Final Capstone Project

KATE JORJOLIANI

MDA 720

APRIL 30, 2024

Table of Contents

Background 2

Objective/Goals of the Project 2

Data Extraction/Collection/Scraping *3*

Data Exploration/Data Visualization *4*

Data Analysis/Data Mining/Text Mining 9

Conclusions/Recommendations 10

Bibliography *11*

Background

The coffee industry is becoming very demanding and popular around the world. As a big coffee lover myself, my first and favorite idea for this project was to open a coffee shop business. Maybe one day I will be able to do it in real life as well. However, opening a coffee shop requires thorough market analysis and an understanding of consumer preferences as it is not an easy task. This project aims to extract and analyze relevant data to support decision-making in launching a successful coffee shop business.

Objective/Goals of the Project

The objectives/goals of this project are as follows:

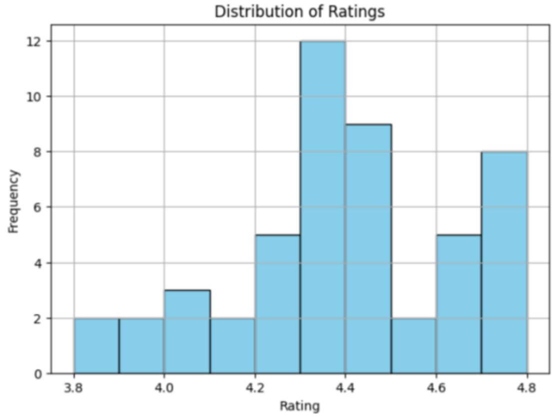
- Extract data related to coffee consumption trends, popular coffee types, and competitor analysis.
- Utilize web scraping techniques and APIs to gather comprehensive data for informed decision-making.
- Explore and visualize the data to identify market opportunities and consumer preferences.
- Analyze sentiments around coffee brands and products to understand consumer perceptions.
- Write a conclusion based on our analysis and suggestions for opening a coffee shop business.

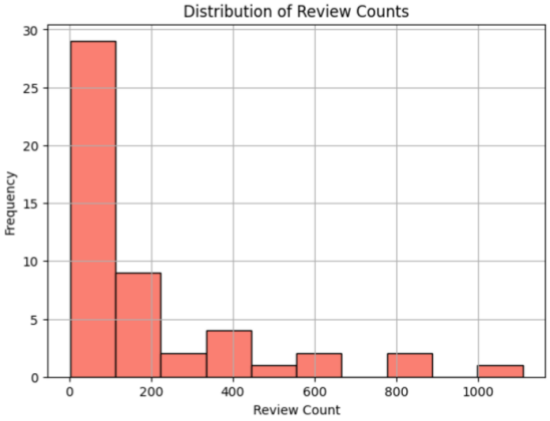
Data Extraction/Collection/Scraping

To start up the business of a coffee shop, we needed a dataset that would help us with analyzing strategic moves to successfully open one. To find a dataset specific to our needs, I used Yelp API keys, which helped me find a dataset of most popular and rated coffee shops. The dataset contained information about various coffee shops in New York City. It included columns such as Name, Rating, Review Count, Address, City, State, and Postal Code. I mostly focused on extracting specific information related to the "Coffee Project New York, East Village" coffee shop, as it was the most highly rated and popular coffee shop in the dataset. To do this, I filtered the dataset based on the name of the coffee shop. After filtering, I obtained a subset of the dataset containing information exclusively about the "Coffee Project New York, East Village" coffee shop. This subset included its name, rating, and review count. Afterwards, I analyzed the extracted data to understand the performance of the coffee shop based on its rating and review count. This analysis provided insights into its customer satisfaction and popularity. Additionally, we can compare the extracted data with similar data from other coffee shops to assess the relative performance of "Coffee Project New York, East Village" compared to its competitors.

Overall, the data extraction process involved filtering the dataset to focus on specific information that was relevant to our analysis, extracting that information, and then analyzing it to gain insights into the performance of the coffee shop in question.

