Kasam Dhakal

CS 598 Data Mining Capstone - Task 1

Exploration of Data Set

Date: 2/21/2024

kasamdh/CS598Capstone (github.com)

kdhakal2@illinois.edu

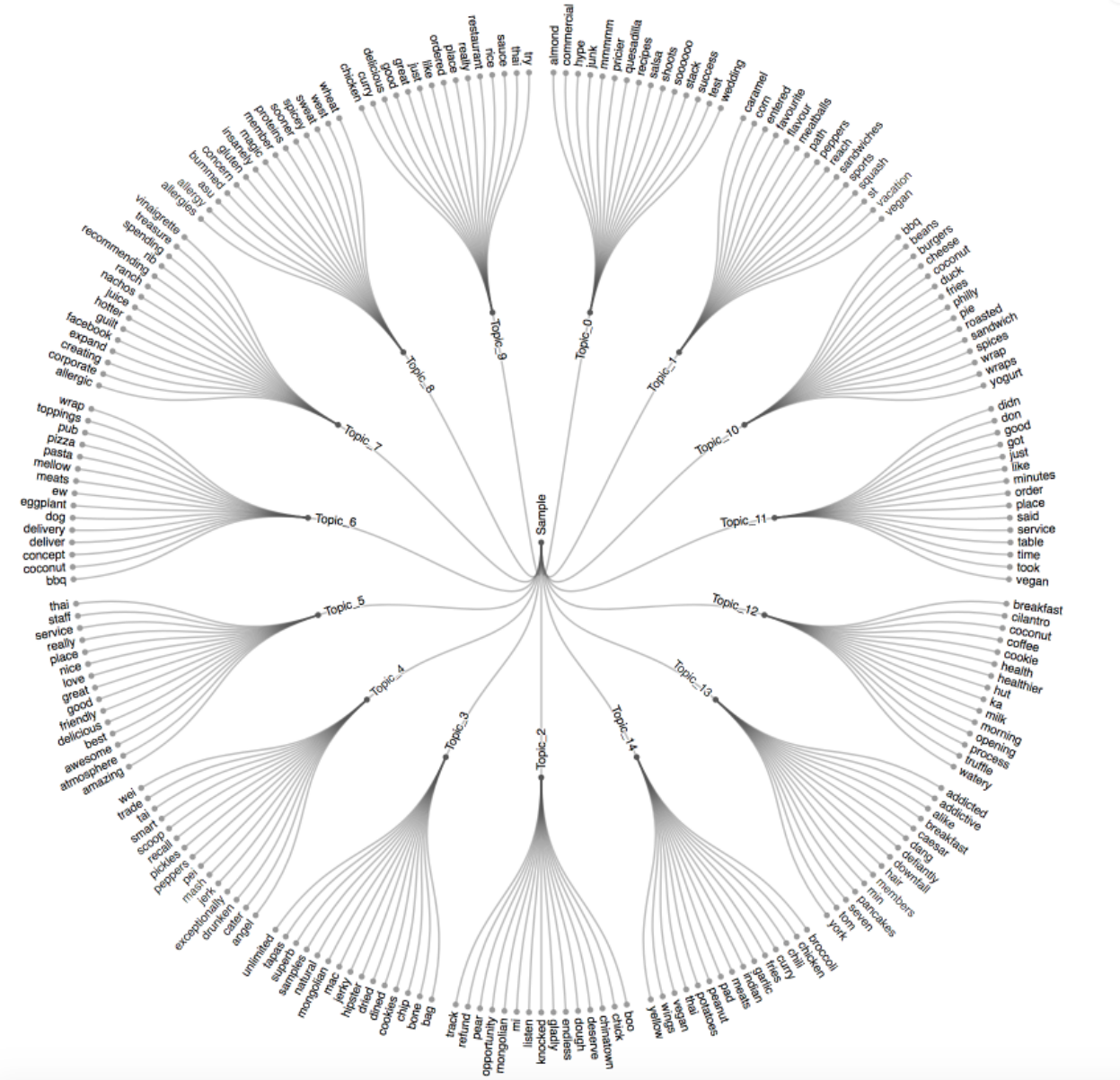
Exploration of Data Set

Objective

The main objective of the Capstone project has been to utilize the knowledge and skills acquired from the Data Mining Specialization to address real-world data mining challenges.

