Statistics Canada's Open Economic Data for Statistics and Data Science Courses: An R Package

Thierry Warin, Professor, Department of International Business, HEC Montréal and Researcher Digital, Data and Design (D^3) Institute (Harvard Business School) and CIRANO[1]

ABSTRACT

This article aims at describing the steps to identify and access data about the Canadian economy through the `{statcanR}` package. With the `{statcanR}` package, we can search for and collect structured geo-located socioeconomic data about the Canadian economy. Canada was ranked 8th in 2017 by Open Data Watch (Government of Canada) for its data accessibility policy. Statistics Canada offers several ways to access data across its over 11,000 data tables. By creating an R package to facilitate the identification and collection of Statistics Canada's open economic data, we aim to simplify access to relevant data for students while minimizing prerequisites to research.

Keywords: dynamic data, real-world data, authentic data, data science, GAISE, R data package, open data

1. OVERVIEW

The pedagogical context of this article is about the twin goals of teaching statistics or data science courses based on real-world data, namely "minimising prerequisites to research while also using real-world data" (Brown and Kass 2009; Cobb 2015; Kim et al. 2018).

These are sometimes difficult goals for educators, who often provide constructed datasets that are readily usable to compute and teach their favorite formulas or problems to their students. Data acquisition and data wrangling do not seem particularly attractive to educators. Indeed, as the saying goes, data collection and cleaning will take 80% of project time, assuming the fun part is in the remaining 20%. Therefore, teachers tend to hide the "messy data" and provide "tidy data" (Wickham 2014). The "modern student" requires courses to adjust to real-world data (Gould 2010).

In this context and spirit, we built a package to allow our students to access Canadian statistics. This article also describes the components of the `{statcanR}` R package allowing users to have access to over 11,000 data tables (CANSIM tables) identified by

[1] The author would like to thank Jeremy Schneider for his comments and help, and express many thanks to Andrew Zieffler, Associate Editor, and two anonymous referees for their corrections and comments on this manuscript. The usual caveats apply.

Product I.D.s (PID) by the new Statistics Canada Web Data Service. Statistics Canada has created a "web data service"[2] providing access to a significant number of databases. The `{statcanR}` package goes one step further by making the data easily accessible and integrated into a research workflow. Students and researchers alike will have access to the latest data available directly in their R Markdown documents or reports. The Github homepage for the `{statcanR}` package can be found at https://warint.github.io/statcanR/.

In this article, we adopt a reconciliatory position between these two conflicting goals. By using an API or functions designed to access real-world data, the data acquisition is simplified, while still allowing students to focus on the data wrangling part, which remains the most interesting aspect. We inscribe ourselves in the principles according to which "there is much benefit in the students seeing how the data were procured" (Grimshaw 2015).

2. EDUCATIONAL LANDSCAPE

2.1 Evidence-Based Education

Students come to introductory statistics, data science, or general quantitative methods courses with different backgrounds and objectives. This is where the two conflicting goals of teaching statistics take their significance (Brown and Kass 2009). Considering student diversity, there is a set of pedagogical principles and approaches that these courses should follow (Kim et al. 2018). These principles are even more relevant in the A.I. new summer (LeCun et al. 2015).

Indeed, the era of "big data" has led to renewed interest by the general public in the discipline of statistics (Kim et al. 2018). The use of big data contributes to the promotion of new and better educational experiences (Reidenberg and Schaub 2018), to an efficient approach in the formative process for a learning group seeking innovative learning due to the study of data (Huda et al. 2018). Therefore, big data and data science offer a new toolbox for the personalization of learning (Dishon 2017). Indeed, "personalized learning has become the most notable application of big data in primary and secondary schools in the United States. The combination of big data and adaptive technological platforms is heralded as a revolution that could transform education, overcoming the outdated classroom model, and realizing the progressive vision of interest-driven and self-initiated learning" (Dishon 2017).

The 2014 ASA Curriculum Guidelines for Undergraduate Programs in Statistical Science (American Statistical Association 2014) insisted on creating a "culminating experience" for students. This experience could take multiple forms in terms of delivery modes, but more importantly, it asks to make the connection between theory, communication, application, and data manipulation. Moreover, the Guidelines for Assessment and

[2] https://www.statcan.gc.ca/eng/developers/wds/user-guide

Instruction in Statistics Education (GAISE, 2016) make various recommendations, including recommendation 3, about teaching real-world examples and contextualized data.

In this context, authentic data experiences can be one of the answers to the ASA guidelines (Çetinkaya-Rundel and Stangl 2013; Lo 2020). In using authentic data, students would realize that data skills are equally as important as analytical skills. Some authors find it best to avoid static datasets and rely on dynamic data (Grimshaw 2015). This is the reason why we built the `{statcanR}` package. The package provides easy identification of and access to complex data and a wealth of variables about the Canadian economy.

At the time of the release of the `{statcanR}` package, the author mainly teaches at the graduate level in a business school environment at HEC Montreal in Canada. The courses this author teaches are varied and include courses such as *Quantitative Methods in International Business*, *Machine Learning for International Finance,* and *Landscape and Challenges: Evidence-Based Analyses*. In every course, he insists on using real-world data. To ease the process, the author exposes students to the foundations of the R language (R Core Team 2020) at the beginning of each course. Students must use RStudio (RStudio Team 2015) and the R Markdown language (Xie et al. 2020). In these courses, there is always a teamwork exercise. It consists of a 20-page report based on a real-world situation written in teams of three students. Students play the role of business analysts and leverage their new skills in R to provide an evidence-based report. In these courses, assignments, handouts, and slides are provided to students in R Markdown format. Students are also expected to only work in this format during the length of the course. The majority of students who learned R and the R ecosystem (e.g., RStudio) through these courses will be recruited as business analysts in consulting firms, public agencies, etc.

2.2. Real-World Data and Open Data

Data Science allows students to explore with dynamic data: "The call for using real data in the classroom has long meant using datasets which are culled, cleaned, and wrangled prior to any student working with the observations. However, an important part of teaching statistics should include actually retrieving data from the Internet. Nowadays, there are many different sources of data that are continually updated by the organization hosting the data website. The R tools to download such dynamic data have improved in such a way that makes accessing the data possible even in an introductory statistics class" (Hardin 2018). This is precisely why we built the `{statcanR}` package.

Another dimension of our conversation is about teaching statistics and data science in a business school context. To ensure that the results are interpretable, domain knowledge is required (He and Lin 2020). It requires students to understand the notion of complexity. Dynamic data and open data initiatives, combined with GAISE to deal with complex data, are essential for economic complexity. Economic complexity is an interdisciplinary analytical framework at the crossroads between evolutionary economics and institutional economics (Cimoli and Dosi 1995; Hirschman 1958; Teece et al. 1994) designed to explain or at least analyze the determinants of economic development. The differences with the previous literature are nearly twofold. This literature proposes considering different

dimensions (industries and product spaces) at the theoretical level, shifting the focus away from only aggregate variables (Hidalgo and Hausmann 2009; Tacchella et al. 2012). The second difference is about economic geography-related data. The increased granularity provided by this new data access enables students and researchers to answer quantitatively a number of policy-relevant questions. In this context, the `{statcanR}` package aims to contribute to this trend by leveraging open data to help with more complex analyses.