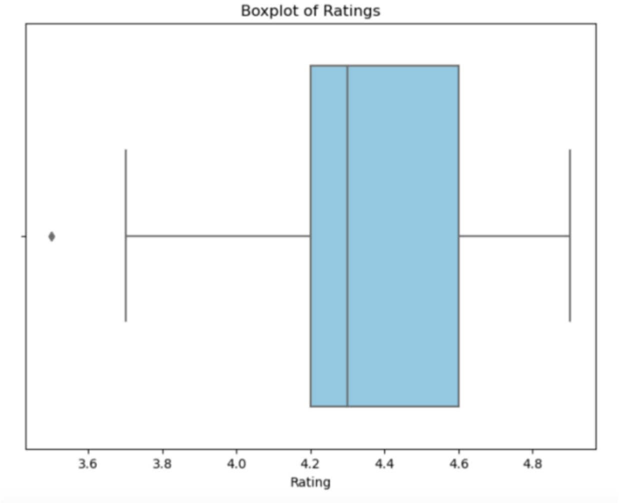
Data Exploration/Data Visualization



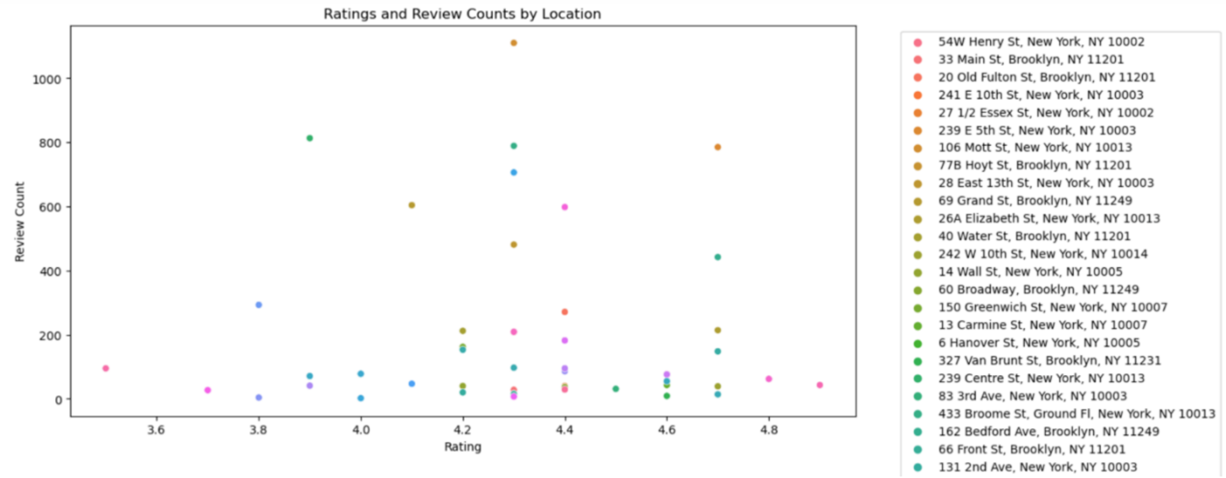


For data exploration, I did some plots to help us interpret the data and its characteristics, which will eventually help us reach proper decision-making toward our goal of opening a coffee shop.

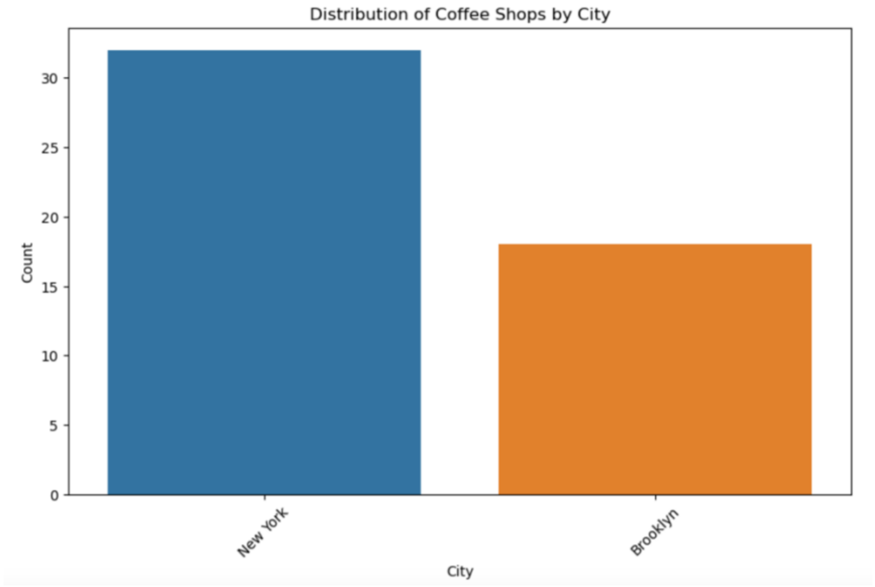
In the plots above, we can see the distribution of ratings and review counts in our dataset. The ratings of coffee shops (left) are distributed unevenly. The most frequent rating is around 4.3 with a frequency of 12, while the most highly distributed review count (right) is around 50-100 counts with a frequency of 30.



