Market Perception of Banks in New York

Oscar J. Urizar

November 1, 2020

1. Introduction

In this section we provide a description of the problem and a discussion of the background

1.1 Background

Retail banking, also known as consumer banking or personal banking, is banking that provides financial services to consumers as individuals not businesses. Retail banking is a way for individual consumers to manage their money, have access to credit, and deposit their money in a secure manner. Services offered by retail banks include checking and savings accounts, mortgages, personal loans, credit cards, and certificates of deposit (CDs).

1.2 Problem

Retail banking remains a highly competitive business, with banks aiming to gain any edge on the competition. Although several services are conducted online, many other services are provided to customers at office branches. The perception of customers towards their bank is highly important in assessing the performance of banks around cities where they operate in order to plan future improvements. We propose a project to gain insights in the market perception of major banks in New York by performing a sentiment analysis on the text feedback provided by customers visiting various banks of this branch.

2. Data Acquisition and Preprocessing

2.1 Data Sources

For this project we are required to find the location of office branches of the various banks operating in New York. Furthermore, we need text feedback/reviews from people who have visited these venues. In summary we require two main data components:

- Location of office branches
- Text reviews from customers

The data provider FourSquare contains a robust database covering the requirements for this project. Firstly, It provides data of multiple types of venues, including banks. Secondly, It also provides tips as plain text provided from people who have actually visited these venues.

2.2 Data Acquisition
To obtain this data we will be using the FourSquare API. The API implements REST operations to query the required data, for example, using the venue category id

Category Id: 4bf58dd8d48988d10a951735
Category Name: Banks

We can implement a REST request to obtain information about the venues. For this project we will use the following data attributes

Attribute	**Description**
id	Unique identifier for this venue
name	Recorded name of the venue
Distance	Distance (m) from the provided location
Lat	Latitude coordinate of the venue
Lng	Longitude coordinate of the venue

Once we have identified the venues of interest, we can use the id to make a REST request and retrieve the following tip's data attributes

Attribute	**Description**
Venue id	Unique identifier for this venue
Id	Unique identifier for this tip
Created at	Datetime stamp of creation
Text	Tip's text
Lang	Language of the tip's text

2.3 Data Cleaning
The data downloaded is stored in CSV files for further use in this project. The files are loaded into pandas dataframes and data is manipulated to obtain a form suitable for this project.

The cleaning process for the venue's data involves checking a valid venue's name is provided, and grouping the venues by the bank they belong to. All the banks with only one venue will be grouped into 'others'.

Cleaning venue's tips encompasses several other steps are performed in the tip's text, including:

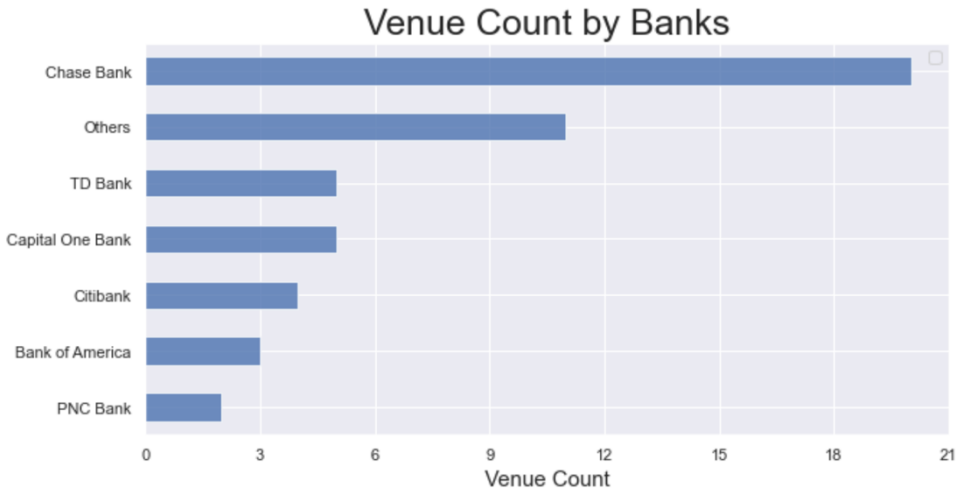
- Remove tips not in English language
- Remove invalid characters
- Remove punctuation
- Remove stopwords

