CSC110 Project Report: A Data-Centric Analysis of the COVID-19 Pandemic's Impact on Different Sectors of the Canadian Economy

Xi Chen, Taymoor Farooq, Se-Eum Kim, Henry Klinck

Tuesday, December 14, 2021

Problem Description and Research Question

The COVID-19 pandemic has undoubtedly had a significant negative impact on Canada's economy. Mandatory lock-downs, new business regulations and the Canadian population's overall social distancing behaviours all contributed to financial strain on many Canadian families and businesses.

Many businesses have suffered due to the lack of customers and many restrictions. Most in-person businesses such as restaurants, retail stores, and movie theatres have been closed. Large events such as sports games and concerts have been cancelled entirely. Travel and tourism have essentially been non-existent.

However, certain businesses have thrived from the pandemic. Most online and digital services including online shopping, streaming services, the video game industry and virtual meeting platforms have prospered thanks to most of the population staying indoors. For example, Netflix added 10 million new customers in the second quarter of 2020 (Kafka). Pharmaceutical and medical companies have also done well.

One factor related to these kinds of economic trends is employment, and this is our motivation for choosing this topic. Unemployment has increased overall and as university students who will be looking for jobs in a few years, it is important for us to know if we are on the right track.

So, our question is: **How has the pandemic affected different economic sectors in Canada?** To answer this question, we want to compare for each sector, the actual data gathered during the pandemic with the predicted values based on pre-pandemic data. Then we can see how certain trends have or have not changed due to the pandemic.

After almost two years, the overall economy has mostly recovered, it has changed significantly. Some industries were able to adjust as remote working has become common. Other industries will probably not recover until the pandemic ends. This has major social consequences for the people who are working in certain industries.

Being able to distinguish between the two branches is necessary to deal with the pandemic. For individuals, this can influence decisions on employment or education. For society, this is the first step to helping the people most affected negatively by the pandemic.

It is also valuable to determine how each economic sector was affected. The four economic sectors that this project will investigate are: the primary (production) sector, the secondary sector (encompasses companies that contribute to producing a finished, usable product or are involved in construction), the tertiary (service) sector, and the quaternary sector (associated with either the intellectual or knowledge-based economy). Based on the results of this project, we will be able to identify which economic sectors were the most impacted by the COVID-19 pandemic.

In this project, linear regression will be used to determine GDP values that could be expected during March, April and May of 2020 if the COVID-19 pandemic never occurred. These expected values will then be compared to the actual GDP values measured in March, April and May of 2020, the first few months of the COVID-19 pandemic. For context, the World Health Organization declared COVID-19 outbreak a pandemic on March 11, 2020.

Dataset Description

The primary dataset we have used for this project comes from the Statistics Canada Covid-19 statistics database and presents the Canadian Gross Domestic Product at seasonally adjusted prices across numerous different industries, ranging from Agriculture to IT, Media and beyond. This dataset consists of GDP values from August 1997 to August 2021, a range of around 22 years of pre-pandemic data and over a year of (post)-pandemic data. We downloaded the dataset to a .csv file (an option available on the Statistics Canada Website) which we have used as the main data source for the data wrangling, computations and visualizations performed by our program.

The rows in this dataset present the GDP values per industry, and the columns present the respective date (in format: "Year Month", eg: "January 2014"). Each row is relevant to the functionality of the program, as it provides the GDP values per industry per date. The number of columns can be adjusted by the user of the program according to how many months of data they want (ie: if the user wants to see the impact of the pandemic from October 2019 to March 2020 inclusive, the dataset would have 6 columns). The next section elaborates upon the specific data wrangling operations performed by our program to adjust the csv file and make it usable in python.

Computational Overview

Data Wrangling

The data wrangling for the program is performed in the Data Module (file: data.py). This module has two main functions:

The first function (**open_and_convert**) converts the dataset (a .csv file) to a dictionary mapping each industry name to a list containing a tuple containing a year-month tuple eg: (2014, 5) == May 2014, and the respective GDP value.

The second function (**open_convert_and_aggregate**) essentially performs the same operation as open_and_convert, yet instead of mapping each industry name, it categorizes each industry into its respective sector, then maps the sector's name (eg: 'Primary Sector'/'Secondary Sector'/'Tertiary Sector'/'Quaternary Sector') to a list containing a tuple containing a year-month tuple and the respective aggregated GDP value.

Both of the main functions make use of various helper functions. Python's built-in csv library is imported to open the dataset in Python using the helper function file_to_list. After the file is opened, the helper function list_to_dict filters the opened csv file by mapping filtering out the top 11 and last 11 rows, as they aren't relevant to any of the computations performed on the dataset (ie: blank rows, redundant strings, etc.). The 0th indexes of the remaining rows contain the industry name (as a string), which are then mapped to the year-month tuple (the helper function month_to_num converts the date in string format to a tuple with type tuple[int, int]) and the respective GDP value.

