|  |
| --- |
| **Impacts of Restaurant Hygiene on Yelp Rating in NYC** |
|  |
| **Chenyue Wang (cw3444)**  **Hanna Gao (qg2205)** |
| **Yi Zhong (yz4330)** |
|  |
|  |
|  |

Introduction

Restaurant hygiene refers to the cleanliness and sanitation practices that are followed in food service establishments, such as restaurants, cafes, and other food outlets, which is a topic that closely relates to our lives. Insufficient practices for restaurant hygiene can easily lead to the spread of harmful bacteria, viruses, and other pathogens that can cause foodborne illnesses. By Hodges et al. (2022), there are approximately 48 million people in the U.S. sickened by foodborne illnesses every year.

Restaurant hygiene is a manifestation of the quality of a restaurant. In any market, if there is no control over the quality, the market can become unstable or unsafe for consumers. Low-quality providers may force high-quality providers to leave the market because providers lack the motivation to improve the quality of their services.

In this paper, we will explore the relationship between customer reviews score and restaurant hygiene scores in New York City. While a few previous research has explored how confounding variables factors affect restaurant hygiene, little has been done to specifically investigate restaurants in New York City. Most of the previous works focus on how the factors impact the hygiene score of restaurants, while our current study seeks to fill the gap in the other direction: how the hygiene score affects online customer ratings for restaurants in the New York City. Finding this link between these two scores will not only provide empirical evidence, but also guide the restaurants that are improving their hygiene scores, delivering quality services, building customer loyalty, and creating a positive reputation to attract new customers.

For our hypothesis, we argue that restaurants in New York City with higher hygiene scores have significantly higher customer review scores, indicating customers are likely to have better perceptions of a restaurant's food and service when the hygiene rates are higher. Moreover, we will examine the effects of potential confounding variables, including price, cuisine type, and location. We argue that given the same hygiene score, restaurants with higher prices will have higher customer review scores. Also, the restaurant's cuisine type and location are crucial factors in determining the review scores. For example, the average customer ratings for restaurants at the same location but with different cuisine types can vary significantly, and vice versa.

We will collect data from two datasets: the New York City Restaurant Inspection result from NYC open data and customer review scores extracted from Yelp using Yelp Fusion. We will utilize Python to preprocess and clean up the data, and then perform correlation analysis, regression analysis, and ANOVA on the data to evaluate the hypothesis.

Related Works

The hygiene and sanitary condition of restaurants can be affected by many factors. According to Laikko-Roto and Nevas (2014), there is a positive relationship between restaurants’ hygiene level and the attributes and hygiene knowledge of the restaurant business. Moreover, Fleetwood (2019) shows evidence of a relationship between restaurant hygiene and public policy of mandatory hygiene level posting on doors. Following the policy of compulsory hygiene score posting on restaurant doors in New York, there was a 35% rise in the likelihood of achieving the top "A" rating for restaurant hygiene. (Fleetwood, 2019).

Nowadays, people rely heavily on online review platforms, such as Yelp and Google Maps, for restaurant selection. This is because people believe user-generated content is more credible than information provided by brand owners (Meek et al., 2021). Farronato and Zervas’s (2022) study analyzed the impact of online reviews from Yelp, which contain hygiene signals, on customer choice. They concluded that the probability of a restaurant being sold out decreased after a negative customer review on hygiene conditions, showing that restaurant hygiene is related to the restaurant's reputation and profit. (Farronato and Zervas, 2022).

Moreover, online reviews might be a predictor of restaurant hygiene. Siering (2021) built a model to predict potential restaurant hygiene violations, which achieved an accuracy of 79.33%.

Data Collection and Preprocessing

The data used in the analysis is a combination of two datasets.

The first dataset is the New York City Restaurant Inspection result from NYC open data collected on 03/20/2023. The raw data contains 206976 inspection records and 32 columns. As the first step, we removed unnecessary columns and dropped records with duplicate restaurants and missing hygiene grades. To achieve the most up-to-date results, we only maintained inspection records after 01/01/2022. Also, because of the Yelp API usage limits and to conserve storage capacity on the local computer, we randomly selected 5000 restaurants for analysis using Python random library. After the cleanup, the dataset contains 5000 inspection records with 9 variables, which are shown in Table.1.

