



THE UNIVERSITY
OF BRITISH COLUMBIA



Statistics
Canada

Statistique
Canada

A Proposal for The Analysis of Rural Business Performance

By

Tingwen Hua, Bowen Yang, Yilin Sun, Song Zhang

Master of Data Science Capstone

University of British Columbia Okanagan Campus



May 8, 2022

Introduction & Background

As 6.1 million Canadian people residing in rural communities and 2.6 million people working in rural areas at the turn of the decade, it is important to know rural business conditions and understand ongoing and future trends of the development of rural business. Analysis on rural business conditions are critical for stimulating the development of rural business.

For this capstone project, our work is based on the Rural Canada Business Profiles (RCBP) database from Statistics Canada. This database contains counts and financial data on small businesses (\$30,000 < yearly revenue < \$5 million) and medium businesses (\$5 million < yearly revenue < \$ 20 million) in Canada, aggregated by geography, industry, incorporation status, and some other relevant variables. There is currently data for 2017, 2018, and 2019. First we will do some visualizations and analysis on the rural business development in Canada based on the dataset. We will conduct a report based on our analysis.

In addition to the analysis report, we will create an interactive dashboard to show the outcome of our analysis. The users can choose different factors they want to see and get the result visualization based on their choices. We will use Dash or Tableau to build the dashboard.

In the end, we will do a cross-analysis if we have enough time. We will use the RCBP data in union with other data to see if there are any interesting patterns or insights that could be found to help understand the rural business better.

For the background research, our clients of Statistics Canada have already done some research based on the dataset¹. In their research, they found that 16.3% share of the Canadian population was living in rural areas at the turn of the decade, which points to the relative vitality of the rural economic scene. In addition, they found that small and medium businesses in rural areas generated, respectively, 17.1% (\$139,300 million) and 14.2% (\$69,300 million) of the revenues of all businesses in Canada in their respective size class in 2019. In their research, they stated that between 2017 and 2019, the numbers of both small and medium businesses increased at a slower rate in the country's rural areas than in its urban areas.

For the cross-analysis, there is a research paper about the rural areas subsidies in the Covid pandemic² published by our client. The research covered the time from January 17 to May 8, 2021 and found that both subsidies and the number of employees supported declined in rural areas over the time. The research also found one in six rural employees supported by CEWS(Canada Emergency Wage Subsidy) in this period. In addition, manufacturing had the highest share of both CEWS payments and supported employees among rural industries.

¹ "A profile of businesses in rural Canada, 2017 to 2019.", Statistics Canada, <https://www150.statcan.gc.ca/n1/pub/21-006-x/21-006-x2022003-eng.htm>

² "Rural insights from the Canada Emergency Wage Subsidy: January 17, 2021 to May 8, 2021.", Statistics Canada, <https://www150.statcan.gc.ca/n1/pub/21-006-x/21-006-x2022001-eng.htm>

Aims & Objectives

Objectives

In this project, we intend to achieve 3 main objectives:

1. To provide an exploratory analysis of the Rural Canada Business Profiles (RCBP) database.
2. To implement an interactive dashboard that allows users to choose dimensions/variables and get related visualizations and detailed links.
3. To perform a cross-analysis of RCBP with other public Statistics Canada data (open ended)

In the process of achieving these three objectives, we will investigate following research questions:

1. How does the Canadian rural business perform according to different dimensions/variables?
2. How can the RCBP data be connected with other StatCan data to produce valuable analysis?

Methods

For the first objective, since the data of the Rural Canada Business Profiles (RCBP) are stored in .csv files, we will load them to Pandas dataframe and perform data cleaning, data wrangling in Python. Through filtering, grouping, sorting and aggregating the original dataframes, we will generate some new dataframes. These new dataframes can be used to calculate statistical values (like average, median, variance, maximum/minimum value), explore data distributions, and generate visualizations. In terms of visualization, we plan to use the Altair library in Python for plotting diagrams. As our client said, the published information of our client (Statistic Canada) needs to be approachable for the majority of Canadians, regardless of educational background. We will try to use simple diagram types (like bar chart, line chart, heat map on geographical map) mostly, although a few more complex analyses might be slightly performed based on specific needs. These diagrams can help us to investigate the relationship between variables/dimensions of the RCBP database, and answer the first research questions. Finally, we will summarize our results of exploration in the analysis report.

For the second objective, we plan to use the Dash and Altair libraries in Python or Tableau to build an interactive dashboard. The dashboard will allow users to select their interested dimensions/variables from tabs and dropdown lists, then generate charts and detailed information page links correspondingly. So users can get an overview of data through the charts on the dashboard and dig deeper by visiting a detailed data information page. Similar to the visualizations described in objective one, we will make charts to be simple and easy-to-understand. We will compare Dash and Altair libraries in Python and Tableau before the implementation and select the most suitable one to work on.

