition

```
pop_data <- get_cansim("17-10-0009")
```

```
Accessing CANSIM NDM product 17-10-0009 from Statistics Canada
```

```
Parsing data
```

```
pop_data %>%
  filter(GEO=="British Columbia") %>%
  select(Date,Population=val_norm) %>%
  head(10)
```

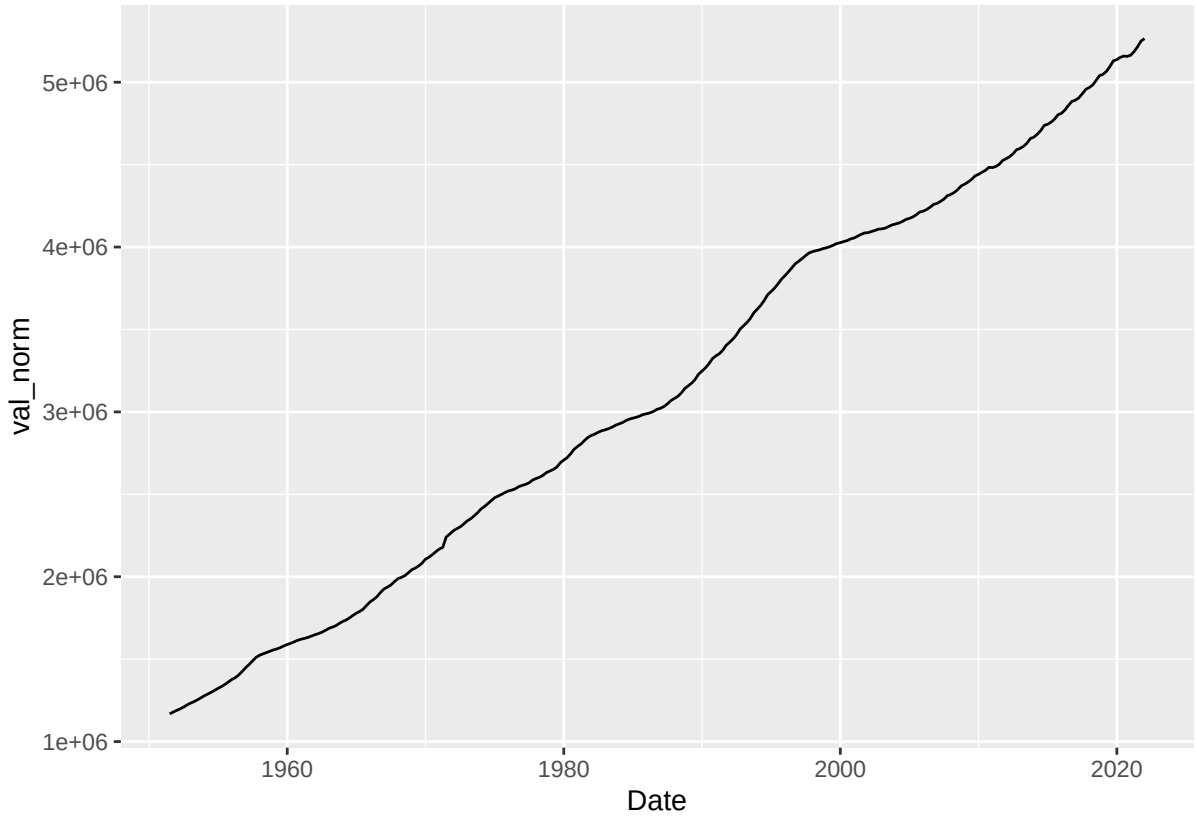
```
# A tibble: 10 x 2
   Date       Population
   <date>          <dbl>
 1 1951-07-01    1168000
 2 1951-10-01    1179000
 3 1952-01-01    1189000
 4 1952-04-01    1198000
 5 1952-07-01    1209000
 6 1952-10-01    1222000
 7 1953-01-01    1233000
 8 1953-04-01    1242000
 9 1953-07-01    1253000
10 1953-10-01    1265000
```

This looks good.