3. THE `{statcanR}` PACKAGE

3.1. Goals

The `{statcanR}` package was developed to facilitate identification and access to Statistics Canada's open data. Indeed, across the world, the question of facilitating access to socioeconomic data is widespread and well established nowadays. While people are more and more connected, the need to provide reliable sources of information and validated data to form better decisions has become crucial and well understood.

In this context, governments at different levels have started to build open data initiatives. One of the goals is to allow researchers, students, and members of civil society to form evidence-based opinions, which will hopefully lead to better individual decisions. In 2017, Canada was ranked the 8th country in the world for its data accessibility. One of the achievements was to propose the “Open Government Portal” initiative, where interested parties could gather geospatial and non-geospatial data. Statistics Canada is a significant contributor to the Open Government Portal, with around 75% of the portal's total non-geospatial content coming from Statistics Canada. In May 2018, to further its data accessibility, Statistics Canada launched its Web Data Service. This Web Data Service provides access to data and metadata that Statistics Canada releases every day.

In the past, Statistics Canada had already taken several steps to make its data more available. In 2012, CANSIM—the agency's socioeconomic database—became free to use. Moreover, the agency adopted an open license and eliminated all royalty and licensing fees. However, Statistics Canada's data may be hard to find. The CANSIM classification system became obsolete, requiring transfer to a new classification system. With this new system in place, the Web Data Service gives access to Statistics Canada data via 12 methods but mainly targets developers. The implicit assumption is that second-party organizations will tap into Statistics Canada and then create secondary data. However, we hypothesize that an open data initiative's real success comes when the largest number of actors have access to primary data. Furthermore, while the agency is moving in the right direction, Open Data Watch noted that some data are difficult to find.

The `{statcanR}` package has been developed to address this concern. Its primary goal is to provide the greatest number of people with access to primary data about Canada's socioeconomic situation. These data, updated daily, provide access to data at different geographical levels: (1) the federal level; (2) the provincial level; and (3) the metropolitan areas.

Having access to data at these three different levels is quite useful. The geographical granularity allows, for instance, researchers to have a new perspective on questions such as regional competitiveness or regional disparities in terms of female labor participation, etc.

Some other packages have been developed to facilitate access to Statistics Canada tables:

- `{CANSIM}` (von Bergmann, Shkolnik, and Jacobs 2021). It is a wrapper to retrieve public Statistics Canada's socioeconomic data. A benefit of `{CANSIM}` over `{statcanR}` is the ability to download "vectors," which are data series prepared by Statistics Canada.
- `{CANSIM2R}` (Lugo 2018). It extracts tables from Statistics Canada and transforms them into panel format. It has not been updated in the last three years.
- `{CANSIM-dataviewer}` (Province of British Columbia 2021). This repository provides access to the code to collect and visualize some of Statistics Canada's socioeconomic data. It is designed for users in British Columbia.

The `{statcanR}` is very similar to `{CANSIM}`. By design, we wanted to be as close as possible to the original tables. When using `{statcanR}`, the user will both find and extract a data matrix with several dimensions into the R environment. We also wanted to allow users to manually select the dimensions of interest for their research in this context. With just basic R commands such as the `unique()` function (see the example below), a user will select the relevant dimensions and extract the corresponding data.

3.2. Interface

The `{statcanR}` package is composed of two functions: `statcan_search()` and `statcan_download_data()`, allowing identification and access to all Canadian statistics open data (CANSIM tables, now identified by product ids (PIDs)) without any limitation and provided by the new Statistics Canada Web Data Service. Below are the complete `statcan_search()` and `statcan_download_data()` functions.

The `statcan_search()` function has two arguments to fulfill to get the desired data: `keywords=` and `lang=`. The `keywords=` argument refers to words that can be found in either the title or the description of the database. For example, inserting the keywords `"economy"`, `"export"`, and `"link"` will bring up the title, table id, description, and release date for databases that include these keywords. In this case, only one data table ("Supply and use tables, link-1997 level") would be returned as it is the only data table containing all three words. Furthermore, words can be put in any order within a vector, or combined within the same quotes if they follow that order in the title / description (see examples in Section 5).

The second argument, `lang=`, specifies the language. As Canada is a bilingual country, Statistics Canada displays all Statistics Data in both languages. Therefore, users can choose to get a statistics data table in French or English by setting the lang argument with `lang=c(fra, eng)`. It is particularly interesting for the author since the courses are in both

languages at HEC Montreal. Students can select the language they feel most comfortable with and write their reports in their language.

The utility of `statcan_search()` can be appreciated by taking the traditional approach of going to the Statistics Canada website to find a given dataset. In this case, we are also able to type keywords and identify datasets that are available. However, it can take far longer to identify the dataframe we are interested in as descriptions are frequently longer than space allows for, and only 10 tables are available on each page. On the other hand, the `statcan_search()` function is used here https://warint.github.io/statcanR/ to produce a data table of the relevant tables based on the inputted keywords. This table has its own search function that immediately allows for further keyword precision if the user realizes their initial search specification was not precise enough.

The `statcan_download_data()` function has two arguments to fulfill to get the desired data: `tableNumber=` and `lang=`. The `tableNumber=` argument simply refers to the Statistics Canada data table the user wants to collect, such as `'14-10-0287-03'` for the *Labour force characteristics by province, monthly, seasonally adjusted* as an example. The second argument, `lang=`, functions in exactly the same manner as in the `statcan_search()` function, namely to specify the language. Students can select English or French depending on the language with which they feel more comfortable. Moreover, the data will be in a tidy format for ease of use, ready to be used by the students (Wickham 2014).

The global process of the `statcan_download_data()` function is the following one. The first step is to clean the table number to match the new official table number typology of Statistics Canada's Web Data Service. The second step is to create a temporary folder where all the following steps can be implemented. The third step is to check and select the correct language asked by the user. The fourth step is to define the correct URL where the statistics data table is stored and download the ZIP file from it. The fifth step is to unzip the previously downloaded ZIP file to get the data and metadata CSV files. The sixth step is to load the data into a data frame called 'data' and add the official table indicator name in the new 'INDICATOR' column. The following section will show how analysis can be done in R easily and efficiently using the `{statcanR}` package.