3. Exploratory Analysis

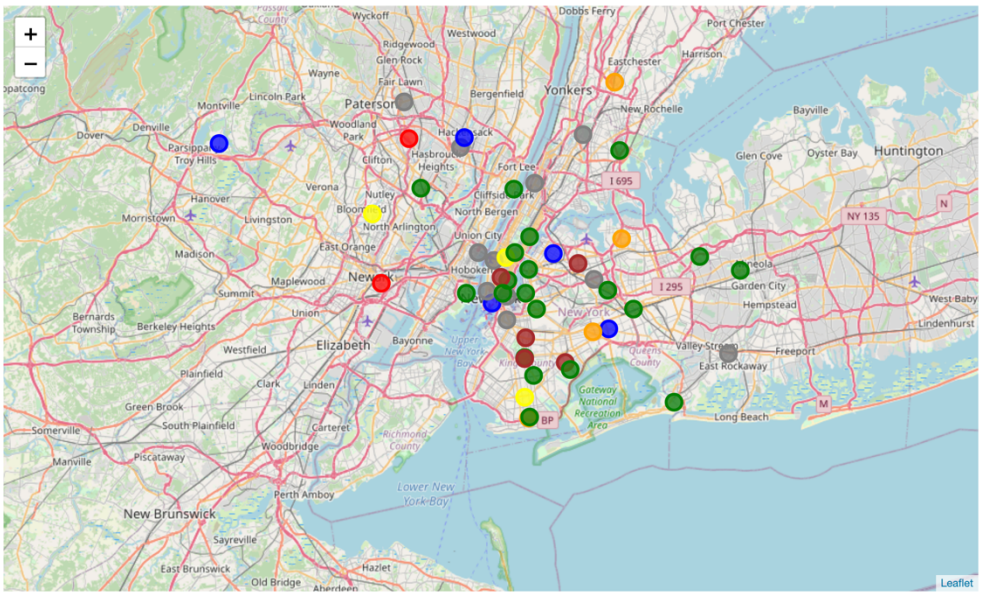
The data collected for both venues and tips is limited, not necessarily due to unavailability but rather to the constrains of the FourSquare account type used for this project.

From the data available, 6 major banks are identified, whereas other banks with minimum presence are grouped into a single category labeled ‘Others’. The predominant banks are Chase Bank, TD Bank, and Capital One Bank.

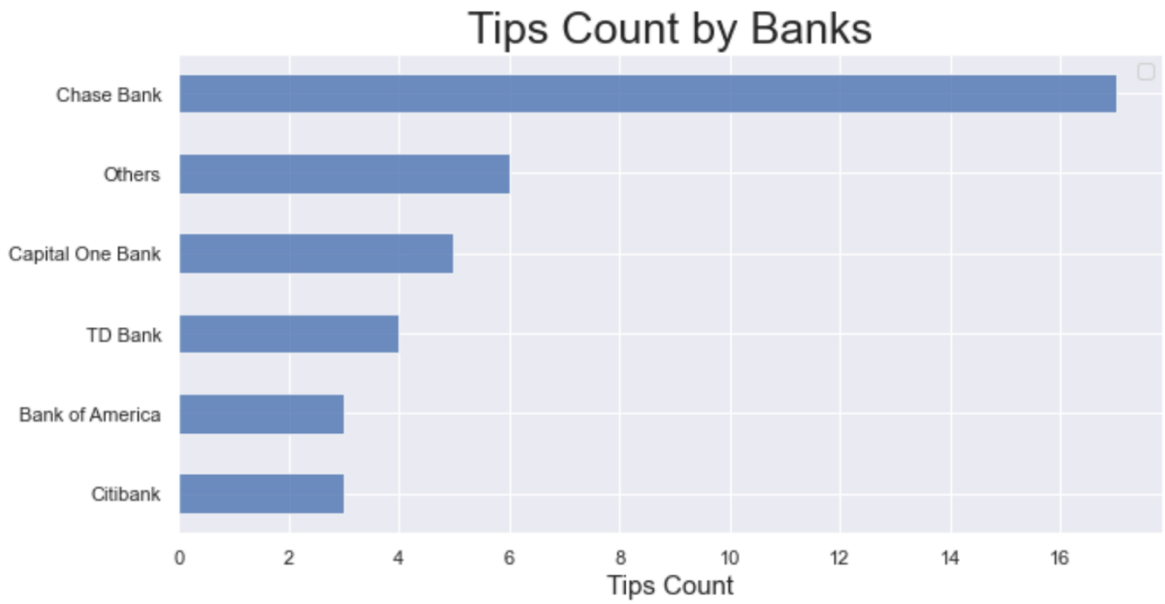
A summary of the venue count by bank



The venues obtained for each bank appear to be reasonably spread across the city of New York, positively impacting the sample significance with relation to the territory covered. We can visually confirm this in the plot below



Thre tips corresponding to each venue are retrieved in small volumes due to the unavailability of data for these venues. In general, Chase Bank reports the highest number of tips, followed by Others, and Capital One Bank. This volume in data closely correlates to the number of venues found in those same banks. A summary of number of tips by banks is presented below



4. Sentiment Analysis

4.1 Approach

Sentiment analysis refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information.

