Canada Emergency Wage Subsidy Exploratory Data Analysis and Interactive Visualization Dashboard
and
API Feasibility Study for the Development of Economic and Social Data Tools

Eric Baxter, Vicens Paneque, Bohan Gao
University of British Columbia
June 2021

Executive Summary

This project was completed by Eric Baxter, Vicens Paneque, and Bohan Gao in collaboration with the Statistics Canada Centre for Special Business Projects. The project includes two distinct components; an exploration of Canada Emergency Wage Subsidy (CEWS) data, and an exploration into the use of APIs for the development of novel data sources.

In the first half of the project, we explore the impact of the CEWS program. CEWS is a subsidy program put in place by the Canadian government to help businesses manage the higher costs associated with operating during the Covid-19 pandemic. By providing wage subsidies, the government's goal was to help businesses maintain and rehire employees during the period of reduced income. Data on the CEWS program is compiled in the CEWS Regional and Community-Level Database, which provides geographic and industry breakdowns of the money given out in the CEWS program. In a data exploratoration done in Python, we examined the effects of geographic regions and industry on given subsidies. We primarily focused on developing insights into the effects of the program on rural Canada, and reported on some of the more interesting results. One of our primary findings was that regions in rural Alberta received disproportionately large subsidies, primarily in the hospitality sector and the oil and gas sector. Furthermore, an interactive visualization dashboard was developed in Microsoft PowerBI to allow users to visually explore the CEWS database. This dashboard includes a number of tools for exploring all dimensions of the data.

During the second half of the project, an inventory of application-programming interfaces (APIs) was taken, with the intent of researching novel sources of economic and social data to be explored at Statistics Canada. Eighteen APIs were researched and reported on. We found that there are major limitations to the useability of most APIs as economic data sources, due to the inaccessibility of aggregate data and the need for user authentication. Of our researched APIs, we considered four APIs "feasible" (usable for our goals), five APIs "somewhat feasible" (limited use), and 9 APIs "practically infeasible" (no obvious use). The most promising use of APIs we found was in social analytics. Through some APIs, we were able to retrieve user posts and comments. Using natural language processing and sentiment analysis methods, we were then able to parse these messages and gain insight into public opinion on certain topics. Preliminary software was developed for three of these APIs; Twitter, Steam, and Reddit. These software are intended as simple test cases to showcase some of what could be done with these APIs, and significant work would be needed to develop the ideas into a usable source of data.

Introduction

This report covers two somewhat distinct projects. The first half is an exploratory data analysis of Statistics Canada's Canada Emergency Wage Subsidy Regional and Community Level Database. The second half is a feasibility study of application-programming interfaces as a source of useful economic and social data.

Canada Emergency Wage Subsidy

The Canada Emergency Wage Subsidy (CEWS) program was introduced as part of Canada's economic response to the Covid-19 pandemic[1] ("Overview of Canada's COVID-19 Economic Response Plan"). It is intended to help businesses cover some of the additional costs due to the pandemic, allowing them to maintain and rehire laid-off workers. Data on the CEWS program is publicly available through the CEWS Regional and Community-level database, which can be found and downloaded on the Statistics Canada website[2]. While this data is public, there are currently limited analyses and no interactive visualization tools available to help the general public better understand the data. In order to make the data more accessible, we have developed an exploratory data analysis of the CEWS database, along with an interactive visualization dashboard.

The goal of the CEWS report was to explore the CEWS Regional and Community-level database across all available dimensions. This database provides information on CEWS subsidies given (and businesses / employees affected), aggregated by industry and geographic areas. Data is given for each of the first 8 "claim periods" of the CEWS program, ranging from March 15 through October 24 2020. The structure and granularity of the dataset allowed for nuanced analysis Particular focus is placed on studying the data for rural regions of Canada, and examining the areas of Canada which received the largest CEWS subsidies. The massive scope and potential of the data forced us to limit our analysis to a handful of topics. While we did not have any specific research questions in mind while we began our analysis, we looked for interesting stories the data could tell.

We began our analysis with high-level analysis of the program by examining the total subsidies given within each industry and within each province. Through this analysis, we learned that certain industries and regions received particularly large subsidies, and chose these topics as the next sources of our analysis. Our most granular analysis involved researching specific industries within rural Alberta.

---

[1] "Overview of Canada's COVID-19 Economic Response Plan," Government of Canada, March 13, 2020, https://www.canada.ca/en/department-finance/services/publications/economic-fiscal-snapshot/overview-economic-response-plan.html

[2] "Canada Emergency Wage Subsidy Regional and Community-Level Database," Statistics Canada, March 18, 2021, https://www150.statcan.gc.ca/n1/pub/11-26-0003/112600032021001-eng.htm.

Our interactive visualization dashboard was designed to allow users to visually explore the CEWS data to find insights of their own. The dashboard is designed to be flexible, and allow for any analysis possible within the scope of the data.

API Feasibility Study

Statistics Canada has historically been reliant on traditional sources of data, including surveys and administrative data. Recently, the agency has been making significant efforts to leverage some of the more technical data sources available in modern times[3]. In this section of the project, we attempted to research the capabilities of application-programming interfaces (APIs) as potential sources of useful data, in particular with regard to studying the economic and social effects of the Covid-19 pandemic. Through APIs, we are able to directly query and retrieve data from various applications. We set out to determine if any of the data accessible through APIs could be valuable to Statistics Canada. We began by researching and studying the documentation of many available APIs. We developed an inventory report outlining the capabilities of some of the APIs we thought may be useful for our goals. For each API, we outlined what they could be used for, the associated costs, and the ease of access.

Using three of the APIs we considered to be feasible, we developed some preliminary software samples. Using natural language processing and sentiment analysis techniques, we attempted to retrieve valuable social data on various topics of interest, including the CEWS program.

Part 1 of this paper contains an explanation of the CEWS components of the project, with the full research paper included as an appendix. In part 2 we discuss our findings on APIs, as well as detail the work we did developing prototype software for API data analysis. Our full API inventory is included as an appendix. For both parts, we have included some background information, a discussion of our methodology, and our results and conclusions.

Part 1. Canada Emergency Wage Subsidy

1.1. Background

1.1.1. Overview of Canada's COVID-19 Economic Response Plan

The COVID-19 pandemic is one of the most serious challenges humanity has faced in recent times. Countries throughout the world are experiencing both a profound health crisis, and an economic collapse that will severely impact the well-being of large segments of the population for the

---

[3] "Statistics Canada Data Strategy," Statistics Canada, March 13, 2020,
https://www.statcan.gc.ca/eng/about/datastrategy.

coming years[4]. Since the World Health Organization declared the COVID-19 outbreak a pandemic in March 2020, most businesses and companies closed temporarily, countries reduced the flows of goods and services, and many countries closed borders to try to avoid contagion[5]. Canada implemented physical and social distancing, temporary restrictions on non-essential businesses, and school closures policies to mitigate the spread of COVID-19. However, these policies also resulted in job losses and lower employment income, having a significant economic impact on the daily lives of Canadians[6].

To alleviate the economic impact of the pandemic, the government of Canada implemented a wide range of financial and monetary measures through Canada's COVID-19 Economic Response Plan (the Plan). By focusing on three areas: *Health Care System and Safety of Canadians*, *Individuals*, and *Businesses*, the Plan used a mix of existing systems and the design and delivery of new programs to provide support to businesses of all sizes across all industry sectors, protect jobs, and help Canadians return to work[7].

Within the Plan, the government introduced CEWS to financially support businesses with the goal of keeping and re-hiring employees. This helped the economy by replacing lost income at a household level, and by preventing higher unemployment rates[8].

1.1.2. Canada Emergency Wage Subsidy (CEWS)

Beginning March 15, 2020, CEWS offers a wage subsidy of 75 per cent, up to $847 per week per employee, for qualifying employers to help cover the costs of maintaining and rehiring workers through periods of reduced profitability during the COVID-19 pandemic[9].

---

[4] Constantino Hevia and Andy Neumeyer, "Un Marco Conceptual Para Analizar El Impacto Económico Del COVID-19 y Sus Repercusiones En Las Políticas [A Conceptual Framework for Analyzing the Economic Impact of COVID-19 and Its Policy Implications]," PNUD LAC C19 PDS 1 (March 20, 2020), https://www.undp.org/content/dam/rblac/Policy%20Papers%20COVID%2019/UNDP-RBLAC-CD19-PDS-Number1-ES-final.pdf.

[5] Cesaire Chiatchoua, Carmen Lozano, and Jesús Macías-Durán, "Análisis de los efectos del COVID-19 en la economía mexicana," Revista del Centro de Investigación de la Universidad la Salle 14, no. 53 (September 18, 2020): 265–90, https://doi.org/10.26457/recein.v14i53.2683.