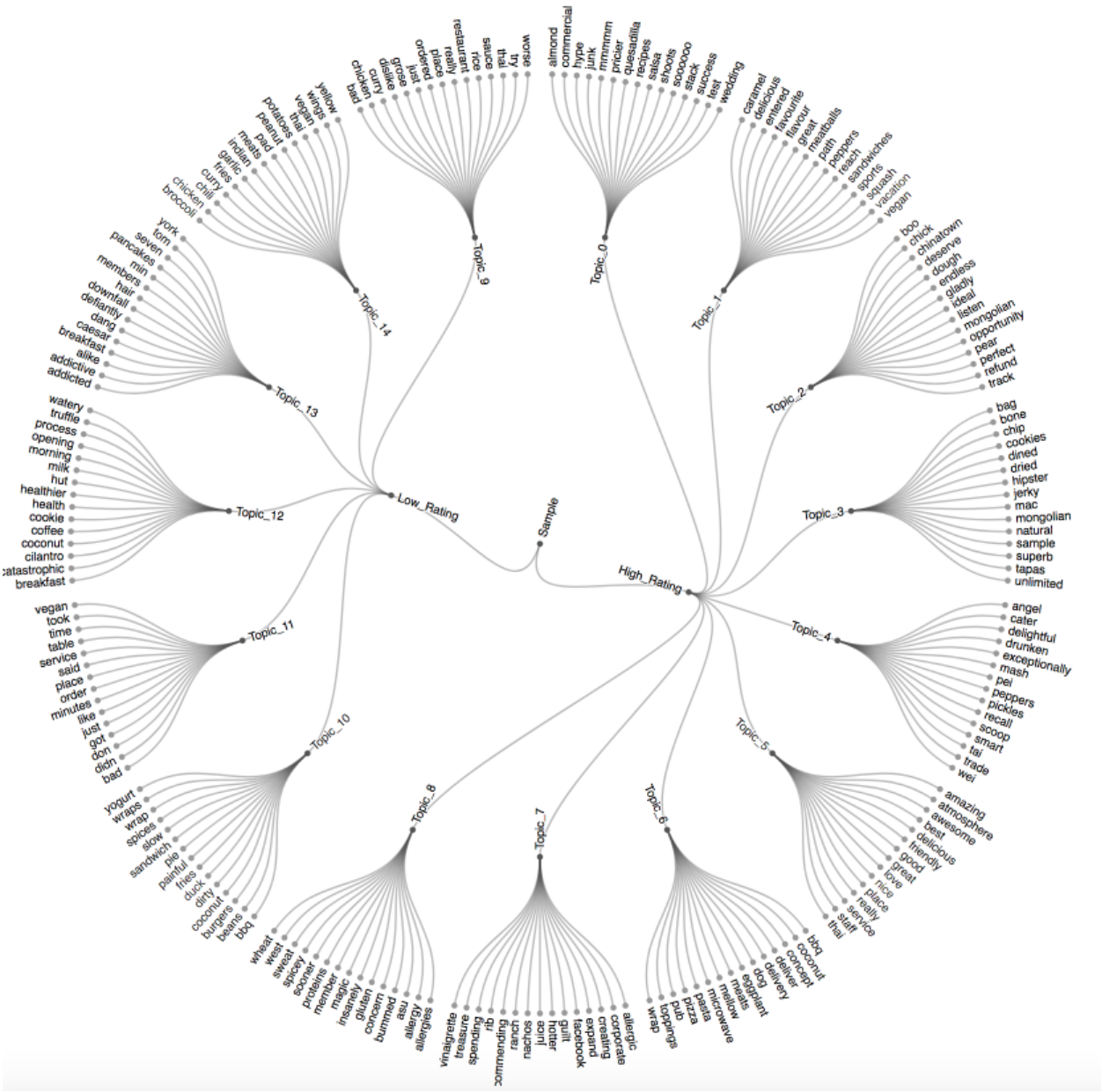
Introduction

In this project, I utilized Latent Dirichlet Allocation (LDA) to analyze topics within the Yelp Dataset Challenge academic dataset. LDA is a generative probabilistic model that identifies latent topics by analyzing the distribution of words across documents. I set the model parameters to 15 topics. The preprocessing steps involved tokenizing the text, converting it to lowercase, removing stop words, and performing stemming or lemmatization to reduce words to their root forms. This processed text was then vectorized into a bag-of-words model, resulting in a document-term matrix. I used the Gensim library in Python to apply the LDA model, preparing the corpus, training the LDA model with the specified parameters, and extracting the top words associated with each topic, as well as the distribution of topics across documents. For visualization, I used the D3.js to generate basic circular dendrogram. I have utilized various topic models and Python libraries such as NLTK, sklearn, gensim, matplotlib, and GraphLab for topic modeling and text processing. Additionally, I have used matplotlib.pyplot and Python's WordCloud to generate word cloud images.



