Week 3 Presentation

Tingwen Hua, Bowen Yang,
Song(Alice) Zhang, Yilin Sun

Client Meeting

Wednesday May 4 9:00am-10:00am

Database variables explanation

Dashboard Expectation Confirmed

Data source Confirmed

Progress

1. Confirmed Dashboard structure:
    a. 1 general overview page including couple datasets on RCBP portal
    b. 1 detailed dashboard based on RCBP dataset
2. Confirmed Cross-analysis Dataset:
    a. 2016 Census (Canada)
3. Completed Exploratory Data Analysis:
    a. Data cleaning and wrangling
    b. Data visualization
4. Initialized Analytical Paper Structure:
    a. Small Business & Medium Business

Data Cleaning and Wrangling Code

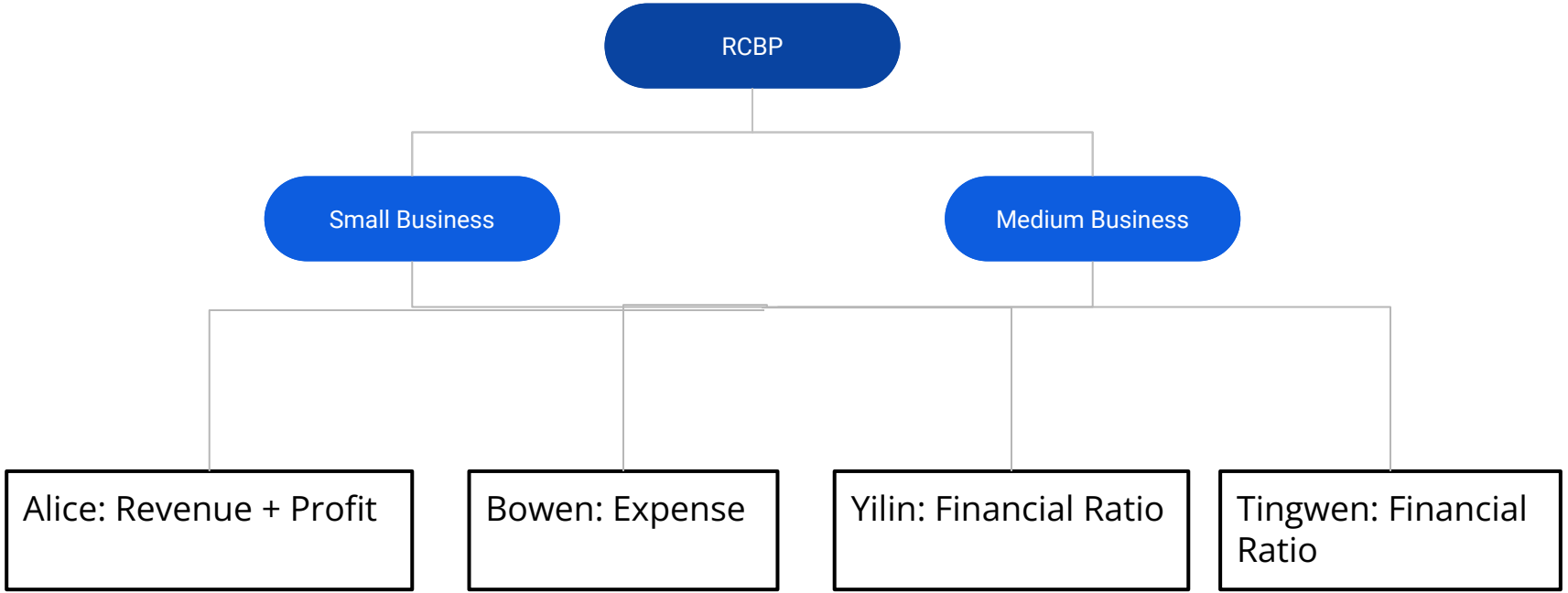
```
def tab_reader(filepathpattern,tabnum,skiprow,low_memory=False):
    tabs = []
    for i in range (tabnum):
        cur_index = i+1
        cur_filepath = filepathpattern.format(cur_index)
        try:
            tab = pd.read_csv(cur_filepath, skiprows=skiprow, low_memory=False)
        except:
            tab = pd.read_csv(cur_filepath, skiprows=skiprow,encoding='latin-1', low_memory=low_memory)
        tabs.append(tab)
    return tabs
```

```
# 2017 data
rcbp_2017_mb_profitMargin_Tabs = tab_reader("2017_csv_eng/2017_Medium businesses_Profit margin based_csv/_2017_Medium businesses_Profit margin_Tab{}.csv",5,5)
rcbp_2017_mb_revenue_Tabs = tab_reader("2017_csv_eng/2017_Medium businesses_Total revenue based_csv/_2017_Eng_Medium_Revenue_Tab{}.csv",5,5)
rcbp_2017_sb_profitMargin_Tabs = tab_reader("2017_csv_eng/2017_Small businesses_Profit margin based_csv/_2017_Small businesses_Profit margin based_Tab{}.csv",7,5)
rcbp_2017_sb_profitMargin_Tabs[0] = tab_reader("2017_csv_eng/2017_Small businesses_Profit margin based_csv/_2017_Small businesses_Profit margin based_Tab{}.csv",1,6)[0]
rcbp_2017_sb_revenue_Tabs = tab_reader("2017_csv_eng/2017_Small businesses_Total revenue based_csv/_2017_Small businesses_Total revenue based_Tab{}.csv",7,5)
rcbp_2017_sb_revenue_Tabs[0] = tab_reader("2017_csv_eng/2017_Small businesses_Total revenue based_csv/_2017_Small businesses_Total revenue based_Tab{}.csv",1,6)[0]
```

```
def selectData(tabs):
    concatTabs = []
    for tab in tabs:
        cols_without_quartile = [x for x in tab.columns if 'quartile' not in x.lower()]
        tab = tab[cols_without_quartile]
        tab = tab.loc[(tab["North American Industry Classification System, NAICS - code"].str.len() < 3) | (tab["North American Industry Classification System, NAICS - code"
        concatTabs.append(tab)
    return pd.concat(concatTabs)
```

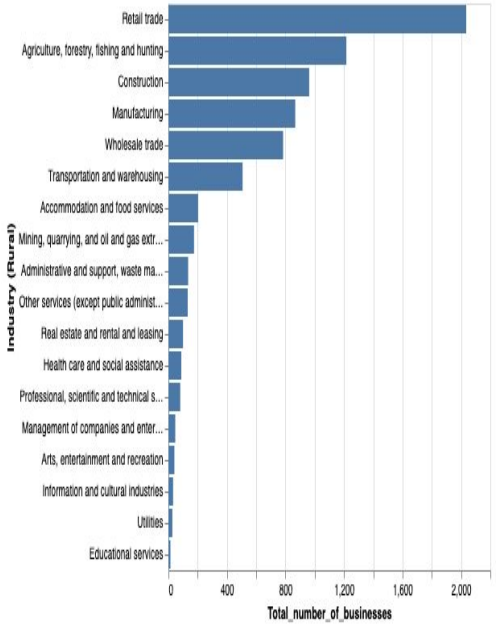
```
mb_2017_revenue_tab2 = rcbp_2017_mb_revenue_Tabs[1]
mb_2018_revenue_tab2 = rcbp_2018_mb_revenue_Tabs[1]
mb_2019_revenue_tab2 = rcbp_2019_mb_revenue_Tabs[1]
mb_revenue_persent_3y = selectData([mb_2017_revenue_tab2,mb_2018_revenue_tab2,mb_2019_revenue_tab2])
```