4. EXAMPLES

In this section, we leverage multiple Statistics Canada datasets to demonstrate the commands available in the `{statcanR}` package as well as their functionality. We use datasets mentioned in some of the most recent publications (which, at the time of writing, is August 3, 2023) of Statistics Canada's "The Daily"[3], a news outlet that publishes short reports of general interest. Providing visual interpretations of the statistics published in "The Daily" ensures that we are providing useful context to reports that are 1) of interest to the general population and 2) coming from the newest available data. Our first example

[3] https://www150.statcan.gc.ca/n1/dai-quo/index-eng.htm

concerns a publication from the above date on profits made by Canadian universities, and how they have evolved since the onset of the COVID-19 pandemic. We will take a close look at both the revenue and expenditure sides and their components in the last eight fiscal years. The second example looks at new motor vehicle registrations in 2023, following a publication on August 2nd, 2023. In particular, we examine the growth of new electric vehicle registrations compared to that of general motor vehicles, as well as the evolution of overall vehicle registrations per year in Canada.

4.1 Question 1: How do Canadian university profits in the pandemic compare to previous years?

One topic that has sparked great interest amidst the COVID-19 pandemic is profit margins of various companies. Some large businesses, such as Canadian grocer Metro Inc., received scrutiny when their financial reports revealed annual profits in 2022 exceeding 10 % while Canadians struggled with higher grocery costs. In the same vein, "The Daily" examined surpluses across another group of organisations, namely Canadian universities, in this same period. Indeed, many pointed out that universities maintained similar tuition fees while no longer providing the same quality of service (i.e., in-person education). Instead of simply reporting the net profit amount from their article, we are interested in exploring the make-up of profits compared to pre-pandemic years.

4.1.1 Data identification: Using the `statcan_search()` command. After loading several packages, we begin by using the `statcan_search()` function. This command allows us to match keywords related to our topic of interest with databases available on the Statistics Canada website. Given our example topic of interest, some keywords could be `"universities"` and `"dollars."`

```
# Search for data table with keywords
library(statcanR)
library(dplyr)
library(DT)
library(curl)
library(data.table)

# Search for data table with keywords
statcan_search(keywords = c("universities", "dollars"), lang = "eng")
```

title	id	description	release_date	lang
Revenue of universities by type of revenues and funds (in current Canadian dollars) (x 1,000)	37-10-0026-01	Financial information of colleges, type of revenues by geography and type of funds.	2023-08-03	eng
Expenditures of universities by type of expenditures and funds (in current Canadian dollars) (x 1,000)	37-10-0027-01	Financial information of universities, type of expenditures by geography and type of funds.	2023-08-03	eng

Figure 1. Results from data table search

4.1.2 Data extraction: Using the `statcan_download_data()` command. We will choose the first dataset, which contains information on all the different sources and amounts of revenues for Canadian universities for each fiscal year, dating back to 2001. Within each

source of revenue, there are several subgroups called sources of funds. For example, one source of revenue could be the Natural Sciences and Engineering Research Council, and within this source, the types of funds could include General operating, sponsored research, and entities consolidated. Revenue sources can also be broken down by province or aggregated at the national level. For our purposes, we only keep the "Total Funds" category for each source of revenue and at the national level, as we don't need such a level of granularity in order to conduct comparative statics analysis on the make-up of university profits. Note that there is already a unique identifier available in this table. We can directly copy this identifier and paste it into the next command: `statcan_download_data()`.

```
# Enter tableNumber and preferred language (here English is specified)
univ_rev <- statcan_download_data(tableNumber = "37-10-0026-01", lang= "eng")

# Wrangle the data
univ_rev |>
  filter(REF_DATE >= "2015-01-01") |>
  filter(GEO == "Canada") |>
  filter(`Types of universities` == "Total universities") |>
  filter(`Types of funds` == "Total funds") |>
  filter(`Types of revenues` != "Total revenues") |>
  group_by(REF_DATE) |>
  mutate(total_value = sum(VALUE)) |>
  mutate(type_percent = VALUE / total_value) |>
  filter(type_percent >= 0.03) |>
  select(c(REF_DATE, `Types of revenues`, VALUE)) |>
  distinct(REF_DATE, `Types of revenues`, .keep_all = TRUE) |>
  mutate(REF_DATE = as.numeric(format(REF_DATE, "%Y"))) |>
  datatable()
```

	REF_DATE	Types of revenues	VALUE
1	2015	Federal	3214050
2	2015	Non-federal	13853384
3	2015	Provincial	13614469
4	2015	Tuition and other fees	9158307
5	2015	Credit course tuition	7611376
6	2015	Investment	2352701
7	2015	Other types of revenues	3810276
8	2015	Sale of services and products	2903056
9	2016	Federal	3244870
10	2016	Non-federal	13837729

Figure 2. Modified table of university revenue sources

4.1.3 Analysis / Exploration. The first part of the analysis will concentrate on the types of revenues that make up total income for universities, and how these revenues have varied with the arrival of the COVID-19 pandemic. To do this, we keep data from 2015 and on at the national level. Furthermore, for simplicity, we limit the types of revenues in our final dataset to those that make up at least 3% of total revenue, as not filtering out

other low-contributing sources leaves around 20 categories. The results are found in Figure 3.

```
library(ggplot2)

# Plot data
ggplot(univ_rev, aes(x = factor(REF_DATE), y = VALUE, fill = `Types of revenues`)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(
    x = "Year",
    y = "Total Value",
    title = "Breakdown of Total Revenue for each Year"
  ) +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
```

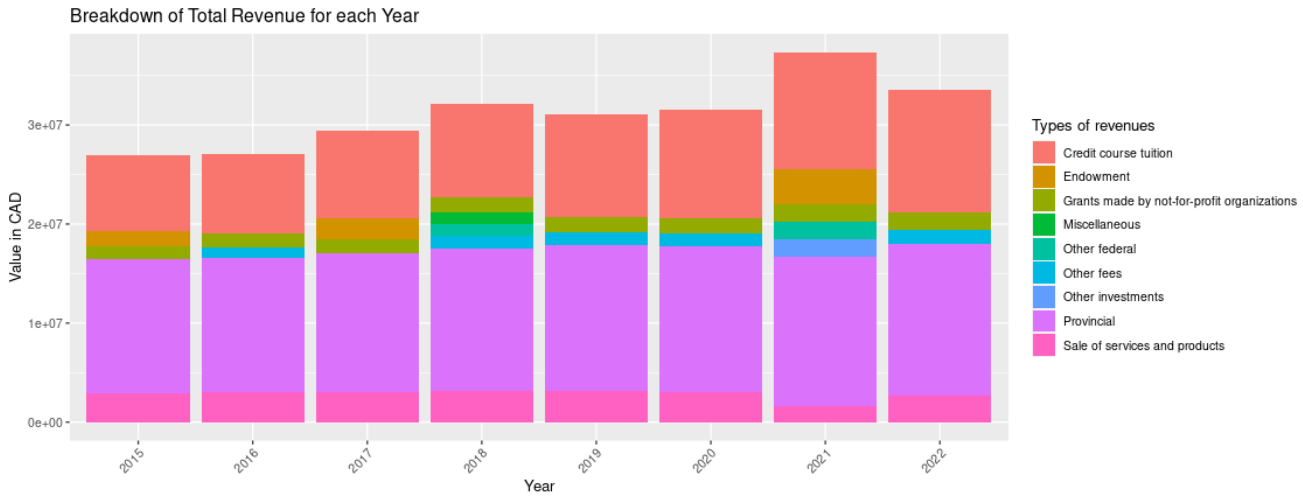

Figure 3. Breakdown of university revenues by year

Interestingly, we see that tuition fees have slightly increased, especially around 2019, and have continued to remain high during and after the pandemic. Furthermore, it is indeed the case that universities saw their overall revenue increase during the COVID-19 crisis, despite courses being given online throughout much of 2020/21. However, to know if this led to higher profits, we need to understand how costs evolved in the same period. We therefore repeat the same analysis, but for the breakdown of university expenditures rather than revenues. Note that by using `"dollars"` as a keyword in the `statcan_search()` function, we were able to locate the relevant table for university expenditures. We therefore take the table number of the second table from Figure 1 and plug it into the `statcan_download_data()` command.