[6] "Overview of Canada's COVID-19 Economic Response Plan," Canada Department of Finance, January 6, 2021, https://www.canada.ca/en/department-finance/services/publications/economic-fiscal-snapshot/overview-economic-response-plan.html.

[7] "Overview of Canada's COVID-19 Economic Response Plan," Canada Department of Finance, January 6, 2021, https://www.canada.ca/en/department-finance/services/publications/economic-fiscal-snapshot/overview-economic-response-plan.html.

[8] "Economic and Fiscal Report Summary," Department of Finance, July 8, 2020, https://www.canada.ca/en/department-finance/services/publications/economic-fiscal-snapshot/summary.html.

[9] "Economic and Fiscal Report Summary," Department of Finance, July 8, 2020, https://www.canada.ca/en/department-finance/services/publications/economic-fiscal-snapshot/summary.html.

According to the Government of Canada CEWS website[10], with a few exceptions, the subsidy is available to businesses of all sizes and from all sectors of the economy from every geography in Canada. To qualify for the subsidy, eligible businesses required an experienced decrease of at least 15 per cent in revenue March 2020, or 30 per cent in April or May 2020.

CEWS is broken down into sixteen 4-week periods with different calculation details and qualifying criteria that are summarized in next bullet-points:

- March 15 - July 4, 2020: To qualify, eligible employers should have experienced a 15 percent decrease in revenue in March 2020 and 30 per cent for the following months of April, May, and June, when compared to their revenue for the same period in 2019.
- July 5 - December 19, 2020: CEWS was available for all eligible employers that experienced a decline in revenue for a claim period. A larger subsidy was available for those employers that were most adversely affected by the COVID-19 crisis.
- December 20, 2020 - March 13, 2021: The rules remained essentially the same as the previous period.
- March 13 - June 5, 2021: The rules remained essentially the same as the previous period, with the addition of an elective alternative baseline remuneration period.

As per the Government of Canada Economic and Fiscal Profile[11], to date, CEWS has provided support by allowing millions of Canadian workers to continue to receive their wages and are ready to return to work as soon as public health measures allow. This program also ensures that supply chains can recover from the crisis in a strong position by safeguarding the employee-employer relation. CEWS has also been significant in increasing consumer spending. By providing affected Canadians with a source of income, CEWS has given many Canadians the confidence to spend money, which would otherwise be kept in savings.

When first launched, the program was intended to last 12 weeks. However, the Government of Canada has kept CEWS to June 30, 2021, and proposed in its Budget 2021[12] to extend the wage subsidy until September 25, 2021. If this extension comes to pass, there will be a gradual decrease in the subsidy rate to ensure a smooth phase-out of the program as vaccinations are completed and the economy reopens.

---

[10] "Canada Emergency Wage Subsidy (CEWS)," Canada Revenue Agency, September 18, 2020, https://www.canada.ca/en/revenue-agency/services/subsidy/emergency-wage-subsidy.html.
[11] "Overview of Canada's COVID-19 Economic Response Plan," Department of Finance, January 6, 2021, https://www.canada.ca/en/department-finance/services/publications/economic-fiscal-snapshot/overview-economic-response-plan.html.
[12] "Budget 2021," Department of Finance, April 19, 2021, https://www.budget.gc.ca/2021/report-rapport/p1-en.html#17.

1.2. Tools, Methodology and Techniques

1.2.1. Data Sources

1.2.1.1. Canada Emergency Wage Subsidy Regional and Community-level Database

The Canada Emergency Wage Subsidy Regional and Community-level Database is "a custom dataset constructed with CEWS microdata and other administrative data sources available at Statistics Canada"[13]. It is publicly available on the Statistics Canada website[14]. This dataset contains information on every subsidy given out through the CEWS program, aggregated to different levels of geography and industry. Geographic aggregation is available at the levels of country, province, census metropolitan area/ census agglomeration (CMA/CA), census subdivision (CSD), and urban/rural splits by province/ canadawide. These geographic classifications were done according to the Standard Geographic Classification 2016 (SGC-2016)[15]. This system provides a hierarchical code to each region, with provinces receiving a 2-digit code, census agglomerations receiving a 5-digit code, and so on. Industry aggregation is available at the level of all industries, as well as by specific industry according to the 2-or-3-digit North American Industry Classification System (NAICS)[16]. 2-digit NAICS are general industries, with 3-digit NAICS more specific subdivisions of these. As such, the most general information available is the total subsidies given in Canada, and the most specific information available is the subsidies given in a particular CSD, within a particular 3-digit NAICS industry. Each row contains data for one claim period. In each row, data is given for the total CEWS supported employees, total supported businesses, and total CEWS claim dollar amounts.

The database contains data for the first 8 four-week claim periods of the CEWS program, between March 15 and October 24 2020. The file contains all approved CEWS claims in these periods as of January 4 2021. It is important to note that as of the time of database creation, businesses still had time to report claims for the last two periods (September and October 2020), and so these periods may contain incomplete data.

Given the high level of granularity provided in the database, data suppression was implemented to preserve confidentiality. Data suppression occurred due to one of two reasons:

- Primary suppression: There are too few businesses contributing to the value of a cell, or one business dominates the value of the cell.

---

[13] "Canada Emergency Wage Subsidy Regional and Community-Level Database," Statistics Canada, March 18, 2021, https://www150.statcan.gc.ca/n1/pub/11-26-0003/112600032021001-eng.htm.
[14] A download link can be found at the bottom of the metadata page.
[15] "Standard Geographical Classification (SGC) 2016," Statistics Canada, November 10, 2016, https://www.statcan.gc.ca/eng/subjects/standard/sgc/2016/index.
[16] "North American Industry Classification System (NAICS) Version 3.0," Statistics Canada, August 17, 2018, https://www23.statcan.gc.ca/imdb/p3VD.pl?Function=getVD&TVD=1181553.

- Secondary suppression: Another cell is suppressed to protect the value of other primary suppressed data during the publication of a higher-level aggregate. For example, if only two cells make up the higher level aggregate and one is primarily suppressed, its value could be found using the value of the aggregate and the value of the other cell.

There is a planned update to extend the data for claim periods up until the end of the CEWS program.

1.2.1.2. Additional Data Sources

In order to allow for per-capita comparisons of different regions and industries, we sourced and incorporated data on within-industry worker counts at the Canada, provincial, CMA/CA, and CSD levels, as of the 2016 census[17], as well as by rural/urban split by province[18].

1.2.2. Tools

Data wrangling, exploration, and visualization for the CEWS report were done in Python. *Pandas* and *Numpy* were used for data structures and wrangling. *Altair* and *Matplotlib* were used for visualizations. The interactive dashboard was created using *Microsoft PowerBI* version 2.93.384.0. The interactive maps in the dashboard were made using *ArcGIS* and *Shape Map*. We considered using alternative, programming-based dashboard tools, but *PowerBI* allowed for easier integration into the Statistics Canada website.

1.2.3. Data Exploration and Wrangling

The first step was to develop an understanding of the structure of the CEWS database, and wrangle it into a format that would be easily usable in both Python and PowerBI. We performed data exploration in the first week of the project and completed most of the data cleaning and wrangling by the second week.

We began with some basic data cleaning by renaming columns for ease of use with our programming language of choice, Python . Suppressed data in the CEWS dataset was represented by an "X", which we changed to a NumPy null datatype (*np.nan)*, which allowed us to then convert the data type of the columns to numeric. We also ensured that "claim period", an attribute indicating a specific four-week period, was of datetime data type.

The multidimensional, hierarchical structure of the database was a challenge. Different rows contained information at different levels of aggregation, both in terms of geography and in terms of

---

[17] This data was sourced from a number of datasets by Statistics Canada, and gives numbers as of the 2016 Census. Statistics Canada, 2016 Census, Catalogue no. 98-400-X2016320. Statistics Canada, 2016 Census, Catalogue no. 98-400-X2016321.

[18] Data from the 2016 census, extrapolated to estimated 2020 numbers. Statistics Canada. Table 14-10-0377-01 Employment by class of worker and industry, annual (x 1,000).

industry. As such, some rows would contain, for example, data for "all industries" throughout all of Ontario, while other rows would contain data for one specific industry in one city in Ontario. Therefore, any sort of naive aggregation would lead to individual subsidies being counted multiple times. Typically, data is presented in this format to be more human-readable, by allowing users to easily see the specific level of aggregation they are interested in. Data in this form would typically be cleaned for machine use by only including the rows at the most granular level of aggregation, and using grouping operations to reconstruct the aggregate data. However, due to the prevalence of suppressed data in the CEWS database, especially in the most granular rows, we were not able to do this without substantial information loss.