3.2.4 Data preparation, analysis and visualization

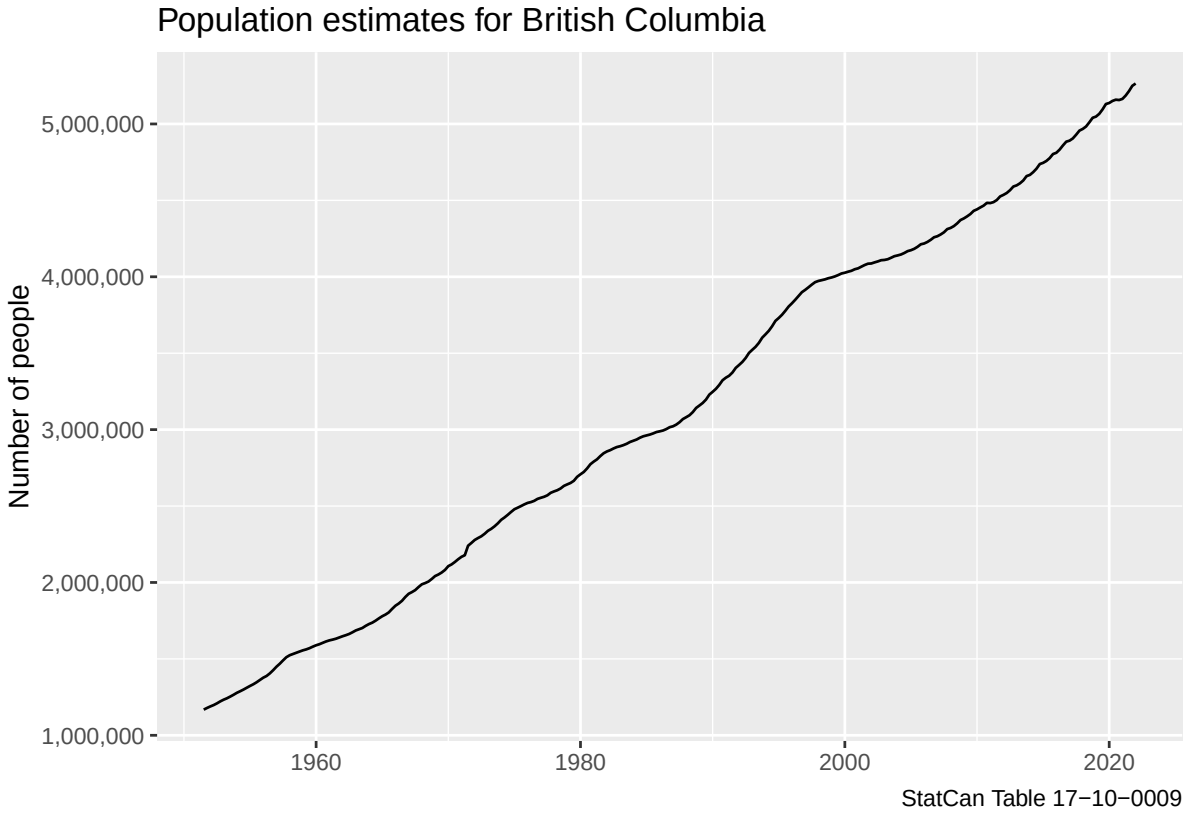
Time to make a graph. We again only take the data for British Columbia and pass it to `ggplot` to make a graph, specifying that we want `Date` on the x-axis, the population value (`val_norm`) on the y-axis and we want a line graph.

```
pop_data %>%
  filter(GEO=="British Columbia") %>%
  ggplot(aes(x=Date,y=val_norm)) +
  geom_line()
```



While this is fine for just getting a quick visual, it's good practice to clean this up a bit and add proper labels so that others looking at the graph will know what it is about. And also so that we understand it if for example we come back to look at this a week later.

```
pop_data %>%
  filter(GEO=="British Columbia") %>%
  ggplot(aes(x=Date,y=val_norm)) +
  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")
```