The industries are combined into economic sectors using the helper function categorize_4_sectors. As the dataset also had industries already categorized into larger categories, this helper function filters them out to ensure that only individual industries are included in the dictionary. The GDP values are then aggregated using the helper function aggregate aggregate_4_sectors, which takes the sum of GDP values across all the industries per sector.

Computational Models

To determine COVID-19's effect on each industry category, we will first generate a line of best fit based on GDP data before March 2020, when COVID-19 was declared a pandemic (World Health Organization). For the time interval from March 2020 to May 2020 (inclusive), we will use the difference between the line's extrapolated GDP value and the actual value to describe the effect. We will extrapolate using a line of best fit as it constructs a better theoretical GDP unaffected by COVID-19 in the months after March 2020 (the start of COVID-19) than using a constant line at the known GDP value in February 2020. This computed data will allow us to compare the effect of COVID-19 on each industry category in the 12 months following the start of the pandemic.

The main function in the Computation Module is run_computations, which took the output dictionary from the Data Module and then used all the subsequent functions to produce computed data points for the Display Module. Of these functions, regress and predict_gdp_values played the most important role.

The regress function used Sci-kit Learn's LinearRegression method. While the GDP values for an industry was used as the regression's y-coordinates, the x-coordinates were a sequence from 0 to the length of the list of x-coordinates, essentially the relative order of the dates. This was possible as the dates were equally spaced apart (by a month). The output coefficients of the regress function is then used by consequent functions to predict GDP values.

The predict_gdp_values function takes the list of tuples (date, gdp) corresponding to a particular economic sector as well as the liner regression model values associated to that particular list of values (derived in regress function). The x-values (index of date in list) corresponding to the months March, April and May of 2020 are inserted into the formula: GDP = (index of date in list) * slope + intercept to determine the expected GDP values for these months, assuming they would follow the the same general GDP growth trends consistent across prior years.

Visualization

Line graphs are an effective method to display the impact of COVID-19 on various Canadian economic sectors over time. Being that this group plans to use linear regression to predict various GDP values of Canadian industries assuming the COVID-19 pandemic never occurred, and create a double line graph comparing the predicted values along with the actual GDP values measured in 2020-2021. The months/years leading up to 2020 will also be included in the line graphs created in order to provide context of normal GDP fluctuation not during a pandemic. Including the GDP of various industries over time before the COVID-19 pandemic will emphasize the significance/insignificance of the impact of the COVID-19 virus on various Canadian industries. Two separate graph will also display the difference and the percentage lost between the predicted and actual values. With these graphs, we can effectively visualize the difference/similarity between the predicted GDP values (if COVID-19 never occurred) and the actual GDP values across the four economic sectors in Canada.

Our group plans to use the Plotly Python Open Source Graphing Library to create the visualizations of our results/findings to the question listed above. Plotly allows programmers to customize elements such as the colour and style of lines, as well as giving users many options to manipulate and interact with the created graph, making the data easy to understand and analyze. Our program also uses the Pandas Library, which gives us access to the DataFrame object, which makes it simpler to use Plotly. There will be two different graphs, with the x axis having dates, and the y axis either having GDP values or the difference between expected and actual values. Each graph has a total of 8 lines representing 4 sectors with 2 lines each for expected and actual values. Each sector will be a different colour, and actual values will be drawn with a solid line while expected value will be drawn with a dotted line. The user can interact with the graph, zooming in and out, adjusting scales and filtering between lines.

To do this, we created a dataclass called Sector. Each Sector object would have a name, and 2 lists of tuples representing where each tuple would represent a point on the graph. Sector objects will be created in the Main Module by using functions from other modules. There are three functions in the Display Module, each displaying a different graph. All of them take a list of Sectors, and converts the data in the Sectors to a DataFrame object, which is then passed to a Plotly function to display the graph.

Program Setup

As mentioned in the project submission page, the first step involves installing the libraries in `requirements.txt` and downloading the dataset. The primary dataset of the monthly Canadian GDP by industry can be downloaded at **https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3610043401** and must be saved to the same folder where the other project-related files are located. If you are unable to access the files using the link above, you can access the datasets using the following link and the information: https://send.utoronto.ca/ Claim ID: UKBx7QPTUBBSW8av Claim Passcode: XSaD7v9kFz2ffG33.

The program can then be ran in the Python console using the `run_program` function in `main.py` with a single parameter, the name of the dataset csv file as a string. The result, opened in a browser window, is a line graph of the predicted and actual GDP values from January 2012 to May 2020 for all 4 sectors. Interactive statistics for each data point is shown upon hovering the cursor over it.