By using regular-expressions (regex) and the hierarchical SGC-2016 and NAICS codes, we were able to augment the dataset with two additional columns. These columns described the level of aggregation of each row, one for geography and the other for industry. So for example, the row describing subsidies given to "accommodation and food services" (NAICS code 72) within "Alberta" (SGC-2016 code 48) would therefore receive values in these columns of "Industry Sector (2-digit NAICS)", and "Province". In this way we solved the multi-counting problem without information loss. By querying a specific level for each of these two "aggregation-level" columns, we ensured that all relevant subsidies were only counted once, and that information loss was minimized by using the most aggregated data possible. This was done by using regex to match all possible cases of geographic and industry aggregation. 2-digit SGC-2016 codes always correspond to a province, 5-digit codes to a CMA/CA, etc. Along the dimension of industry, the data was partitioned into 3 levels of aggregation ("All industries, "Level 1 (2-digit NAICS)", and "Level 2 (3-digit NAICS)"). Along the dimension of geography, the data was partitioned into 7 levels of aggregation ("Canada", "Province", "CMA/CA", "CSD", "Urban/rural by province", "urban/rural (Canadawide)", and "Unknown rural[19]". This work can be found in the file EricWrangling.ipynb.

Further wrangling was carried out later to further augment the dataset. Using regex and the SGC-2016 codes, a "Province" column was added. This allowed us to color-code individual census subdivisions by province, and allowed a new dimension of exploration on the interactive dashboard. We also augmented our dataset with information on the number of workers within regions and industries[20], which allowed for normalized comparisons between regions and industries of varying size. We were able to source data at the "All industries" and "2-digit NAICS" industry aggregation levels, and at the "Canada", "Province", "CMA/CA", and "CSD" levels of geographic aggregation, however the "2-digit NAICS" breakdown of these data proved to use slightly different definitions than

---

[19] "Unknown rural" corresponded to three regions in the territories which had 5-digit NAICS codes (typically representative of CMA/CA's), but were described as being rural. As a region cannot be both rural and a CMA/CA, we considered these anomalous regions as their own group and did not include them in analyses.
[20] This data was sourced from a number of datasets by Statistics Canada, and gives numbers as of the 2016 Census. Statistics Canada, 2016 Census, Catalogue no. 98-400-X2016320. Statistics Canada, 2016 Census, Catalogue no. 98-400-X2016321.

are found in the CEWS database, and were therefore unreliable. This work is found in cews_wrangling_v2.ipynb.

One further wrangling file was created, which made minor changes to the terminology found in the data file in order to make the interactive dashboard appear more professional, polished, and user-friendly. This work is found in cews_wrangling_for_powerbi.ipynb. A few minor terminology changes were then done in PowerBI itself.

	Period	RegionCode	Region	GeographicClassification	CensusLevel	IndustryCode	Industry	BusinessLocations	Subsidy	SupportedEmployees	RehiredEmployees	GeoAggregation	IndustryAggregation	ProvinceCode	Province	TotalEmployees
475914	2020-08-30	3543017	Innisfil	URBAN	Not applicable	713	Amusement, gambling and recreation industries	5.0	164000.0	141.0	0.0	CSD	Industry Subsector (3 Digit NIACS)	35.0	Ontario	NaN
457807	2020-08-30	2452055	Saint-Barthélemy	RURAL	Not applicable	44-45	Retail trade	0.0	NaN	NaN	0.0	CSD	Industry Sector (2 Digit NIACS)	24.0	Quebec	80.0
483975	2020-08-30	4705034	Whitewood	RURAL	Not applicable	419	Business-to-business electronic markets, and a...	0.0	NaN	NaN	0.0	CSD	Industry Subsector (3 Digit NIACS)	47.0	Saskatchewan	NaN
107898	2020-04-12	3518017	Clarington	URBAN	Not applicable	451	Sporting goods, hobby, book and music stores	0.0	NaN	NaN	0.0	CSD	Industry Subsector (3 Digit NIACS)	35.0	Ontario	NaN
354433	2020-07-05	4811052	Strathcona County	URBAN	Not applicable	52	Finance and insurance	10.0	NaN	NaN	0.0	CSD	Industry Sector (2 Digit NIACS)	48.0	Alberta	1800.0
458943	2020-08-30	2457040	Beloeil	URBAN	Not applicable	416	Building material and supplies merchant wholes...	0.0	NaN	NaN	0.0	CSD	Industry Subsector (3 Digit NIACS)	24.0	Quebec	NaN
422365	2020-08-02	4806021	Airdrie	URBAN	Not applicable	416	Building material and supplies merchant wholes...	0.0	NaN	NaN	0.0	CSD	Industry Subsector (3 Digit NIACS)	48.0	Alberta	NaN
498815	2020-08-30	5935802	Tsinstikeptum 9	URBAN	Not applicable	621	Ambulatory health care services	0.0	NaN	NaN	0.0	CSD	Industry Subsector (3 Digit NIACS)	59.0	British Columbia	NaN
516878	2020-09-27	2437067	Trois-Rivières	URBAN	Not applicable	517	Telecommunications	10.0	NaN	NaN	0.0	CSD	Industry Subsector (3 Digit NIACS)	24.0	Quebec	NaN
491812	2020-08-30	4817029	Slave Lake	RURAL	Not applicable	441	Motor vehicle and parts dealers	5.0	NaN	NaN	0.0	CSD	Industry Subsector (3 Digit NIACS)	48.0	Alberta	NaN

Figure 1: Sample from the final wrangled version of the CEWS dataset used in the report and dashboard.

1.2.4. Methodology - CEWS Report

The CEWS report was written during the 2nd, 3rd, and 4th weeks of the project. During the initial stages of exploration, the focus was to investigate the geographic and industry dimensions independently. In the first few iterations of the exploratory analysis, we examined high-level, general trends (for example, total subsidies given to each province). However, Statistics Canada had already done most of these analyses themselves, and over the course of a number of revisions and meetings with them, we developed a handful of more nuanced and interesting narratives to explore. Some of the topics we explored are the following:

- Within the CMA/CA's or CSDs which received the largest per-capita subsidies, where did that money go? Which industries are prevalent in these areas that lead to the necessity of large subsidies?
- How do the subsidies given differ between the urban and rural parts of Canada? What are the urban and rural regions in each province/territory that received the most CEWS subsidies?
- Are all heavily-subsidized rural areas subsidized in the same industries, or are there multiple distinct groups?
- Are there any provinces or territories with a disproportionate number of highly subsidized census subdivisions? Do these CSDs all have similar industry breakdowns, or are they distinct?

The intent of the report is to be accessible to general audiences (high school education), and as such simple methodology was used. Complex statistical methods were not used, and visualizations

were kept simple and easily understandable (bar charts, pie charts, and line charts). Analyses were kept simple. We reported on general trends and differences, and attempted to make any finding clear with a couple sentences and a visualization. The report was designed to be modular, so any section could be used independently as a short "daily report" on the Statistics Canada website.

1.2.5. Methodology - Interactive Dashboard

Work on the GIS component of the dashboard was started in week 1, as it was the most technical part of the dashboard project and required a significant time investment. ArcGIS extensions and the Shape Map tool for PowerBI were used to develop interactive heat maps, allowing users to visually explore the distribution of CEWS subsidies throughout Canada, at varying levels of geographic aggregation. Since the official shape files are already available in ArcGIS, we were able to draw the geographical heat map once we upload the subsidy distribution and the supported employee data with different geographical aggregation to PowerBI. However, due to the access limitations on PowerBI, we were unable to change the projection style to orthographic, which our client prefers. Therefore, on the cover page of our dashboard, instead of using the embedded ArcGIS, we used the shape map function in PowerBI (see Figure 2). Compared to ArcGIS, the shape map function allows us to manually upload the boundary files that enable preferred projection style, although we lost all the appealing labels and legends.

Once the CEWS report was complete in week 4, more of the team began working on the dashboard. Over the next two weeks, additional dashboard pages were developed. The first allows users to compare subsidies, affected businesses, or affected employees between any two geographic regions by breaking down the total subsidies given to each by industry.

In all dashboard pages, functionality was built to allow users to choose the levels of geographic and industry aggregation they are interested in (using SGC-2016 for geography, and NAICS for industry). Users can also use any of the four dependent variables as their metric; subsidy amount ($), supported employees, supported businesses, or rehired employees.

1.3. Results, Analysis, Interpretation, and Conclusions

1.3.1. CEWS Report

1.3.1.1. Conclusions

The full exploratory report is included as **Appendix A**. Some of the most interesting results and conclusions are:

- CMA/CAs which received the largest subsidies per-capita received their largest subsidies in the "mining, quarrying, and oil and gas extraction" industry. We conclude that this industry

was heavily affected by the pandemic, and required a lot of additional money in subsidies to remain operational. Areas that are very dependent on this industry particularly struggled during the pandemic.