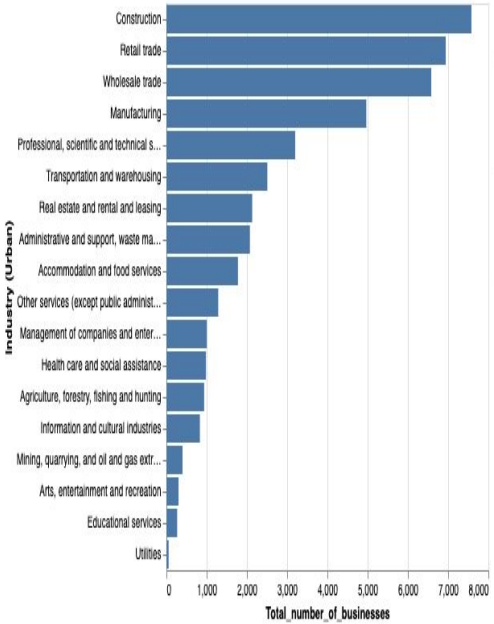
Database Structure

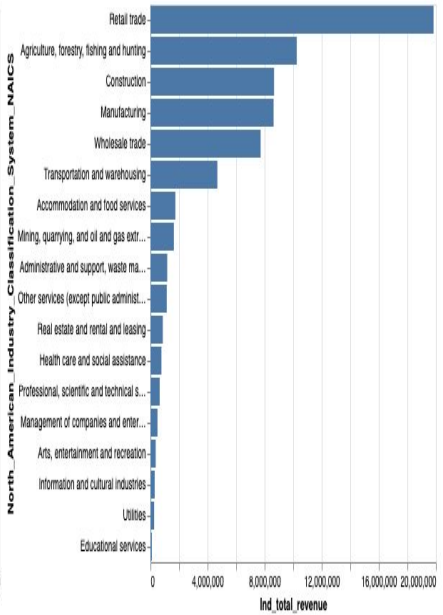


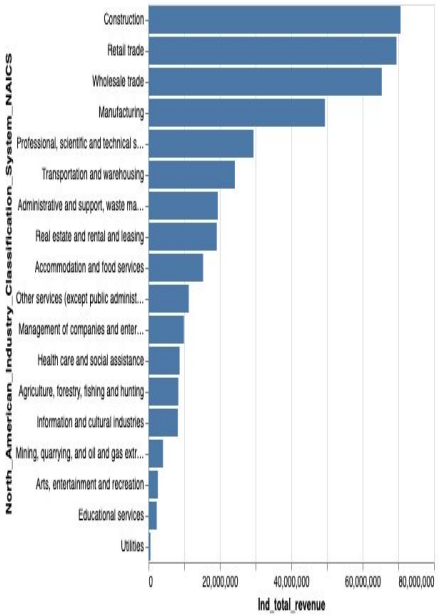
Some Data Visualizations

Tab1: Medium Business Revenue（Total number, Total revenue）



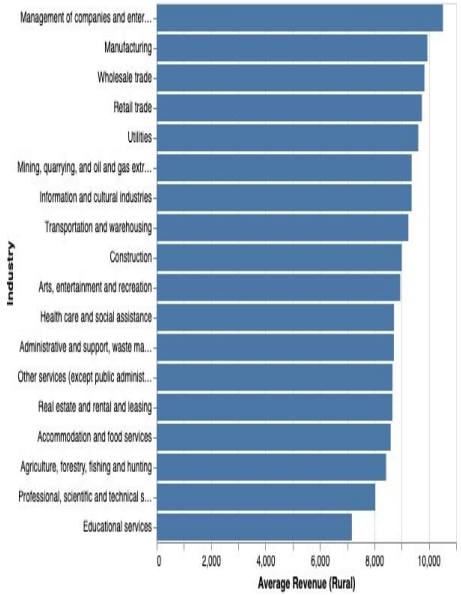


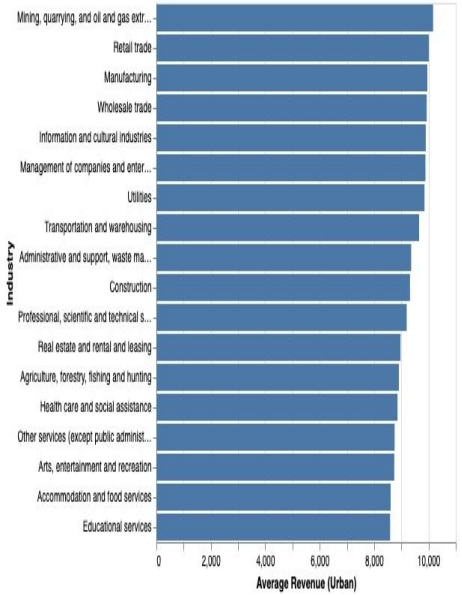


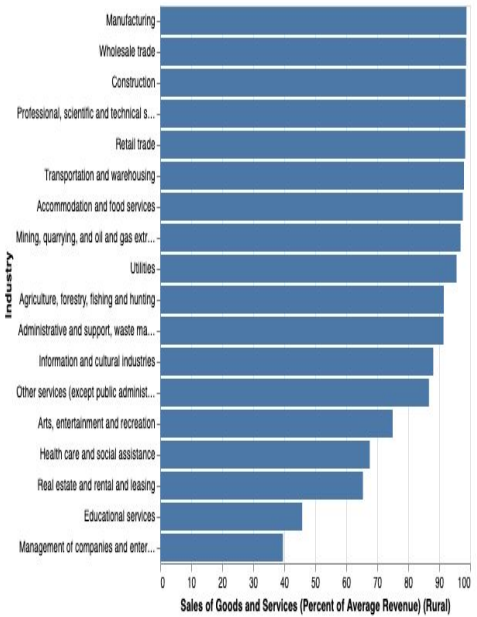


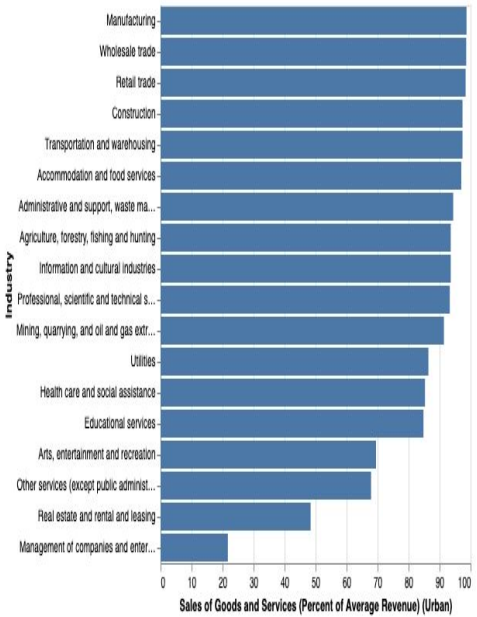
Some Data Visualizations

Tab1: Medium Business Revenue (Average revenue, Sales of goods and services percent to average revenue)



