Data 604 - Working with Data at Scale - Group 7

Analysis of the Canadian Banking Industry

By Luke Bramfield, Ethan Burke, Riki Enquist and Niza Ngwira

Table of Contents

Introduction

Our personal finances and the financial systems in which we operate are incredibly impactful in determining our quality of life and the opportunities that are available to us. A robust and well-functioning financial and banking system is one of the most important factors that contributes to a country's global economic competitiveness and the health of its society and citizens. In Canada, the “Big 5” Banks along with the Bank of Canada form the foundation of our financial system. The Big 5 banks are Royal Bank of Canada (RBC), Toronto-Dominion Bank (TD), Bank of Montreal (BMO), Canadian Imperial Bank of Commerce (CIBC), and Scotiabank (BNS). These are the largest chartered banks in Canada by market capitalization.

In this report we will be assessing the banking industry from a multitude of diverse business operations that will help foster meaningful insights into our individual analyses and the connections between them. More specifically we will be reviewing transaction data between the Bank of Canada and the Receiver General, lending conditions from reports generated by the Bank of Canada, geo-spatial location and demographic data from the City of Calgary and consumer sentiment data from the subreddit PersonalFinanceCanada. We highlight findings made by the analysis of each dataset and connect the datasets at applicable points of convergence. Lastly we conclude our report with a simple yet impactful perspective on the importance of the interaction between the Bank of Canada, Canadian banks, and Canadian Citizens and how the push and pull between these key forces dictate the state of the Canadian financial system that we all participate in.

Summary of Guiding Questions

1. What is overall consumer sentiment surrounding the banking industry and how do the banks compare
2. How do the volume and types of transactions processed by the Big 5 vary seasonally over the last 5 years?

3. How do lending conditions translate into reality and how do they affect the consumer?
4. How does the location of the big 5 Chartered Banks (BMO, CIBC, TD, RBC, BNS) in Calgary relate to median income, country of birth of immigrant, homeownership count, and population count?

Individual Data Explorations

Customer Sentiment Analysis - Luke

For this portion of our Canadian Banking Industry project, we will be performing some customer sentiment analysis with python libraries such as NLTK, Chat GPT 4 API and transformers libraries. We want to assess how people feel toward the Canadian Banking Industry as a whole, how sentiment towards the Big 5 Banks compare to one another and how their sentiment compares to other financial institutions such as no-fee online banks and Credit Unions.

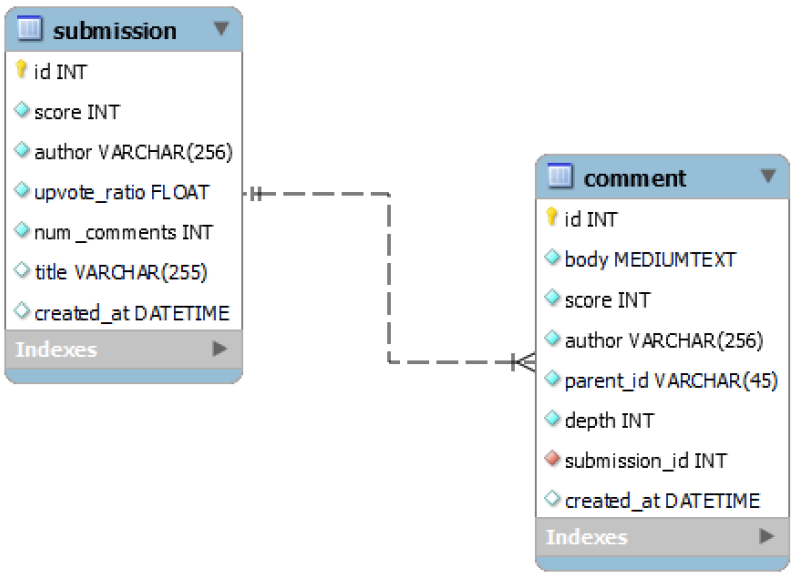
Description of Data Source

To undergo this analysis we have parsed our own data from the subreddit "r/PersonalFinanceCanada '' using the Python Reddit API Wrapper (PRAW). We created our own Reddit application, followed their API process and were granted a licence to retrieve and use their data for research purposes (see appendix).

“PersonalFinanceCanada” is a vibrant subreddit in the reddit community where people go to ask questions, comment, seek advice and share thoughts and opinions on anything related to personal finance in Canada. It has been around since 2012 and there are approximately 1.3M members in the community representing a significant subpopulation of Canadian banking customers.

Description of Dataset

This image provides a visual summary of the relational database used to store the data retrieved from the reddit api, the database structure and its associated attributes.



In terms of data dimensionality the submission table contains 89 “title” entries which represent the subreddit thread titles, an example would be “Big Banks vs online banking”. The comment table includes the features related to that specific title. The “body” attribute which contains all the comments extracted from the thread is the main input variable used for the sentiment analysis. There are a total of 21,721 comments in the database.

Cleaning Steps

The first step of obtaining the required data for the sentiment analysis was to perform an SQL query to retrieve any comments that contained at least one of the names of the banks of interest, less any “[deleted]” comments. The query returned a total of 2,807 comments. A

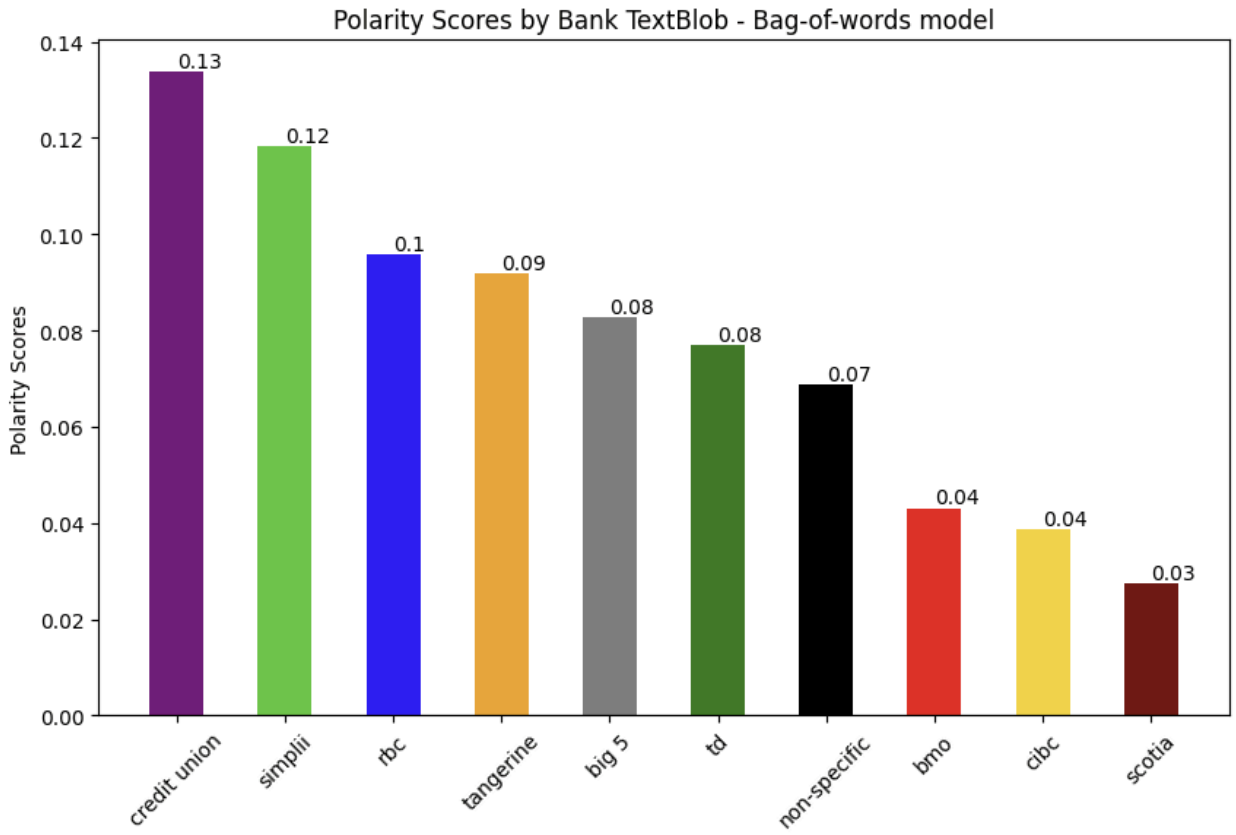
major problem encountered was that many comments were complex and contained the names of multiple banks, with multiple sentiments towards different banks. To separate and isolate sentiments for each bank within each comment we utilized the Chat GPT 4 API and looped through all comments, giving it detailed instructions on how to transform and structure the data (see appendix). This data was written to a text file and loaded into a python dataframe that enabled us to perform the analysis.

Chat GPT 4 has impressive accuracy but still makes mistakes so we had to perform some additional cleaning. We sliced the dataframe to only return banks of interest and to map and combine variations of the bank names into one common label. Additionally we removed rows containing missing, empty or incomplete data. Lastly we sliced this dataframe into smaller data frames for the categories you will see in the visualization of our results. This process resulted in 5080 comments that would be used for the calculation of sentiment scores.

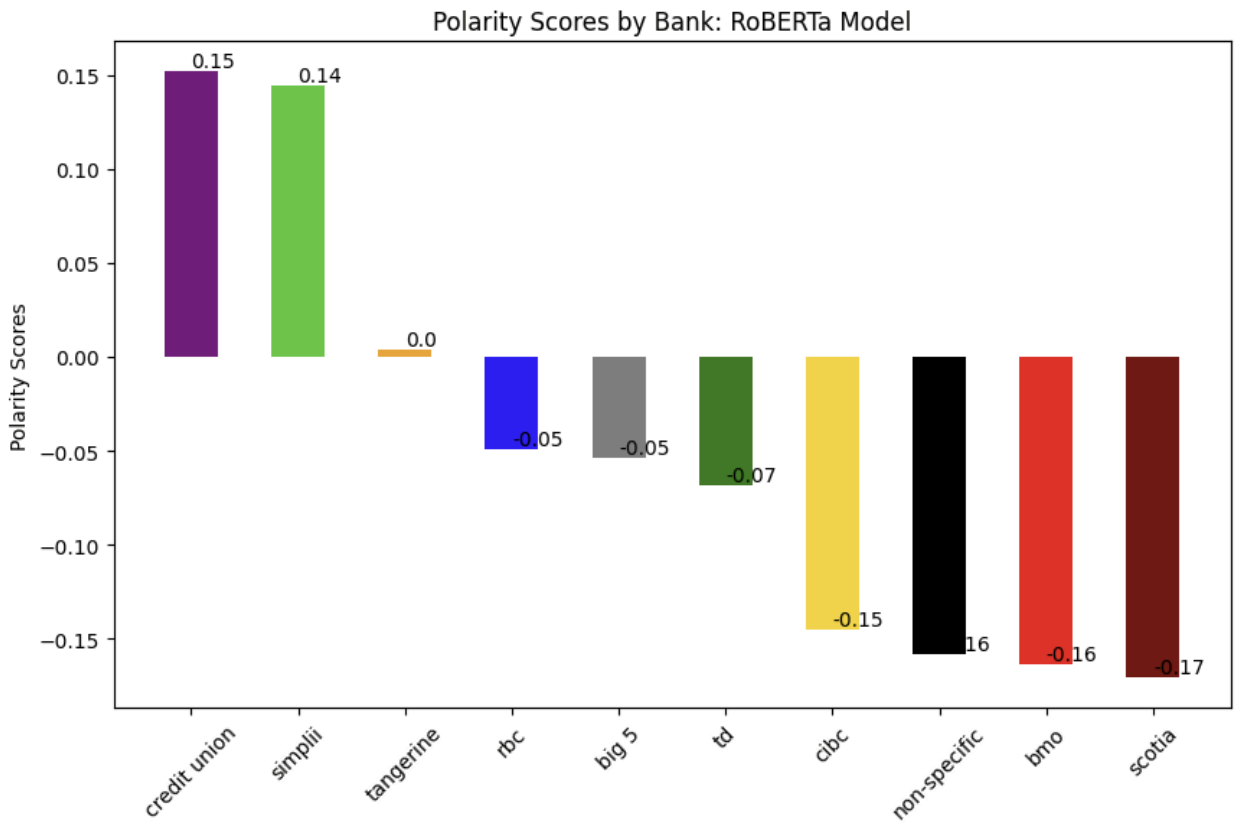
Analysis Framework

To compute sentiment scores we deployed three different NLP models. This way we were able to assess the scores using multiple techniques that evaluate language in their own unique way. The models used were the TextBlob bag-of-words model, the Chat GPT 4 model, and the Hugging Face Twitter-roBERTa-base model. The TextBlob model essentially calculates sentiment based on the mix of words found (positive or negative) in a comment and has no ability to understand context in human language. The roBERTa and Chat GPT model contain the ability to understand the nuances of human language such as context and sarcasm. Note that the roBERTa model and TextBlob model assign a “polarity score” between negative one to positive one. Negative one being the maximum negative sentiment value and positive one being the maximum positive value. For the Chat GPT 4 model we asked it to label comments as positive, negative, neutral, or mixed and took the ratio of positive to negative comments to give a similar comparison of sentiment to the other two methods. By taking this approach we are able to cross-validate results giving us more confidence in the conclusion.