Task 1.1

The initial radial dendrogram presents the top 15 topics identified by Latent Dirichlet Allocation (LDA) and showcases the most frequently occurring words in each of these topics.



Task 1.2

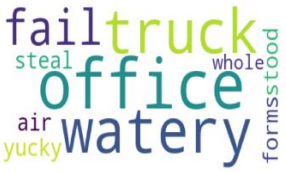
The second radial dendrogram illustrates the top topics in two categories: low and high ratings. To achieve this, I divided the dataset into two subsets: containing reviews with 1 and 5 stars, reviews. From this analysis, we can clearly see that the most frequent words in low-rating topics tend to have negative connotations, while those in high-rating topics are positive.

- Word clouds were generated from all the collected reviews.
- A fourth group of subsets was formed by dividing all reviews into positive and negative categories.
- The dendrograms were generated using data exclusively from restaurant reviews.
- Visualizations were also created for subsets of both negative and positive restaurant reviews.

Positive Word Cloud:

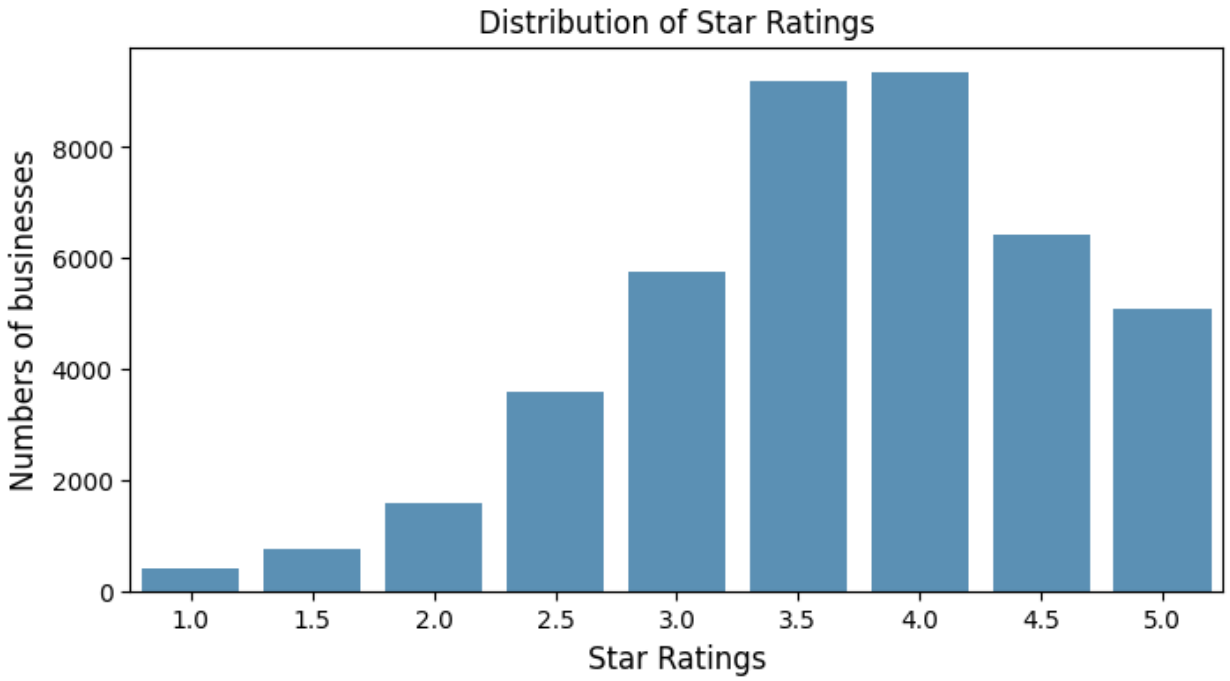
CS598Capstone/data.ipynb at main · kasamdh/CS598Capstone (github.com)