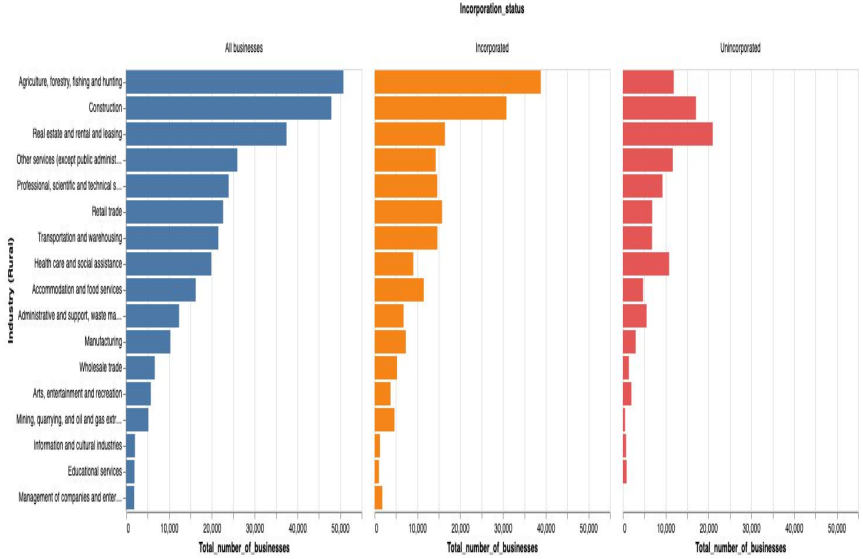


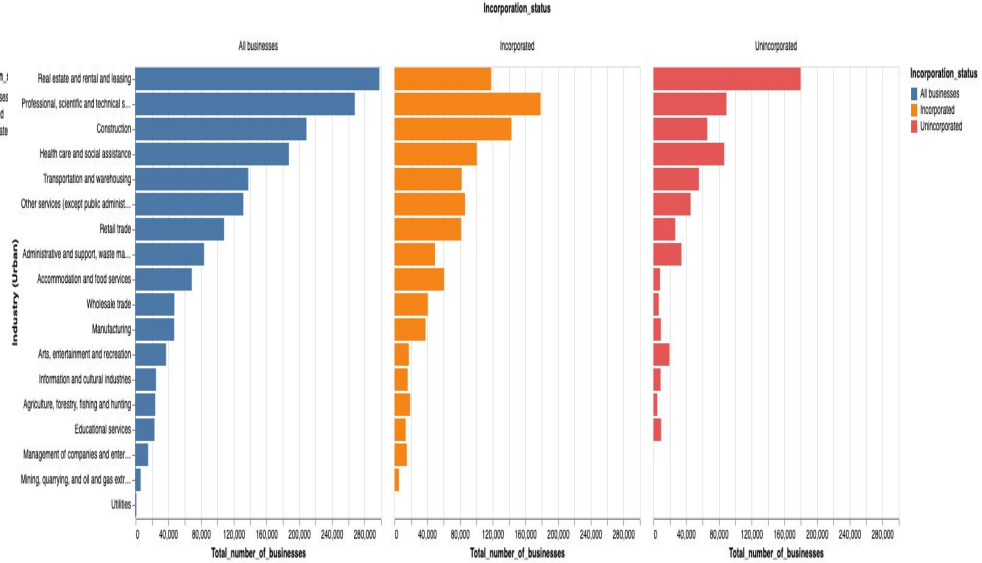




Some Data Visualizations

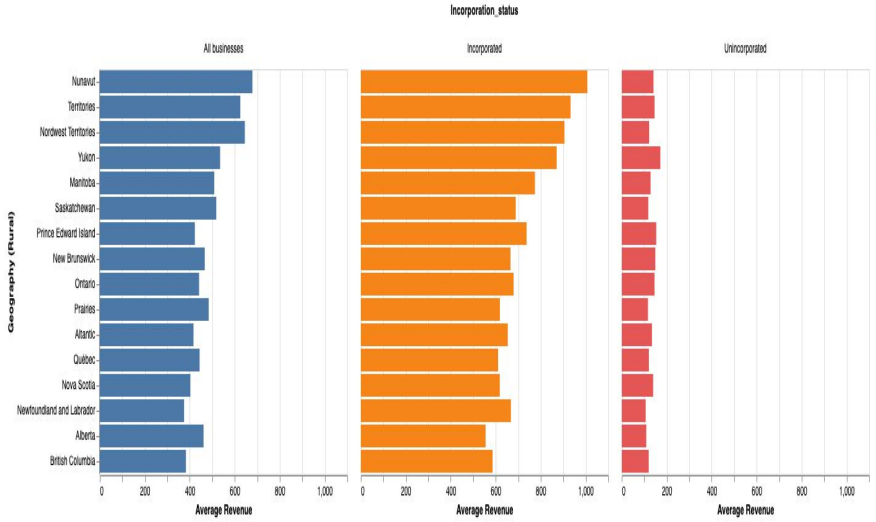
Tab1: Small Business Revenue (Total number)