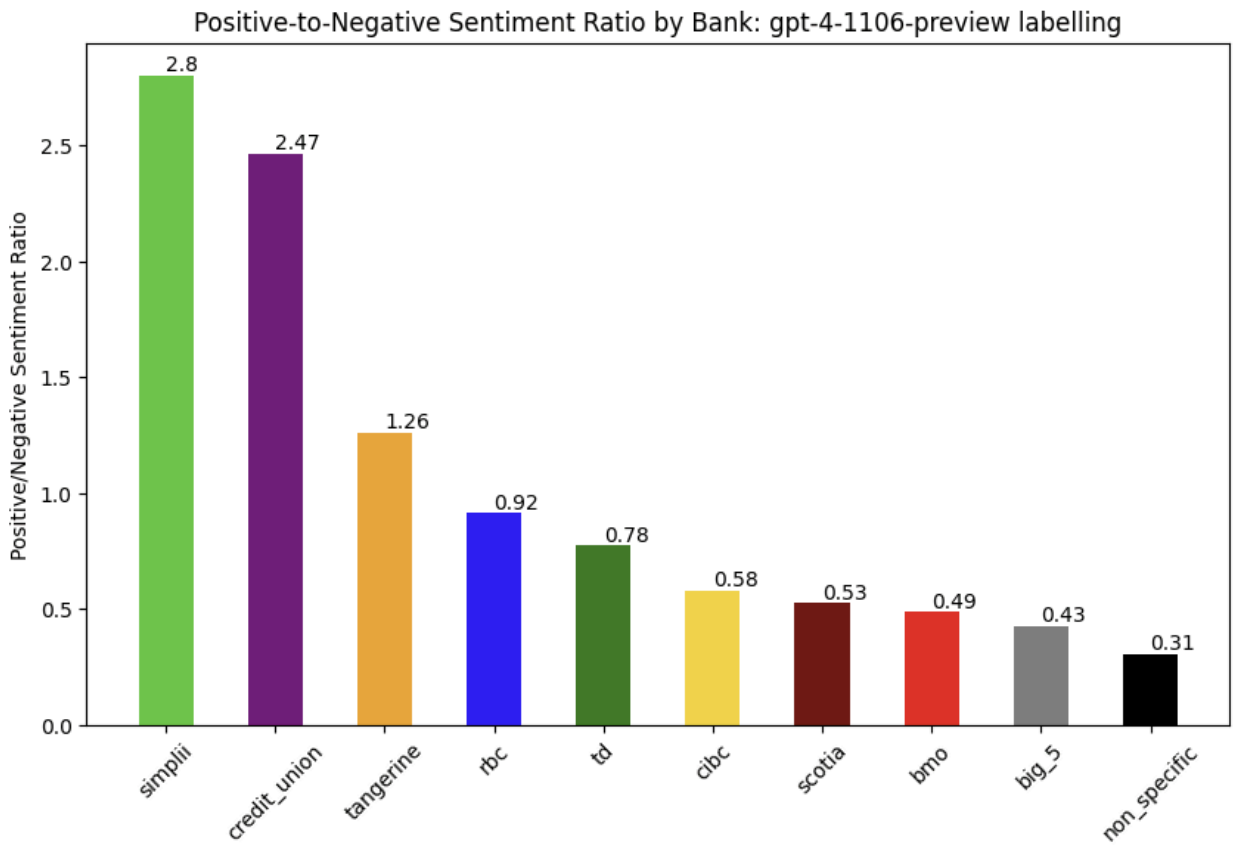
Results and Visualization



From the TextBlob model, notice that the overall distribution of scores indicates the sentiment is neutral for the banks as a whole. Clearly Credit Unions and Simplii financial have the highest polarity scores followed by RBC and tangerine.



The roBERTa model tells a slightly different story. Notice that overall the sentiment scores in general are now more skewed to the negative. However, the ranking of the banks is similar to the previous graph with Credit Union and Simplii leading in scores, trailed again by Tangerine and RBC. While the remaining Big 5 are clustered together with very comparable scores.



The Chat GPT 4 results are more similar to roBERTA's with overall sentiment also skewed to the negative and exhibiting a very similar pattern in terms of ranking.

Summary

Summarizing the results of the sentiment analysis we can see that the general sentiment towards banks seems to be on the negative end of the spectrum. Under all three models credit unions had the highest scores indicating that PersonalFinanceCanada Subredditors have better experiences with these institutions. RBC is the clear leader in terms of sentiment when comparing amongst the Big 5 banks. The remaining Big 5 are closely related in terms of sentiment.

Transaction Analysis - Riki

Our next guiding question targets the analysis of seasonal transaction trends processed by the Big 5 banks, using datasets from the Government of Canada spanning the last five years. This time frame was chosen to capture the most up-to-date economic cycle, including the impact of a global pandemic, providing a current perspective on financial trends. The datasets mirror the rapid evolution of the financial industry, shaped by technological advances, regulatory shifts, and changes in consumer behavior. This analysis is particularly insightful for its revelations about banks' financial behavior, playing a critical role in economic planning, policy-making, and influencing banks' market reputations. A key aspect of our study is assessing the impact of the COVID-19 pandemic on the banks' risk management strategies.

Dataset Explanation

The datasets, provided by the Department of Public Services and Procurement Canada and managed through the Government Banking System (GBS), offer detailed information on financial transactions, including deposit dates and amounts across various institutions. Organized annually from April 1st to March 31st and available in CSV format, these datasets are well-suited for data analysis tools manipulation.

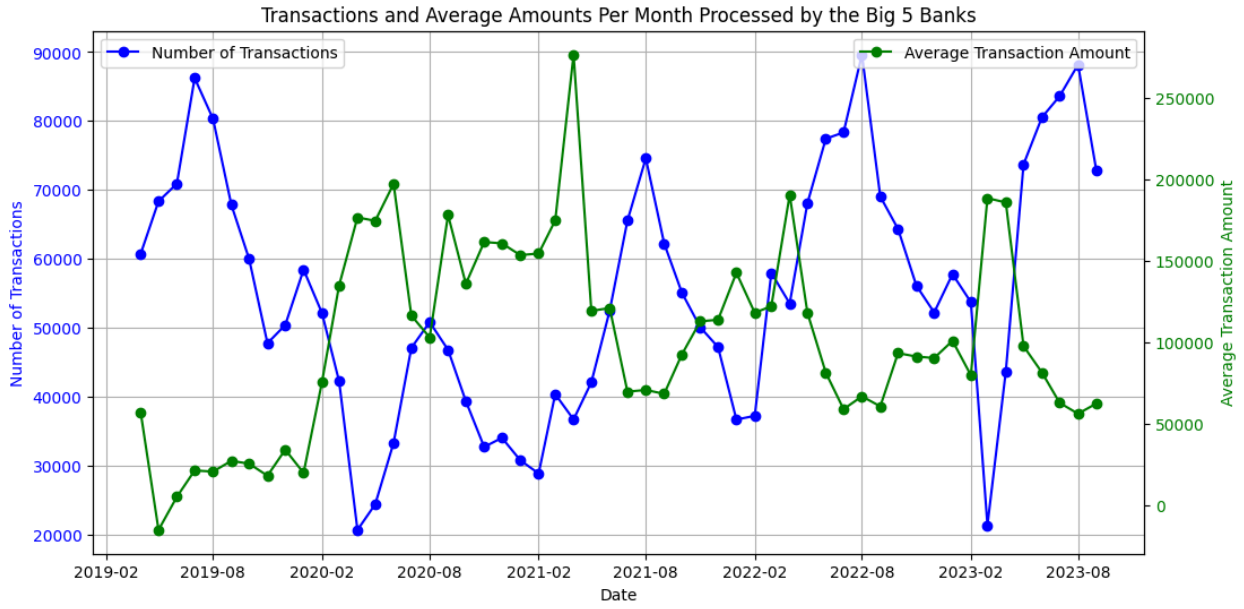
Analysis Steps

By utilizing MySQL for storage, the project will efficiently manage large data volumes, enabling us to identify and analyze transaction trends. Our methodology includes data cleaning, transformation, exploratory, and time-series data analysis, aiming to provide actionable insights for banks, government bodies, and researchers. This analysis is vital for government bodies in resource allocation and budget forecasting, for financial institutions in risk management and operational efficiency, and for researchers in understanding market indicators and economic health.

Cleaning Steps (Important Queries)

1. Filtering the Big 5 Banks: This query retrieves all records from a table and filters them to include only those associated with the Big 5 Banks (identified by specific Financial-Institution-Numbers). The filtered data is then efficiently written back to the database in chunks.
2. Joining All Tables: This query merges records from tables filtered for the Big 5 Banks into a new table, creating an aggregated dataset of transactions from April 2019 to September 2023.
3. Counting Number of Transactions by Month: This query extracts the year and month from transaction data and counts the number of transactions for each year-month combination, organizing the results in a DataFrame with columns for 'Year', 'Month', and 'NumberOfTransactions'.
4. Calculating Average Monthly Transaction Amounts: The average_amount_query calculates the average transaction amount for each month and year, aggregating and sorting the data by year and month, and then presenting it in a DataFrame with columns labeled 'Year', 'Month', and 'AverageTransactionAmount'.
5. Transaction Types: The transaction_type_query categorizes and counts transactions based on their type codes, converting these codes into descriptive names, and organizes the results by the frequency of each transaction type.
6. Renaming Big 5 Banks and Counting Transaction Types: The transaction_type_and_bank_query categorizes and counts transactions by both type and financial institution, associating each transaction with a bank name, and sorting the results by each bank and the number of transactions for each type.

Total Number of Transactions and Average Transaction Amounts Per Month Processed by the Big 5 Banks