- Rural Alberta received disproportionately large CEWS subsidies. The majority of the most subsidized rural CSDs (both per-capita and total) are in Alberta.

- Heavily-subsidized rural CSDs in Alberta are typically dependent on either oil and gas or tourism, and one of those two industries received the largest CEWS subsidy in every one of the 13 most subsidized CSDs in rural Alberta.

1.3.1.2. Setbacks and Limitations

The main limitation in our exploratory analysis was the suppressed data. Some industries had almost all subsidies suppressed (for example, the utilities industry was totally suppressed in almost every region). In small CSDs especially, there are often only one or two businesses within an industry sector or subsector, forcing the data to be suppressed. As such, the analysis we were able to do at the most granular levels was severely limited, and there is significant error introduced into the analysis we have done. We avoided looking at the industry subsectors (3-digit NAICS) entirely, as there was a lot of suppression in this data for most regions.

Furthermore, the dataset is incomplete. The final two claim periods present in this data do not contain complete data, which introduced systematic error into our analyses (especially when comparing claim periods). This work would also be far more useful if we were to analyze the totality of the CEWS program, instead of just the first 8 claim periods. While we had initially intended to update the analysis when the updated CEWS data was released, delays in the release of this data rendered this impossible.

1.3.1.3. Next Steps

The most immediate next step would be to expand the findings to include data from the full duration of the CEWS program, once that data is released publicly. Currently, the report only includes data until October 24 2020, but CEWS has been ongoing throughout the entire duration of the pandemic. We attempted to write all of our code such that the final dataset can be fed in once it is released, and will produce the complete visuals.

There are also many topics left to explore in the CEWS dataset. We focused our granular analysis on a handful of industries and locations (oil and gas and tourism, rural Alberta), and there is much more insight that could be extracted by looking at other regions and industries. We attempted to develop our interactive dashboard in such a way that inquisitive users could replicate much of our analysis for other industries and regions of their interest.

1.3.2. Interactive Dashboard

1.3.2.1. Conclusions

In total, we released 7 dashboard pages. The next 7 figures are screenshots of these pages.
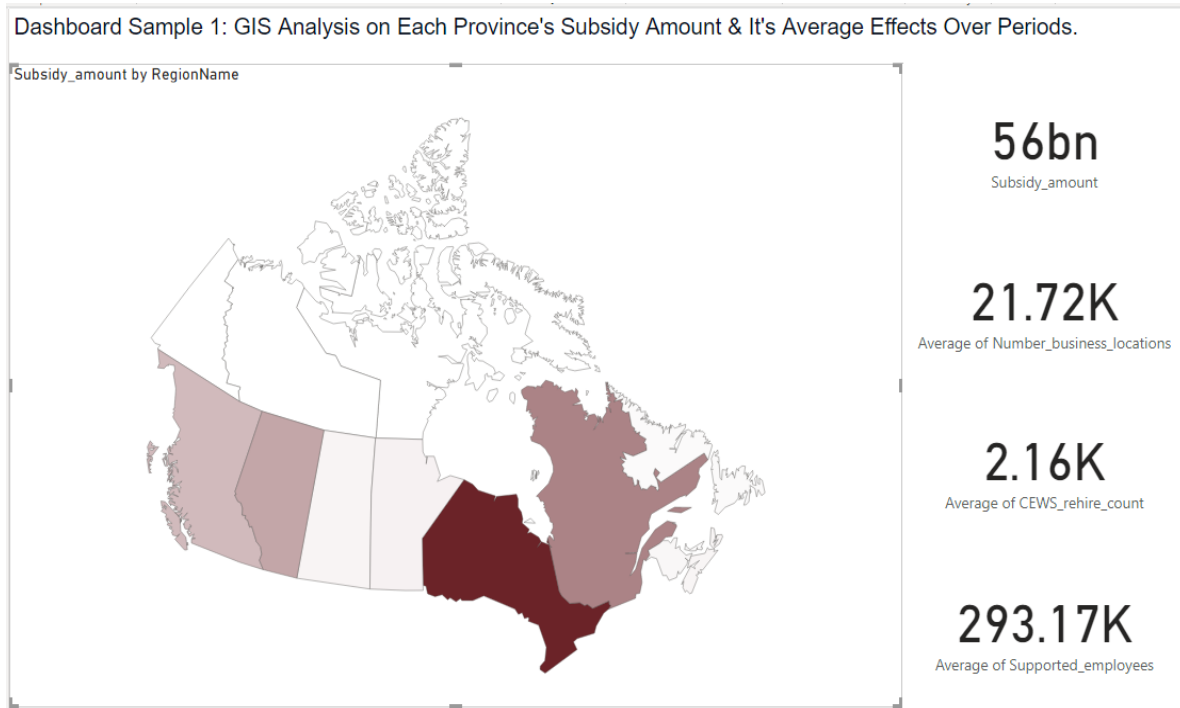

Figure 2: This visualization is a cover page of our dashboard, and it demonstrates the subsidy amount by Canadian provinces. (created using the shape map function in power BI)
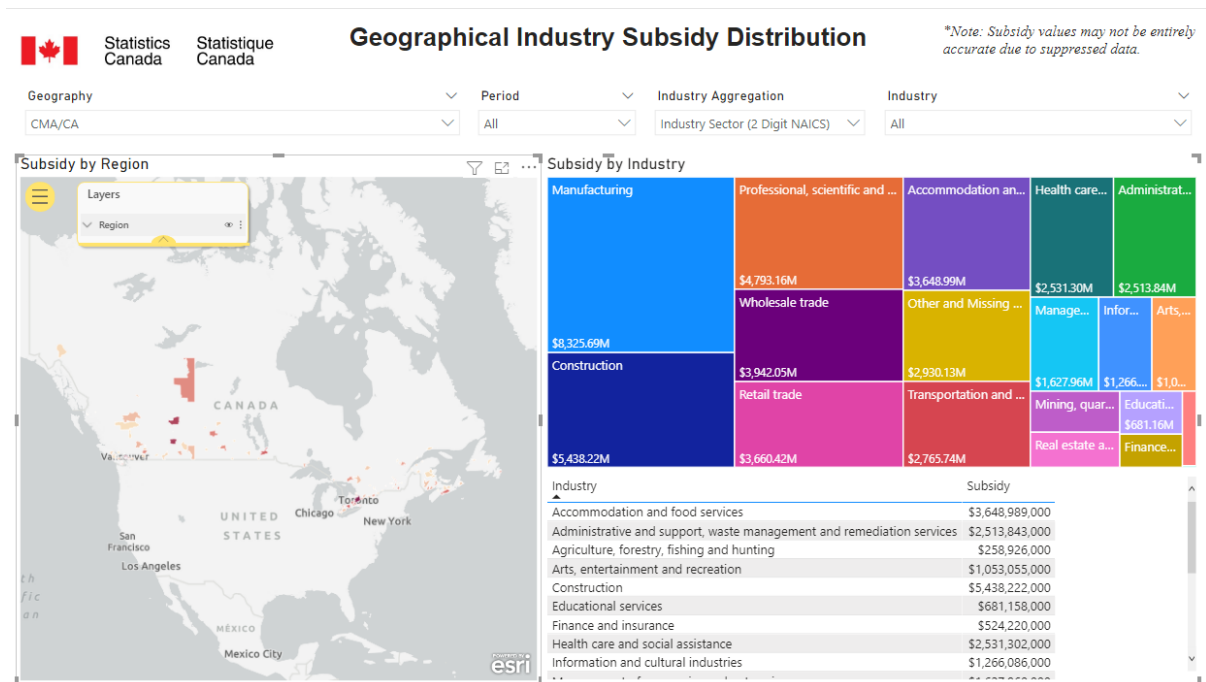
Figure 3: This dashboard explains how the subsidy is distributed among industry sectors in each region. (Created using ArcGIS)
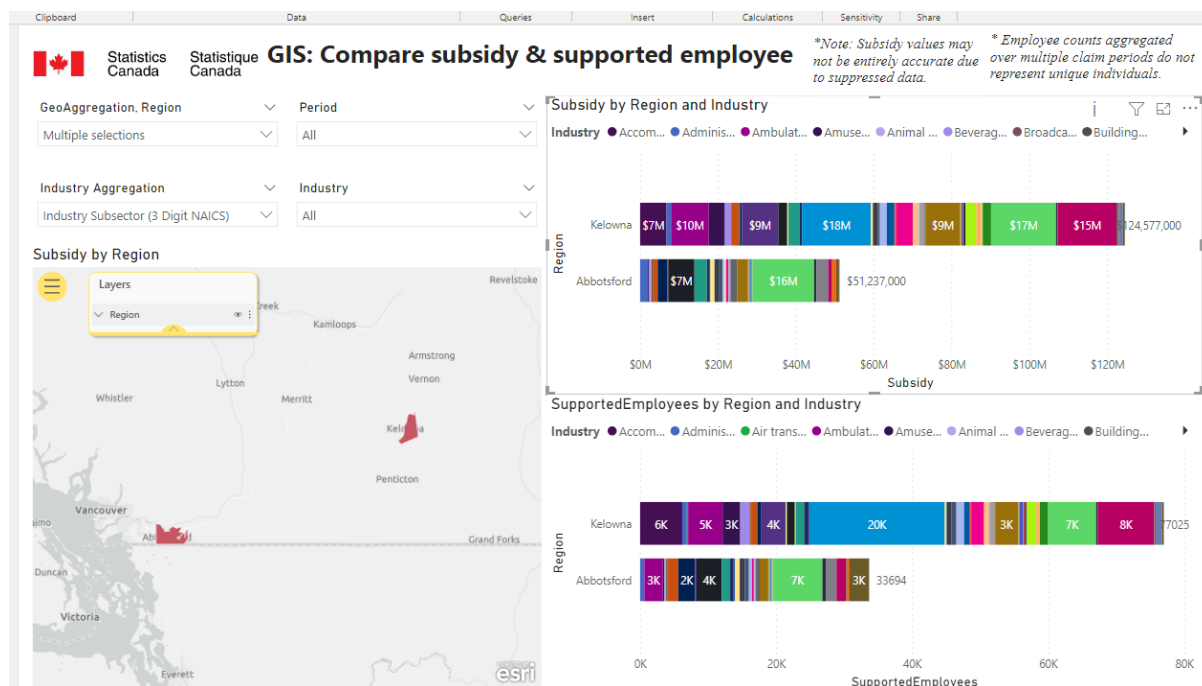