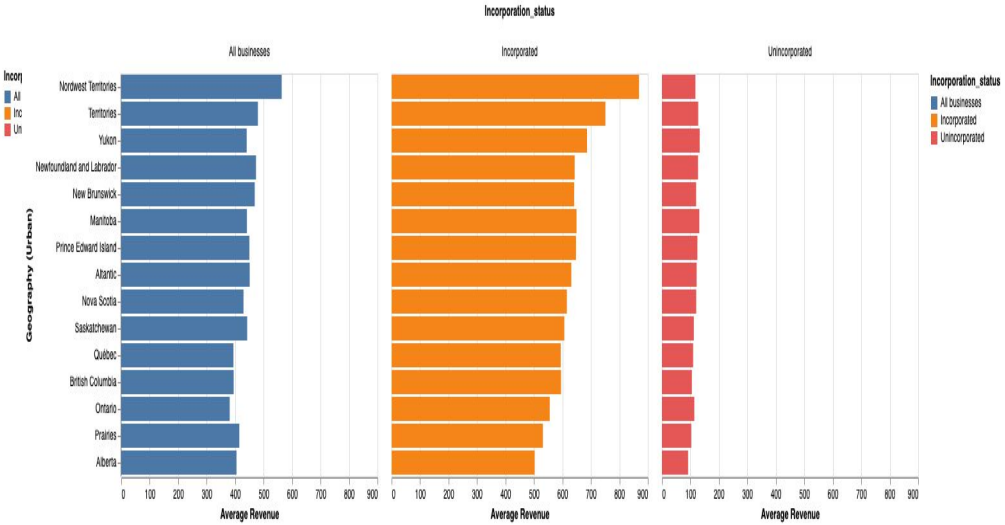




Some Data Visualizations

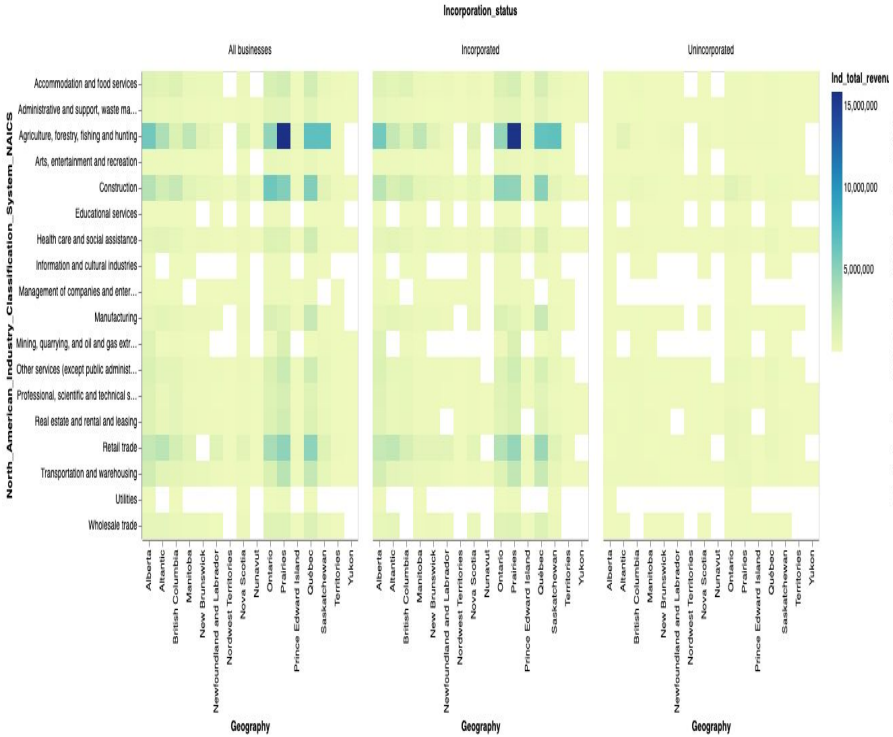
Tab1: Small Business Revenue (Average revenue)

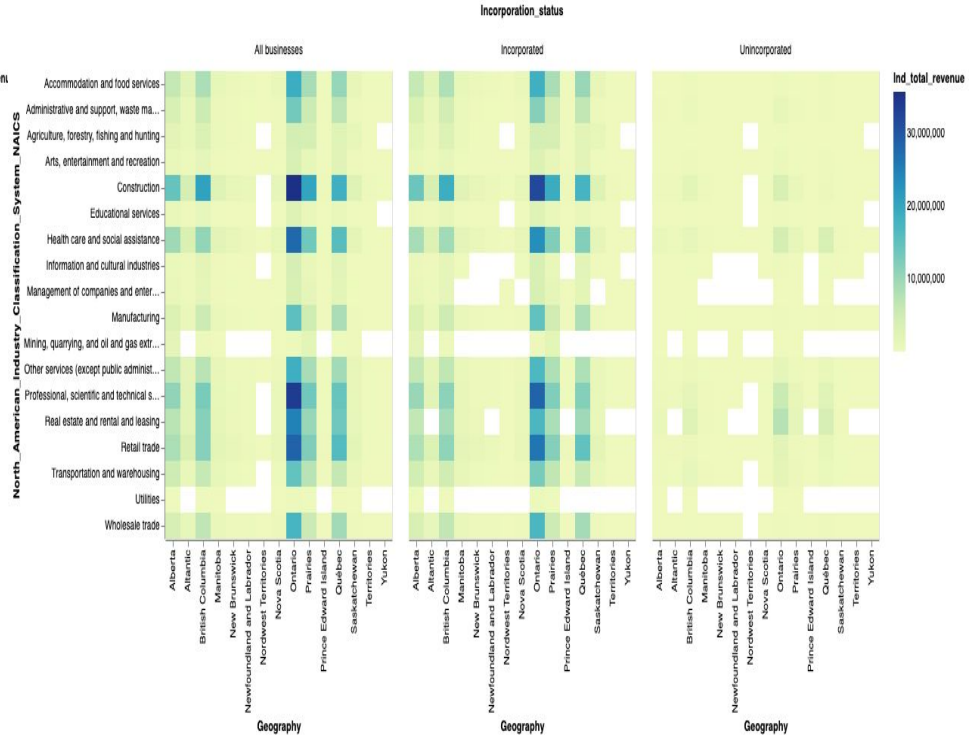




Some Data Visualizations

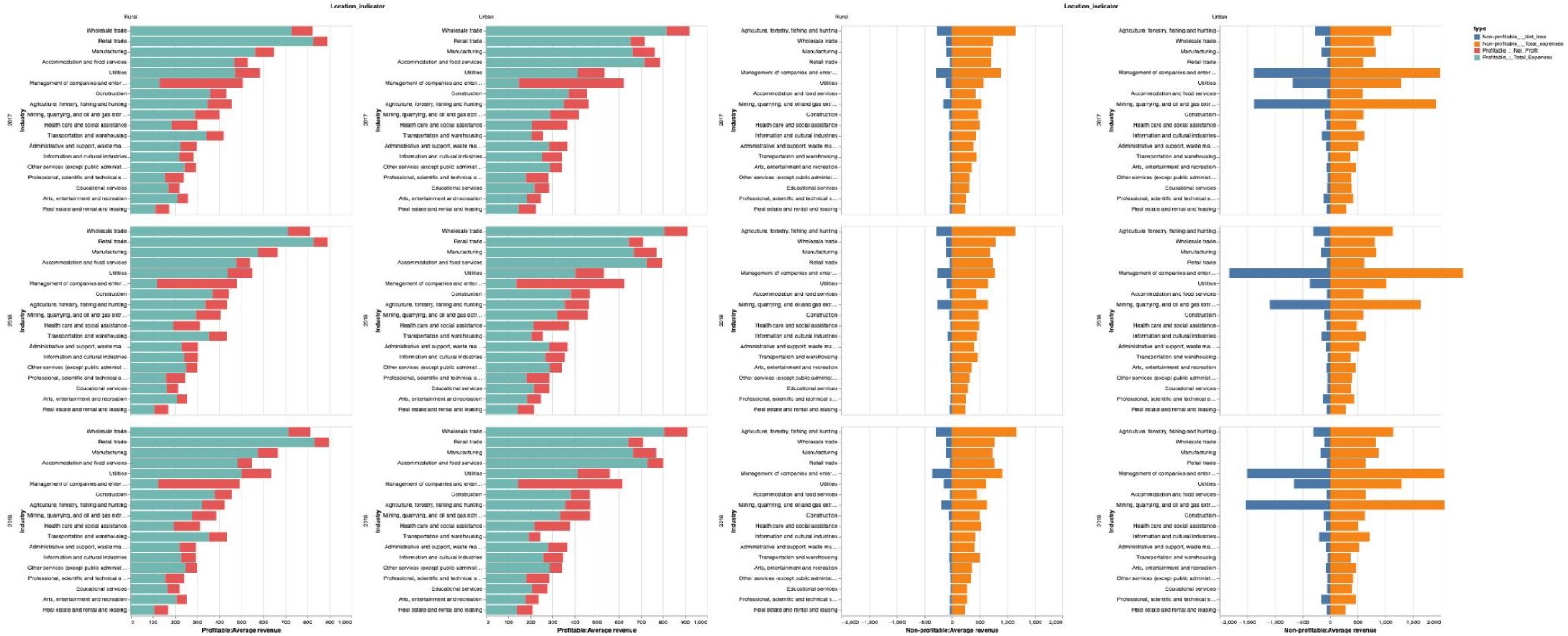
Tab1: Small Business Revenue (Total revenue)