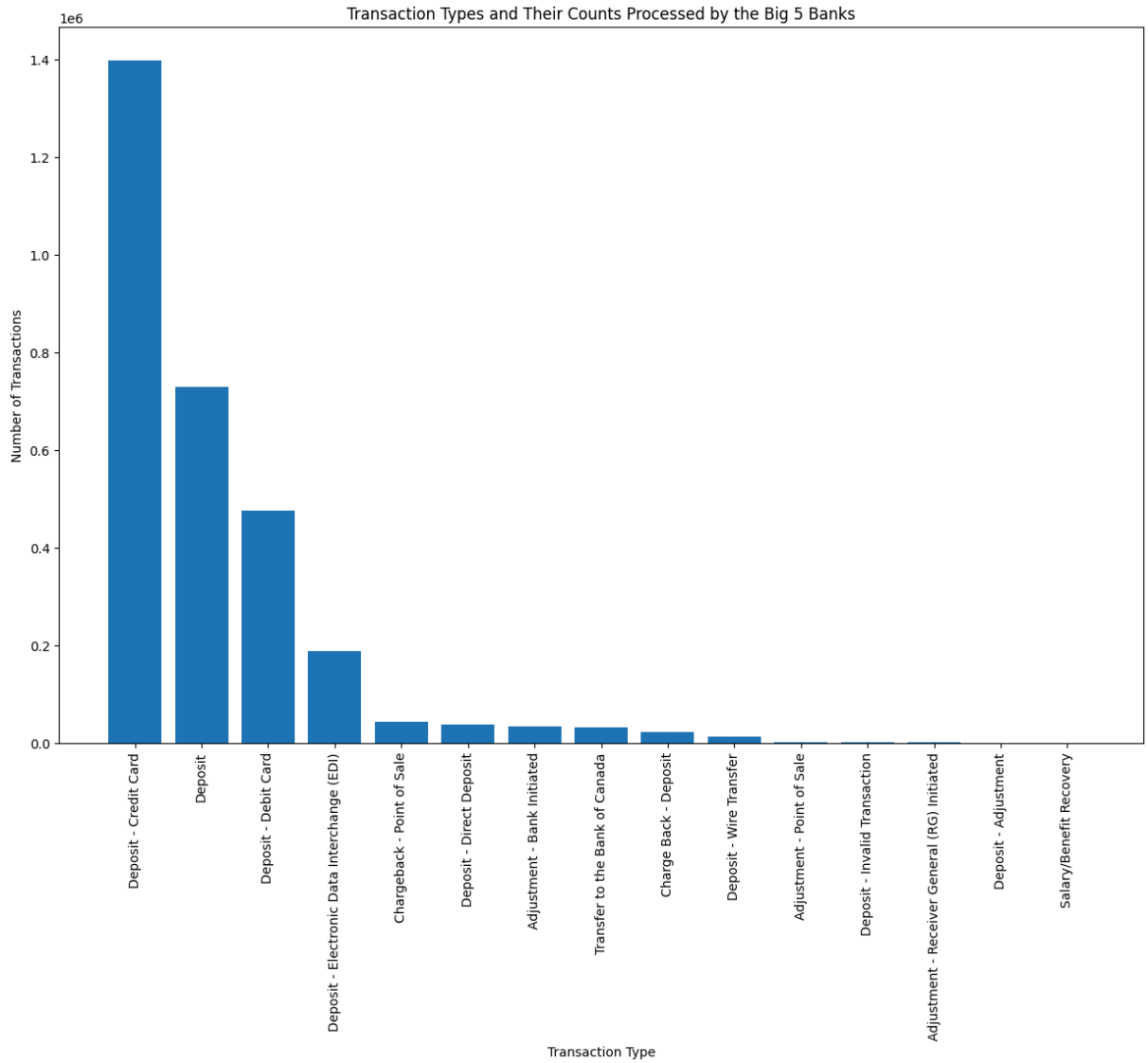


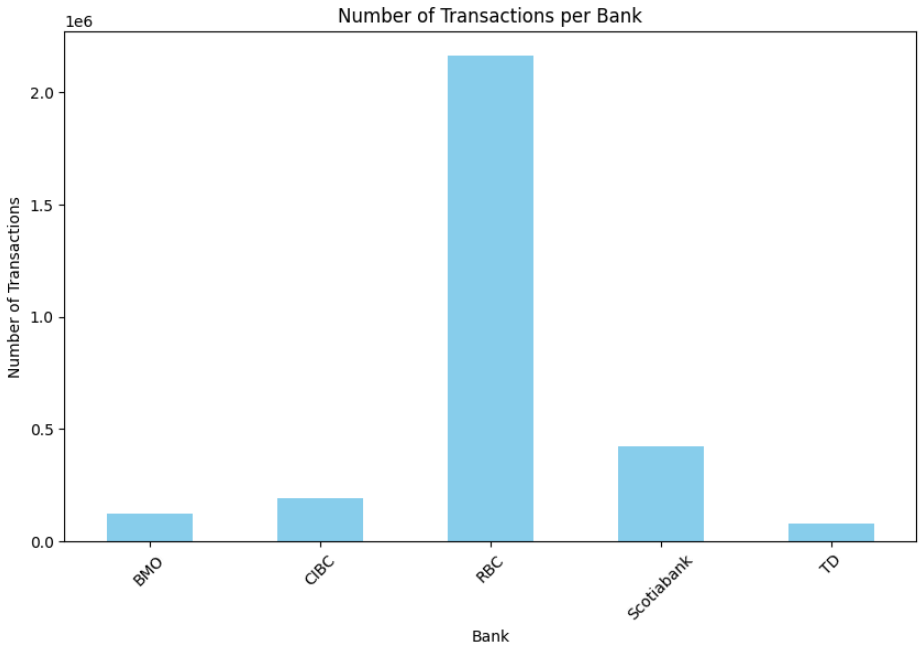
The transaction data from the Big 5 banks exhibit distinct seasonal business cycles, characterized by consistent spikes and dips throughout the year. In March, leading up to Canada's tax deadline, transaction volumes decreased, likely because individuals and businesses are finalizing financial records and tax submissions, usually through direct channels to the government, rather than the Big 5 banks. In contrast, transaction volumes rise during the summer months, potentially due to increased government spending on public projects and tax refunds after the spring's budget approvals. This pattern suggests that the dataset primarily reflects public fund flows managed by the Receiver General, rather than standard consumer banking activities.

The COVID-19 pandemic has notably influenced these trends, with a more pronounced low during the pandemic year and a comparatively lower peak, attributed to reduced economic activity from business closures and the implementation of financial support programs like the Canada Emergency Response Benefit (CERB). Additionally, the spring months show fewer but larger transactions, likely due to significant fiscal activities such as tax settlements, while the summer months display a higher number of smaller transactions,

reflecting routine financial activities like tax refunds and tourism-related revenues. This data highlights a seasonal shift in financial patterns, with spring focusing on larger, less frequent transactions and summer on more frequent but smaller transactions.

Transaction Type Counts and Totals per Big 5 Bank



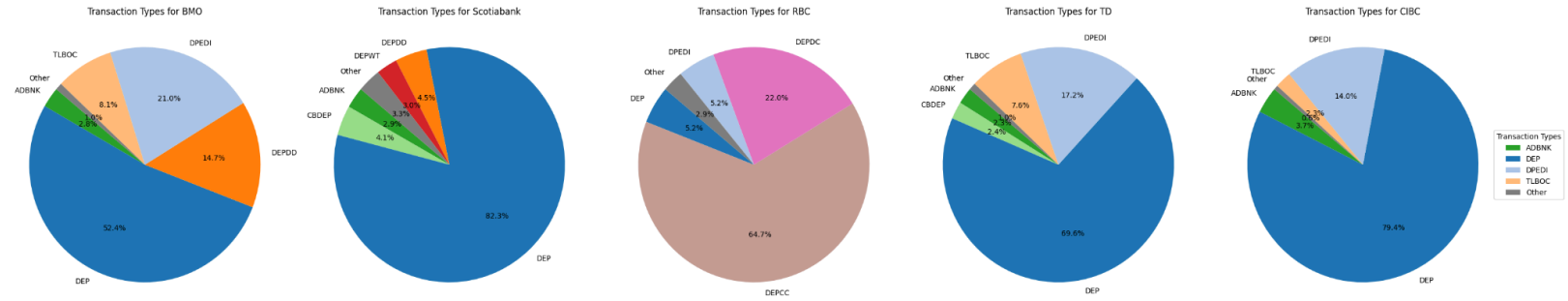


The bar graph displaying transaction types processed by the Big 5 Banks reveals that 'Deposit - Credit Card' transactions dominate, reflecting the prevalent use of credit cards for government service payments and aligning with the Receiver General's role in managing payments to the government. Following closely are general 'Deposits' and 'Deposit - Debit Card', indicating a significant volume of direct deposits, commonly used for tax payments and other government services. Other categories such as 'Adjustment - Bank Initiated' and 'Transfer to the Bank of Canada' appear less frequently, associated with specific banking adjustments or governmental fiscal transfers. These lower-count transaction types might represent less frequent but higher-value activities requiring meticulous management.

Interestingly, the majority of transactions are processed through RBC, possibly due to its larger customer base, extensive range of transactional services, or a more significant role in government banking. In contrast, other banks like BMO, CIBC, Scotiabank, and TD show fewer transactions, with TD having the least. These variances likely reflect each bank's scale

of involvement with government financial operations, influenced by factors such as the number and size of government contracts, their role in handling public funds, and their capacity to process large transaction volumes.

Transaction Type Distribution per Big 5 Bank



The pie charts display the distribution of transaction types among the Big 5 Banks, revealing that regular deposits ("DEP") constitute the majority of transactions for four out of the five banks. This predominance of deposits likely includes tax payments and fees for government services, reflecting the Receiver General's responsibility for managing public funds. Additionally, the "TLBOC" slice, representing transfers to the Bank of Canada, though smaller, is significant as it illustrates the movement of funds in the government's cash management process. Other slices like "RG" for Receiver General initiated adjustments and "DEPDD" for direct deposits underscore the variety of transactions overseen by the Receiver General for precise and secure financial management.

Notably, RBC demonstrates a substantial proportion of "DEPCC" transactions, indicating a higher volume of credit card-related government payments processed through it. This could be attributed to several factors: RBC might have a preferential partnership or a specific contract with the government for processing these transactions, or it could be the primary choice for credit card payments due to convenience, promotional strategies, or its processing capabilities. Additionally, the public's preference to use RBC for credit card payments to the government might be influenced by perceived benefits, rewards, or existing banking relationships.

Summary

The analysis of transaction data from the Big 5 banks reveals a clear seasonal pattern in transaction volumes, with peaks during the summer months and dips in March, suggesting a correlation with government spending cycles and tax activities managed by the Receiver General. The COVID-19 pandemic has accentuated these trends, showing more pronounced lows and muted peaks, reflecting the economic impact and government interventions. Credit card deposits are the most frequent transaction type, particularly through RBC, which may suggest a preferred relationship with the government or consumer choice. Regular deposits form the bulk of transactions for other banks, indicating routine government service payments. The data suggests that the Big 5's transaction volumes and types are influenced by their role in government banking, the scale of their operations, and consumer behavior in response to economic conditions and bank offerings.

Lending Conditions - Ethan

Introduction

In this section, we investigate the practical implications of lending conditions within the Canadian banking industry, focusing on their translation into real-world effects on consumers. Understanding these dynamics could be helpful to the consumer as access to credit is a significant driver of the economy, particularly in the housing market. For most individuals, purchasing a home is often the largest purchase of an individual's life and would be impossible without access to affordable credit.

Dataset

Our data comes from the Bank of Canada’s Senior Loan Officer Survey conducted quarterly which gauges the balance of opinion for lending behaviour of financial institutions. This survey highlights two main aspects in relation to Mortgage Lending: price conditions and non-price conditions. Price conditions refer to the actual cost of the loan, encompassing

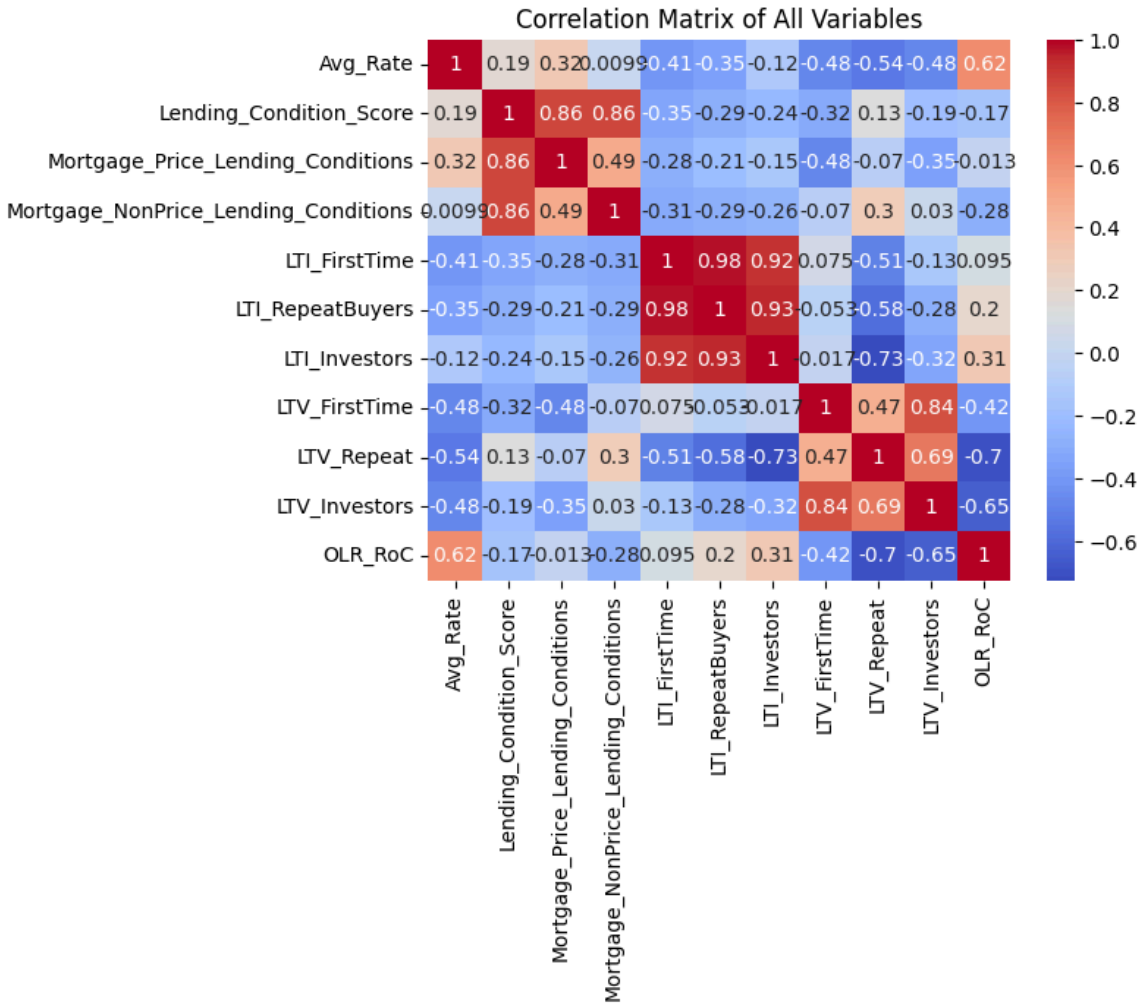
interest rates, fees, and other financial aspects. Essentially, this is what dictates the dollar amount consumers pay.

Non-price conditions, however, reflect the criteria consumers need to meet to secure and maintain your loan. This includes factors such as credit score, debt-service ratio, income, and employment status.