Negative Word Cloud:



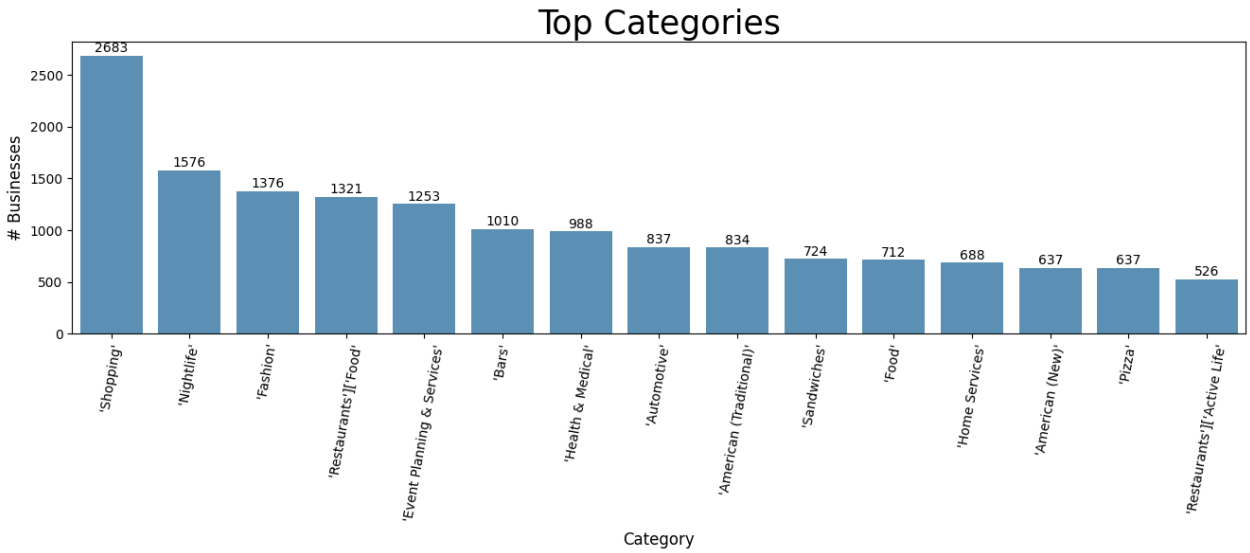
Distribution of Star Ratings (1-5):

CS598Capstone/data.ipynb at main · kasamdh/CS598Capstone (github.com)



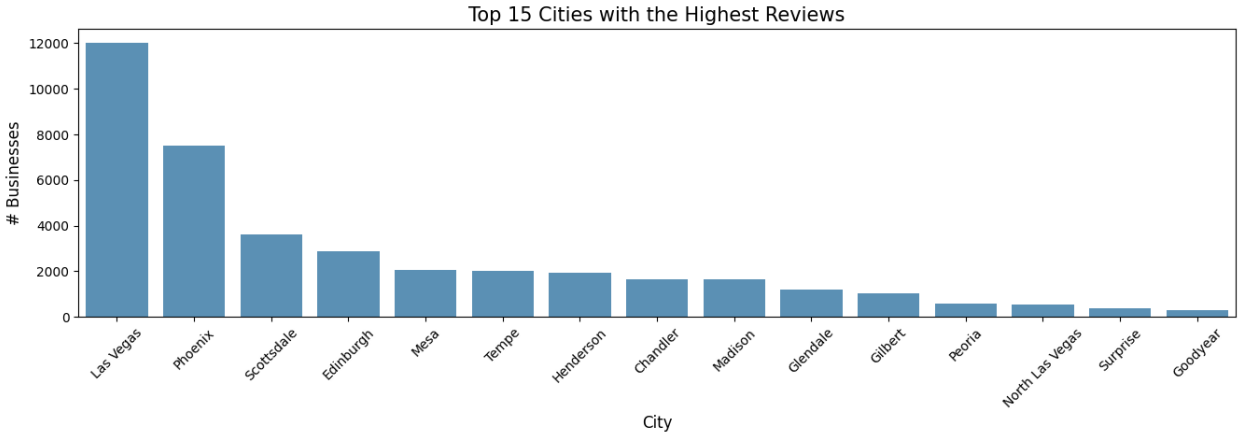
Initially, analyze the distribution of ratings to fulfill a key objective in our data analysis. It becomes apparent that most reviews lean towards the positive end of the spectrum, as per Yelp's 1–5-star rating system.