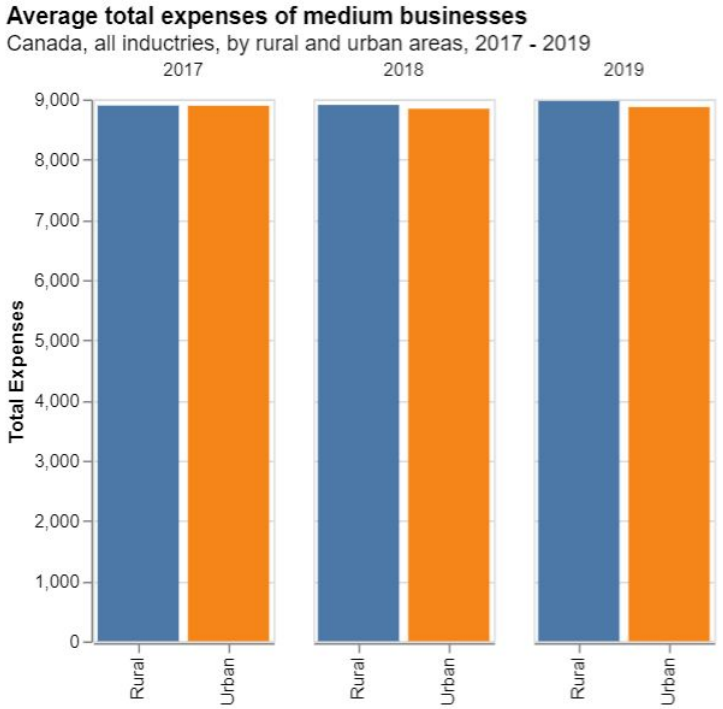
Some Data Visualizations

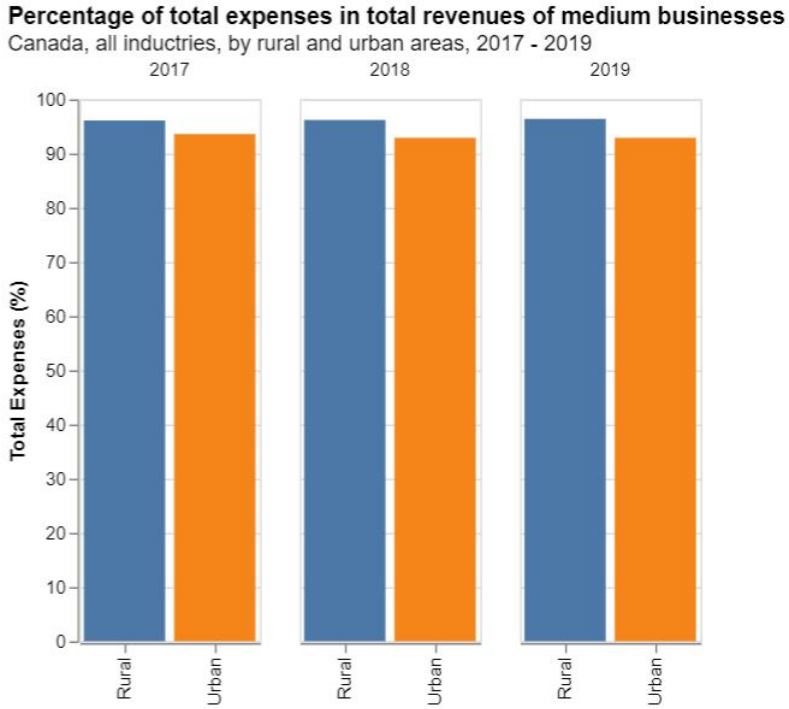
Tab7: Small Business Revenue (Average revenue)



Some Data Visualizations

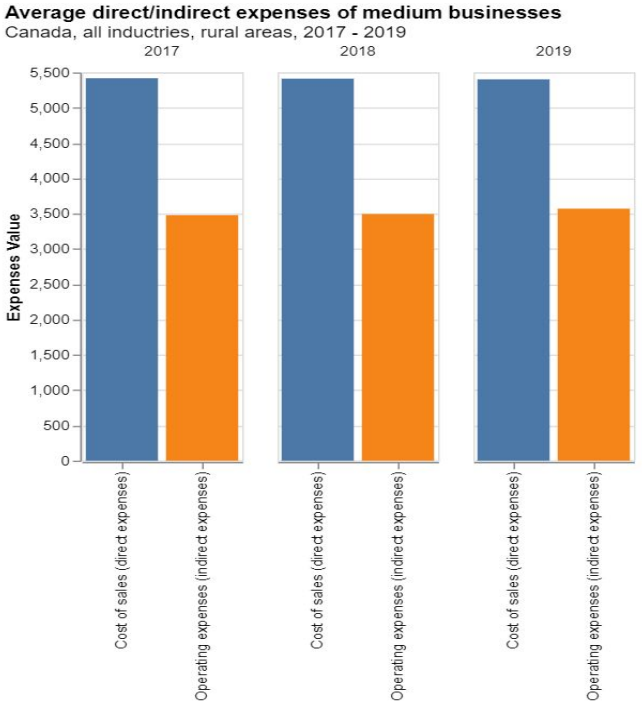
Tab 2 - Selected expense items as a percent of total revenue and Tab 3 - Selected expense items values

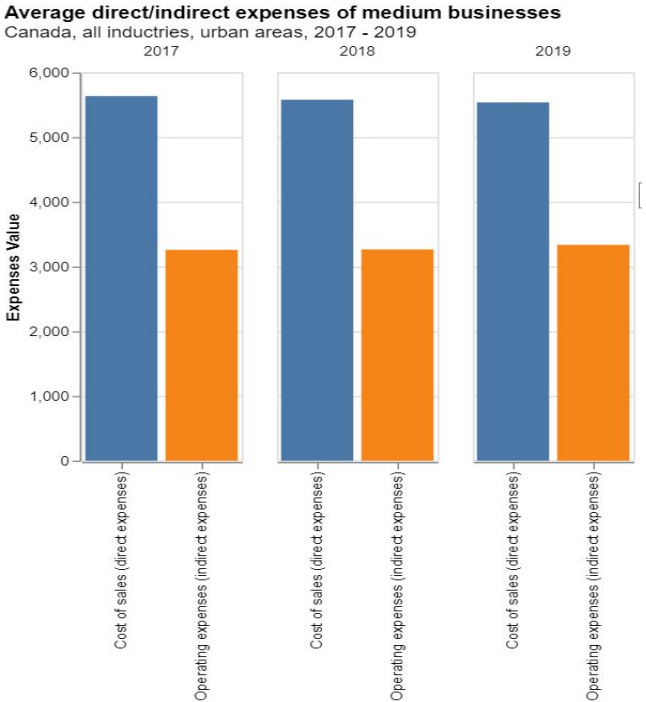




Some Data Visualizations

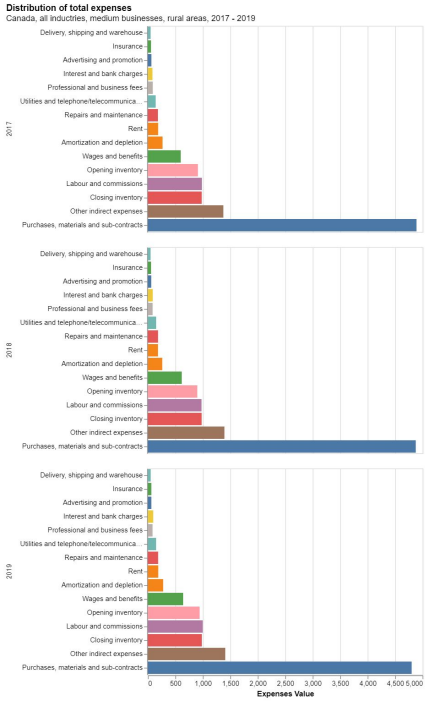
Tab 2 - Selected expense items as a percent of total revenue and Tab 3 - Selected expense items values