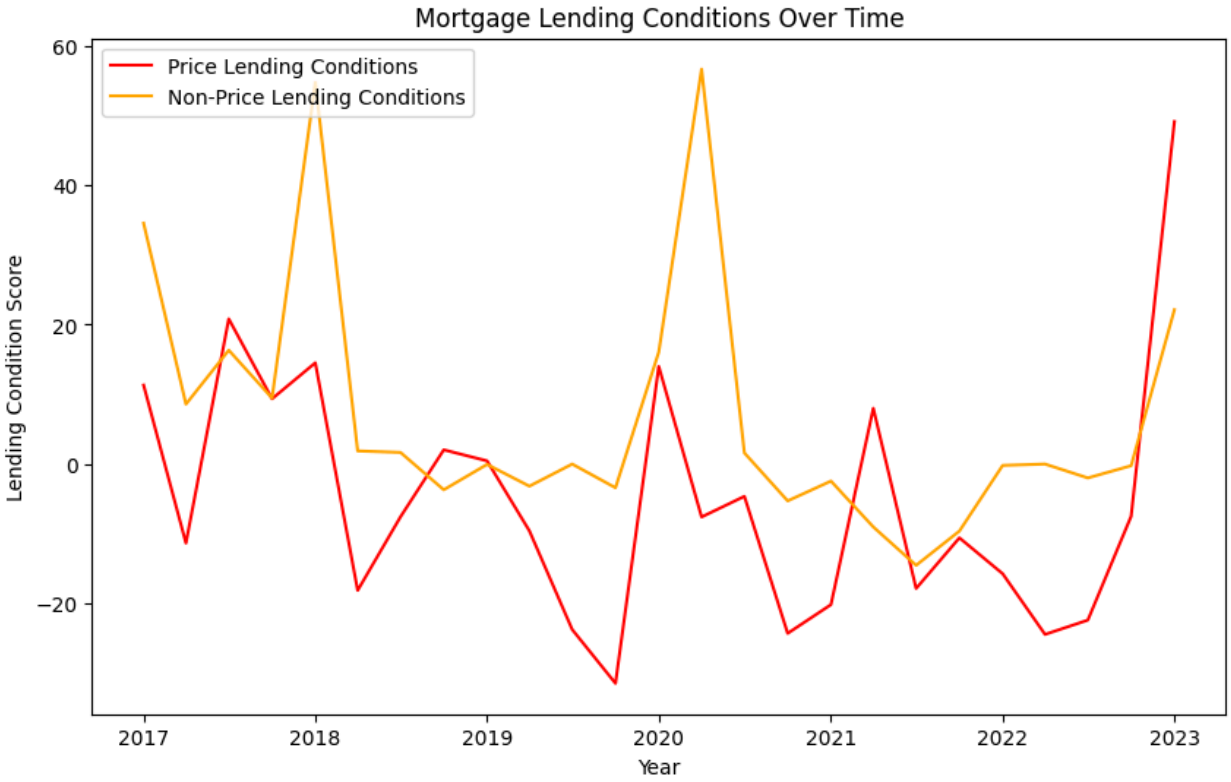
Pulling from the Bank of Canada's Valet API, we brought fields from the Financial Vulnerability Indicator such as Loan to Income Ratio and Loan to Value ratio to compare against the lending condition data as well as the Overnight Lending Rate. All these were brought into python in a JSON format and inserted into my database using the *sqlalchemy* library.

Analysis

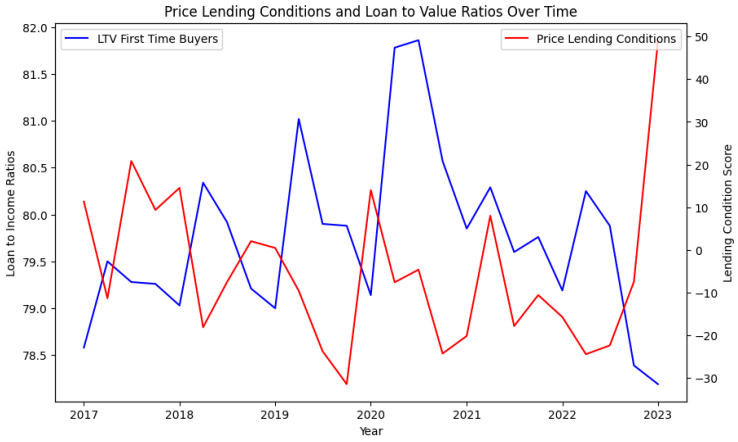
To begin our analysis, we first established a correlation matrix containing all variables.

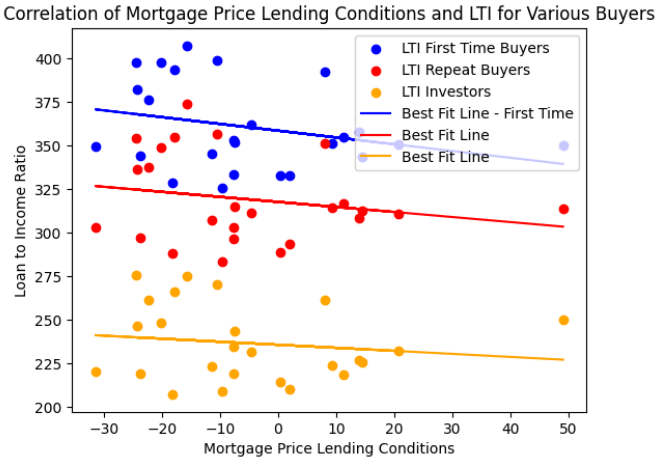


This led us to investigate relationships between variables with significant and interesing correlations. We decided to examine Price vs. Non-Price Lending conditions, The Loan to Value Ratio vs. Price Conditions, Loan to Income vs. Non-Price conditions, Overnight Lending Rate vs. Price Conditions and finally we look at the behaviour of different homebuyers.

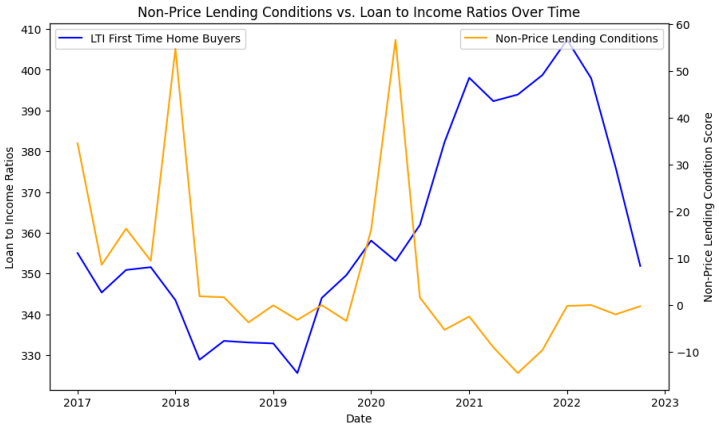


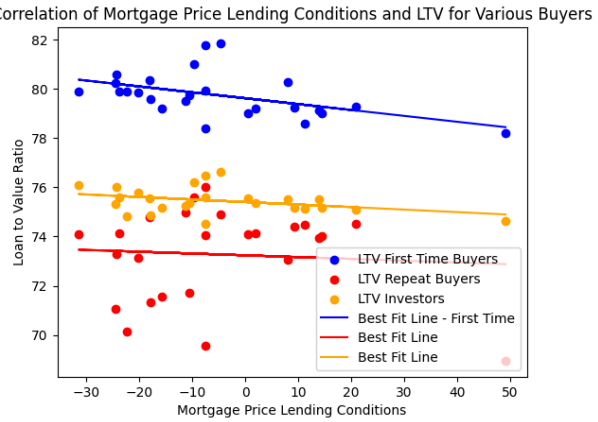
Starting with the overall lending conditions, this chart plots price vs. non-price conditions in the context of mortgage lending. Price conditions move sharply and exhibit periodic fluctuations. In contrast, non-price conditions show stability, then sharp rises, reflecting regulatory shifts or economic shocks. The 2018 increase aligns with Guideline B-20's revision, mandating lenders to assess borrowers' resilience to interest rate or income changes, thus tightening non-price conditions. The 2020 spike relates to heightened default risk amid the pandemic's onset, impacting income and increasing economic uncertainty.





These two graphs investigate Loan to Value ratio plotted against Price Conditions. The loan to value ratio is a metric comparing loan size to property value. It is correlated more with price conditions due to its direct link with loan risk and cost. A Higher LTV ratio represents a greater risk for the bank and thus is often associated with increased interest rates, a core aspect of price conditions.



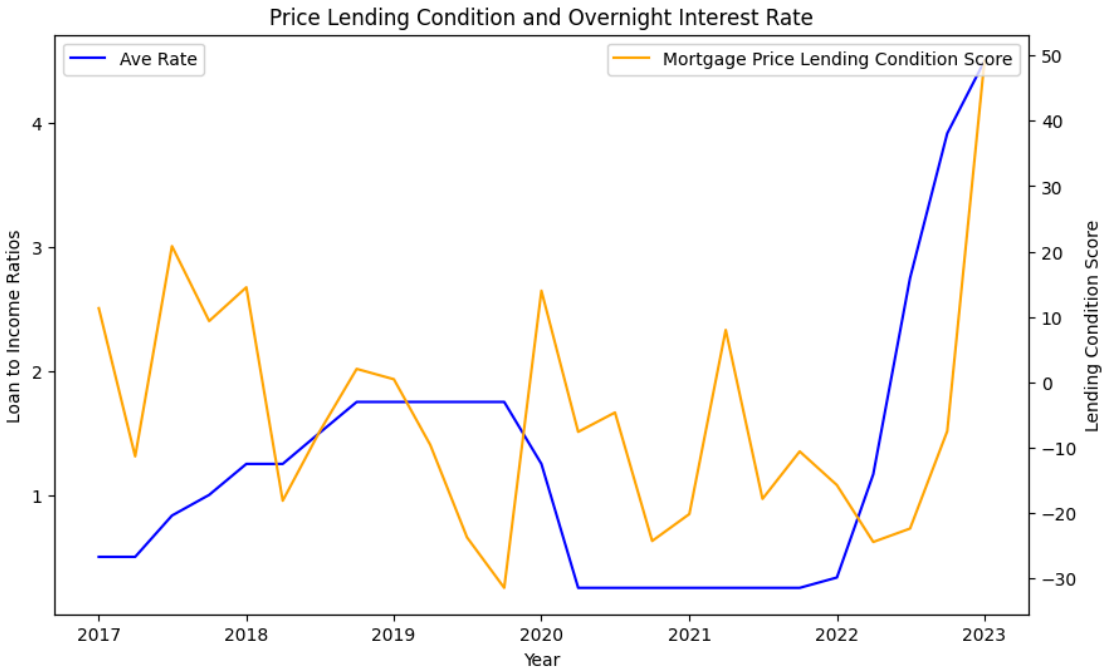


In this final graph, we observe a mild inverse correlation of non-price conditions with the loan-to-income ratio. This ratio, comparing a borrower's mortgage to their income. The negative correlation suggests that as the lending conditions ease, the loan-to-income ratio increases - that is to say, one can buy more house at the same income level.

In the period of the lowest interest rates, many borrowers in Canada opted for variable-rate mortgages with fixed payments. This option has saved many borrowers thousands of dollars during low interest rate eras, however it leaves them at an increased risk, when interest rates increase. In this scenario a larger and larger portion of the fixed payment

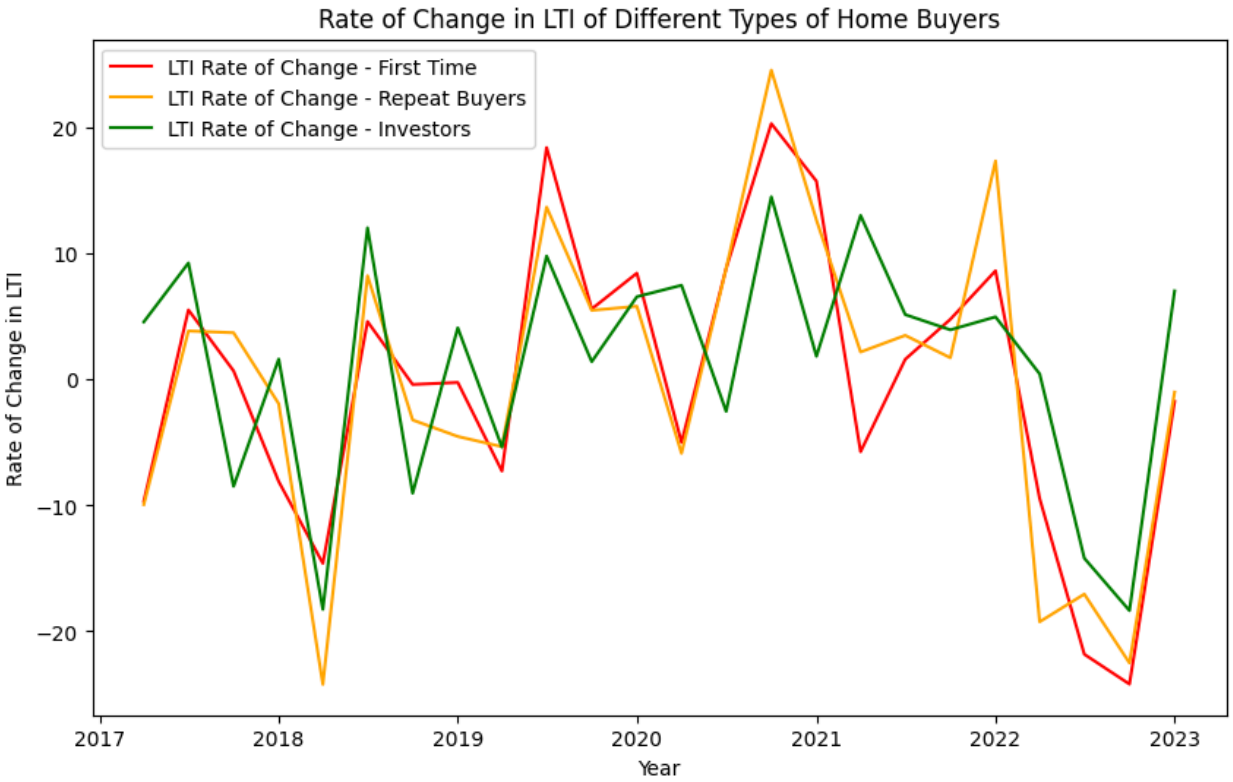
goes towards interest, instead of reducing the principal. This phenomenon became particularly significant in 2022, as we saw substantial interest rate hikes by the Bank of Canada. As a result, for many with these types of mortgages, more of their payment was consumed by interest, leaving less for principal repayment.

A significant number of borrowers reached the 'trigger rate', where borrowers' payments were only covering the interest, not reducing the loan principal or in some instances, decreasing the principal in the home. Consequently, this situation led to the Loan-to-Income (LTI) ratio dipping because loan balances weren't decreasing as expected, even while incomes may have remained relatively stable.