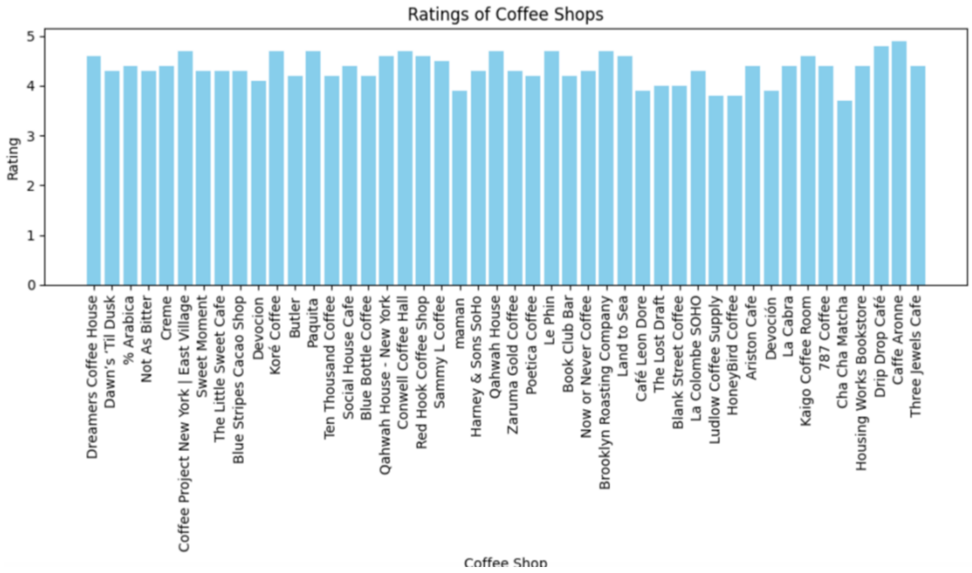
This boxplot shows us that there is an outlier such as missing values in the dataset. The data is not evenly distributed, and most ratings are around 4.3, which is what we also saw on the histogram prior. I removed missing values from the data afterwards.



This scatterplot provides the address of all the coffee shops in New York that we have in the data (not every address is visible here). Each address has its color and is represented with that color on the plot, which shows the rating and review count of the coffee shops. Analyzing the distribution of points can provide insights into the relationship between these two variables. For example: Are there any noticeable trends or patterns in the distribution of ratings and review counts? Do coffee shops with higher ratings tend to have more reviews, indicating higher popularity or customer engagement? Based on this plot Mott St. has the highest review count with over a thousand and Henry St. has the highest rating coffee shops, although the review count is small. This means it did not get a lot of reviews, but the ones it got were very good. We could potentially investigate the coffee shops on these streets, analyze the reason for their high ratings and review counts, and implement similar techniques in our coffee shop operations. We could also steer clear of opening a shop on these streets due to high competition and consider doing so in areas that have less and lower-rated coffee shops so that the competition is not as high.

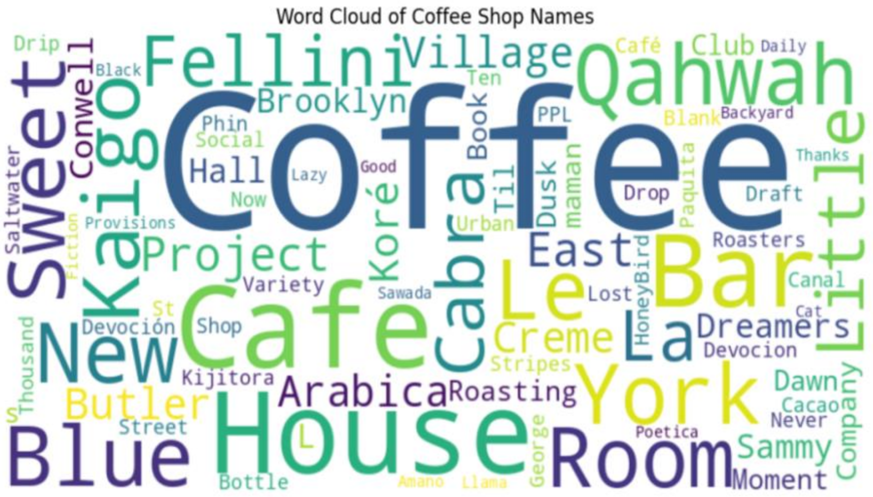


This bar plot provides us with the distribution of coffee shops by city: New York (Manhattan) and Brooklyn. As seen, Manhattan seems to have a higher distribution meaning there are more coffee shops in that area, thus a higher competition as well. Therefore, we could consider Brooklyn's locations for our café.