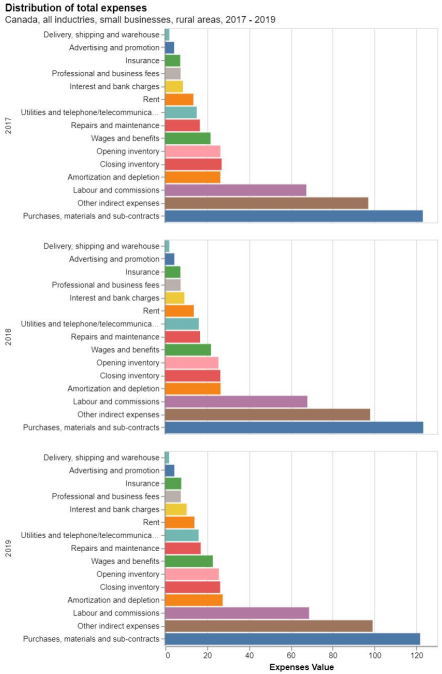


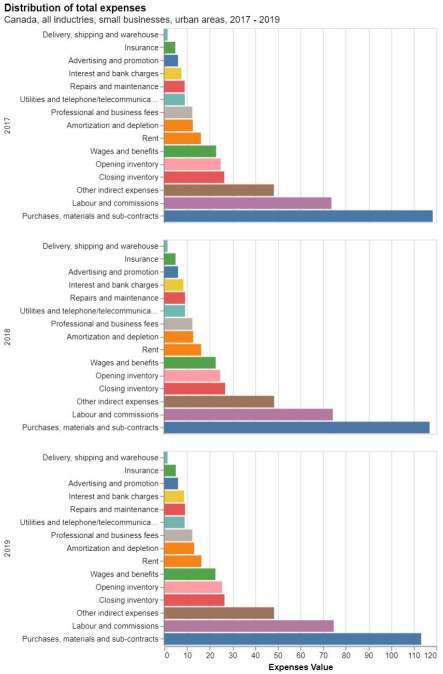


Some Data Visualizations

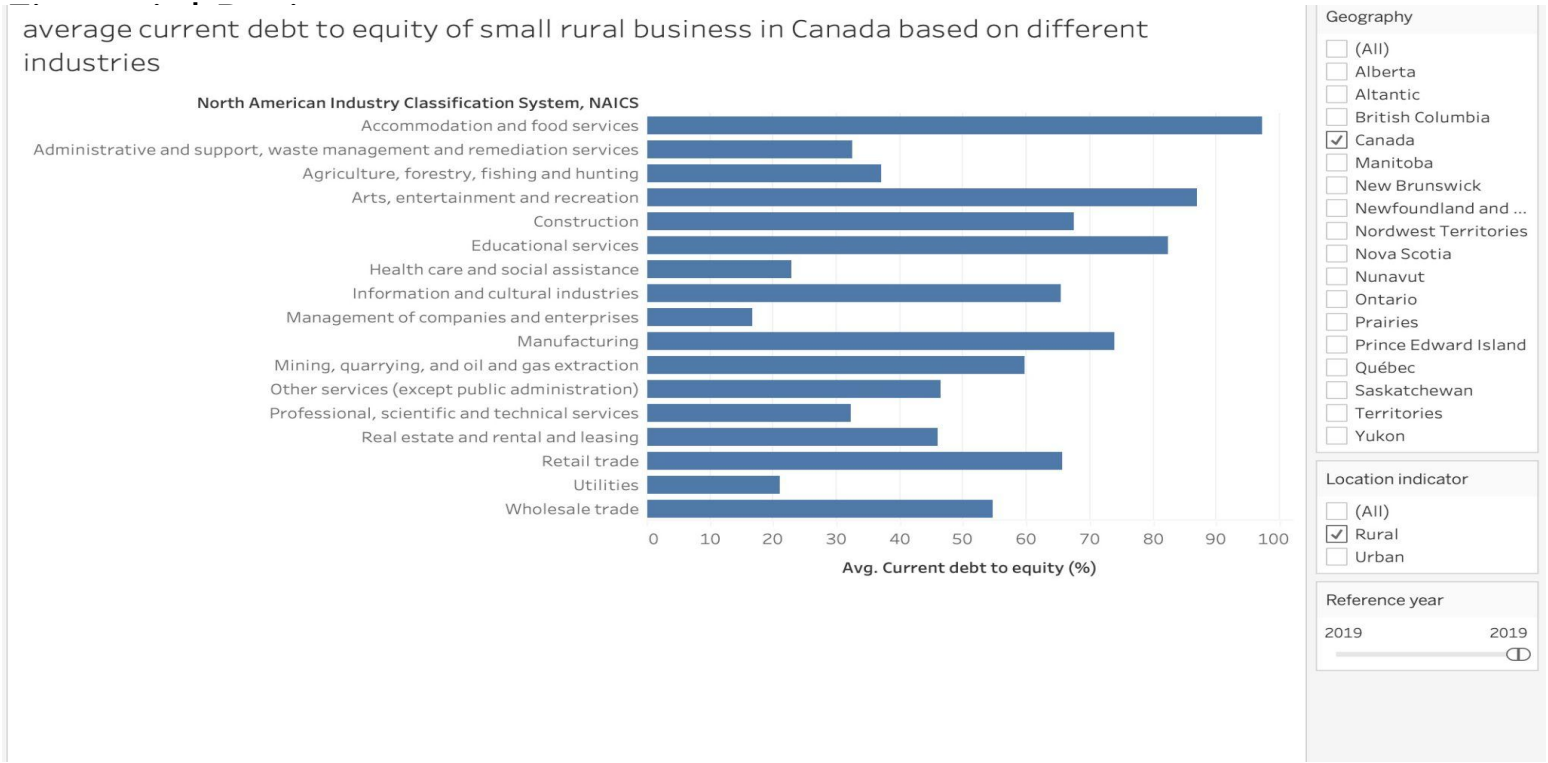
Tab 2 - Selected expense items as a percent of total revenue and Tab 3 - Selected expense items values