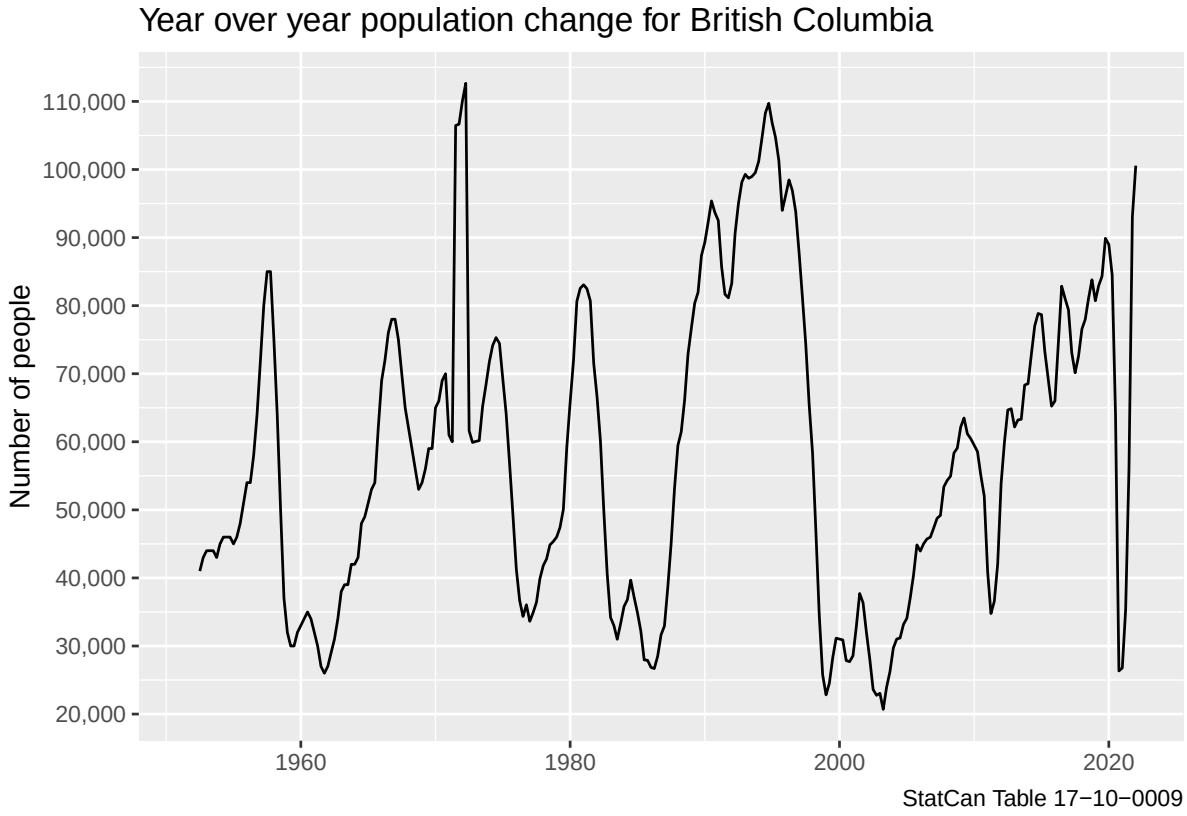


We formated the y-axis values by adding thousand separators, gave it a title, a y-axis label, and noted the data source in the caption.

So far so good, but our question was about population change, not total population. To compute population change we have to take the population estimate and subtract the value from 4 quarters prior. We can do that with the `lag` function.

```r
pop_data %>%
  filter(GEO=="British Columbia") %>%
  mutate(change=val_norm-lag(val_norm, order_by = Date, n=4)) %>%
  ggplot(aes(x=Date,y=change)) +
  geom_line() +
  scale_y_continuous(labels=scales::comma, breaks=seq(0,200000,10000)) +
  labs(title="Year over year population change for British Columbia",
       y="Number of people",
       x=NULL,
       caption="StatCan Table 17-10-0009")
```

```
Warning: Removed 4 row(s) containing missing values (geom_path).
```



Here we also specified manual `breaks` for the y-axis that should be used for labelling, putting a mark for every 10k.

3.2.5 Interpretation

This answers the first part of our question, the latest year over year population change edges just over the 100,000 people mark. But the second part of the question, if this is the most in 60 years, is definitely false. We see that year over year population growth was higher on several occasions during the past 60 years.

Nonetheless, it is interesting to see the fairly continuous increase in year over year population growth since 2000, with several major dips. The most recent one lines up with the COVID pandemic. Other than that our current growth looks pretty much on target with past trends. It will be interesting to see how it continues, if there will be some over-shoot to make up for the deep dip during the pandemic, or if it will roughly follow the previous path.