Figure 4: This dashboard compares the amount of subsidy and the number of supported employees in each industry sector for two or more selected regions. (Created using ArcGIS)
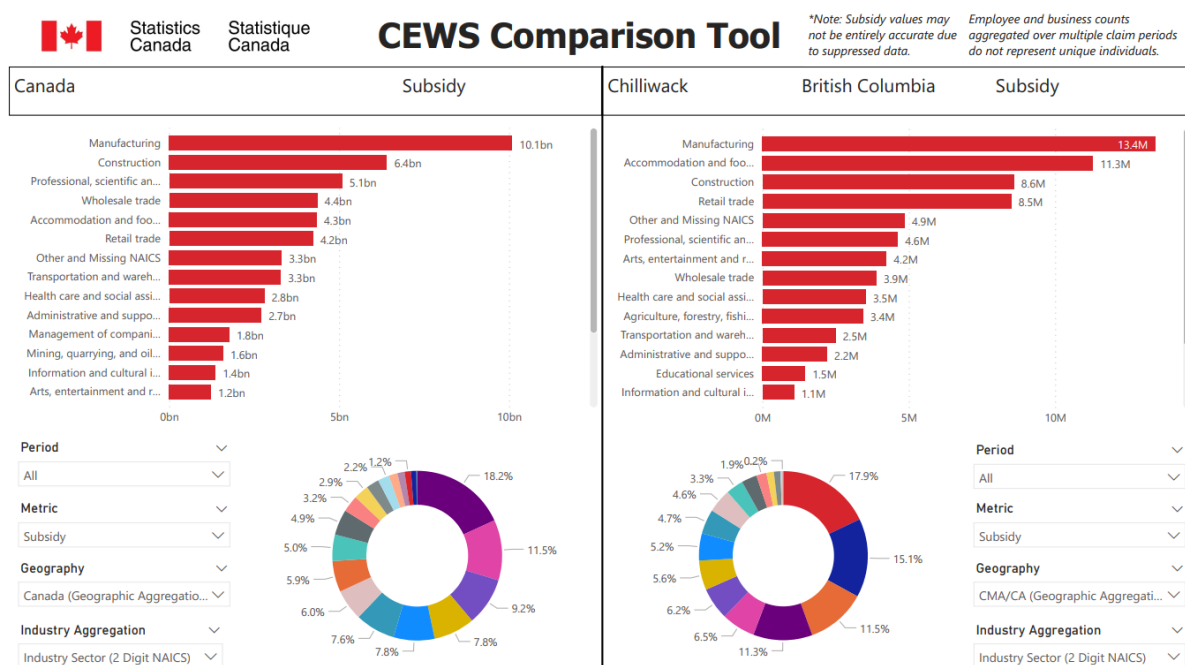

Figure 5: CEWS comparison tool, allowing the comparison of industry subsidies between any two Canadian regions.
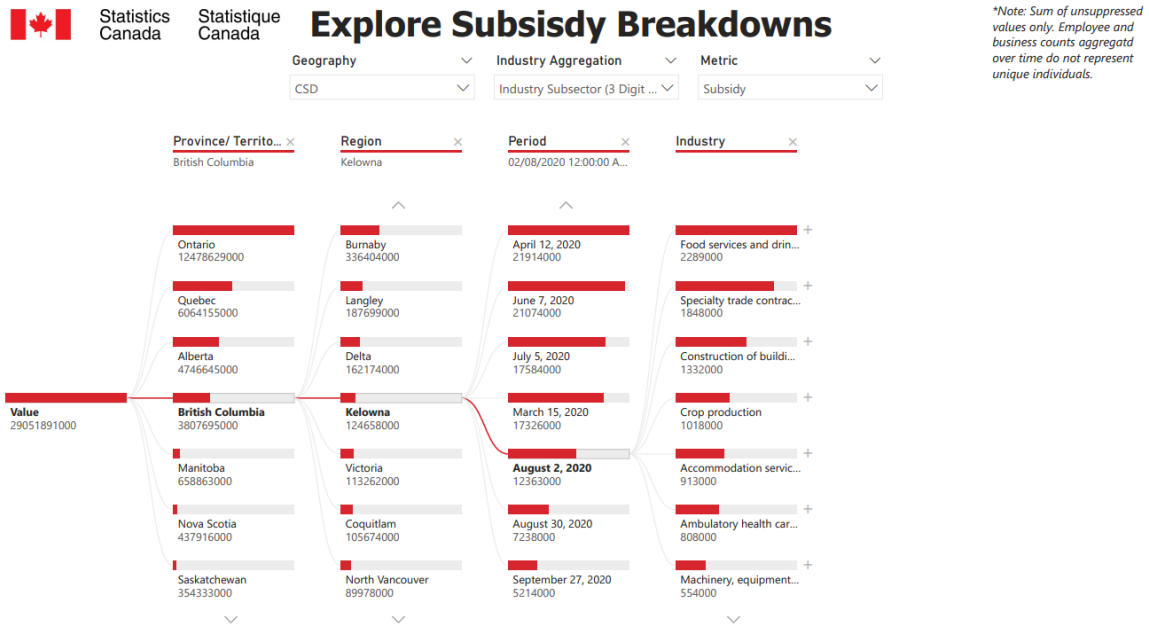
Figure 6: Subsidy breakdown tool, allowing users to subset total CEWS subsidies by any dimension of interest.
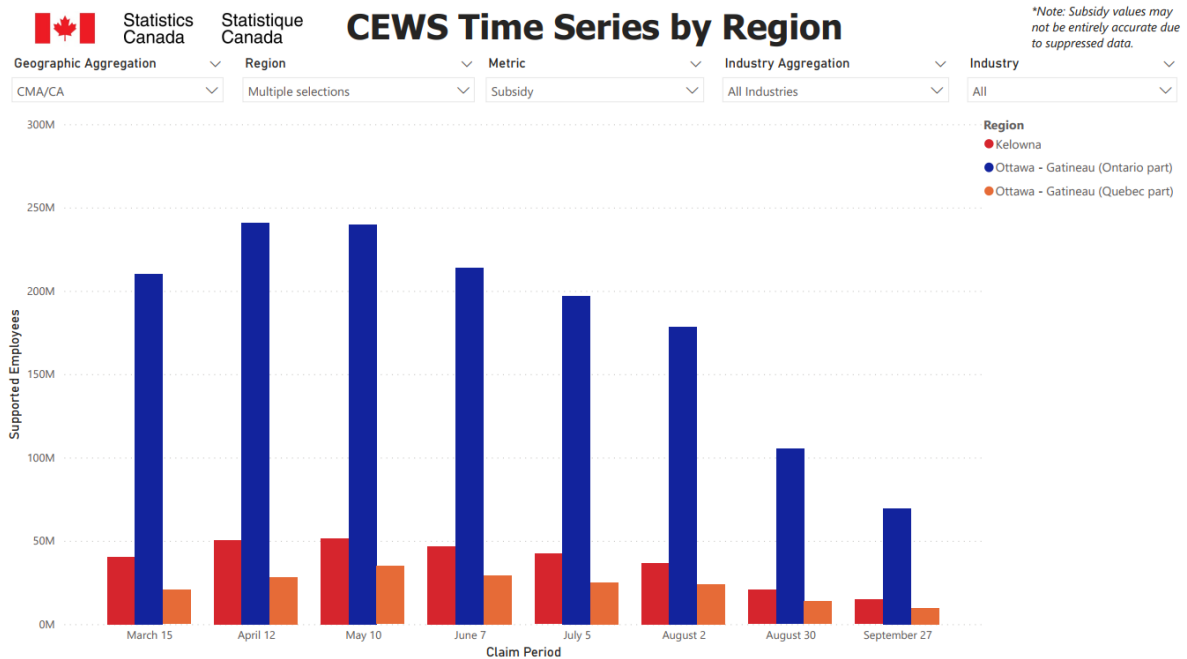