Structurally, the table of university expenditures is identical to the table of university revenues; it contains annual information on total expenditures, which can be broken down by types of universities (meaning either those universities that are or are not members of the Canadian Association of University Business Officers); types of expenditures

(including salaries and benefits, materials and supplies, and scholarships); and types of funds and functions, such as student services and external relations. For simplicity, we omit the equivalent of Figure 2 on the cost side and proceed directly to the breakdown of expenditures over time.

```
# Enter tableNumber and preferred language (here English is specified)
univ_costs <- statcan_download_data(tableNumber = "37-10-0027-01", lang = "eng")

# Wrangle the data
univ_costs |>
  filter(REF_DATE >= "2015-01-01") |>
  filter(GEO == "Canada") |>
  filter(`Types of universities` == "Total universities") |>
  filter(`Types of funds and functions` == "Total funds") |>
  filter(`Types of expenditures ` != "Total expenditures") |>
  group_by(REF_DATE) |>
  mutate(total_value = sum(VALUE)) |>
  mutate(type_percent = VALUE / total_value) |>
  filter(type_percent >= 0.03) |>
  select(c(REF_DATE, `Types of expenditures`, VALUE)) |>
  distinct(REF_DATE, `Types of expenditures`, .keep_all = TRUE) |>
  mutate(REF_DATE = as.numeric(format(REF_DATE, "%Y"))) |>
  datatable()

# Plot data
ggplot(univ_costs,
   aes(x = factor(REF_DATE), y = VALUE, fill = `Types of expenditures `)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(
    x = "Year",
    y = "Total Value",
    title = "Breakdown of Total Expenditures for each Year"
  ) +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
```

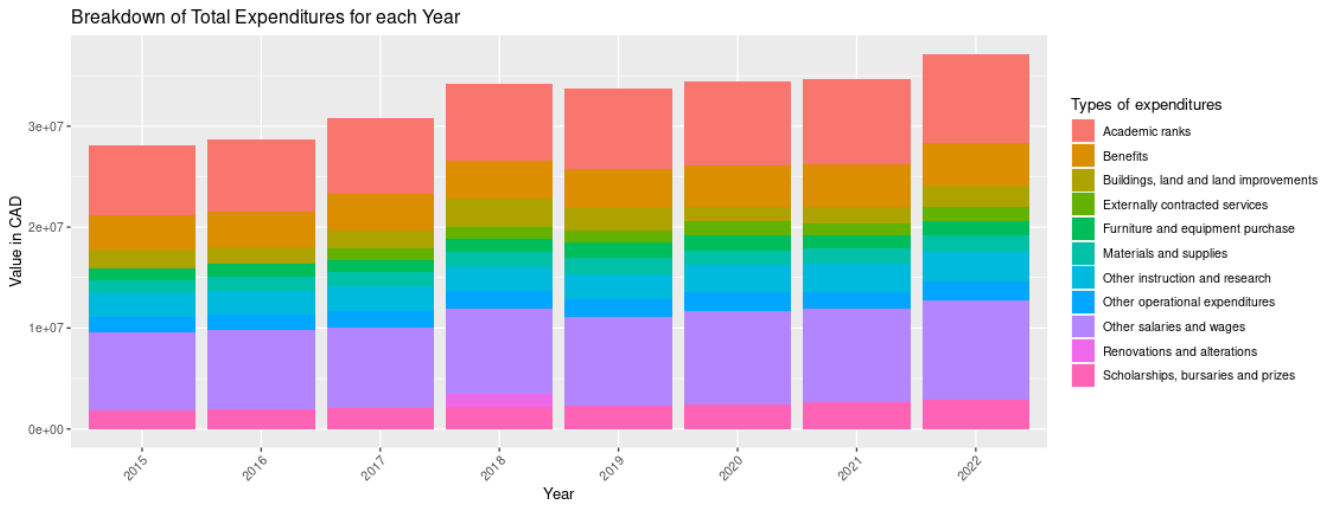

Figure 4. Breakdown of university expenditures by year

Similar to the revenue side, there is no obvious fluctuation in the expenditure make-up that universities face from before the pandemic to after. Indeed, it seems that certain categories, especially salaries and wages of non-academic employees, have continued to grow throughout the last seven years while others have remained stable. However, there is no sharp jump of any given expenditure category, as overall costs seem to largely mirror the stable university growth of revenues.

4.1.4 Comparison with Excel workflow. In this subsection we briefly compare the process of using Excel to conduct data analysis with the process outlined above (i.e., with the `{statcanR}` package). The first step is to extract the same datasets on university revenues and expenditures. These datasets can be found by going to the relevant date of the publication from "The Daily". The table numbers, which are unique identifiers for datasets, are available below the table titles, as shown below in Figure 5.

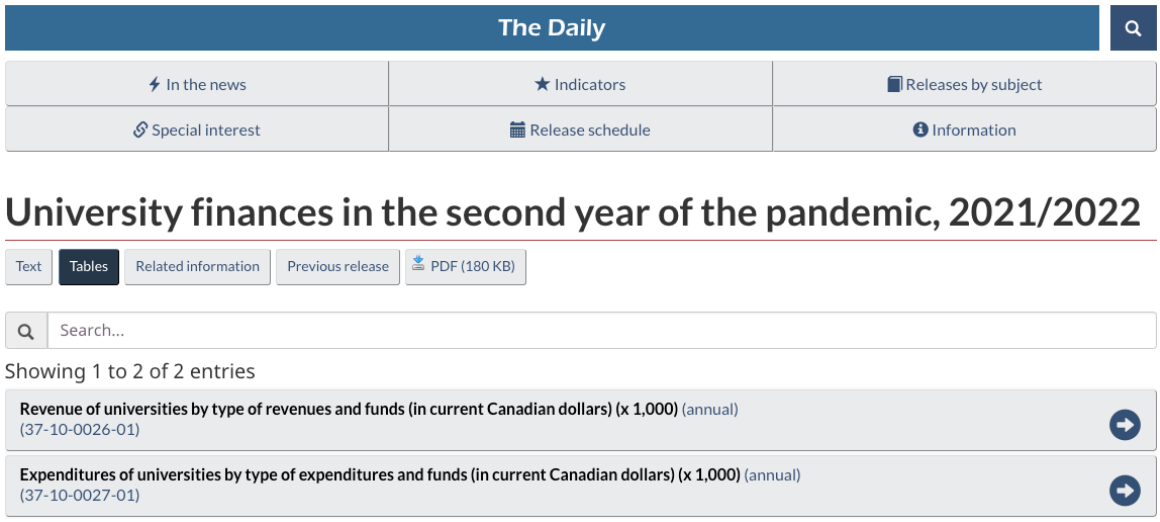

Figure 5. Print screen of table names and their identifiers in "The Daily"