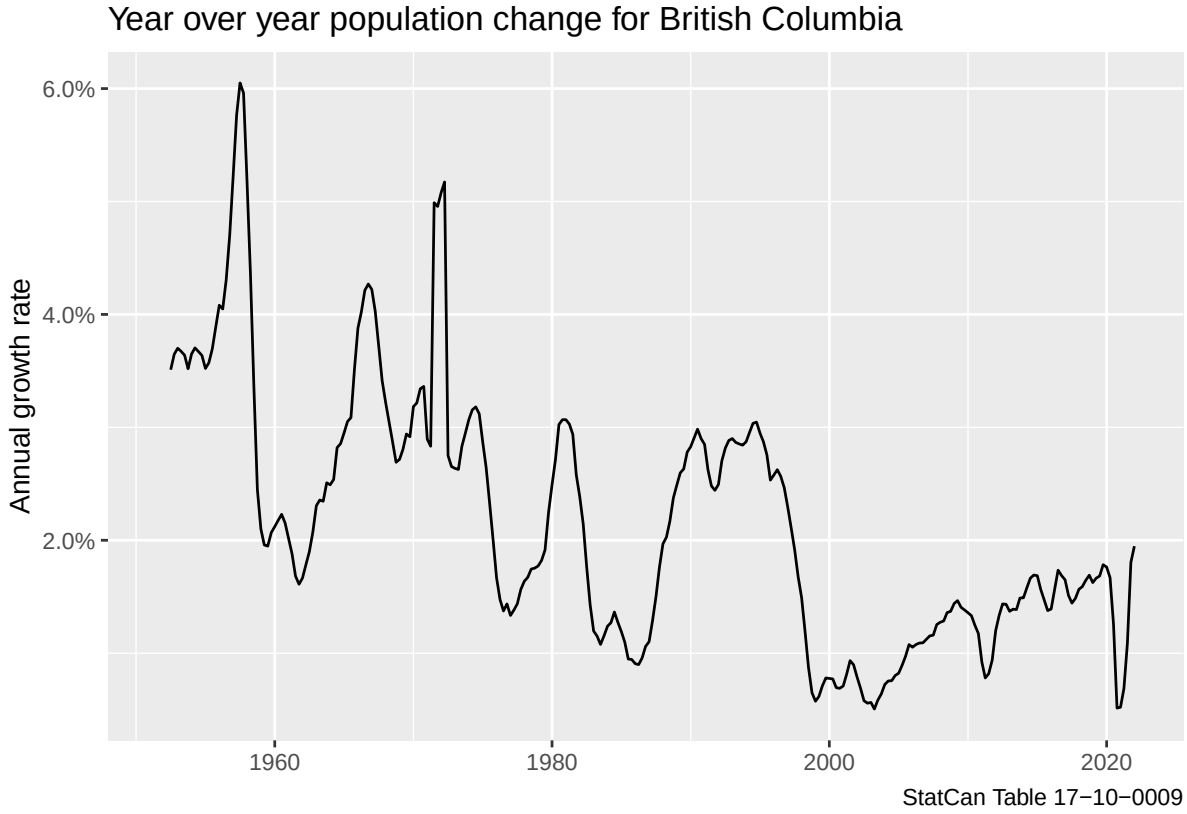
When it comes to interpreting the data, looking back up to the first graph, we notice that British Columbia's population has quintupled over the time frame we are looking at. That's a huge change, the British Columbia from the 50s is very different from the British Columbia today. Does it really make sense to use the total number BC gained year over year when comparing over such long time frames? A more meaningful metric would be to look at the percent change of the population. Let's do that.

3.2.6 Back to analysis

We only need minimal adjustments, to compute the percentage change we take the quotient (and subtract 1) instead of taking year over year differences.

```
pop_data %>%
  filter(GEO=="British Columbia") %>%
  mutate(change=val_norm/lag(val_norm, order_by = Date, n=4)-1) %>%
  ggplot(aes(x=Date,y=change)) +
  geom_line() +
  scale_y_continuous(labels=scales::percent) +
  labs(title="Year over year population change for British Columbia",
       y="Annual growth rate",
       x=NULL,
       caption="StatCan Table 17-10-0009")
```

```
Warning: Removed 4 row(s) containing missing values (geom_path).
```



This makes our recent growth much less impressive, although we still see that trajectory of increasing growth since 2000.

3.2.7 More interpretation

So what is our interpretation of this statistic? While historically over the past 60 years British Columbia has seen higher absolute growth and higher growth rates than currently, our present population growth rate is the highest it has been since the late 90s, and has been slowly increasing since 2000. The current record (since 2000) high fits well into this trend. It is preceded by a record drop in growth rate during the COVID-19 pandemic, which makes it difficult to say how much of the current high is simply a bounce-back from the COVID-19 dip, or part of an ongoing trend to higher growth.

References