For the third objective, We will use the RCBP data in union with other StatCan data to explore patterns or insights that could be found to help answer our last research questions. Since Statistic Canada does not use data to predict the future, it seems better to choose a model that has better performance on inferential aspects. Therefore, we plan to try linear regression and multiple linear regression to investigate the relationship between variables in these two datasets. Since there are many variables available in these datasets, we might attempt to use Python/R to apply predictors selection methods (Forward Selection/Backward Selection/ Mixed Selection), Ridge Regression and LASSO for selecting “more significant variables”.

Dataset: Rural Canada Business Profiles 2017-2019

Rural Canada Business Profiles(RCBP) is a newly launched database which contains the financial reports of the Small and Medium-Sized Businesses(SMB) in Canada from 2017-2019. In this database, SMB are defined as companies whose total annual revenue are from \$30,000 to \$5,000,000 and \$5,000,001 to \$20,000,000 respectively.³

For each business size, it provides the following financial report as tables:

- Counts of businesses
- Revenue breakdowns
- Expense breakdowns
- Balance sheet items (assets, liabilities, equity breakdowns)
- Financial ratios⁴

In each table, the data are managed by the following variables:

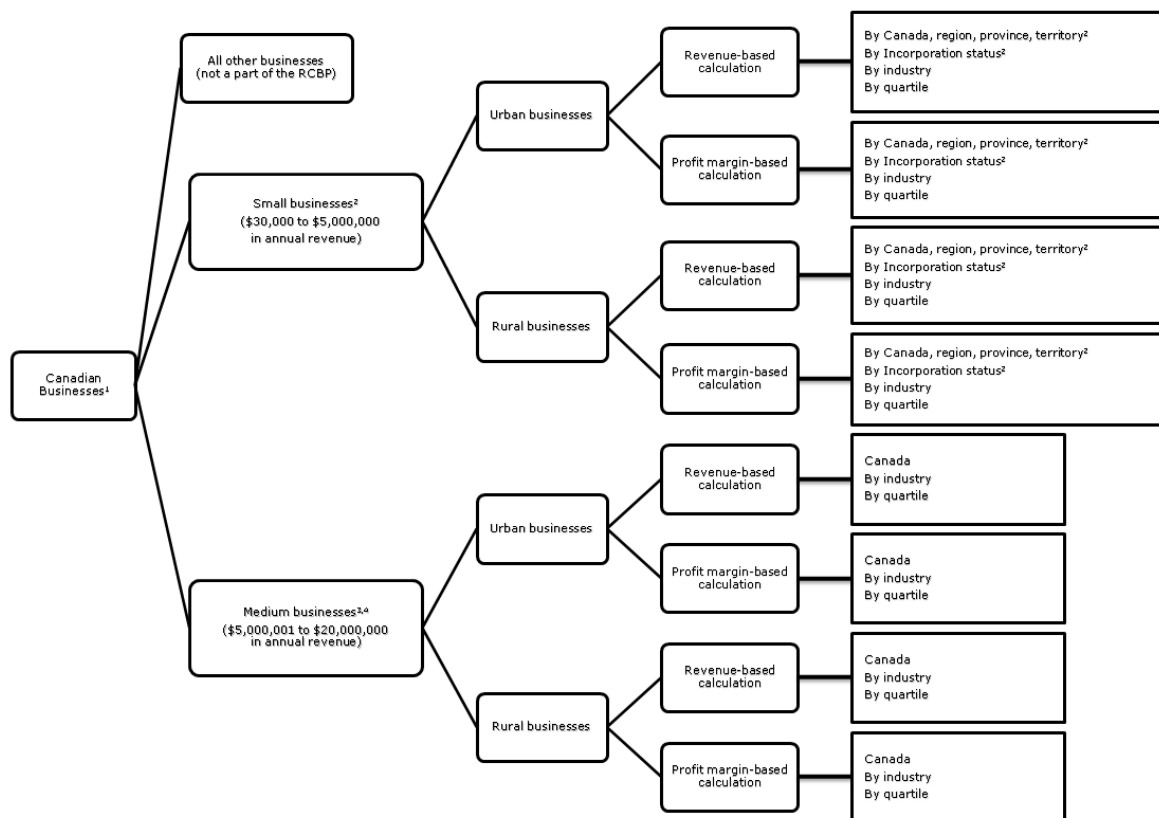
- Small or medium size of business by annual revenues
- Various levels of geography (Canada, region, province/territory)
- Rural and urban areas
- Industry
- Incorporation status
- Profitable and non-profitable businesses⁵

In our project, we are going to find out how these variables may influence the financial reports. The following picture gives a more detailed database structure:

³ “Rural Statistics in Canada Rural Canada Business Profiles, 2017 to 2019.” Statistic Canada,

⁴ “Rural Canada Business Profiles Database 2017-2019 Metadata and user guide.” Statistic Canada.

⁵ “Rural Canada Business Profiles Database 2017-2019 Metadata and user guide.” Statistic Canada. <https://www150.statcan.gc.ca/n1/pub/45-20-0004/452000042022001-eng.htm>.