Changes between the Proposal and the final Project

One of the most important change was the decision to use relative deviation as a measure of COVID-19's impact rather than deviation. The measure of relative deviation is deviation proportional to the "size" of each aggregated sector, which allows a fairer comparison between them.

In terms of the computational model, while the proposal sought to predict GDP values 12 months into COVID-19, the final program predicted only 3 values. This shifted the focus of the report to more cross-sectional analysis rather than a longitudinal one. This comes as using simple regression to predict a complex mathematical index like the GDP was only practical in the short. Additionally, while the amount of pre-pandemic data points used to generate the line of best fit was not specified in the proposal, the final program used 74 data points for this purpose. This amount is large enough for statistical calculations but also does not contain excess amounts of older data (e.g. data from four decades before COVID-19).

Discussion

First, to summarize the results, in terms of the total change in GDP, the quaternary sector was the least affected (<$6000M), then the primary sector (<$23000M), then the secondary sector (<$90000M), and finally the tertiary sector (<$240000M). However, percentage wise, the secondary sector lost the largest proportion (<23%), then the tertiary sector (<18%), then primary (<11%) and quaternary (<9%). Overall, the secondary and tertiary sectors were affected the most and the primary and quaternary sectors were affected the least.

The GDP values started to fall in March, corresponding with quarantine and other measures being implemented in mid-March. The GDP loss peaked in April, and then slowly recovered from May onward. Our graphs clearly display the significant negative impact of the COVID-19 pandemic on various Canadian economic sectors during the first few months of the pandemic. The fact that the secondary and tertiary sectors were affected the most means that manufacturing and service-based industries were heavily affected by COVID-19. This finding makes logical sense because manufacturing and service-based industries likely have the most person-to-person contact which led to a decrease in customer interest and an increase in government regulations during the pandemic. Most of the findings of this project were expected by group members. However, the "Percentage Lost Between Expected and Actual GDP Values" graph was particularly interesting to all group members. We had not known how much more the secondary and tertiary sectors were negatively impacted by the pandemic than the primary and quaternary sectors.

Ultimately, our project certainly helps answer the question of "How has the pandemic affected different economic sectors in Canada?". To answer this question directly, it is clear that, during the first few months of the pandemic, COVID-19 had a extremely negative affect on all economic sectors in Canada. The secondary and tertiary sectors seemed to be most negatively affected by the pandemic.

Primary Sector

The primary sector involves any industry that harvests, extracts, or produces raw materials and resources, such as farming, mining, fishing, forestry, etc. The relatively small affect of COVID on the primary sector might by explained by the fact that many industries and businesses in the primary sector are located in isolated or rural areas and that there is little human interaction involved. This makes in harder for the pandemic to spread and reduce the impact of restrictions put in place.

Secondary Sector

The secondary sector involves industries that manufactures products or supplies, including construction and energy and was effected the most by the pandemic. This is somewhat surprising as the pandemic does not seem to directly affect the secondary sector compared to many service industries in the tertiary sector. There could be a few reasons why this is the case. Almost all secondary sector industries are located in urban areas, so the workforce could have been affected the virus. Another reason is that changes in supply and demand have disrupted production lines.

Tertiary Sector

In the sectors included in the dataset, the tertiary sector was the largest aggregated sector, including the areas of supply chain, business to customer, financial, technical, and social services. As mentioned in the introduction, lockdown measures prevented many organizations from delivering in-person services, including some restaurants and hotels that only offer in-person services.

Quaternary Sector

The quaternary aggregated sector, which included the fewest number of original sectors from the dataset, is focused on information and cultural industries. Specifically, this area involves the production and distribution of cultural and informational products (except by retail or wholesale methods) (Government of Canada). Because production is involved, COVID-19's effect on it may be similar to the secondary sector.

Statistical Methods

Various statistics can be generated by the program. Some statistics for a sample file of the data from January 2014 to August 2021 (although only data up to May 2021 was used) is shown in the table below. As mentioned previously, the independent "x-value" used in the regression is a sequence of integers from 0 to 91 representing the 92 months from January 2014 to August 2021.