Figure 7: Time series breakdown tool, aggregated by industry and split by region.
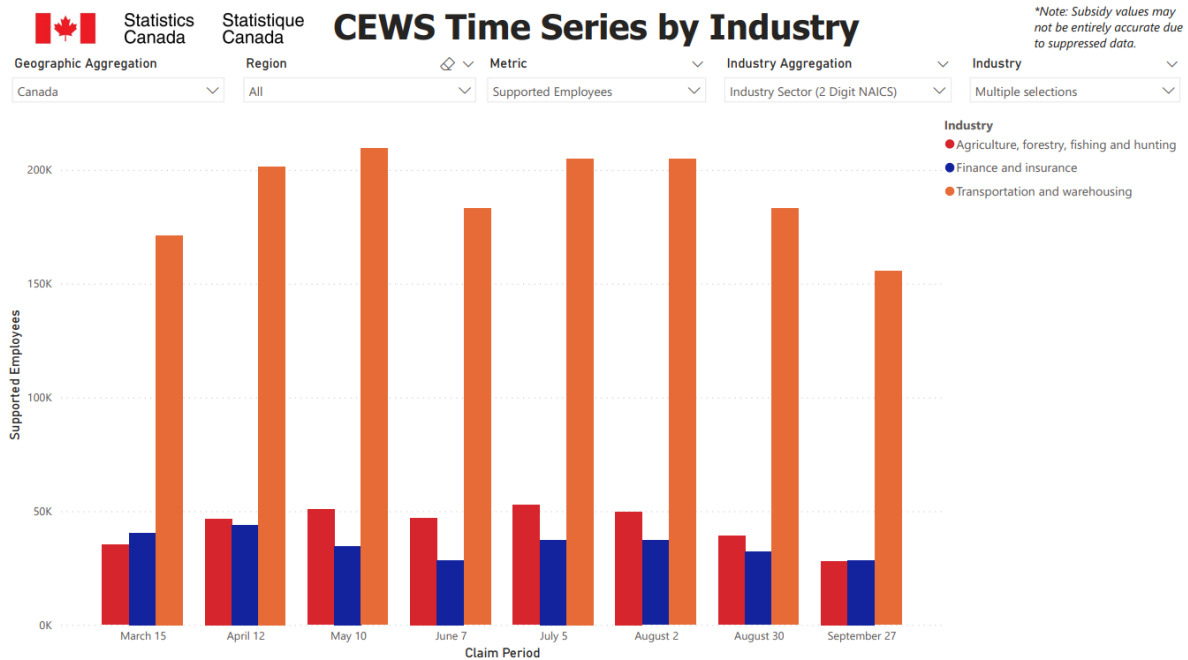
Figure 8: Time series breakdown tool, aggregated by region and split by industry.

We deployed all pages online using the PowerBI website[21][22]. Unfortunately the GIS maps do not show on the online version, due to institutional limitations imposed on our PowerBI account.

1.3.2.2. Setbacks and Limitations

There were a number of issues encountered during this part of the project. First, there is no PowerBI software available for Apple computers. Our teammate who uses an Apple computer was therefore unable to work on our dashboard. While we experimented with virtual machines, we decided it would be more time efficient for him to simply move on to the API study and leave the dashboard to the rest of the team. Second, we were not granted a PowerBI license through either Statistics Canada or UBC. While we could use the desktop version without a license, we were locked out of a number of premium features. Many useful GIS features were unavailable, and the online version of PowerBI was unavailable, along with the deployment and sharing tools included with that version. Fortunately, a member of our team was able to procure a premium trial through his undergraduate university, the University of Ottawa. Strangely, the institution disabled the use of ArcGIS maps, which prevented the full functionality of our GIS maps on our web-deployed dashboard. Finally, the version of PowerBI

---

[21] The GIS pages can be viewed at
https://app.powerbi.com/view?r=eyJrIjoiN2NlNDRjMWEtYTU2OS00ODE5LWE2YmMtZTI2ZGVkNWZjNTI4IiwidC I6ImQ0MWZkYWIxLTdlMTUtNGNmZC1iNWZhLTcyMDBlNTRkZWI2YiJ9&pageName=ReportSection1a2d35faa52 a0c3db256

[22] The graph pages can be viewed at
https://app.powerbi.com/view?r=eyJrIjoiOWZkNzZlODItZDRiNC00ZWYzLWJlMWEtY2MyMTBjOTI0NzIzIiwidCI6I mQ0MWZkYWIxLTdlMTUtNGNmZC1iNWZhLTcyMDBlNTRkZWI2YiJ9&pageName=ReportSection14b8a5c54e363 d2aa0ad

used at StatCan is 2 years outdated. We developed our dashboards on modern software, which we later discovered was not backward compatible with the version used at Statistics Canada. Statistics Canada was therefore unable to open the .pbix files. We are unable to make the dashboards backward-compatible without redoing the entire project on the older version of PowerBI used at Statistics Canada. We are currently working with the IT team at Statistics Canada to come up with a solution.

1.3.2.3. Next Steps

The dashboard currently only uses the overall numbers, and cannot visualize the per-worker trends we reported in our paper. Due to the flexibility of the dashboard, we would have to ensure that the presented data is accurate, regardless of the analysis done. Since the per-worker numbers we sourced were inaccurate across the industry dimension (CEWS and the worker count survey used slightly different industry definitions), we could not use this data without restricting the flexibility of the dashboard. It would have been challenging to incorporate this data without the potential of introducing inaccuracies and misleading users. An important next step would be to find a way to use this per-worker data in an interactive dashboard.

Part 2. API Feasibility Study

2.1. Background

Data and the digital economy are fueling innovation, as well as economic and social opportunities, in Canada and around the world. Statistics Canada is currently making significant effort to leverage the growing selection of novel data sources to enhance its analyses. Statistics Canada, in its 2019 to 2022 Data Strategy (SCDS), indicates that its "…data strategy is embedded in capitalizing on this data revolution and leveraging the explosion in the volume of data and data analytics technologies by actively researching, developing, and implementing solutions to provide Canadians and decision-makers with increased access to high-quality, relevant and timely information on issues affecting the society"[23].

SCDS also indicates that "Statistics Canada is currently modernizing its statistical programs to respond to a rapidly changing and increasingly complex economy and society, a proliferation of data and providers, and increased user expectations for "real-time" and micro-detailed data… This includes creating strategic partnerships with other organizations and researching and discovering data inputs that can be used by statistical programs, such as administrative data, open data, found data,

---

[23] "Statistics Canada Data Strategy," Statistics Canada, March 13, 2020, https://www.statcan.gc.ca/eng/about/datastrategy.

commercial data, crowdsourced data and web-scraped data, while respecting privacy and maintaining public trust"[24].

Statistics Canada has historically been reliant on traditional surveys and administrative data sources[25]. Leveraging the digital data economy will allow for the development of novel techniques for measuring economic activity and studying social behavior at high frequency and in real time. Social media platforms, sharing economy enterprises, and crowdsourcing data companies are some sources of potentially valuable data, which may be accessible via APIs.

APIs are tools which allow individuals and businesses to access and use data collected by other applications and companies. Many data companies have APIs available for public use, often with an associated cost. The data accessible from these APIs is often collected real-time, and learning how to harness these programs for Statistics Canada would allow for more efficient policy and decision making. Standard survey and administrative data sources necessarily have a significant lag between the time the need for data is established and the usable data is collected, and APIs and other technical data sources can greatly reduce or eliminate this lag. APIs could potentially be used to learn about the spending habits and quality of life of Canadians as the country recovers from the pandemic. Furthermore, APIs could allow for the study at far more granular levels than traditional measures allow, by allowing the real-time analysis of individual neighborhoods, streets, or businesses.