Initially, it was presumed that the overnight lending rate would significantly influence mortgage pricing conditions. Contrary to this assumption, the data revealed only a weak correlation (0.32) with the overnight lending rate. Considering that interest rates are a major component of mortgage costs, the low correlation presents an unexpected deviation from the anticipated impact on lending conditions.

To analyze the behaviors of various homebuyer categories, we computed the derivative—or rate of change—of the Loan-to-Income (LTI) ratio for different types of homebuyers. By evaluating the slope at distinct intervals, we compared the reactivity among first-time



buyers, repeat buyers, and investors. The graph clearly shows that investors exhibit a consistently lower responsiveness compared to both repeat and first-time buyers, suggesting that the motivation of homebuyers purchasing for personal use differs markedly from those acquiring properties for investment purposes.

Results

Our analysis identified key correlations linking lending conditions, consumer behavior, and credit access. We found a moderate correlation (0.48) between Loan-to-Value (LTV) ratios and mortgage price conditions, suggesting a link between property values and loan pricing. A similar pattern was observed between LTV ratios and the overnight lending rate.

Additionally, a moderate correlation between Loan-to-Income (LTI) ratios and non-price lending conditions highlighted the impact of borrower attributes on credit access. Notably, First time and Repeat home buyers displayed high correlation in both LTI and LTV, contrasting with investors which exhibited less reactivity to market conditions. This

difference potentially reflects divergent motivations between personal and investment property purchases.

With the understanding that the overnight lending rate is a key driver of price lending conditions, it would follow that the two would exhibit some collinearity. However, the overnight lending rate, a primary influencer of price conditions, showed a weaker correlation (0.32). Investigating changes in this rate, we calculated its slope at discrete points and compared it with other variables. Interestingly, a strong inverse correlation (0.7) was particularly evident in Repeat Homebuyers' LTV ratios, indicating a heightened sensitivity to monetary policy in this group.

In summary, these findings offer a more granular view of the interactions between lending conditions, monetary policy and the behaviours and ability of consumers to access credit.

Geo-Spatial Bank Location and Demographic Analysis - Niza

This guiding question is intriguing because it will reveal patterns about how the Big 5 targets different demographics or how the presence of these banks correlates with economic indicators in certain neighbourhoods. Understanding these patterns can provide insights into economic development strategies, urban planning, and financial accessibility for different communities within Calgary. The data from Google provides geographical information for the bank branches and starred reviews. Open Calgary offers demographic datasets as CVS’s for Calgary's population broken down into wards. There is information on each ward such as income statistics, country of birth of immigrant, homeownership rates, and population density.

The proposed methodology involves geospatial and statistical analysis to explore the relationship between bank locations, demographic factors and overall performance. The project will be considered successful if it can identify significant correlations or trends between the presence of the Big 5 banks and the socio-economic characteristics of the areas where they are located.

Dataset Explanation

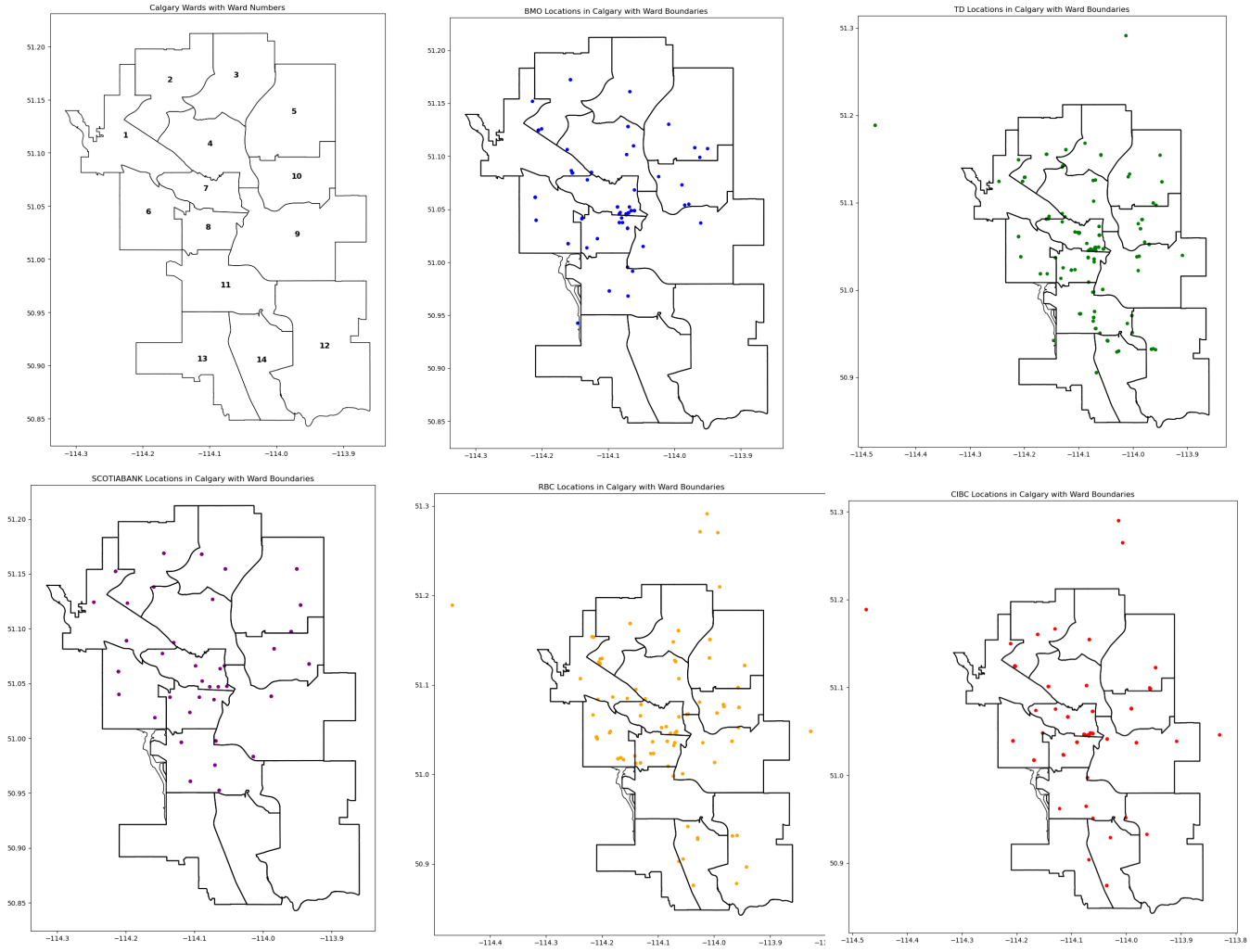
The location dataset and google review stars dataset is freely available from Google. The geolocation data is a table with rows of url, bank (Big 5), longitude and latitude and what kind of location. Type of location is organized by a banking center, financial adviser, and ATM. This data was last updated November 2023. Additionally, the star data is tabular with two columns: first being bank and second being star rating and this was last updated December 2023. The star data for CIBC was not able to be collected. From this data we will map each type of bank location as a point on a map of Calgary and relate performance by the average star rating.

The boundary data and demographic data was sourced from an open Calgary database. This is a public domain and requires no permission to use as stated on their website. All datasets were downloaded in CSV format. The ward and city boundary are both tabular data with multipolygon information regarding city of Calgary boundary. The city boundary data has two columns of city and multipolygon and consists of 1 row. The Calgary ward data has columns of multipolygon, councillor, and ward number. The boundary information is from 2018. The Calgary census data is in a tabular format with data that was collected from the wards in 2018. It contains 100 columns and 14 rows. In this data set we are only concerned with resident count and ward number, since we will be creating a heat map of resident count by wards. The household income data is in tabular format and was collected from the 2016 Canadian census. It contains 12 columns and 14 rows. There is a column for ward, household income divided into groups and multipolygon. The household income is divided into groups of “Under $20,000'', “$20,000 to $39,999”, “$40,000 to $59,999”, “$60,000 to $79,999”, “$80,000 to $99,999”, “$100,000 to $124,999”, “$125,000 to $149,999”, “$150,000 to $199,999”, and “$200,000 and over”. For simplicity we will be grouping ‘Under $20,000’, ‘$60,000 to 99,999’, and ‘$150,000 and over’ to create three heatmaps of Calgary involving household income. Lastly, the country of birth of immigrant data is in tabular format and was collected from the 2016 Canadian census. It contains 15 columns and 14 rows. There is a column for ward, total number of immigrants, non-immigrants, Top four countries of birth and the total number of people from each stated country. For simplicity we will only look at the ward and top two places of birth to create a map describing origin.

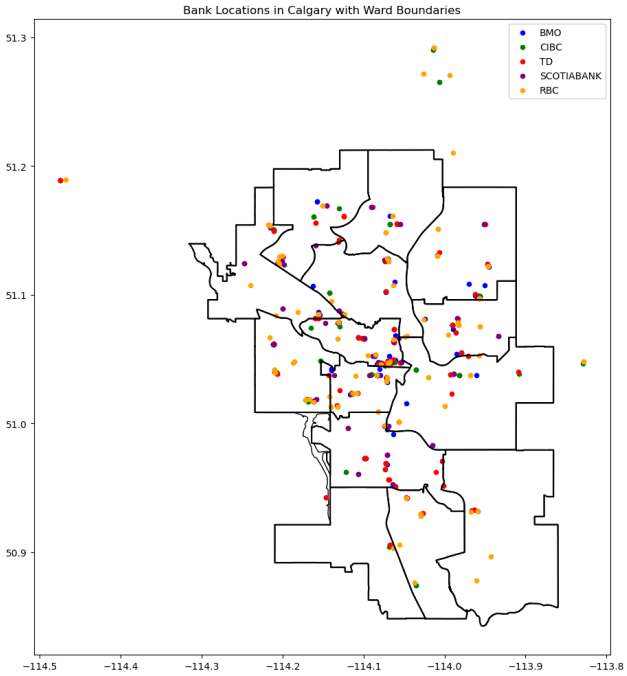
Analysis and Cleaning Steps

To initialize our analysis, using Python, the Google data will be collected for geolocation information and star rating. Further cleaning steps will be conducted using Python and the dataset will be then uploaded into SQL for better group access. Next, data from Open Calgary will be acquired with no further steps necessary, since the data is clean. On a map of Calgary bank locations will be plotted, followed by demographic data, and a complete analysis will be done on a comprehensive map of both variables. In a final analysis, the insights gained from the geospatial analysis will be compared with overall success (google rating).

Bank Location Analysis

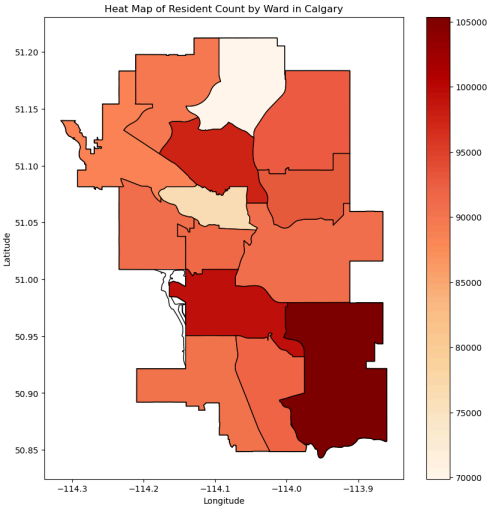


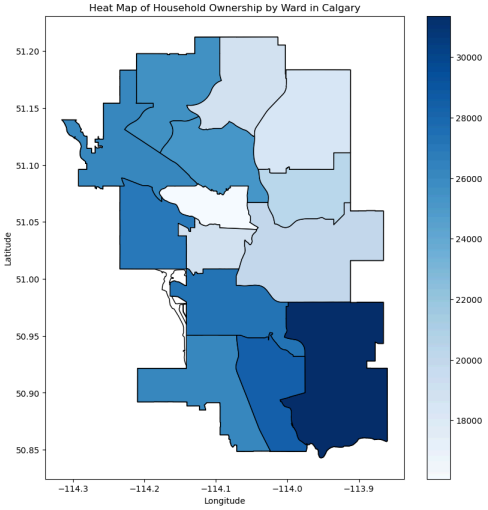
To begin geospatial analysis is done to uncover patterns by analyzing Calgary by its 14 wards. First, we looked at the distribution of the banks in Calgary. In the BMO and Scotiabank map we can see that the locations are mostly centralized but many bank locations are absent in the southern wards. The CIBC and TD locations are better distributed in the city, but we can still see most locations appear in the city center and northern wards. Lastly, RBC has many locations throughout the city but is notably absent from WARD 11.

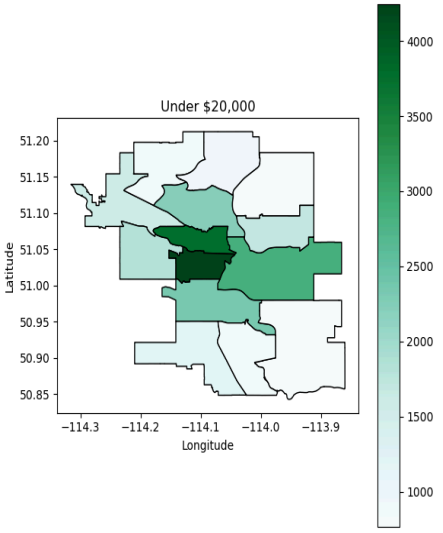


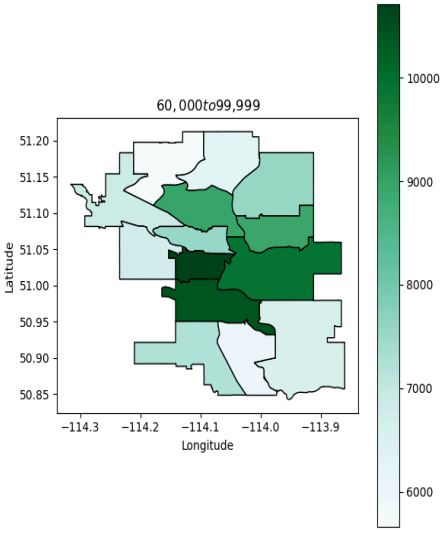
Now in a complete map with all the banks we can see the complete distribution. All banks seem to have a significant presence in the central areas, likely due to higher commercial activity and population density. However, the spread in suburban areas varies, with some banks having a wider presence in the outskirts than others.