15 top Categories:



I have filtered down to 15 top categories for yelp with reviews.

Top 15 Cities with Highest Reviews:

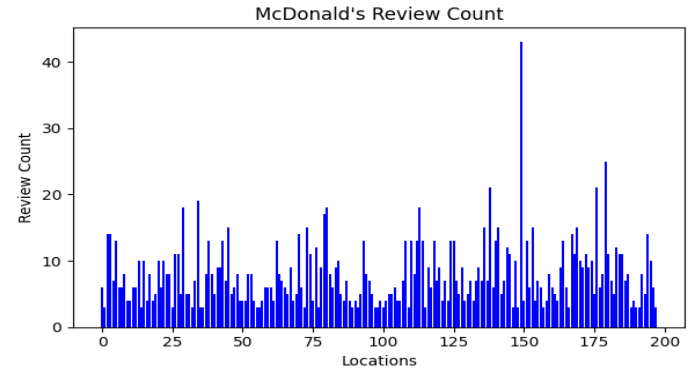


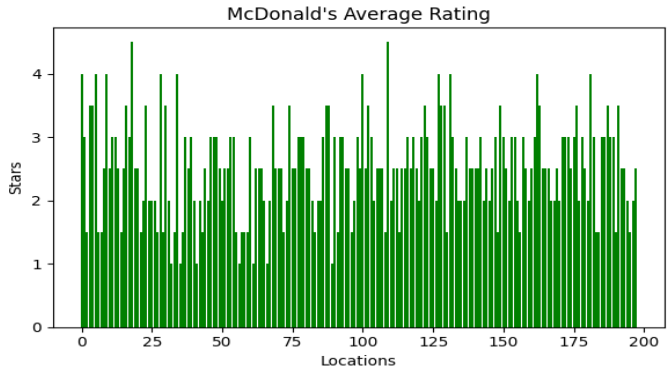
I have filtered down to 15 top Cities with the highest reviews.

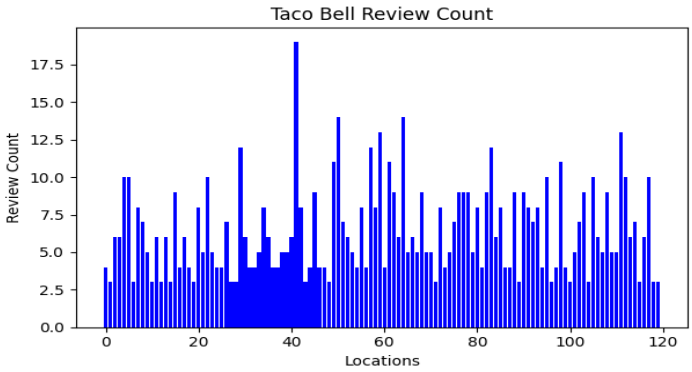
Restaurant with Highest and Lowest Star Ratings:

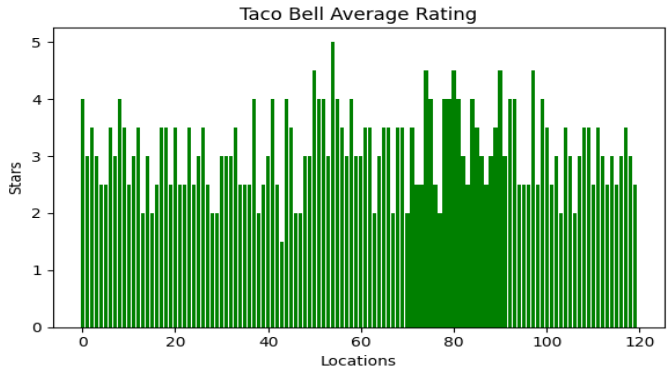
CS598Capstone/review.ipynb at main · kasamdh/CS598Capstone (github.com)