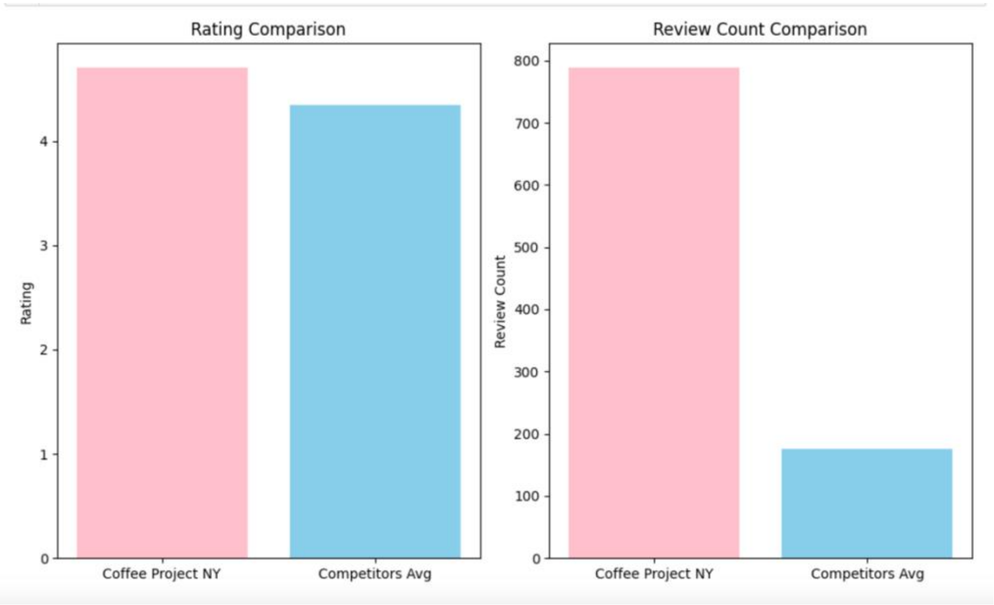


The bar plot on the left showcases ratings of certain coffee shops in NY. Based on the plot, "Cafe Aronne" 'Coffee Project New York | East Village,' 'Drip Drop Café' and a few others have the highest ratings.

However, when I performed data analysis on the best coffee shop with the highest rating was "Coffee Project NY | East Village."