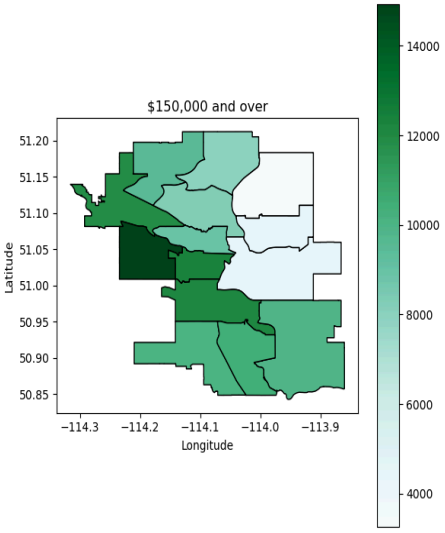
Demographic Analysis

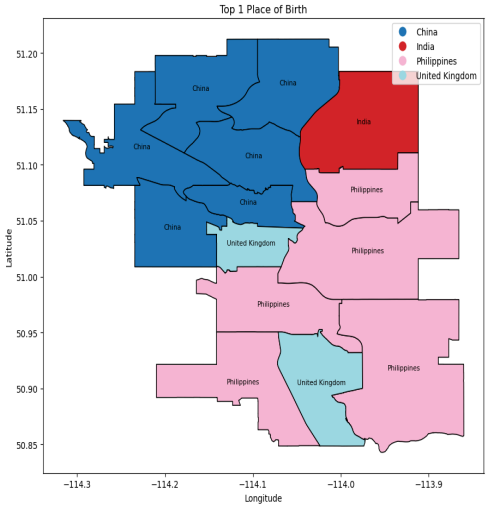


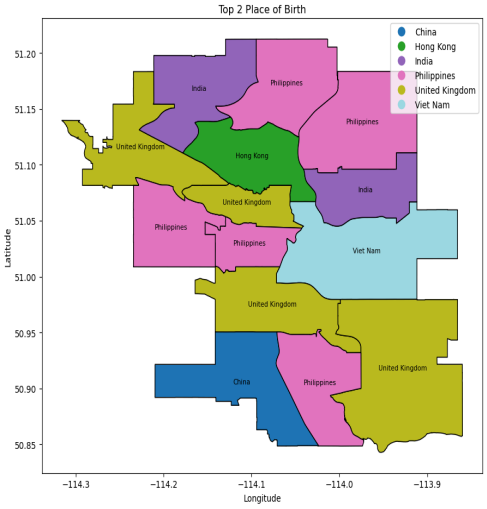






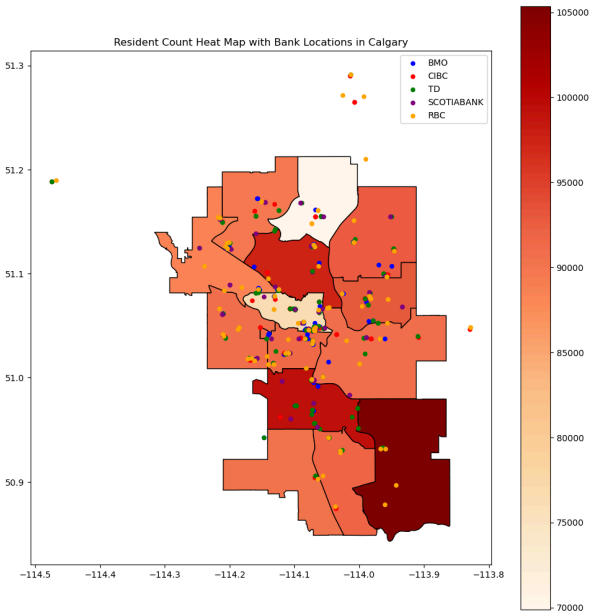


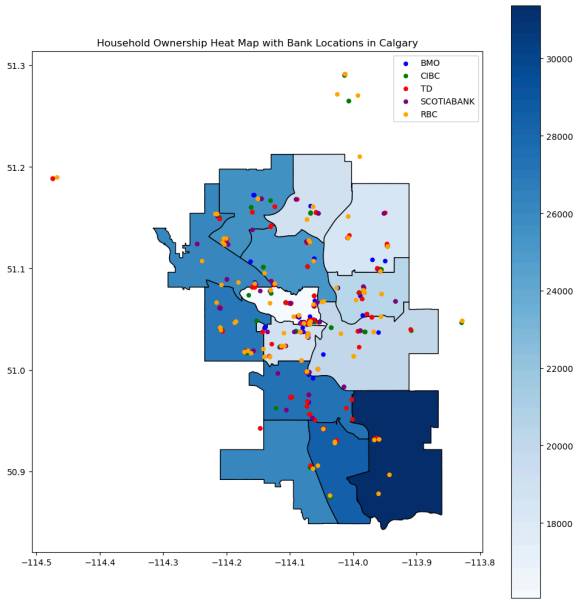


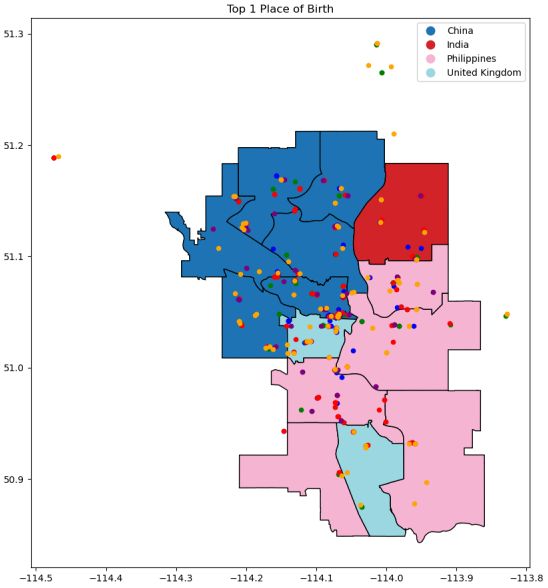


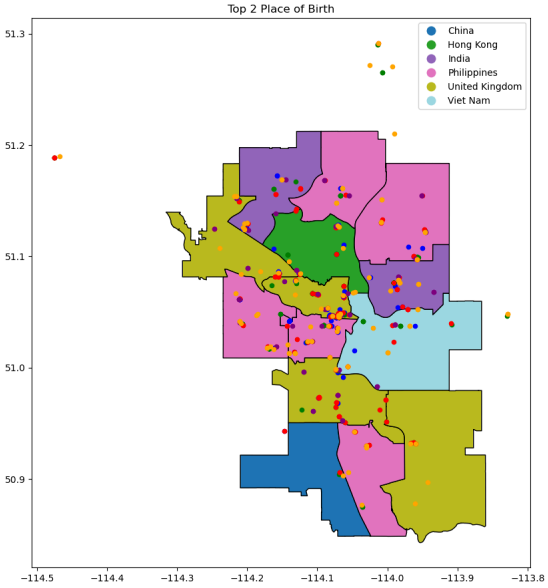
Next, we looked at demographic information. The heat map of resident count by ward, shows a higher concentration of residents in the southern and central wards in dark red. This could correlate with factors such as availability of housing or economic opportunities. Areas with higher populations may have greater demands for services, including banking. Second, household ownership with dark blue showing wards with the most homeowners. Thirdly, we looked at household income with dark green showing the most households with that income range. Lastly, we looked at immigrant data with the top two places of birth for each ward. The maps detailing the top places of birth for immigrants show diverse origins, with China, India, the Philippines, the United Kingdom, and others represented. This diversity is indicative of Calgary's multicultural landscape and suggests a need for culturally sensitive services and language support, especially in banking. For the financial sector, understanding the interplay between resident count, household ownership, and immigrant origins can guide the strategic placement of new branches and the development of tailored financial products. For example, a ward with a high number of immigrants from a country with a strong entrepreneurial culture may benefit from business banking services. By analyzing these maps together, one can see how the fabric of Calgary is shaped by its people — where they live, where they come from, and their economic status. This information is essential for anyone involved in business strategies.

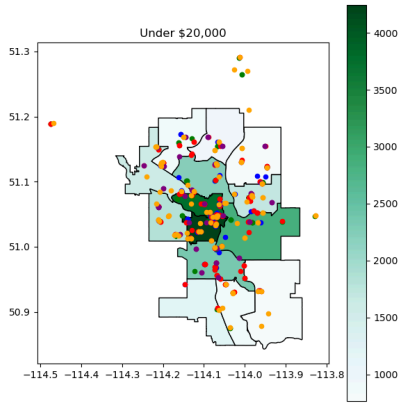
Overall Performance

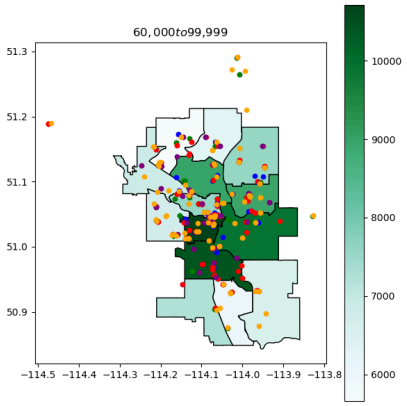


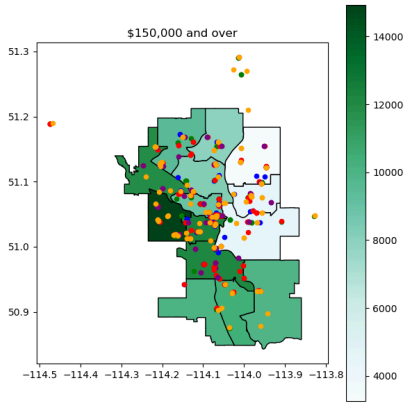






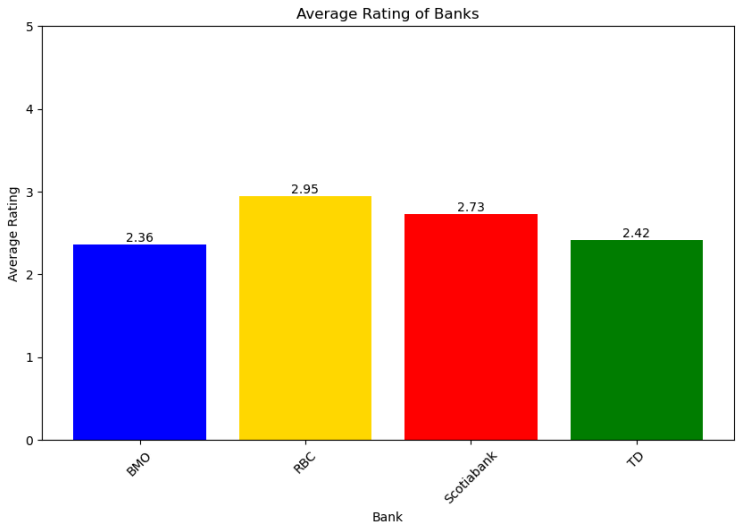






For the resident count and bank map together, we see that although there is a higher population density in the southeast there are very few bank locations in that area. This would go against our initial idea that banks would try to target places that simply have more people. But this makes sense from a business perspective, central locations have higher foot traffic and greater accessibility for a larger number of potential customers. Since these metropolitan areas have a diverse range of people, this could indicate a focus on serving a diverse range of customers, from individuals to families, highlighting that a broad consumer base is a priority for banks. From a competitive standpoint, this finding suggests that banks are actively vying for a wide range of customers. As consumer needs evolve and the demographic landscape changes, it will be interesting to see how these placement strategies adapt. When analyzing the homeowner and banking graph, we can note the scant distribution of banks although ward 11 is dense in homeowners. Many homeowners rely on many banking services however the Big 5 do not seem to prioritize targeting just homeowners.

Next for the income maps, we see that RBC has the most locations in high income areas when compared to the other banks. RBC the largest of the Big 5, RBC targeting higher income areas could have something to do with their size or possibly making strategic location placement could be a clue into their success. This could indicate a strategic focus on people with more liquid assets and who may have different banking needs, such as mortgages, home equity loans, and insurance services. Lastly, there does not seem to be any real pattern when analyzing the immigrant place of birth map.



Overall the banks have average review scores of BMO =2.36 ,RBC =2.95 , Scotiabank = 2.73, TD =2.42. This displays RBC as having superior performance when compared to the other banks. In the geospatial analysis RBC had better distribution in the city and had more location in the southern wards which other banks typically were absent from. These differences could be the reason for their success. However, we can not state for certain if it is the decisions RBC makes when placing locations that can be attributed to its performance or their success is due to other factors.