4.1.5 Downloading and analysing the data. We select each of the above tables, click on "Download Options", and separately download them in CSV format. From the extracted table, we can see a list of revenue sources. To visualize this data, we can use a simple funnel chart or any other relevant chart in Excel. Just like in the above analysis, we reduce the number of sources to keep the chart comprehensible. Here, we consider only the top six sources of university revenue for 2015/16, by applying the ordering filter to the data. Note that this can be done for any of the given years.

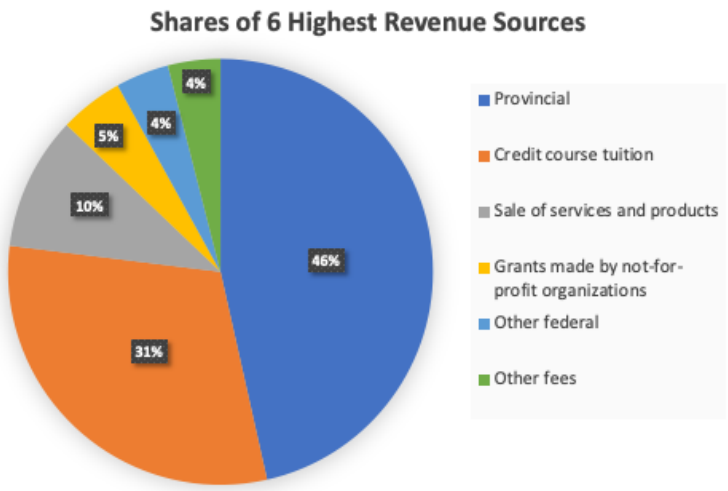

Figure 6. Visualisations of highest revenue sources for 2015/16

4.1.6 Conclusion of Example 1. Since the onset of the COVID-19 pandemic, the contrast between businesses continuing to make high profits and consumers facing financial challenges has been frequently documented. In this example, we examined a group of businesses, namely universities, and whether the evolution of their revenues and expenditures also reflected the aforementioned narrative. While tuition fees may have gradually increased, there is also no evidence of lower university expenditures. Rather, it seems that previous trends persisted throughout recent years, where tuition represents an increasingly important source of revenue, and non-academic salaries represent and increasingly important source of costs.

4.2 Question 2: What do new electric vehicle registrations look like in 2023, and how has their growth compared to general motor vehicle registrations?

In this second example vignette, we leverage a Statistics Canada dataset to demonstrate the commands available in the `{statcanR}` package and their functionality. We then proceed with some brief data analysis to demonstrate the ease with which the `{statcanR}` functions allow for a rapid transition between data cleaning and analysis. For the sake of brevity, we refrain from demonstrating further examples in Excel in this section.

"The Daily" published a brief article on new motor vehicle registrations in Canada on August 2nd, 2023. This is a topic that is certainly of general interest when one considers the efforts of countries to transform their fleet of vehicles from diesel and gas to electric. We want to explore new motor vehicle registrations for the first quarter of 2023. More specifically, we would like to draw some conclusions about the comparative evolution of electric vehicle registrations compared to general motor vehicles. It is well known that electric vehicles have become increasingly popular, but how much does this growth represent?

4.2.1 Data identification: Using the `statcan_search()` command. In order to identify pertinent datasets, we start by using the `statcan_search()` command. This command allows us to match keywords related to our topic of interest with a database or databases available on the Statistics Canada website. Given our example topic of interest, some keywords could be `"motor vehicle"` and `"registrations"`. A useful feature of the `statcan_search()` command is that it can search for several words in a row. For example, instead of needing to write out `"motor"` and `"vehicle"` as separate terms in a vector, we could simply write `"motor vehicle"`. This is demonstrated below:

```
#Enter keywords and preferred language (here English is specified)
statcan_search(keywords = c("motor vehicle", "registrations"), lang = "eng")
```

title	id	description	release_date	lang
New motor vehicle registrations, quarterly	20-10-0024-01	The monthly Building Permits Survey collects data on the value of permits issued by Canadian municipalities for both residential and non-residential buildings, and the number of residential dwellings authorized. The survey also measures the number of dwelling units demolished.	2023-08-02	eng
New motor vehicle registrations	20-10-0021-01	Annual data on new motor vehicle registration by fuel type, vehicle type and number of vehicles, for Canada and provinces.	2022-04-21	eng
Road Motor Vehicles, Registrations	53-219-X	Data on registration of motor vehicles by type including passenger automobiles, trucks, motorcycles, buses, trailers and others are presented in this publication. A historical table of total registrations is provided. Motor vehicle registrations are shown by census divisions and municipalities where available. Data definitions, analysis, the methodology employed, an explanation of data quality and a bibliography are included.	1999-11-04	eng

Figure 7. Results from data table search

4.2.2 Data extraction: Using the `statcan_download_data()` command. We will choose the first dataset in the list and conduct some comparative statics analysis on motor vehicle registrations. Following the same process in Example 1, we use the table number found from the data table produced by the `statcan_search()` command in order to download the dataset, which we call "motor_reg". This dataset contains quarterly data on new motor vehicle registrations at both the provincial and the national level. Furthermore, it contains more granular information on the types of newly registered vehicles (i.e., trucks, cars, etc.) and types of fuel (gas, diesel, electric) that the vehicles are powered by.

```
# Enter tableNumber and preferred language (here English is specified)
motor_reg <- statcan_download_data(tableNumber = "20-10-0024-01", lang = "eng")

# Wrangle the data
motor_reg |>
  filter(REF_DATE >= "2021-10-01" & GEO != "Canada" & !is.na(VALUE) &
         `Fuel type` != "All fuel types" & `Vehicle type` == "Total, vehicle type") |>
  select(c(REF_DATE, GEO, `Fuel type`, VALUE))
```

4.2.3 Analysis / Exploration. The above code yields a dataset that contains information on new motor vehicle registrations over the last six quarters for the majority of Canadian provinces. Furthermore, the registrations are broken down by type of vehicle. This allows us to observe not only how much electric vehicles are growing compared to diesel and gas vehicles, but also how fleets of vehicles are evolving across provinces. Figure 8 shows these results for six provinces.

```
# Plot data
ggplot(motor_reg, aes(x = factor(REF_DATE), y = VALUE, fill = `Fuel type`)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(
    x = "Year",
    y = " Number of Vehicles",
    title = " Newly Registered Motor Vehicles, by Type and Province"
  ) +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1)
  ) +
  facet_wrap(~ GEO)
```