Applied to this project we make use of the customer's feedback voiced via tips in the FourSquare database. A score is to be computed for each tip in order to asses the subjective affective impression of customers. The scale employed is a continuous-valued number in the range presented below:

Sentiment Score	Affective State
-1	Negative
0	Neutral
1	Positive

4.2 TextBlob Library

To implement the sentiment analysis, we make use of a pre-trained model provided by the Python library TextBlob.

TextBlob is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

4.3 Data Preprocessing

It is essential to clean the text to be used before feeding this data to the sentiment analysis model. Several steps are included in this process:

Language Selection

It is important to make sure the same language is used across the text to be used. FourSquare provides this attribute in the data, and for convenience, we focus on using text only in English language.

Remove Invalid Characters

Due to multiple reasons, the data collected may contain characters not accepted by the model, significantly impacting the final sentiment score. Hence it is important to remove invalid characters for the language we require.

Remove Punctuation

Symbols other than letters are not necessarily relevant to the inference of the affective state, and they can even negatively impact the results provided. For this reason, any punctuation mark is removed from the text.

Remove Stop words

stop words are words which are filtered out before or after processing of natural language data (text). Though "stop words" usually refers to the most common words in a language, there is no single universal list of stop words used by all natural language processing tools, and indeed not all tools even use such a list. These stop words are removed as they don’t contribute any semantic substance to the sentiment of the text.

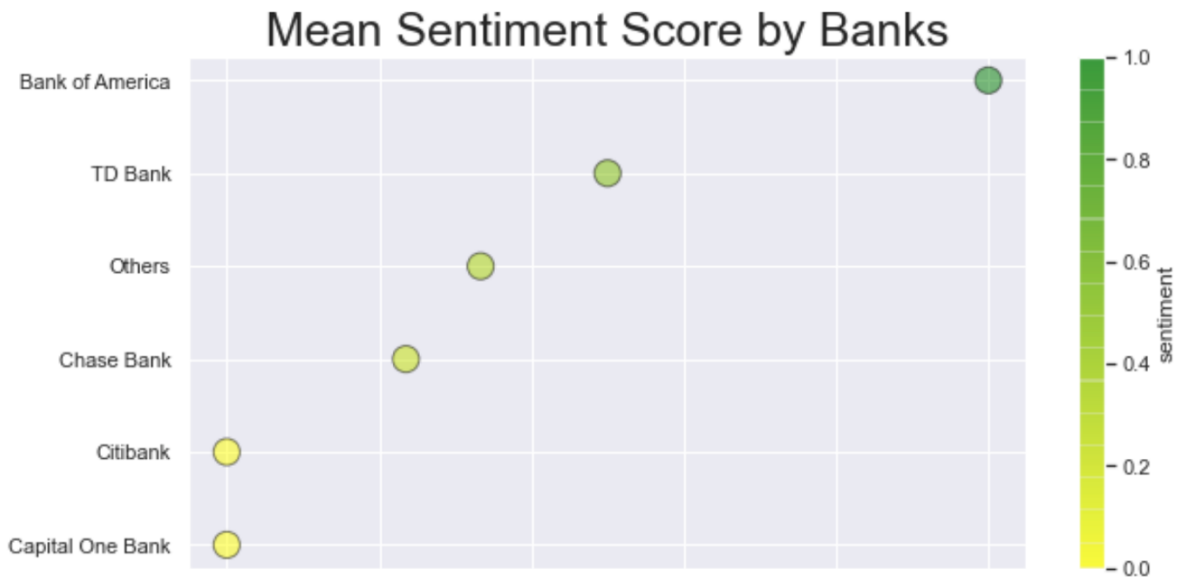
4.4 Sentiment Analysis Results

Once the data is ready, we can proceed to compute the sentiment score. TextBlob provides an easy to use interface that requires few lines of code to be implemented.

The result returned for each text analysed, contains a sentiment property that returns a named tuple of the form Sentiment(polarity, subjectivity). The polarity score is a float within the range [-1.0, 1.0]. The subjectivity is a float within the range [0.0, 1.0] where 0.0 is very objective and 1.0 is very subjective.For the purpose of this project, we will only use the polarity as it refers to the actual sentiment score.

We compute a sentiment score for each tip, and the mean sentiment score is computed for each bank previously identified.

The mean sentiment score is presented below:



5. Conclusions

The sentiment analysis conducted in this project provides us with several valuable insights:

- Based on the collected data, Bank of America has a mean sentiment score of 1, placing it as the bank with best market perception in New York. This comes with the caveat that Bank of America is at the bottom half in number of venues and number of tips provided by customers. In short, the reported market perception for Bank of America is positive but very representative due to the low amount of data available for this bank

- Chase Bank has the most venues and tips, making its sentiment score the most significant to report. Their sentiment score is at 0.23, a score slightly above neutral. - None of the evaluated banks report a negative market perception. However, the results are not statistically significant due to the modest volume of data available.

- This approach is a viable way to gain insights into the market perception of banks in any given location, given that sufficient data is collected.