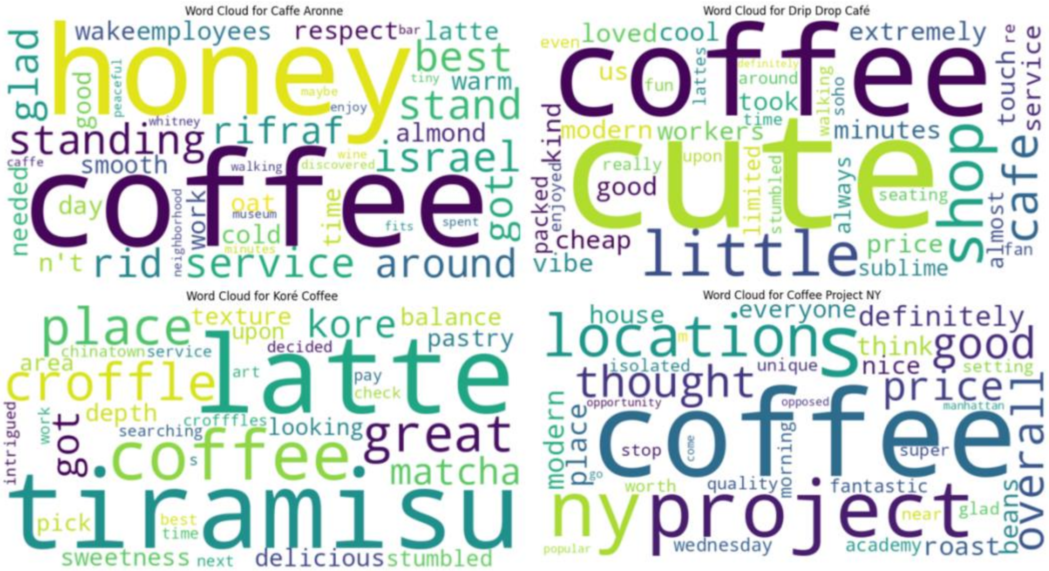


Above, we can observe the world cloud of coffee shop names that we got using text mining, where the size of each word corresponds to its frequency in the text. The bigger the word, the higher its frequency. This could help us in choosing a unique name for our business, to make it stand out from other coffee shops in the location, or on the contrary, follow the trends of coffee shop names to make it more relatable. And as we can see, to no surprise, the word 'Coffee' was used the most in the names.



Lastly, I performed an analysis to identify the biggest competitor by multiplying the rating of coffee shops by the log of the review count. As a result, we got ‘Coffee Project New York | West Village.’ I then visualized it with a bar plot that shows the coffee shop’s ratings and review counts vs competitors’ averages. By comparing the rating of ‘Coffee Project New York…’ with the average rating of its competitors, we can see if it stands out positively or negatively in terms of customer satisfaction. A higher rating compared to the average suggests that customers perceive it more favorably. And a higher review count might indicate greater customer engagement and interest.

Lastly, I extracted more data from reviews of some of the coffee shops with the highest ratings in the bar plot we got earlier. After, I did text mining to find the most frequent words used by reviewers, which will help us with understanding trends and what kind of coffee, coffee shop design, etc. people love in these cafes.



As seen above, some of the most frequently used words in the reviews are: coffee, honey, tiramisu, latte, modern place, cheap, great matcha, roast beans, etc. Based on this analysis, we could implement some of these characteristics and qualities in our own business, which will make people love it more.

Data Analysis/Data Mining/Text Mining

Data analysis includes examining, cleaning, transforming, and modeling data to uncover insights, make conclusions, and decision-making. For this project, I analyzed the dataset containing information about various coffee shops in New York City. We then moved on to data mining, which is the process of discovering patterns, correlations, and trends within large datasets. In this project, I performed data mining by:

- Identifying the most common words in the coffee shop names to understand common themes or keywords within coffee shops in NYC.
- Calculating summary statistics such as average rating and review count to evaluate the overall performance of the coffee shops.
- Visualizing the distribution of ratings and review counts to identify trends and outliers.

Next, I did text mining. It involves extracting useful information, patterns, and insights from unstructured text data. I performed text mining by:

- Tokenizing the names of coffee shops to break them down into individual words.
- Removing stopwords (common words like "and", "the"…) to focus on meaningful words.
- Analyzing word frequencies to identify the most common words used in coffee shop names.
- Extracting sentiment from customer reviews to gauge overall satisfaction with coffee shops.

These steps allowed us to gain insights into the characteristics, performance, and customer sentiment associated with the coffee shops in our dataset. By analyzing both structured and unstructured data, we were able to provide a comprehensive understanding of the New York City coffee shops.

Conclusions/Recommendations

In conclusion, the analysis conducted on the dataset of New York City coffee shops provided us with valuable insights into the coffee industry landscape and consumer preferences. Here are the key findings:

Market Trends: The distribution of ratings and review counts revealed that most coffee shops in New York City have ratings of around 4.3, with different review counts. This suggests a competitive market with a diverse range of coffee shops satisfying different consumer preferences.

Geographic Analysis: Manhattan appeared as the area with the highest concentration of coffee shops, indicating higher competition compared to Brooklyn. Therefore, exploring locations in Brooklyn might offer opportunities with possibly lower competition.

Naming Trends: Text mining of coffee shop names highlighted common themes such as the frequent use of the word "Coffee" and other similar terms. This information can be used to craft a unique and memorable name for a new coffee shop that stands out in the market.

Competitor Analysis: Through competitor analysis, we identified "Coffee Project New York | West Village" as the biggest competitor based on its rating and review count. Comparing the coffee shop's performance to its competitors can inform strategic decisions and underline areas for improvement.

Bibliography

Ekaterine Jorjoliani. (2024). Yelp API Data Extraction Script (Version 1.0). Self-published.

[https://api.yelp.com/v3/businesses/search]