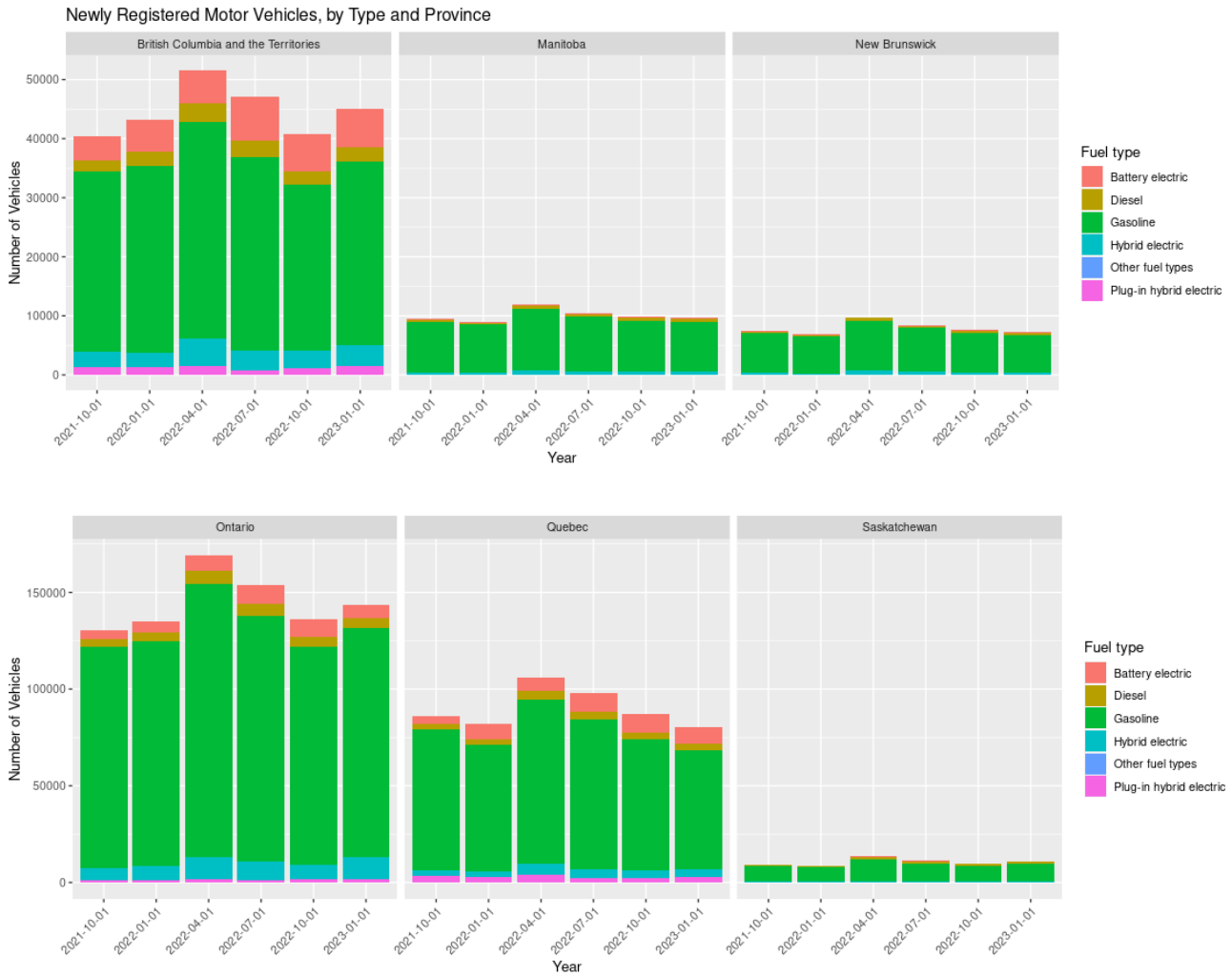

Figure 8. New motor vehicle registrations, by type and province-quarter

Several patterns worth noting are visible in the above graphs. First, it appears that nearly all of electric vehicle registrations are in the three largest provinces (Quebec, Ontario, and British Columbia). Furthermore, these same three provinces seem to have comparatively higher proportions of electric vehicles making up their total new vehicle registrations. Indeed, the red, light blue, and purple bars (i.e., those that make up the electric vehicle registrations) are hardly visible in the other three provinces for which there are data.

From a temporal perspective, the growth in electric vehicle registration seems to have stagnated. While one can see a difference in the above graphs between the most recent quarter of electric vehicle registrations (Q1 of 2023) and the earliest quarter (Q4 of 2021), the latest quarter-to-quarter changes are negligible. More generally, it appears that overall vehicle registrations are not increasing year over year. To see this, we take a more macro perspective and look at total vehicle registrations in Canada over the last five years, only considering one unique quarter across years to avoid seasonal fluctuations. Figure 9 shows this output below.

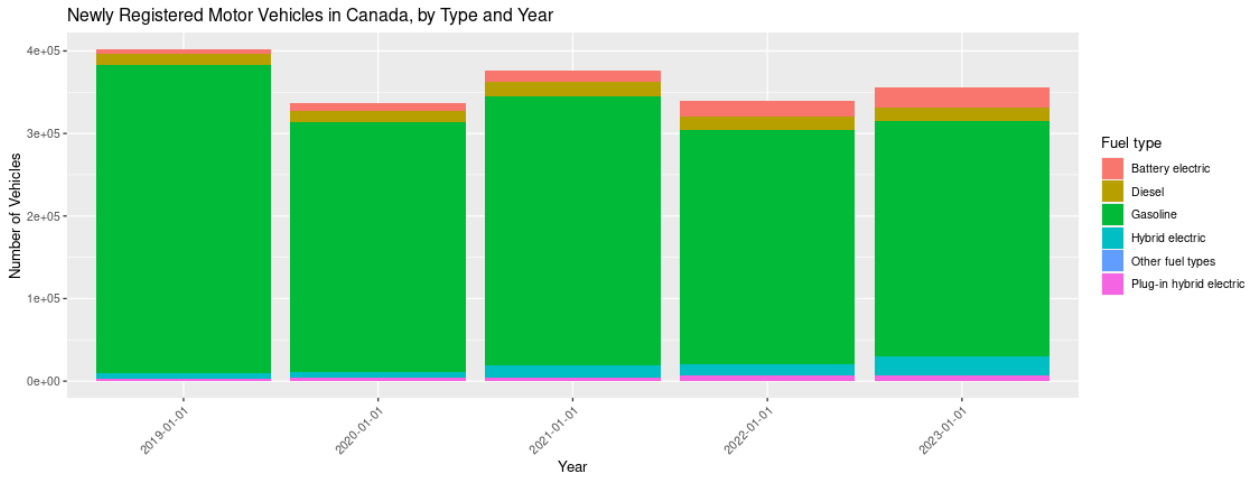

Figure 9. New motor vehicle registrations, by type and year

4.2.4 Conclusion from Example 2. From the above graphics, we can see that within Canada, not only are electric vehicles representing an increasing share of the total new vehicle registrations, but that overall vehicle registrations are slightly declining (or at least, not increasing). Given that Canada hasn't seen any population decline, this evolution is potentially in part due to a substitution from vehicles towards other means of transportation, especially in provinces that have large cities with a network of public transportation. Regardless, a declining amount of new vehicles combined with a larger share of electric vehicle registrations represents the continuing transition towards a more sustainable economy.