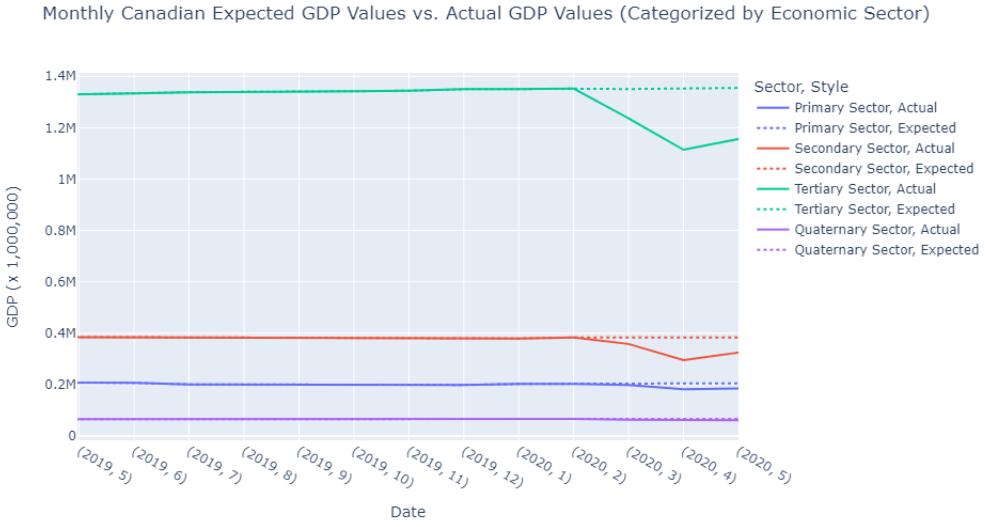
Sector	Slope (10^6 Dollars/Month)	Intercept (10^6 Dollars)
Primary	464.36	169641
Secondary	225.63	366649
Tertiary	2352.62	1176210
Quaternary	150.67	55044

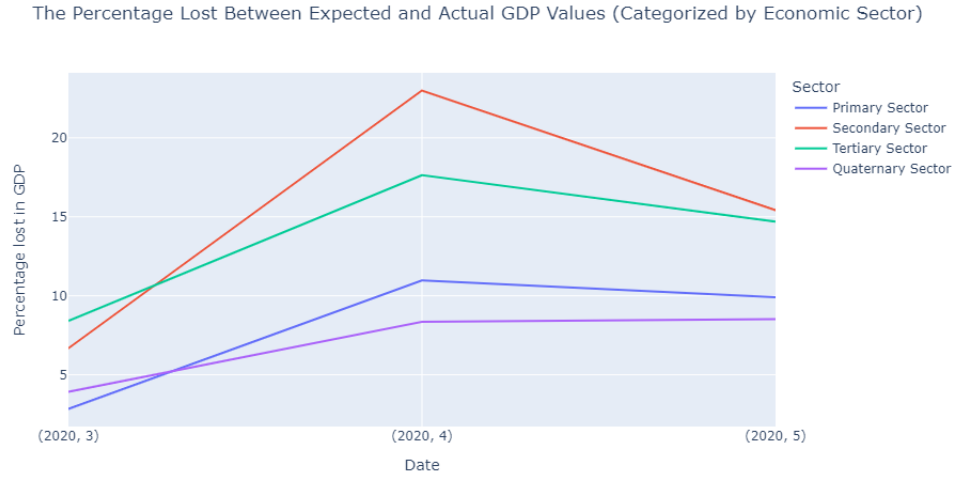
Fig. The coefficients of the lines of best fit for the four sectors.

Further Exploration

With so much data relating to the Canadian economy publicly available, there are many opportunities for further exploration in this topic. For example, a similar project to ours could be developed that explores the affect of the COVID-19 pandemic on various Canadian industries instead of economic sectors, which we explored in this project. Also, an analysis of which Canadian economic sectors experienced the most lasting negative affects as a result of the pandemic could provide an an interesting insight of which economic sectors suffered for the most amount of time.

Graphs





References

Armstrong, Peter. "Where the Lopsided Economic Impact of Covid-19 in Canada Goes from Here — CBC News." CBC News, CBC/Radio Canada, 26 Jan. 2021, https://www.cbc.ca/news/business/covid-coronavirus-economic-recovery-inequality-1.5886384.

Drolet, Mike. "Coronavirus: Covid-19 and the Canadian Economy, One Year Later - National." Global News, Global News, 9 Mar. 2021, https://globalnews.ca/news/7685042/coronavirus-canadian-economy-one-year-later/.

"Gross Domestic Product (GDP) at Basic Prices, by Industry, Monthly (x 1,000,000)." Covid-19, 29 Oct. 2021, https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3610043401. Accessed 4 Nov. 2021.

Kafka, P. (2020, July 16). The pandemic has been great for Netflix. Vox. https://www.vox.com/recode/2020/7/16/21327451/n
covid-earnings-subscribers-q2.

"Linear and Non-Linear Trendlines." Plotly, plotly.com/python/linear-fits/.

"WHO Director-General's Opening Remarks at the Media Briefing on COVID-19 - 11 March 2020." World Health Organization, World Health Organization, www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020.