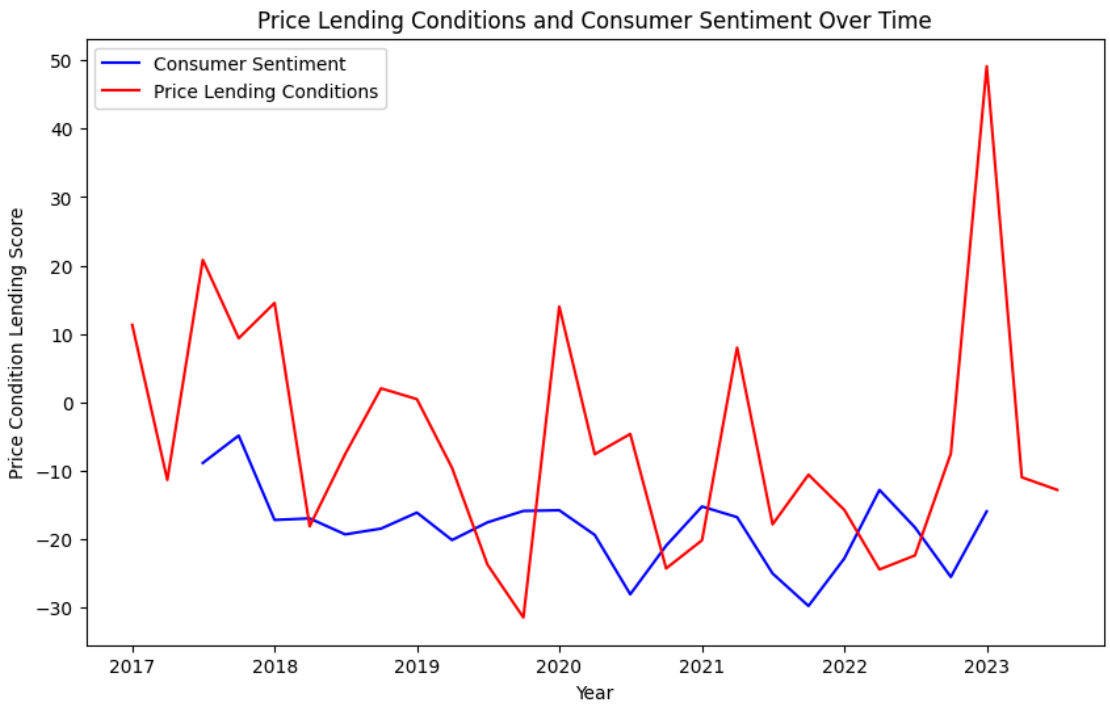
Summary

From a strategic business perspective, banks could use this data to optimize their branch network, improve customer accessibility, and adjust to evolving market demands. Such as the shift towards online banking which may influence the future of physical branch locations. Further insights could be gained by overlaying demographic data, public transport lines, and economic activity indicators to understand the strategic placement of branches better. However, we would like to note that the demographic data is from 2016 and 2018. Calgary is a fast growing city with ever changing demographics, therefore some of the data may be out of data affecting our overall insights.

Connecting Datasets

Connecting Consumer Sentiment with Lending Conditions

To evaluate the impact on consumers, we combined the sentiment analysis and the lending condition data to examine its correlation. Integrating these datasets provides a practical perspective on the assessments made by the Bank of Canada, allowing us to better understand the effect lending conditions may have on individuals.



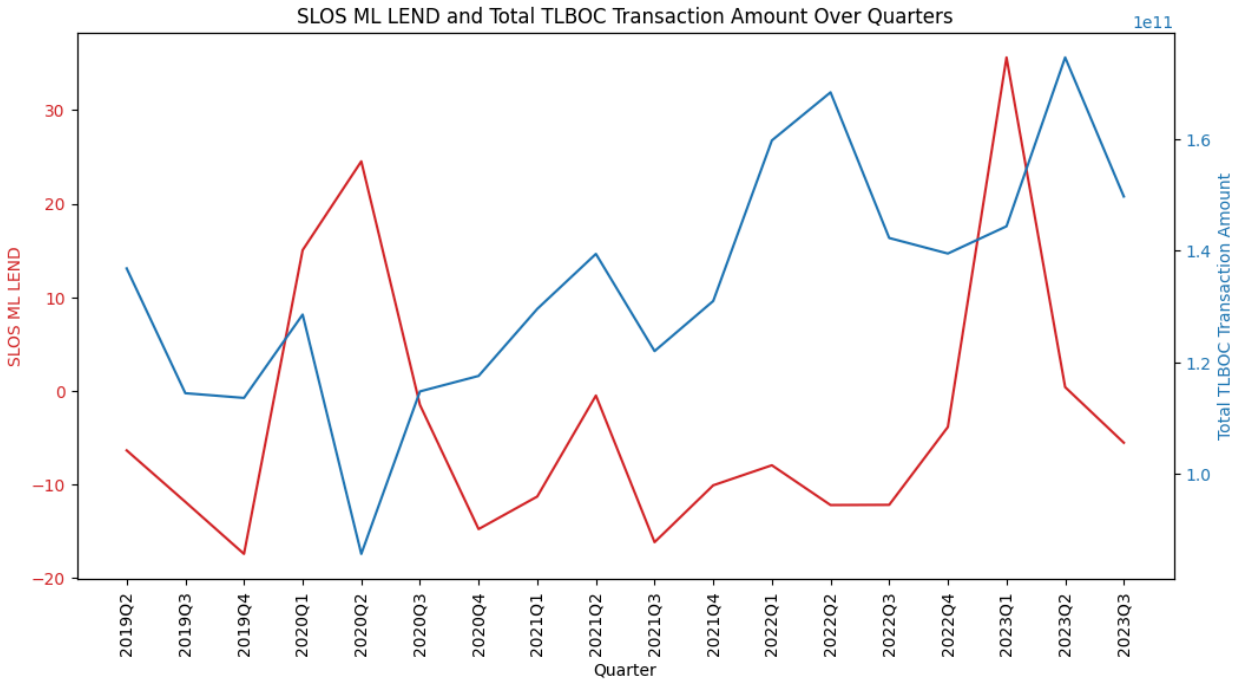
The observed correlation coefficient of 0.27 between consumer sentiment and price conditions indicates a modestly positive relationship. However, it is imperative to note that the data on consumer sentiment is uniformly negative. This trend may be attributed to the nature of discussions on platforms like Reddit, where discourse often gravitates towards critical, negative and sarcastic views, particularly in the context of banking services. Alternatively, this pattern could reflect a prevailing negative sentiment among consumers towards the banking sector. Considering the correlation between price conditions and consumer sentiment, it can be inferred that factors such as fees and interest rates are pivotal in influencing consumer attitudes. This observation further validates the significance of costs in shaping consumer perceptions of banking institutions.

Connecting Transactions with Lending Behaviour

Filtering "Transfers to Bank of Canada" Transaction Types and Merging Totals with Senior Loan Officer Survey (SLOS) Quarterly Data 2019-2023

This query merges quarterly aggregated transaction data with senior loan officer survey data on mortgage lending conditions. It retrieves and converts total transaction amounts associated with the Bank of Canada's 'TLBOC' code from negative to positive for consistency. The query calculates the sum of transactions per quarter and joins these totals with corresponding data from the slos_2 table, which includes various indicators of mortgage lending conditions. The join aligns quarters starting from the second quarter of 2019, enabling a time series analysis that compares financial transactions with mortgage lending trends, providing insights into their relationship over time.

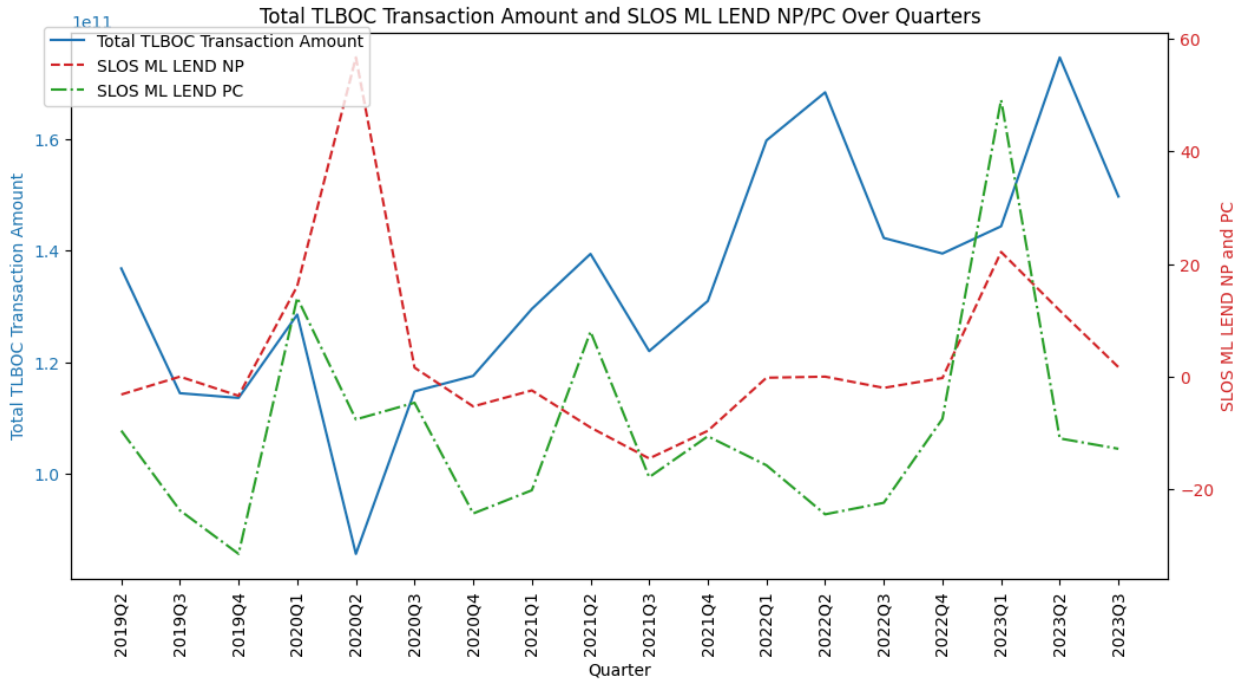
Senior Loan Officer Survey Mortgage Lending and Total "Transfers to the Bank of Canada" Transaction Amounts Over Quarters 2019 to 2023



The SLOS ML LEND series (in red) and the Total TLBOC Transaction Amount series (in blue) exhibit fluctuating trends over time, with the former showing changes in mortgage lending conditions and the latter displaying a general upward trajectory with occasional sharp declines and recoveries. These series show a loosely similar trend for several quarters,

suggesting a potential relationship or shared influencing factors. Notably, before the second quarter of 2020, both trends moved in tandem, but around the onset of COVID-19, a significant divergence occurred. The SLOS ML LEND series spiked, reflecting eased mortgage lending conditions due to the Bank of Canada's drastic reduction in the overnight lending rate. Conversely, the Total TLBOC Transaction Amount sharply decreased, possibly due to economic uncertainty or changes in central bank policies. After 2020Q2, the trends realigned, possibly stabilizing from the initial pandemic shock as financial institutions and markets adjusted to the new economic landscape. Although these observations indicate similar trends between the two lines, except for the notable divergence in 2020Q2, establishing a definitive link or fully explaining their movements requires more detailed data and analysis.

Senior Loan Officer Survey Mortgage Lending Conditions (Non-Price and Price Condition) and Total "Transfer to the Bank of Canada" Transaction Amount Over Quarters from 2019 to 2023

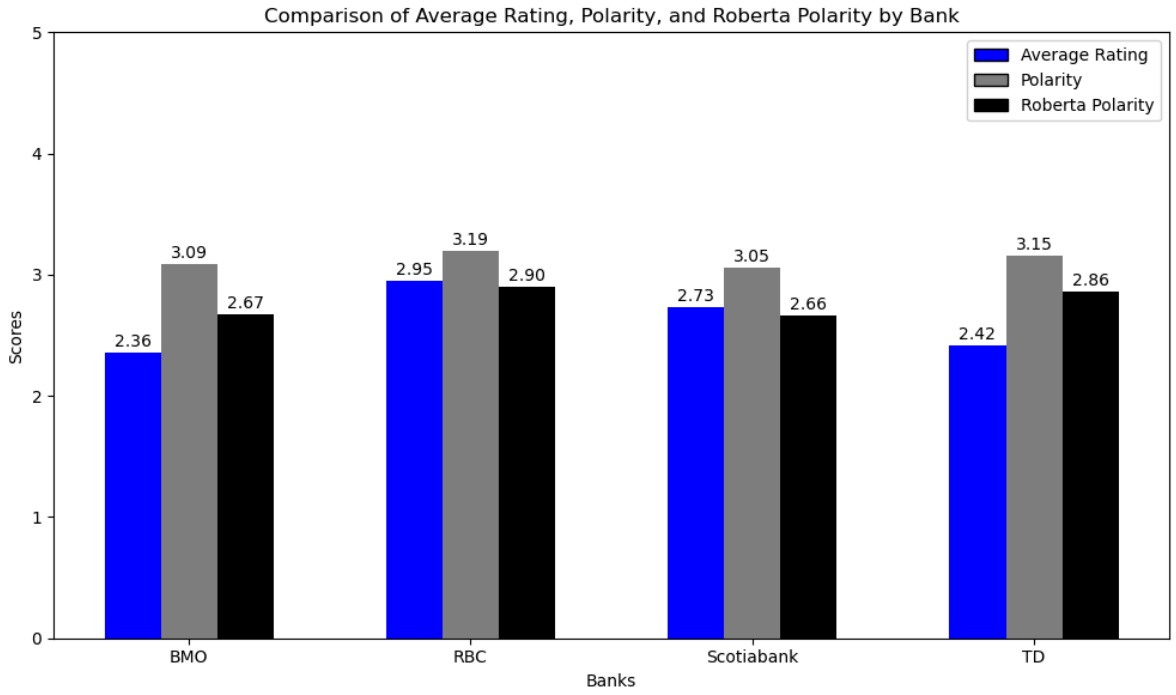


The graph compares the Total TLBOC Transaction Amount with two metrics from the Senior Loan Officer Survey (SLOS) - Non-Price (NP) and Price-Condition (PC) for Mortgage Lending (ML LEND) over several quarters. The dashed red line, indicating non-price conditions for mortgage lending, suggests a tightening of conditions when below zero and a loosening when above. The dashed green line reflects changes in price-related conditions, such as interest rates and fees, where values above zero indicate increasing costs for borrowers and below zero, decreasing costs.