For other visualizations, the user may want to read the Github vignette here: https://warint.github.io/statcanR/.

6 IMPACT AND CONCLUSION

Like all the countries, Canada faces different transitions simultaneously: from the digital transformation to the transition towards an environmentally sustainable economy. Developing new technologies, enhancing products and services based on - also - new ways of financing while training human capital equipped with the right skills are amongst the biggest challenges for countries' attractiveness and competitiveness. In this context, the reasons why such packages are applicable are—at least—threefold.

First, having easy access to data has a remarkable impact on education and research. New research questions can be answered thanks to the new level of data availability. Indeed, by leveraging its open data policy, the Canadian government offers researchers in Canada and the world the opportunity to work on economic and societal research questions using empirical information from Canada. It is a real boost for researchers at all stages of their academic careers. It capitalizes on the principles from the "power of the crowd" (Avasilcai and Galateanu 2018; Cai et al. 2019; De C. Wang et al. 2019; Gal 2019). The use of these new data, structured and unstructured, is changing and will continue to change our education systems (Quezada-Sarmiento et al. 2020; Samuelsen et al. 2019).

Second, developing easy access to data is of great importance for policymaking. In the past, there were lots of areas in public policymaking where data were not accessible. As a result, decisions were made on assumptions coming from theoretical foundations or benchmarks from other sources. With more and more access to data globally, being open data initiatives or not, evidence-based decisions are more and more possible in our day and age. Numerous authors have demonstrated the role of data in informing better evidence-based policies (Giménez-bertomeu et al. 2019; Payán and Lewis 2019; Villumsen et al. 2019; Wolffe et al. 2019).

Third, having access to Canadian data is extremely useful for business analytics. It is a real boost for our students who will be future analysts in companies. Having access to this Canadian economic and societal data will help companies devise better strategies and contextualize their data. Business analytics (Cui et al. 2020; Hindle et al. 2020; Rialti et al. 2019; Vidgen et al. 2020) improves firms' performance (Elhoseny et al. 2020; Ferraris et al. 2019), fosters innovation in a country (Duan et al. 2020), and helps reinforce competitiveness, by reducing strategic uncertainty (Ganesan and Gopalsamy 2019).

Based on these three reasons, with better access to data comes the possibility to assess the level of economic complexity of a country. Complexity is at the root of economic development. Until recently, because of the absence of adequate complexity measures, the macroeconomic literature has developed around national aggregates and has tended to forget the search for detailed capabilities and their patterns of complementarity, hoping that aggregate measures of physical or human capital would provide sufficient policy guidance. Recent developments in computing power and data accessibility offer new tools to develop policies to promote new capabilities or enhance existing capabilities to encourage the further coevolution of new capabilities, echoing ideas put forward by Albert Hirschman more than 50 years ago (Hirschman 1958). The difference is that students,

researchers, policymakers, and business analysts can now analyze them in practice. The `{statcanR}` package leverages Statistics Canada's open data initiative in this context.

Acknowledgments

The author is grateful to Jeremy Schneider, Marine Leroi, Martin Paquette at CIRANO (Montreal), and Thibault Senegas, Shruti Jhunjhunwala, and Romain Le Duc for their help and comments. The usual caveats apply.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

American Statistical Association (2014), "Curriculum guidelines for undergraduate programs in statistical science," Available at https://www.amstat.org/asa/education/home.aspx/curriculumguidelines.cfm

Avasilcai, S., and Galateanu, E. (2018), "Co-creators in innovation ecosystems. Part II: Crowdsprings 'Crowd in action." https://doi.org/10.1088/1757-899X/400/6/062001

Brown, E. N., and Kass, R. E. (2009), "What Is Statistics?," *The American Statistician*, Taylor & Francis, 63, 105–110. https://doi.org/10.1198/tast.2009.0019

Cai, C. W., Gippel, J., Zhu, Y., and Singh, A. K. (2019), "The power of crowds: Grand challenges in the Asia-Pacific region," *Australian Journal of Management*, 44, 551–570. https://doi.org/10.1177/0312896219871979

Çetinkaya-Rundel, M., and Stangl, D. (2013), "A Celebration of Data," *CHANCE*, Available at https://chance.amstat.org/2013/09/classroom_26-3/

Cimoli, M., and Dosi, G. (1995), "Technological Paradigms, Patterns of Learning and Development: An Introductory Roadmap," *Journal of Evolutionary Economics*, 5, 243–68. https://doi.org/10.1007/BF01198306

Cobb, G. (2015), "Mere Renovation is Too Little Too Late: We Need to Rethink our Undergraduate Curriculum from the Ground Up," *The American Statistician*, Taylor & Francis, 69, 266–282. https://doi.org/10.1080/00031305.2015.1093029

Cui, Y., Kara, S., and Chan, K. C. (2020), "Manufacturing big data ecosystem: A systematic literature review," *Robotics and Computer-Integrated Manufacturing*, 62. https://doi.org/10.1016/j.rcim.2019.101861

De C. Wang, P., Soares, V. S., De Souza, J. M., Esteves, M. G. P., Schots, N. C. L., and Duarte, F. R. (2019), "A crowd science framework to support the construction of a gold standard corpus for plagiarism detection," pp. 440–445. https://doi.org/10.1109/CSCWD.2019.8791853

Dishon, G. (2017), "New data, old tensions: Big data, personalized learning, and the challenges of progressive education," *Theory and Research in Education*, SAGE Publications, 15, 272–289. https://doi.org/10.1177/1477878517735233

Duan, Y., Cao, G., and Edwards, J. S. (2020), "Understanding the impact of business analytics on innovation," *European Journal of Operational Research*, 281, 673–686. https://doi.org/10.1016/j.ejor.2018.06.021

Elhoseny, M., Kabir Hassan, M., and Kumar Singh, A. (2020), "Special issue on cognitive big data analytics for business intelligence applications: Towards performance improvement," *International Journal of Information Management*, 50, 413–415. https://doi.org/10.1016/j.ijinfomgt.2019.08.004

Ferraris, A., Mazzoleni, A., Devalle, A., and Couturier, J. (2019), "Big data analytics capabilities and knowledge management: impact on firm performance," *Management Decision*, 57, 1923–1936. https://doi.org/10.1108/MD-07-2018-0825