2.2. Tools, Methodology and Techniques

2.2.1. Tools

Our API inventory report was written in Google Docs. Demo software was developed in Python, using the *requests* package, along with the *Tweepy* Twitter API wrapper. *Pandas* and *Regex* were used for data structures and wrangling. *Altair* and *Matplotlib* were used for visualization; and *String*, *NLTK* (natural language toolkit) and *TextBlob* were used for sentiment classification. We used the official Reddit, Steam, and Twitter APIs, as weel as the third-party SteamSpy APi.

2.2.2. Methodology - API Inventory

We reviewed the documentation of about 20 APIs which we thought might provide some value in procuring new sources of data which could be used to study the economic effects of the pandemic, as well as economic recovery after the fact. For each API, we provided a brief background on the company, a potential use case for our goals, an overview of the useful endpoints, a discussion

---

[24] "Statistics Canada Data Strategy," Statistics Canada, March 13, 2020, https://www.statcan.gc.ca/eng/about/datastrategy.
[25] "Statistics Canada Data Strategy," Statistics Canada, March 13, 2020, https://www.statcan.gc.ca/eng/about/datastrategy.

on how those endpoints could be used to meet our goals, some of the access requirements and limitations, and a conclusion on whether or not the particular API is worth further investigation.

2.2.3. Methodology - Demo Software

We performed preliminary natural language processing and sentiment analysis using the Twitter API. The first step to achieve this was to get a Twitter developer account. To access Twitter data, we used Tweepy. Tweepy allowed us to get tweets by querying keywords or locations, and filtering by language and date. This functionality allowed us to search tweets in Canada and its provinces and territories during the COVID-19 pandemic to discover and assess the conversation happening on Twitter.

The next step was to clean the tweets to make them machine-readable for further modeling. We removed URLs and usernames from tweets with regular expressions. We also removed standard English stop words using NLTK. We also used NLTK for tokenizing and lemmatizing (splitting words by whitespaces and breaking words to its root).

For the sentiment analysis, we used NaivesBayesAnalyzer, which is an NLTK model trained on a movie reviews corpus to classify tweets as having a positive or negative sentiment. We also used TextBlob to get sentiment scores on polarity and subjectivity of the tweets. TextBlob sentiments are defined based on semantic relations and the frequency of each word in an input sentence.

Finally, we visualized our results using Altair and a word cloud visualization, in which the size of each word indicates its frequency in a text.

To access the Steam data, we used the official API of the company who developed Steam, *Valve*, as well as a third-party website called steamSpy. The Valve API can be used to collect comments under any game. The steamSpy can be used for looking up various game's identity information and statistical data. In our demo it is used to look up the game ID, which is required for specifying a certain game in the Valve API's query string. Although there is not a good API wrapper like Tweepy for the Twitter API, we can use a python library called *requests* to access the APIs using python. In our demo, due to high relevance to the pandemic, we choose to analyze the comments under a game called Plague Inc. Firstly, we began by looking up the game ID and substituting it in the query string for collecting comments. Secondly, we filtered out the comments that included any words related to COVID-19 or feelings during the lockdown. Lastly, we passed these comments into a pre-trained sentiment intensity analyzer developed in the NLTK library to get an idea on whether playing this game has helped people to relieve their stress.

The Reddit demo used the official Reddit API and the *requests* package to retrieve data from a number of API endpoints. We were able to retrieve posts within specific subreddits of interest (we looked at subreddits devoted to specific Canadian regions, such as r/toronto, r/calgary, and r/newfoundland). By filtering these retrieved posts by subject using lists of keywords, we were able to

look at the upvote ratios to develop an understanding of how people in specific regions feel about our subject of interest.

2.3. Results, Analysis, Interpretation, and Conclusions

2.3.1. API Inventory

2.3.1.1. Conclusions

Unfortunately, the majority of the APIs researched had limited or no feasibility for tracking economic data. None of the APIs offered the aggregate data that would be most useful, and most required user authorization to access an individual's data. The most promising data available through APIs tended to be the publicly facing comments and posts that users of the platform make. Using natural language processing and sentiment analysis techniques to extract meaning from these comments was the primary use for the majority of usable APIs. While the full inventory report is included as **Appendix B,** the table below summarizes our findings.

API	Use Case	Feasibility
Twitter	NLP and sentiment analysis of Tweets about specific topics.	Feasible.
Reddit	NLP and sentiment analysis of posts within region-specific forums.	Feasible.
Foursquare	Using device-tracking location data to estimate economic activity within regions.	Somewhat feasible. API has limited use, but Foursquare's paid services could work.
Yelp	Gather information on businesses (customer base and covid-compliance) using NLP on user reviews.	Somewhat feasible. API has limited use, but Yelp's paid services could work.
Ticketmaster	Estimate social and economic activity through the purchases of event tickets.	Somewhat feasible. Information on events/venues can be gathered, but not information on the number of tickets sold.
Indeed	Gain insight into the job market by looking at job postings.	Feasible.
Eventbrite	Similar to ticketmaster.	Somewhat feasible, but less information is available than

		through Ticketmaster's API.
Airbnb	Study the recovery of the tourism industry in various regions through the number of short term accommodation rentals.	Infeasible. The Airbnb API is intended for commercial and business use only.
Facebook/Instagram	Study posts on businesses' social media pages to gain insight into how various business types were affected by the pandemic, and how they are recovering.	Infeasible. Both APIs require user authorization by the businesses for us to access any useful data.
Tripadvisor	Estimate the number of tourists planning to visit Canadian cities.	Infeasible. The Tripadvisor API is intended for commercial and business use only.
Skyscanner	Estimate tourism and the recovery of the transportation sector through the number of passengers travelling to various areas.	Infeasible. There is no data available on passenger count.
Youtube	Use the change in aggregate watch time data to estimate how much people are going out and doing other things post-pandemic.	Infeasible. This data is not available.
Netflix	Use the change in aggregate watch time data to estimate how much people are going out and doing other things post-pandemic.	Infeasible. This data is not available.
Uber Eats	Study long term changes in the food service industry. Determine whether people are going out to restaurants or continue to order in.	Infeasible. The API is intended for app integration, and not for data collection.
Steam	Study the impact of lockdown on the gaming industry.	Somewhat feasible. Data only available for the past 2 weeks
Google	Google has many APIs available. Use the *Places API* to retrieve data on businesses, or the *Knowledge Graph API* to gain insight into Google Search frequency.	Feasible, but gaining insight into the 300+ Google APIs would require a significant time investment.
Fitness APIs (Under Armor, Garmin, Runkeeper, Strava,	Examine the effects of the pandemic on exercise	Infeasible. User authorization required to access data, and no

	frequency and general health.	aggregate data available.
etc.)		

Table 1: API inventory summary.

2.3.1.2. Setbacks and Limitations

Many of the APIs we researched had incomplete documentation, especially regarding rate limits and associated costs. Some of the tools we discovered required consultations with the company in order to learn anything meaningful. For example, *Yelp Knowledge* seemed like a potentially useful tool, but the only information available online was a basic, 3 sentence description.

2.3.1.3. Next Steps

The 18 APIs we have researched is a good first step, but is certainly not a full inventory of everything that exists. There may very well be APIs that we did not research which have the potential of being a revolutionary source of data. Furthermore, correspondence with an economist would be very useful to better understand what particular questions are challenging to answer with typical data sources and could benefit from more technical data sources, such as APIs. We believe that our research can help determine if APIs could be a useful source of data for a particular research topic, but more specific goals are necessary to precisely determine the feasibility of any particular tool.

2.3.2. Demo Software

2.3.2.1. Conclusions

We chose three of the above APIs and developed some basic test-case functionality. We looked at Twitter, Reddit, and Steam.