1. Includes all businesses that have submitted a Canadian tax return for the reference tax year.
2. Small businesses are classified by incorporation status into two groups: unincorporated and incorporated businesses. Data are also provided for the sum of these two categories.
3. As all medium businesses are incorporated, there are no breakdowns by incorporation status.
4. Province/territory breakdowns are not provided for medium businesses to protect confidentiality.

Figure 1. The structure of the Rural Canada Business Profiles Database⁶

In the analysis, our team is going to focus on the rural business datasets by filtering the business type and cleaning the dataset to keep the most useful and interesting variables. In addition, we will do some analysis based on business industries, geography and incorporation status. While we are processing the data, we will keep in mind that this dataset is suppressed and we will keep the data confidential. Since the data are organized very well, there is not much data cleaning work or preparation.

⁶ "Rural Canada Business Profiles Database 2017-2019 Metadata and user guide." Statistic Canada.
<https://www150.statcan.gc.ca/n1/pub/45-20-0004/2022001/meta-eng.htm>.

Deliverables & Schedule/Timeline

The duration of the project will last for 7 weeks, from 1st May to 21st June with a mid-project presentation on 31st May and the final presentation on 21st June. We will follow the capstone project schedule, report our progress to the instructors and TAs and remain in correspondence with Haaris Jafri, Eric Baxter and Mahdia Khodja in Statistics Canada via Microsoft Teams meeting on a weekly basis. At the end of the project, we will complete an analytical paper and develop a dashboard for our client and finish the final report of the project.

Timeline

The high-level schedule of weekly goals is as follows:

Week	Theme	Goals
1 (01 – 09 May)	Proposal	Establish regular connection with client, understand the problem, develop the work plan, deliver project proposal and presentation
2 (10 – 17 May)	EDA and data visualization	Understand the RCBP dataset, data wrangling, do EDA and data visualization, develop analytical aspects
3 (18 – 24 May)	Data visualization and Analytical paper (first draft)	Data visualization, deliver the first draft of the analytical paper
4 (25 – 31 May)	Dashboard Mid-project presentation	Develop the dashboard, make the mid-project presentation
5 (01 – 07 June)	Dashboard	Develop the dashboard, debugging
6 (08 – 14 June)	Open-ended Question	Explore other related public StatCan data, do the cross-analysis with RCBP
7 (15 – 21 June)	Final report Final Presentation	Deliver the Analytical paper, final report and make the final presentation

Deliverables

1. Analytical paper

The goal of the analytical paper is to explore more aspects of the RCBP database based on the previous work our client has done. It will analyze more dimensions/variables of the database, find more characteristics, patterns, and relationships for Canadian rural business, and deliver with visualizations.

2. Dashboard

We aim to develop a dashboard as an overview page for the Rural Canada Statistics Portal. The dashboard is intended to be designed as an interactive visualization tool that enables users to choose the dimensions/variables they are interested in and generate visualizations, without them needing to go through the individual datasets and analysis papers. If they want to dig deeper, the dashboard could link them to the location of the data that generates the visualizations.

3. Final report

The final report will be the integration of our work, including the analysis of RCBP database, the display of the dashboard, and the insight analysis of the open-ended questions about cross-analysis of RCBP with other public StatCan data, like the Canada Emergency Wage Subsidy data.

Member roles and responsibilities

Our team will work collaboratively and make the most of everyone's strengths. We will work together with developing the project ideas, solving the problems, dividing the tasks, setting the schedule and so on. We will meet on zoom very often, whenever we need, and take the meeting minutes in turn. At the team meeting, the sub-problems will be identified together and assigned (mainly volunteer) to work on individually by team members. We have agreed to Tingwen Hua as our client liaison, and every member will take their part on the weekly/mid-project/final presentations. According to our backgrounds, during the project Yilin Sun and Bowen Yang will do more coding work, and Song Zhang will do more modeling work, Tingwen Hua will contribute more on business ideas.

References

- Khodja, Mahdia, Eric Baxter, and Haaris Jafri. 2022. "Rural and Small Town Canada Analysis Bulletin A profile of businesses in rural Canada, 2017 to 2019." Statistic Canada.
<https://www150.statcan.gc.ca/n1/pub/21-006-x/21-006-x2022003-eng.htm>.
- Munro, Ash, Eric Baxter, Peter Timusk, Andrew Balcom, and John Baker. 2022. "Rural insights from the Canada Emergency Wage Subsidy: January 17, 2021 to May 8, 2021." Statistic Canada.
<https://www150.statcan.gc.ca/n1/pub/21-006-x/21-006-x2022001-eng.htm>.
- Statistics Canada. 2022. "Statistic Canada." Rural Canada statistics.
https://www.statcan.gc.ca/en/subjects-start/society_and_community/rural_canada.
- University of British Columbia. n.d. "UBC Logos | UBC Brand." UBC Brand. Accessed May 5, 2022.
<https://brand.ubc.ca/guidelines/downloads/logos-signatures-and-visual-identity-assets/>.
- Wikimedia Common. n.d. File:Statistics Canada logo.svg.
https://commons.wikimedia.org/wiki/File:Statistics_Canada_logo.svg.