In 2020Q2, there's a notable upward spike in SLOS ML LEND NP, signifying a significant loosening of non-price lending conditions, potentially as part of economic stimulus measures during the early COVID-19 crisis. This period might have seen more flexible lending standards to encourage borrowing. Concurrently, the dip in SLOS ML LEND PC suggests tightening price conditions, perhaps due to heightened risk perception or economic uncertainty. The sharp drop in the Total TLBOC Transaction Amount during this time could reflect these dynamics, showcasing the complex interaction between market liquidity, lending standards, and policy responses to the pandemic.

This trend indicates that during the pandemic's initial economic uncertainty, there was likely an effort to ease borrowing (non-price loosening) while managing risk through price conditions (price tightening). This highlights the nuanced strategies employed by financial institutions and policymakers to balance stimulating the economy and risk management in response to economic shocks.

Connecting Overall Sentiment

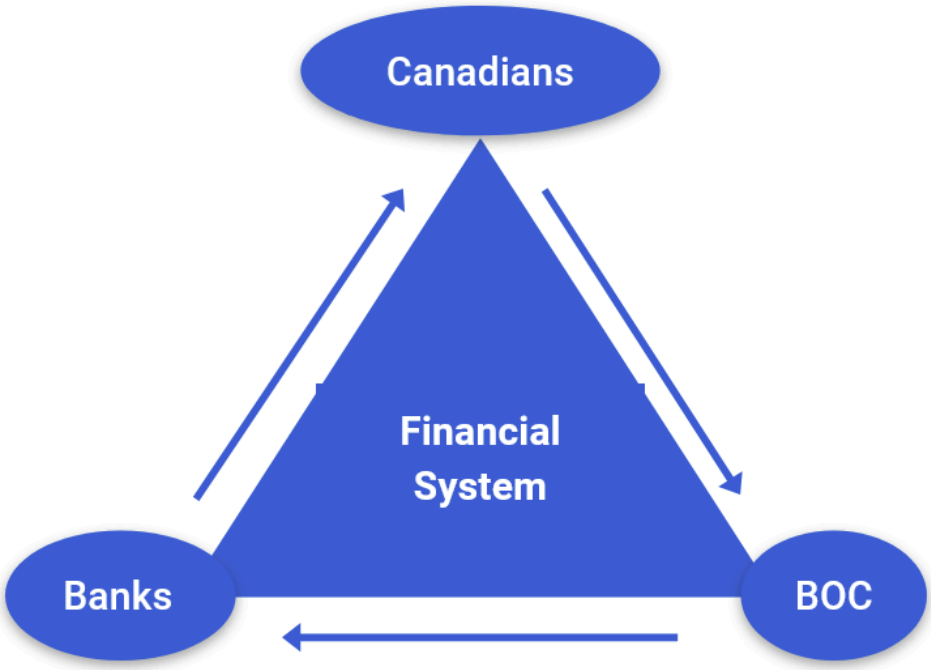


Comparing sentiment scores for RBC, BMO, TD, and Scotiabank from Google review stars (range from 1 to 5) with the roBERTa and TextBlob polarity scores adds another layer of depth to our analysis. We converted the polarity scores to the same scale as Google review stars for ease of comparison. Using SQL the two data sets were joined on the common key, Banks. Reviewing the graph we observe similar patterns emerging to those seen in the previous sentiment analysis. With average ratings ranging anywhere from 2.36 to 3.15, customer reviews of banks overall are considerably low. Also notice that the language models used to compute sentiment scores from comments very closely mirror those obtained by the google review stars. Especially the roBERTa model, which is expected given its superior ability to understand context, compared to the TextBlob model. This finding helps to confirm that the language models used did a good job accurately evaluating sentiment and that the conclusions drawn earlier are probably not localized to reddit and

apply more broadly to the canadian audience. For example, notice that once again RBC, while by a very small margin, has the highest average scores for Google star reviews.

Additionally, this bar graph shows us how these comments from across Canada are reflected on the local scale. What people are saying about banks online is reflected by the genuine experiences they have in real life.

Conclusion



The key takeaway from this project is this concept of what we call the “triangle of compromise”. From our analysis and research, we have realized these three entities, the

Bank of Canada, Banks, and citizens are like the three pillars that form the foundation of the Canadian Financial System. While these parties do have common interests at many times their incentives and priorities are in competition and these conflicts of interests can create a lot of friction in the socio-economic state of the country. This is an incredibly challenging balancing act where each party has to make adjustments to their business, their policies or their lives to accommodate what the other two parties are doing and settle for an equilibrium that each party can tolerate. For example, banks want your deposits, to extend credit to low risk borrowers, full-service customers that have their entire suite of products, to sell you their investments and so on and they do all this to appease shareholders and achieve sustainable profitable growth. Contrast this with customers they want to know their money is secure, to have easy access to credit when they need it, they want good efficient service and to feel important, they want affordable living and to preserve their wealth and the Bank of Canada well their mandate is to keep inflation steady and facilitate the functioning of a healthy growing economy.

The various topics that we investigated for example geo-spatial analysis and the accessibility that consumers have to banks like their location and hours of operation contribute to the relationship between the bank and customer. The transaction data and the interplay between the receiver general and the banks is another layer of the operation that influences the overall system. These components of a much larger system influence the experience of each party and how customers feel and think about the industry. Even though we have one of the best financial systems in the world from our sentiment analysis, people don't seem to think highly of banks. With that said, the question becomes how we could improve the system? Perhaps the banks need to rethink their strategy and not try to oversell clients on their products and services. Perhaps citizens need to take some responsibility for their own financial literacy and know their role in the triangle of compromise. Maybe we can find a way to convert this triangle of compromise into a triangle of prosperity where everyone walks away happy. Or perhaps the status quo will remain in place. Only time will tell.

What did we learn?

During this project as a group we have all dramatically increased our breadth of knowledge of the Canadian banking industry.

Luke: By conducting the Sentiment Analysis I have learned that in general overall sentiment towards the banking industry and the Big 5 Banks is low based on comments from the subreddit PersonalFinanceCanada. People seem to appreciate online no-fee banks that do not charge account fees and prefer the service received from credit unions. The google review star data also indicated that the results obtained from reddit may be representative of the overall Canadian population as the two levels of measurement coincided with one another. From a more technical perspective I learned how to interact with an API, set up a database schema, query the database and got an introduction to machine learning and NLP libraries and techniques.

Riki: Through analysis, I learned that transactions processed by the Big 5 Banks display seasonal trends with transaction volumes typically peaking during the summer months and dipping in March, ahead of the tax deadline. I also learned how these trends changed during the midst of a global pandemic, as it caused more pronounced lows and subdued peaks, which could be attributed to changes in economic activity and the implementation of financial support programs. Through further querying, I also learned the most common transaction types processed by the Big 5 Banks and the distribution amongst them. As for SQL, I learned how to construct complex SQL queries to extract meaningful insights from large banking datasets. Many functions like GROUPBY, CASE and JOIN were important elements in my learning.

Ethan: By comparing the Senior Loan Officer Survey results from the Bank of Canada with financial vulnerability indicators, I discovered that despite expectations, market and monetary changes, which are thought to drive mortgage lending price conditions, are not adequately explained by the main macroeconomic indicator, the Overnight Lending Rate. Moreover, an analysis of Loan-to-Income and Loan-to-Value ratios highlighted distinct motivations between investors and residential homebuyers. An examination of non-price lending conditions revealed the significant influence of legislation on the loan amounts

borrowers can obtain relative to their income. These insights together, enhance our understanding of the distribution of credit among banks, the Bank of Canada, and consumers.

Niza: In my geographical analysis, I have learned the power of joining dataset to create insights. The google review data on Calgary banks and the reddit sentiment data on the banking industry are both dataset that would not be typically joined. However, when we do so we are able to confirm that what people across Canada are saying is reflected in the local scale. This was interesting to learn and I would not typically think of joining datasets that are so different. Additionally, by connecting this data I was able to leverage new tools such as SQL to handle big data.

Further Research Opportunities

Some opportunities for future research involve looking at employee sentiment and how that relates to bank performance. It would be interesting to see how employee satisfaction translates into customer satisfaction as many management books often promote the idea that happy employees get the best results. Employee reviews could provide great insights into the culture and management of the company which is a key aspect of any long lasting and sustainable business. Perhaps many of the complaints of everyday consumers would be similar to those of employees. As for the transactions, there are several avenues for further research that could yield valuable insights in the Canadian banking industry. A predictive analytics approach could leverage the historical transaction data to forecast future banking trends. This could be valuable for strategic planning. Additionally, we could explore the monetary factors behind the variance in the price lending conditions. With the overnight lending rate only having a 0.32 correlation with the price conditions, further investigation is needed. Further geospatial work could involve linking the google star review to the exact bank location in Calgary. Insights gained from this would allow us to see how exactly performance relates to location.

References

cardiffnlp/twitter-roberta-base-sentiment · Hugging Face. (2001b, June 1). https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

Deposits at financial institutions (Bank Statement Data). Open Government Portal. (n.d.). https://open.canada.ca/data/en/dataset/2061075a-55fe-47a5-be2f-984f12fc0d40

Nicholas Renotte. (2020, October 4). Presidential Debate Twitter Sentiment Analysis using

OpenAI Platform. (n.d.). Platform.openai.com. https://platform.openai.com/docs/guides/text-generation

Python and NLTK [Video]. YouTube. https://www.youtube.com/watch?v=_EgqxIoUE7U

Python, F. (2023, April 11). Use ChatGPT API for sentiment analysis in Python - Financial Python - Medium. Medium. https://medium.com/@financial_python/use-chatgpt-api-for-sentiment-analysis-in-python-5a152ddb3238

Rob Mulla. (2022, May 5). Python Sentiment Analysis Project with NLTK and 🤗 Transformers. Classify Amazon Reviews!! [Video]. YouTube. https://www.youtube.com/watch?v=QpzMWQvxXWk

TextBlob: Simplified Text Processing — TextBlob 0.16.0 documentation. (n.d.). https://textblob.readthedocs.io/en/dev/

Bank of Canada. (n.d.). Indicators of financial vulnerabilities. Retrieved from https://www.bankofcanada.ca/rates/indicators/indicators-of-financial-vulnerabilities/#Loantoincome-ratio

Bank of Canada. (n.d.). Posted interest rates offered by chartered banks. Retrieved from https://www.bankofcanada.ca/rates/banking-and-financial-statistics/posted-interest-rates-offered-by-chartered-banks/

Bank of Canada. (n.d.). Senior Loan Officer Survey. Retrieved from https://www.bankofcanada.ca/publications/slos/

Appendix

Chat GPT 4 API Text Processing “System” instructions

" “ “ your job is to analyze the comments I provide and structure the data like the following in csv form:

bank | comment | sentiment | created_at

Revised Sentiment Categories:

Positive: Clearly positive statements about a specific bank.

Negative: Clearly negative statements about a specific bank.

Mixed: Statements that contain both positive and negative sentiments about specific banks, or where the sentiment towards each mentioned bank is ambiguous.

Neutral: Factual statements or opinions that don't convey a clear sentiment about any specific bank.

Non-specific: General statements about banks as a whole, big banks, or "the big 5" without direct reference to a specific bank.

In many cases multiple banks appear within one comment. Try and separate out the comments to ISOLATE the particular bank being discussed. If it is not possible to do so disregard the comment.As I am calculating sentiment scores to compare the banks among one another. Also be mindful to not break comments up too much into separate components especially when they are for the same bank/share the same sentiment.

I will be extracting the text from your response and writing to a csv file so do not provide a response outside of the parameters I have set." “ “

Reddit API Licence

Reddit Support (Reddit Support)
Nov 4, 2023, 21:12 PDT

Thanks for contacting us!

You can use the Reddit API, subject to our Developer Terms and Data API terms. We will reach out if there are any concerns.

If you are interested in getting approval for commercialization, or getting bulk access to Reddit data for commercial, academic, or other purposes, please fill out a separate helpdesk ticket with "I'm an Enterprise/Business Partner".

For assistance with the Reddit Ads API please reach out to your sales rep or fill out the contact form linked here.

We encourage our API users to sign up for the Developer Platform beta, where you will have access to additional capabilities: https://developers.reddit.com/waitlist.

Please respond to this message if you are in need of further assistance.

Happy hacking!

Cheers,
Your friends at Reddit