We performed preliminary NLP and sentiment analysis on tweets containing the word "CEWS" to measure the overall sentiment of the subsidy program in the conversation happening in Twitter. Twitter Standard API only allows retrieval of conversations from the last 7 days. This allowed us to retrieve a total of 280 tweets in Canada for the period between June 11 to June 19, 2021. Of those tweets, 79 showed enough information for us to interpret their sentiment.
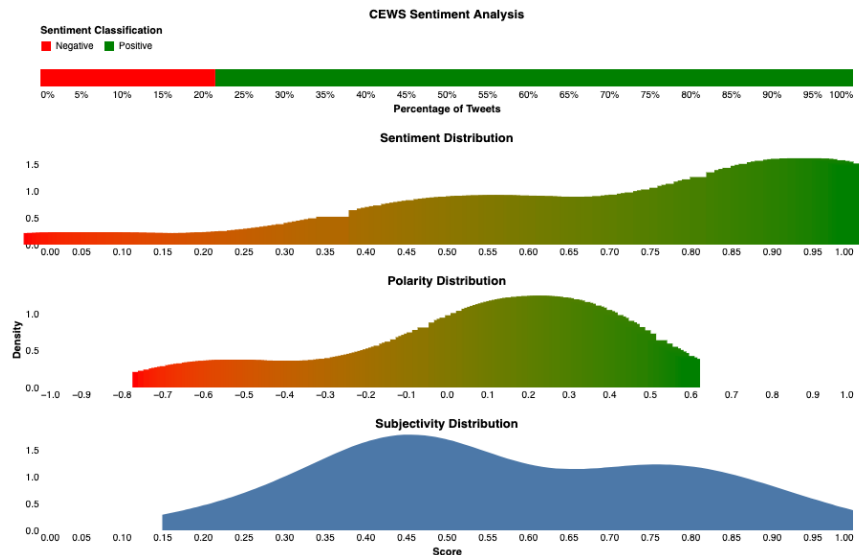
Figure 9: Summary of sentiment analysis in tweets containing the word "CEWS".

80% of the analyzed tweets were classified as having a positive sentiment and 20% were classified as negative. Overall, tweets were more likely to have a positive polarity and express a subjective opinion.

The code for the CEWS sentiment analysis is included as **Appendix C**.

In the steam demo, we looked at user reviews of Plague Inc. to collect comments related to the pandemic and lockdowns. However, most of the comments are related to the game itself, and only a few reference player's feelings before and after playing this game during lockdown and quarantine. Since we only checked if the comments included a certain word to determine whether the comments are on-topic or off-topic, we might have missed some of the comments that should be collected. A better solution would be implementing a machine learning and nlp model to filter out the on topic comments. However, developing a model like this is beyond the scope of this capstone project.

We also looked at other angles other than the game comments, for example, which game has the most average daily active players during the entire pandemic lockdown. However, statistical data like this is only available for the past two months. Despite the limitations mentioned above, the steam APIs can still be used to collect data on players' preferences after certain holidays (like Christmas, Valentine's day, etc.) if it would be of our partner's interests.

With Reddit, we cannot search individual subreddits for posts containing keywords of interest. Reddit however does not have the strict rate limits that Twitter imposes, and it may be possible to get a large enough sample of posts from relevant subreddits where filtering by keyword in post-processing would still yield a significant sample. The Reddit API also allows us to get the "upvote ratio" for each post, which acts as a sentiment indicator for how the population of the subreddit feels about each post. In our quick test demo, we retrieved 20 "hot" posts for the day (June 10) from each of r/toronto, r/calgary, r/vancouver, and r/newfoundland. We then filtered these lists to only contain posts with one of a list of covid-relevant keywords in their title. By looking at the upvote

ratio of each post, we were able to get a general sense of each region's opinion toward various topics. For example, the upvote ratio of posts in r/calgary was much smaller than in r/toronto, which could indicate that Calgarians are more averse to covid news than Torontonians.

2.3.2.2. Setbacks and Limitations

For all of the APIs we looked at, we would need to pull data continuously to gain insight into changes over time. Since we had very limited time to work on this part of the project, it was not feasible to do so. As such, all of our demo software only included data from small time frames (one day to one week, depending on the demo), and we could not accurately analyze the time dimension. Furthermore, the restrictive rate limits of free tier APIs greatly inhibited the sample sizes generalizability and statistical significance of our findings.

Our preliminary NLP and sentiment analysis using Twitter shows promise but is somewhat limited by the capabilities of the Twitter API. By accessing Twitter with the Tweepy wrapper, we were able to expand the Twitter API's capabilities. However, Tweets cannot be searched using both keyword and geography simultaneously, which greatly limits our use cases. Furthermore, the Twitter API has strict rate limits of 500,000 Tweets per month and a retrieval rate of 450 tweets per 15 minute window. This limit makes it challenging to receive a large, representative sample from all regions of Canada, and filtering the stream of Tweets in post for only ones with relevant keywords would mean a substantial loss in this already small sample.

2.3.2.3. Next Steps

None of these demos are finished products as-is, and would require significant work to ensure the data retrieved is useful, unbiased, and interpretable. We would be required to develop a complete list of keywords for any topic we wish to analyze, to ensure we are precisely filtering relevant content out of the data stream. This in and of itself is a challenging task, as more abstract topics may have many keywords which could be associated, but may not. For example, posts with the word *reopening* may be about the end of the pandemic, but could also be about many other topics. A list of keywords may not be sufficient for some topics, and machine learning algorithms may be required to predict the topic of some posts.

Regarding our Twitter sentiment classification model we could segment the tweets using their locations to assess the sentiment by Provinces and Territories of Canada. Another potential use case would be to cluster texts by topic, for example, economy, life-style, politics, and so on, and assess the overall sentiment. To overcome the Twitter API rates, we could also look at wrappers other than Tweepy to evaluate if they allow retrieval of older tweets and show expanded functionality.

Finally, work would be required to develop useful indicators, in order to provide actionable recommendations from the data we gather and analyze through APIs. This would be a very significant

amount of work, and would require a team of people with diverse skills and backgrounds to implement.

References

Statistics Canada. "2016 Census, Catalogue No. 98-400-X2016320," November 29, 2017. https://www150.statcan.gc.ca/n1/en/catalogue/98-400-X2016320.

Statistics Canada. "2016 Census, Catalogue No. 98-400-X2016321," November 29, 2017. https://www150.statcan.gc.ca/n1/en/catalogue/98-400-X2016321.

Department of Finance. "Budget 2021," April 19, 2021. https://www.budget.gc.ca/2021/report-rapport/p1-en.html#17.

Canada Revenue Agency. "Canada Emergency Wage Subsidy (CEWS)," September 18, 2020. https://www.canada.ca/en/revenue-agency/services/subsidy/emergency-wage-subsidy.html.

Statistics Canada. "Canada Emergency Wage Subsidy Regional and Community-Level Database," March 18, 2021. https://www150.statcan.gc.ca/n1/pub/11-26-0003/112600032021001-eng.htm.

Chiatchoua, Cesaire, Carmen Lozano, and Jesús Macías-Durán. "Análisis de los efectos del COVID-19 en la economía mexicana." Revista del Centro de Investigación de la Universidad la Salle 14, no. 53 (September 18, 2020): 265–90. https://doi.org/10.26457/recein.v14i53.2683.

Department of Finance. "Economic and Fiscal Report Summary," July 8, 2020. https://www.canada.ca/en/department-finance/services/publications/economic-fiscal-snapshot/summary.html.

Statistics Canada. "Employment by Class of Worker and Industry, Annual, Population Centres and Rural Areas, Inactive," January 8, 2021. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1410010801.

Hevia, Constantino, and Andy Neumeyer. "Un Marco Conceptual Para Analizar El Impacto Económico Del COVID-19 y Sus Repercusiones En Las Políticas [A Conceptual Framework for Analyzing the Economic Impact of COVID-19 and Its Policy Implications]." PNUD LAC C19 PDS 1 (March 20, 2020). https://www.undp.org/content/dam/rblac/Policy%20Papers%20COVID%2019/UNDP-RBLAC-CD19-PDS-Number1-ES-final.pdf.

Statistics Canada. "North American Industry Classification System (NAICS) Version 3.0," August 17, 2018. https://www23.statcan.gc.ca/imdb/p3VD.pl?Function=getVD&TVD=1181553.

Department of Finance. "Overview of Canada's COVID-19 Economic Response Plan," January 6, 2021. https://www.canada.ca/en/department-finance/services/publications/economic-fiscal-snapshot/overview-economic-response-plan.html.

Canada Department of Finance. "Overview of Canada's COVID-19 Economic Response Plan," January 6, 2021. https://www.canada.ca/en/department-finance/services/publications/economic-fiscal-snapshot/overview-economic-response-plan.html.

Statistics Canada. "Standard Geographical Classification (SGC) 2016," November 10, 2016. https://www.statcan.gc.ca/eng/subjects/standard/sgc/2016/index.

Statistics Canada. "Statistics Canada Data Strategy," March 13, 2020. https://www.statcan.gc.ca/eng/about/datastrategy.

Appendix A: CEWS Exploratory Report

The full CEWS report can be read and downloaded with the following link:

📄 final_cews_report_16Jun21

It is also available in our GitHub project repository.

Appendix B: API Inventory Report

The full API report can be read and downloaded with the following link:

📄 APIFeasibility.docx

It is also available in our GitHub project repository.

Appendix C: CEWS Sentiment Analysis

The code for the CEWS sentiment analysis can be downloaded with the following link:

CEWS_SentimentAnalysis.html

It is also available in our GitHub project repository.