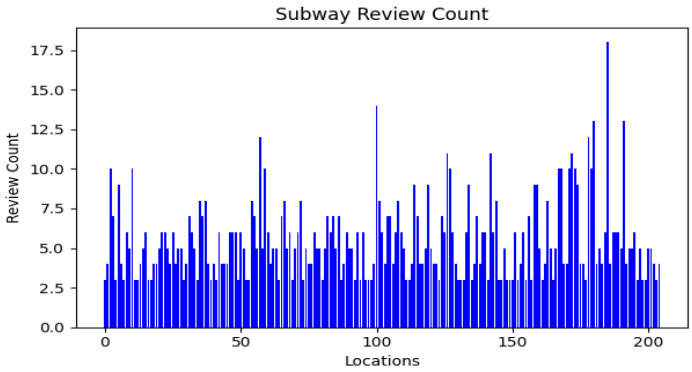
The restaurant with highest review count is McDonald and Lowest review count is Subway, but the Highest Star Ratings is Taco Bell, and the Lowest Star Ratings is McDonald.

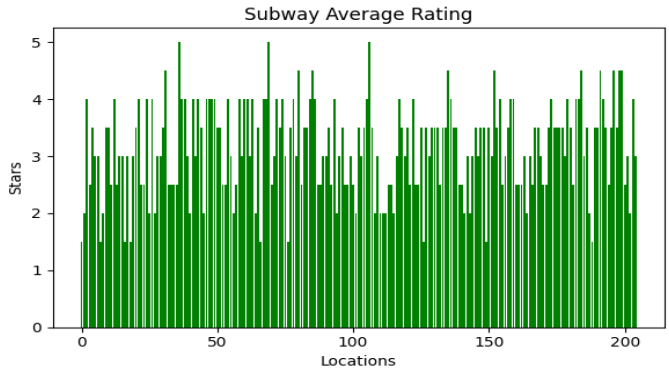

McDonald's Review Count
Review Count
Locations

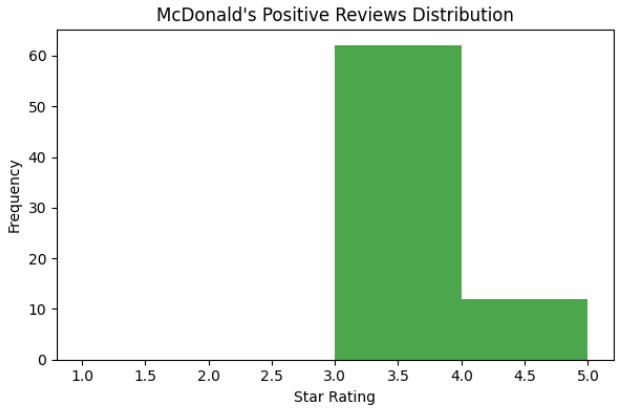

McDonald's Average Rating
Stars
Locations

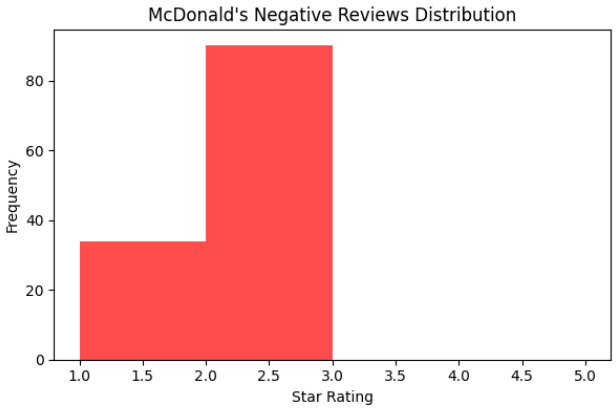

Taco Bell Review Count
Review Count
Locations

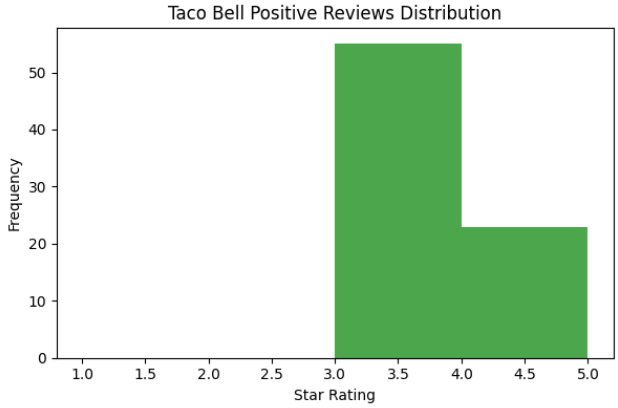

Taco Bell Average Rating
Stars
Locations

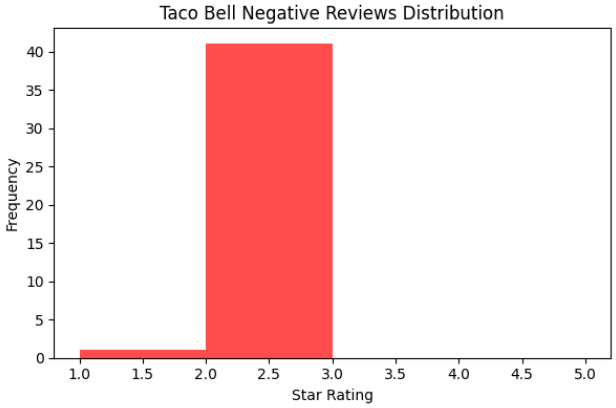

Subway Review Count
Review Count
Locations