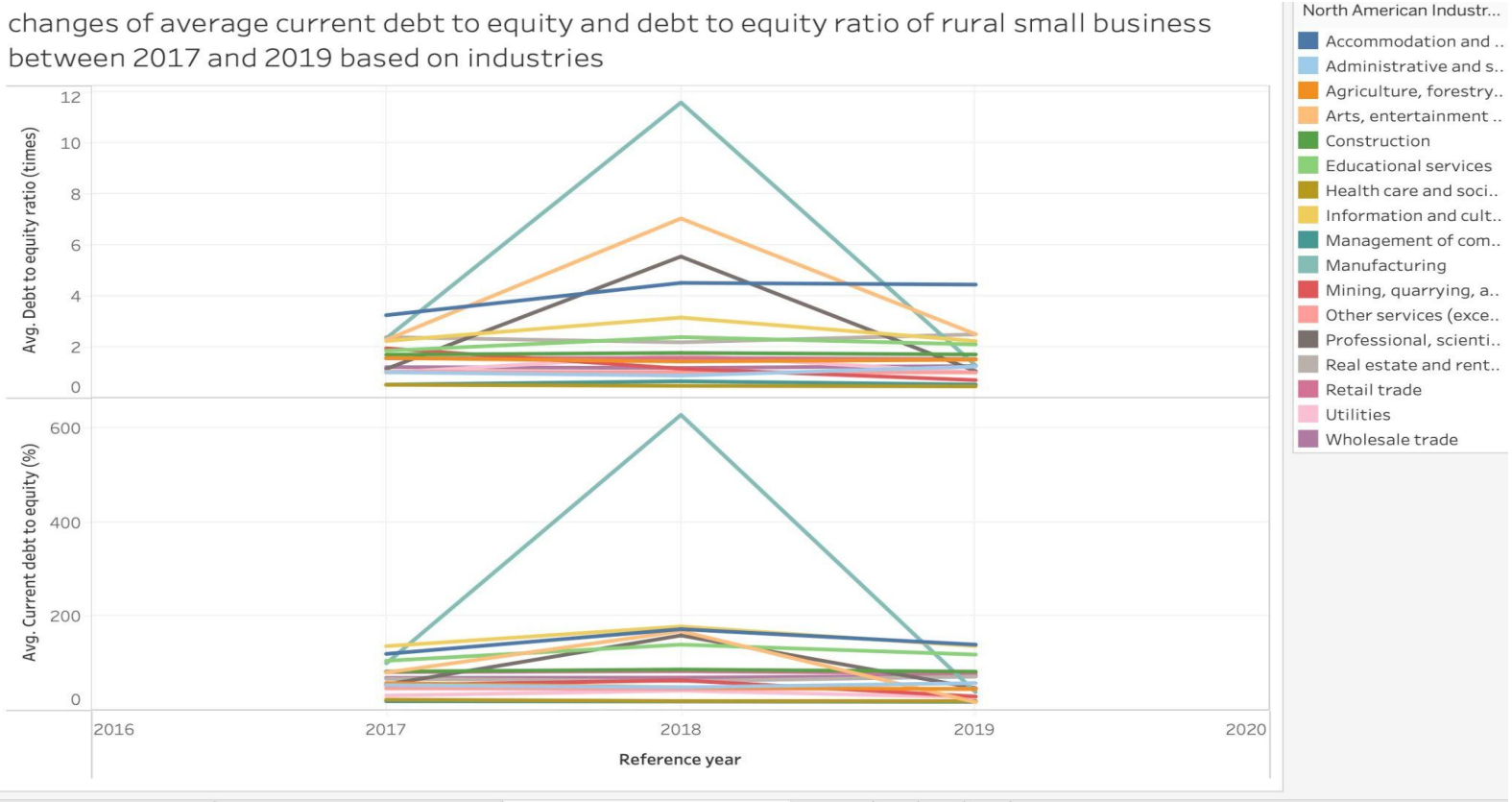




Some Data Visualizations



Some Data Visualizations



Some Visualizations

Financial Ratios:

Gross Margin

Net Profit to equity

Return on total assets

Revenue to equity ratio

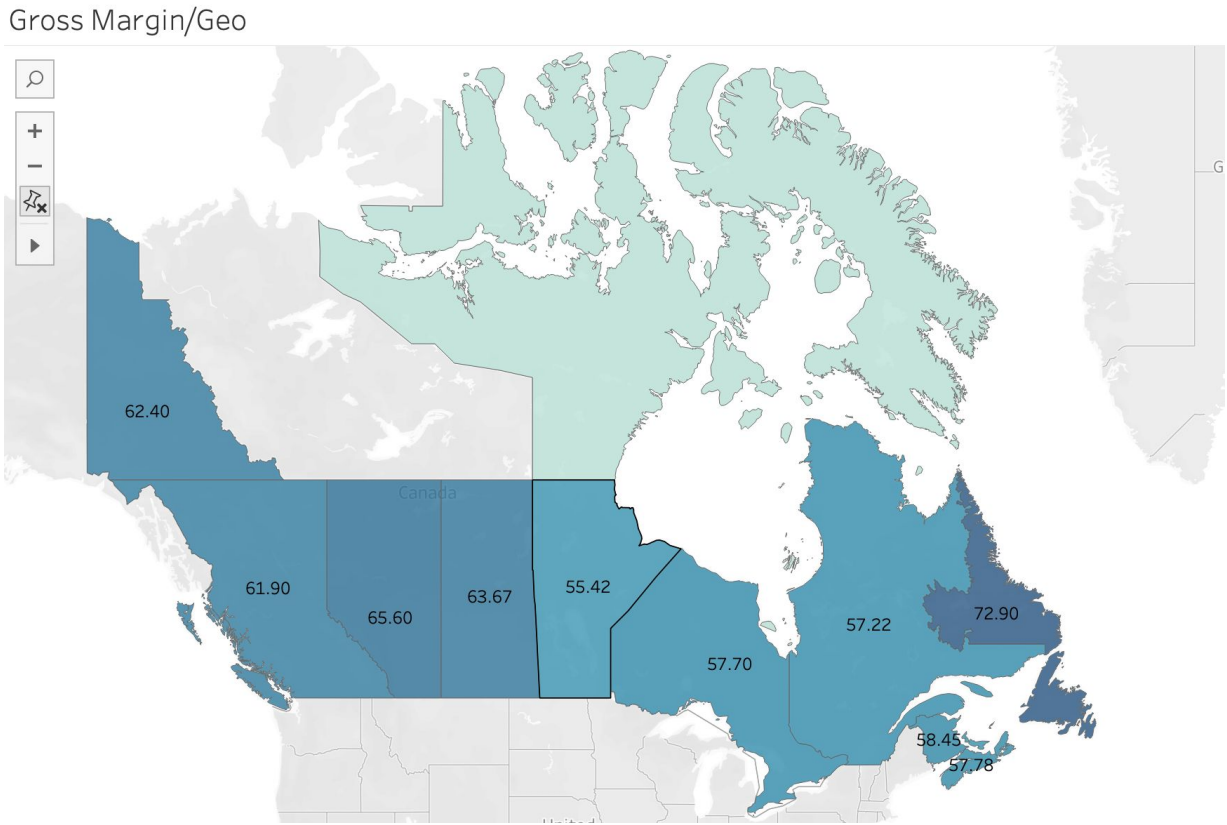
Finanical Ratio in different industries

North American Industry Classification System, NAICS	Avg. Gross margin (%)	Avg. Net profit to equity (%)	Avg. Return on total assets (%)	Avg. Revenue to equity ratio (times)
Professional, scientific and technical services	68.3	74.1	35.0	2.7
Construction	34.3	37.5	14.5	4.5
Health care and social assistance	76.4	36.6	10.2	5.7
Real estate and rental and leasing	63.8	26.2	11.1	1.5
Mining, quarrying, and oil and gas extraction	57.7	23.7	10.6	2.4
Other services (except public administration)	39.5	23.0	14.5	2.0
Educational services	93.3	22.1	7.4	3.5
Manufacturing	27.0	20.5	9.8	3.2
Administrative and support, waste management and remediation servic..	55.0	20.0	8.9	2.4
Retail trade	20.7	17.4	8.0	6.3
Accommodation and food services	57.3	17.2	7.6	2.7
Wholesale trade	20.0	15.0	7.3	4.4