GAISE College Report ASA Revision Committee, "Guidelines for Assessment and Instruction in Statistics Education College Report 2016," http://www.amstat.org/education/gaise

Gal, M. S. (2019), "The Power of the Crowd in the Sharing Economy," *Law and Ethics of Human Rights*, 13, 29–59. https://doi.org/10.1515/lehr-2019-0002

Ganesan, S., and Gopalsamy, S. (2019), "Business intelligence and advanced analytics: Impact and behavior of business decision making process," *International Journal of Recent Technology and Engineering*, 8, 375–379. https://doi.org/10.35940/ijrte.C1080.1083S19

Giménez-bertomeu, V. M., Domenech-lópez, Y., Mateo-pérez, M. A., and De-alfonseti-hartmann, N. (2019), "Empirical evidence for professional practice and public policies: An exploratory study on social exclusion in users of primary care social services in Spain," *International Journal of Environmental Research and Public Health*, 16. https://doi.org/10.3390/ijerph16234600

Gould, R. (2010), "Statistics and the Modern Student," *International Statistical Review*, 78, 297–315. https://doi.org/10.1111/j.1751-5823.2010.00117.x

Grimshaw, S. D. (2015), "A Framework for Infusing Authentic Data Experiences Within Statistics Courses," *arXiv:1507.08934 [stat]*. https://doi.org/10.48550/arXiv.1507.08934

Hardin, J. (2018), "Dynamic Data in the Statistics Classroom," *Technology Innovations in Statistics Education*, 11. https://doi.org/10.5070/T5111031079

He, X., and Lin, X. (2020), "Challenges and Opportunities in Statistics and Data Science: Ten Research Areas," *Harvard Data Science Review*, PubPub. https://doi.org/10.1162/99608f92.95388fcb

Hidalgo, C. A., and Hausmann, R. (2009), "The building blocks of economic complexity," *Proceedings of the National Academy of Sciences*, 106, 10570–10575. https://doi.org/10.1073/pnas.0900943106

Hindle, G., Kunc, M., Mortensen, M., Oztekin, A., and Vidgen, R. (2020), "Business analytics: Defining the field and identifying a research agenda," *European Journal of Operational Research*, 281, 483–490. https://doi.org/10.1016/j.ejor.2019.10.001

Hirschman, A. O. (1958), *The strategy of economic development*, Yale paperbound, Yale University Press. ISBN: 0300001177

Huda, M., Maseleno, A., Atmotiyoso, P., Siregar, M., Ahmad, R., Jasmi, K. A., and Muhamad, N. H. N. (2018), "Big Data Emerging Technology: Insights into

Innovative Environment for Online Learning Resources," *International Journal of Emerging Technologies in Learning (iJET)*, 13, 23–36. https://doi.org/10.3991/ijet.v13i01.6990

Kim, A. Y., Ismay, C., and Chunn, J. (2018), "The fivethirtyeight R Package: 'Tame Data' Principles for Introductory Statistics and Data Science Courses," *Technology Innovations in Statistics Education*, 11. https://doi.org/10.5070/T5111035892

LeCun, Y., Bengio, Y., and Hinton, G. (2015), "Deep learning," *Nature*, Nature Publishing Group, 521, 436–444. https://doi.org/10.1038/nature14539

Lo, V. S. Y. (2020), "Top 10 Essential Data Science Topics to Real-World Application from the Industry Perspectives," *Harvard Data Science Review*, PubPub. https://doi.org/10.1162/99608f92.4ff28438

Lugo, M. (2018), *CANSIM2R: Directly Extracts Complete CANSIM Data Tables*.

Payán, D. D., and Lewis, L. B. (2019), "Use of research evidence in state health policymaking: Menu labeling policy in California," *Preventive Medicine Reports*, 16. https://doi.org/10.1016/j.pmedr.2019.101004

Province of British Columbia (2021), *bcgov/CANSIM-dataviewer*, R, Province of British Columbia.

Quezada-Sarmiento, P. A., Enciso, L., Conde, L., Mayorga-Diaz, M. P., Guaigua-Vizcaino, M. E., Hernandez, W., and Washizaki, H. (2020), "Body of Knowledge Model and Linked Data Applied in Development of Higher Education Curriculum," *Advances in Intelligent Systems and Computing*, 943, 758–773. https://doi.org/10.1007/978-3-030-17795-9_57

R Core Team (2020), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing.

Reidenberg, J. R., and Schaub, F. (2018), "Achieving big data privacy in education:," *Theory and Research in Education*, SAGE PublicationsSage UK: London, England. https://doi.org/10.1177/1477878518805308

Rialti, R., Marzi, G., Ciappei, C., and Busso, D. (2019), "Big data and dynamic capabilities: a bibliometric analysis and systematic literature review," *Management Decision*, 57, 2052–2068. https://doi.org/10.1108/MD-07-2018-0821

RStudio Team (2015), "RStudio | Open source & professional software for data science teams," Available at https://posit.co/.

Samuelsen, J., Chen, W., and Wasson, B. (2019), "Integrating multiple data sources for learning analytics—review of literature," *Research and Practice in Technology Enhanced Learning*, 14. https://doi.org/10.1186/s41039-019-0105-4

Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A., and Pietronero, L. (2012), "A New Metrics for Countries' Fitness and Products' Complexity," *Scientific Reports*, 2, 1–7. https://doi.org/10.1038/srep00723

Teece, D. J., Rumelt, R., Dosi, G., and Winter, S. (1994), "Understanding corporate coherence: Theory and evidence," *Journal of Economic Behavior & Organization*, 23, 1–30. https://doi.org/10.1016/0167-2681(94)90094-9

Vidgen, R., Hindle, G., and Randolph, I. (2020), "Exploring the ethical implications of business analytics with a business ethics canvas," *European Journal of Operational Research*, 281, 491–501. https://doi.org/10.1016/j.ejor.2019.04.036

Villumsen, S., Faxvaag, A., and Nøhr, C. (2019), "Development and progression in Danish eHealth policies: Towards evidence-based policy making," *Studies in Health Technology and Informatics*, 264, 1075–1079. https://doi.org/10.3233/SHTI190390

von Bergmann, J., Shkolnik, D., and Jacobs, A. (2021). cancensus: R package to access, retrieve, and work with Canadian Census data and geography. v0.4.2. Available at https://mountainmath.github.io/cancensus/index.html.

Wickham, H. (2014), "Tidy Data," *Journal of Statistical Software*, 59, 1–23. https://doi.org/10.18637/jss.v059.i10

Wolffe, T. A. M., Whaley, P., Halsall, C., Rooney, A. A., and Walker, V. R. (2019), "Systematic evidence maps as a novel tool to support evidence-based decision-making in chemicals policy and risk management," *Environment International*, 130. https://doi.org/10.1016/j.envint.2019.05.065

Xie, Y., Allaire, J.J., & Grolemund, G. (2018). R Markdown: The Definitive Guide (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781138359444