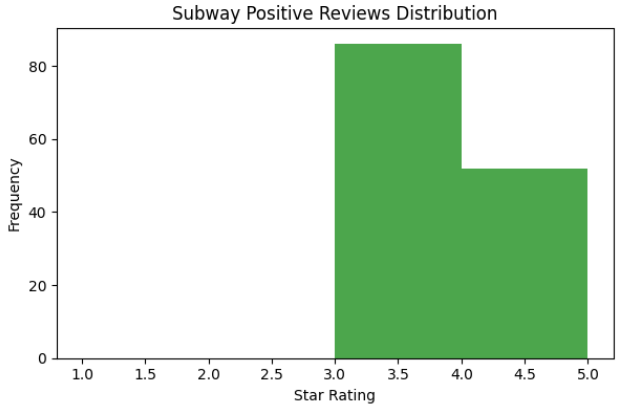

Subway Average Rating
Stars
Locations

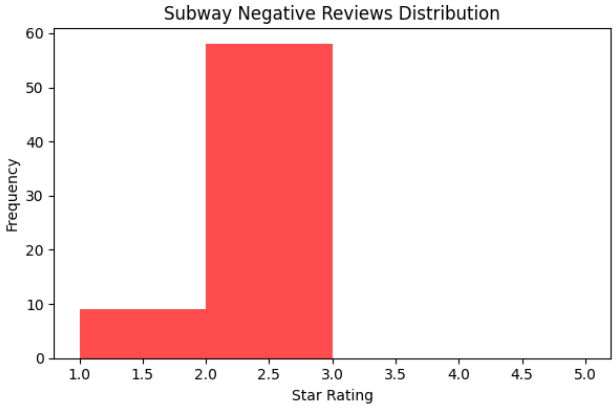

McDonald's Positive Reviews Distribution
Frequency
Star Rating


McDonald's Negative Reviews Distribution
Frequency
Star Rating


Taco Bell Positive Reviews Distribution
Frequency
Star Rating


Taco Bell Negative Reviews Distribution
Frequency
Star Rating


Subway Positive Reviews Distribution
Frequency
Star Rating


Subway Negative Reviews Distribution
Frequency
Star Rating

References:

1. gensim · PyPI
2. Most basic radial dendrogram in d3.js (d3-graph-gallery.com)
3. D3 by Observable | The JavaScript library for bespoke data visualization (d3js.org)