Some Visualizations

Financial Ratios:

Gross Margin



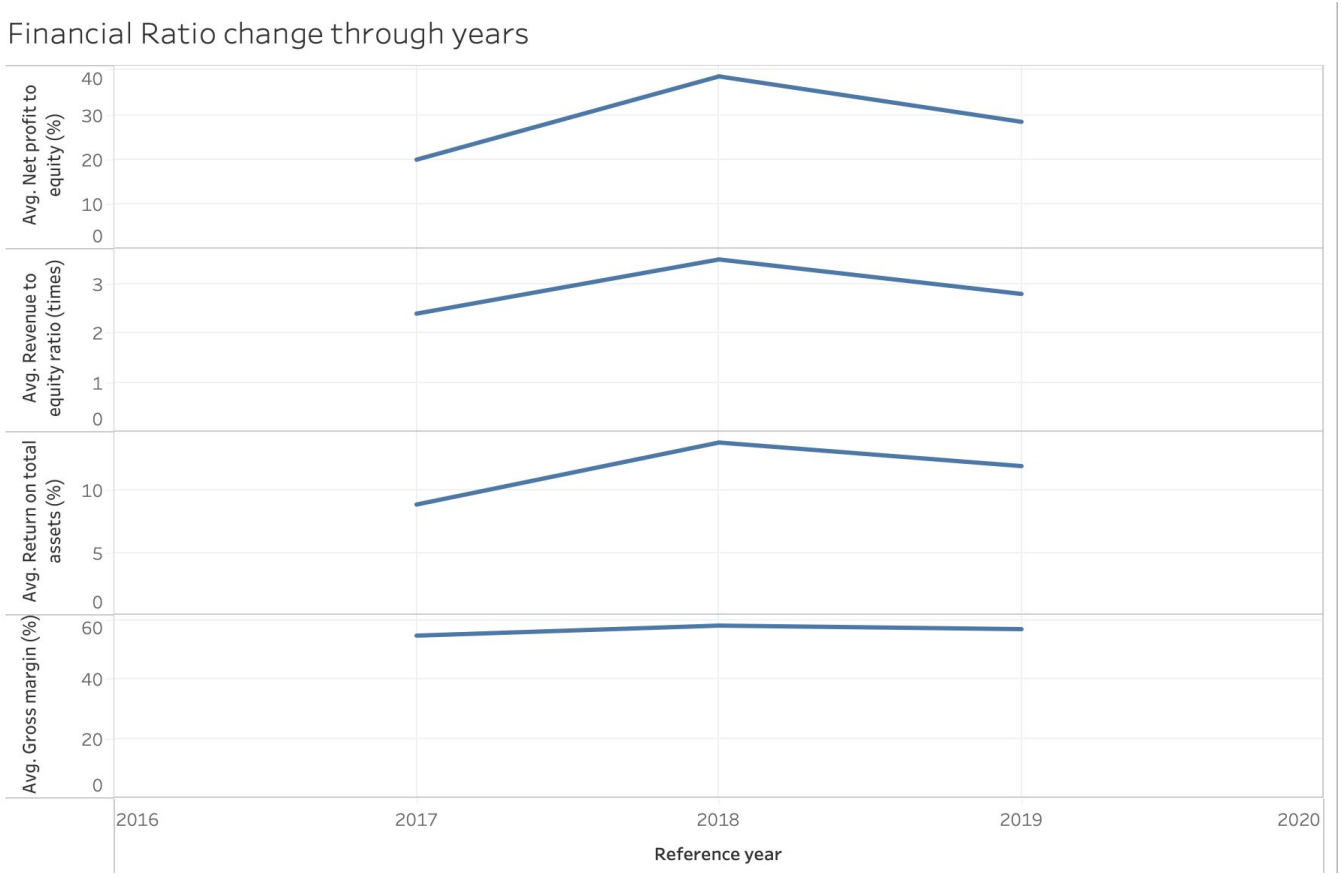
Some Visualizations

Financial Ratios:

Net Profit to equity

Revenue to equity ratio

Return on total assets



Compare to original plan

2 (10 – 17 May)	EDA and data visualization	Understand the RCBP dataset, data wrangling, do EDA and data visualization, develop analytical aspects

RCBP Database

Data wrangling and cleaning

Data visualization

We are on track with our original timeline.

Team Minutes

Meeting: Team + Clients 9 hr

1. 5/9 Team:Database Distribution 1 hr
2. 5/10 Team:Database discussion 2 hr
3. 5/11 Client: Q&A 1 hr
4. 5/11 Team: Distribute EDA 1 hr
5. 5/13 Team: Progress share 2 hr
6. 5/14 Team: EDA complete & presentation 2 hr

Individual Logs

Tingwen: 24 hr + 9 hr (team)

1. Database exploration and concept understanding: 4 hr
2. Tableau Learning: 4 hr
3. Data Cleaning and Wrangling: 5 hr
4. Data Visualization: 6 hr
5. Comments on the visualization: 3 hr
6. Presentation slides and preparation: 2 hr

Individual Logs

Bowen Yang: 36 hr

1. Upload data and write loading code for .csv files: 2.5 hr
2. Write meeting minutes: 0.5 hr
3. Explore data table 2 and 3: 9hr
4. Data wrangling: 6 hr
5. Data visualization: 10 hr
6. Edit data wrangling, visualization script and comment on visualization: 8 hr
7. Presentation slides and preparation: 1 hr

Individual Logs

Song(Alice) Zhang: 34.5 hr

May 9 Prepare proposal presentation and team meeting (3h)
May 10 Database exploration and team meeting (6h)
May 11 Client meeting, team meeting, organize work in trello (4h)
May 12 Tab1 and Tab7 EDA: data cleaning, data wrangling, data visualization (7.5h)
May 13 Tab1 and Tab7 data visualization, team meeting (4.5h)
May 14 Medium business Tab1 EDA (6.5h)
May 15 Small business Tab1 EDA and weekly slides (7h)

Individual Logs

Yilin Sun: 25 hr + 9 hr (team minutes)

May 9	Go through the dataset and learn the background knowledge for the dataset (5 hours)
May 10	Explore the dataset and find some features for future analysis (5 hours)
May 12	Write my own EDA about debt ratio and make some visualizations using altair (8 hours)
May 13	Learn how to use tableau and make some visualizations in tableau (6 hours)
May 14	Make slides and write the documentations for this week (1 hour)

Next Week Plan

Discuss current progress with client

Data visualization upgrade

Deliver the first draft of the analytical paper