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| CAMIS | Unique identifier for the restaurant |
| DBA | Restaurant name |
| BORO | Borough in which the restaurant is located |
| ZIPCODE | ZIPCODE of restaurant location |
| Latitude | Latitude of restaurant location |
| Longitude | Longitude of restaurant location |
| Cuisine Description | The restaurant cuisine |
| Score | Total score for inspection |
| Grade | Grade associated with the inspection (N = Not Yet Graded; A = Grade A; B = Grade B; C = Grade C; P = Grade Pending) |
| Grade Date | The date when the current grade is issued |

Table.1 Variables Description of NYC Restaurant Inspection Results

The second dataset is the customer review scores of each restaurant collected using Yelp Fusion API on 03/20/2023. We first used the restaurant’s name, altitude, and longitude to search for each restaurant in the first dataset. Then, we retrieved 30 customer reviews of each restaurant. To line up the dates, we only maintained customer reviews that were posted after 01/01/2022.

After cleaning up, matching, and concatenating first and second dataset, the dataset contains 1946 records with 16 attributes for each restaurant. The attributes are CAMIS, BORO, ZIPCODE, Latitude, Longitude, Cuisine Description, Score, Grade, Grade Date, restaurant id, restaurant name, review\_count, rating, price, location, Yelp url.

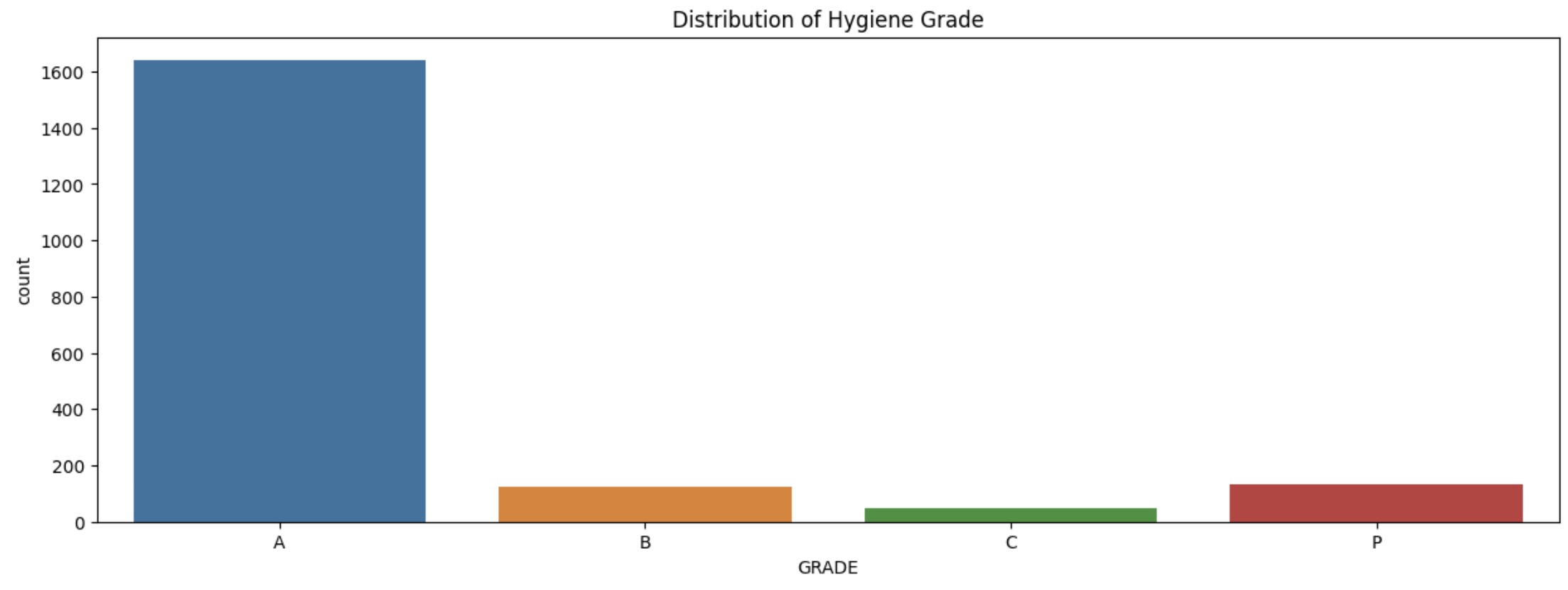


Figure.1 Distribution of Hygiene Grades of the restaurants

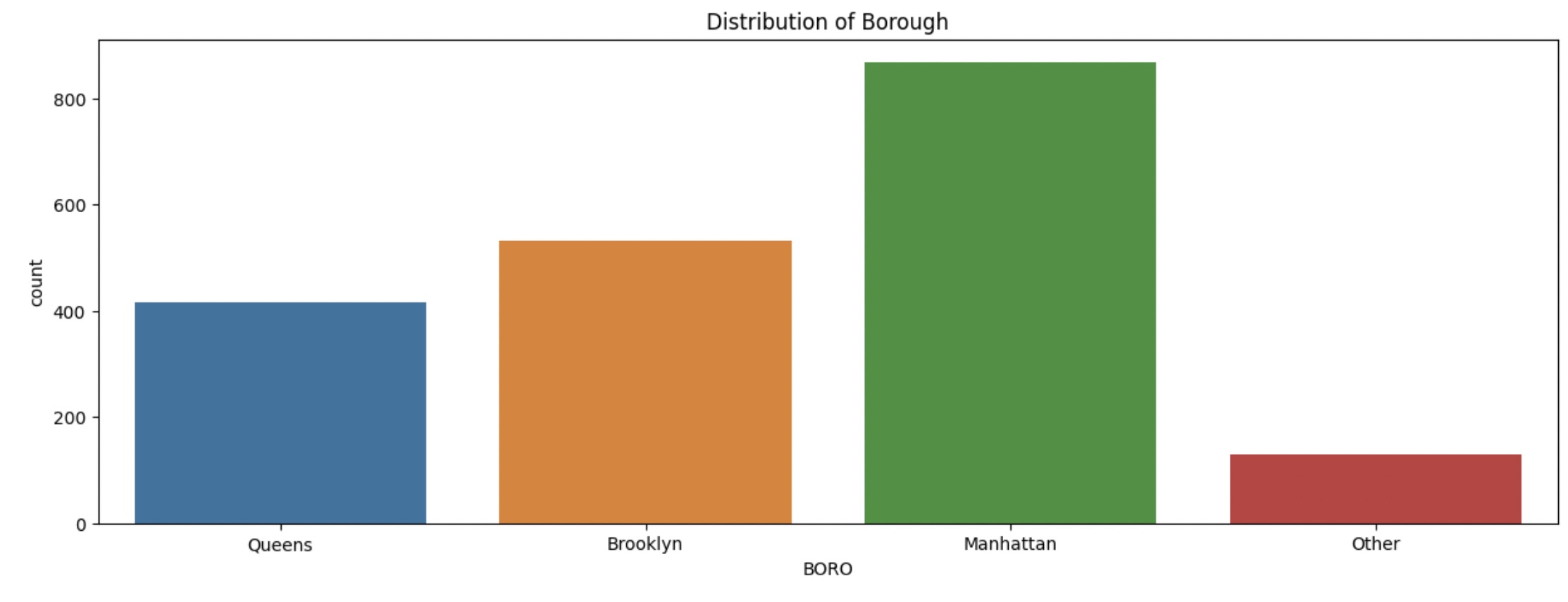


Figure.2 Distribution of Borough of the Restaurants

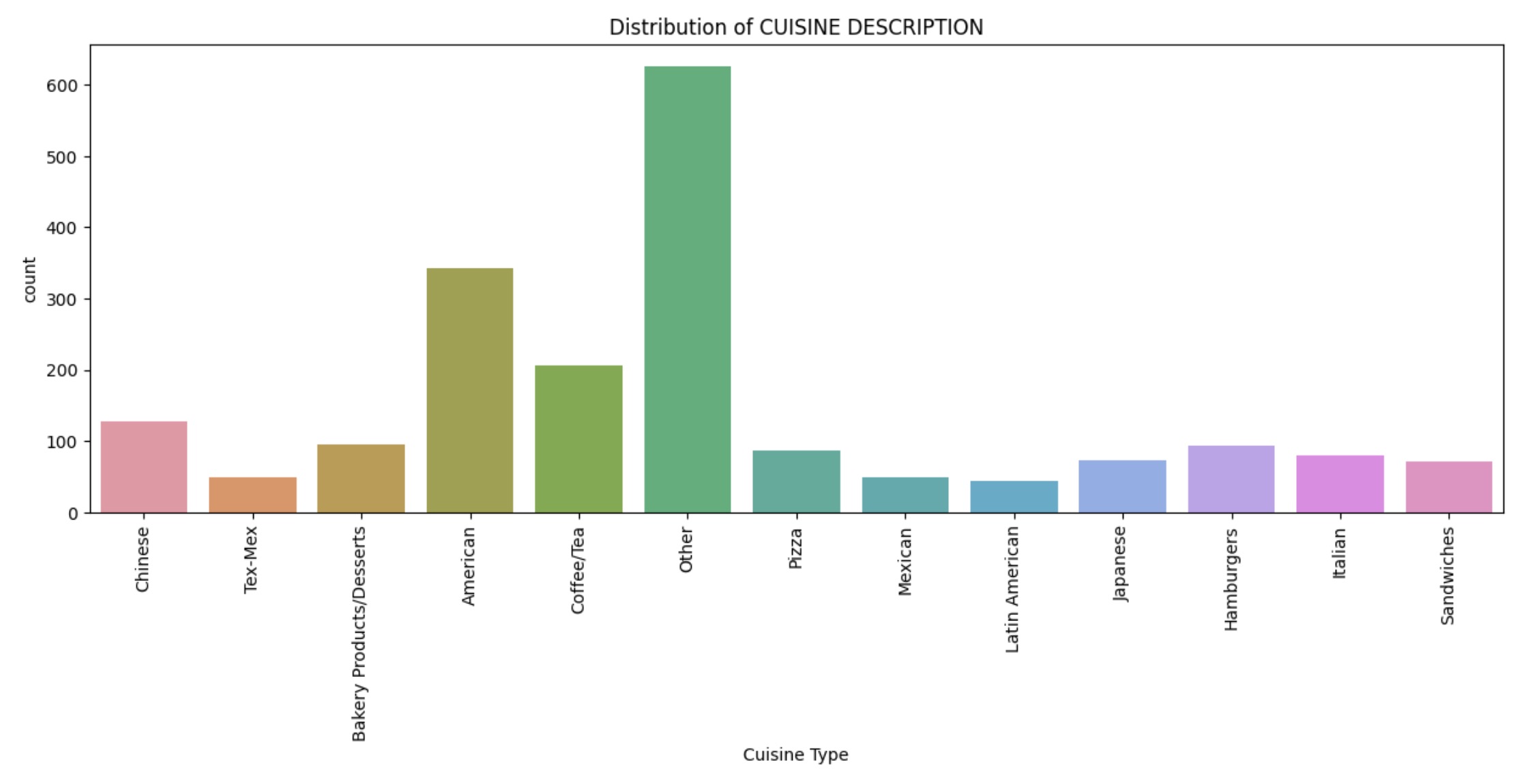


Figure.3 Distributions of Cuisine Types of the Restaurants

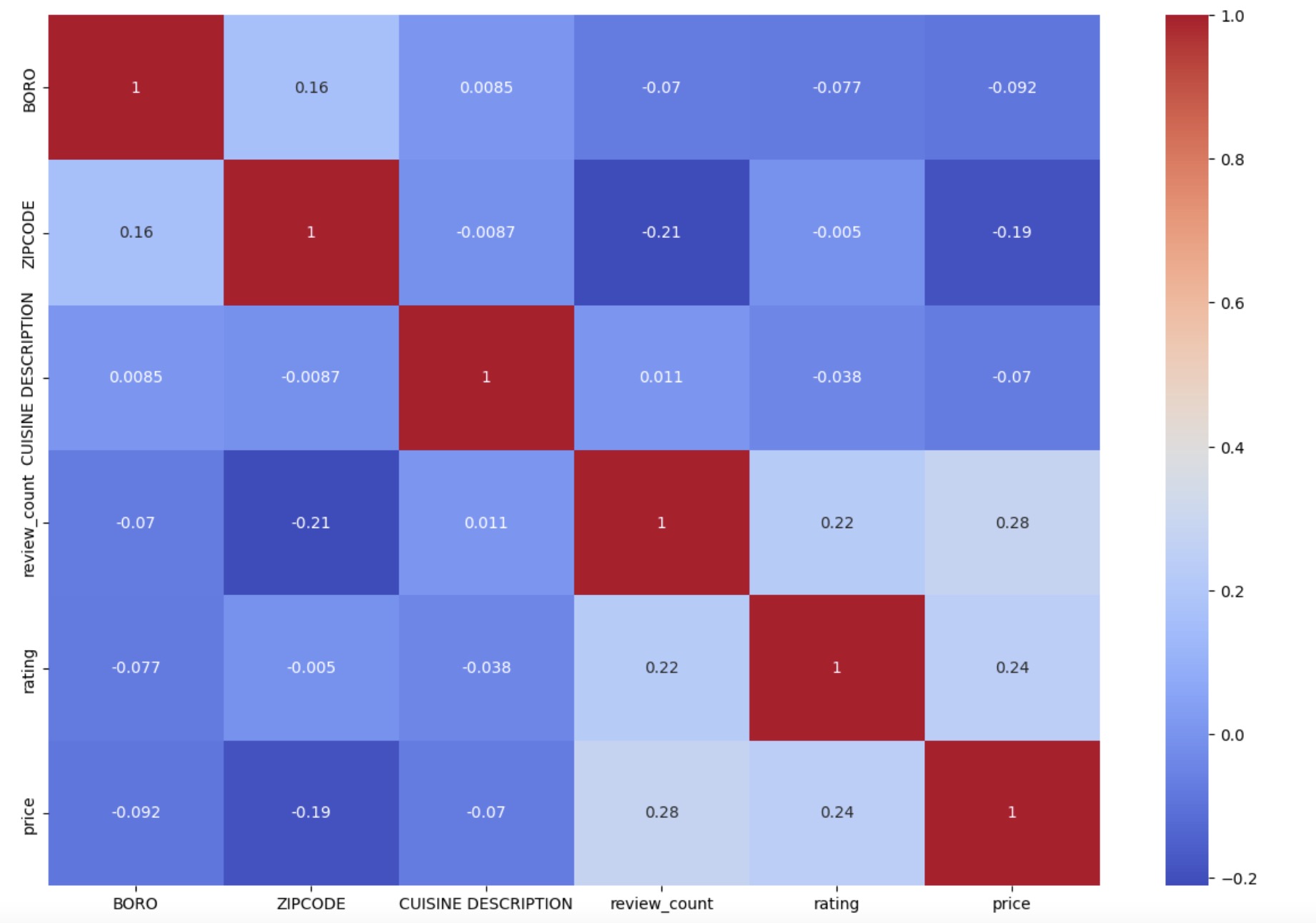


Figure.4 Correlation of restaurant features

(To Do: Perform filtering to only select comments that relate to hygiene and were made after 01/01/2022, and calculate the average customer rating of those comments; perform EDA on the final second dataset)

1. Methods (To Do: revise after finishing the statistical analysis)

We’ll perform several statistical analyses to test our hypotheses using external tools such as Python and R, along with relevant packages and librarians. These tools can help us create accurate data visualization and conduct effective statistical tests on datasets to investigate the relationship between customer review scores and restaurant hygiene ratings.

The first potential statistical analysis would be correlation analysis. We could calculate the correlation coefficient between the hygiene score and the customer review score. If we obtain a positive and statistically significant correlation coefficient, then it would serve as evidence to support our hypothesis. We can also plot a scatter plot to visualize the relationship between these two variables.

The second option would be regression analysis. We could conduct a linear regression to model the relationship between review scores and hygiene ratings, controlling for any possible effects of potential confounding variables. For instance, we can set the hygiene score as the independent variable and the review score as the dependent variable. If the regression coefficient is positive and statistically significant, then it would provide us with support for our hypothesis. We can plot a residual plot to check these two variables’ relationships.

The third analysis is to consider other confounding variables that can potentially affect customer review score, e.g. cuisine type, location, and price. For instance, for location, we will plot their rating data on NYC map to show distribution of ratings w.r.t. boroughs. For price, we will use a density plot to represent the distribution of ratings w.r.t. price of each restaurant. The price is a categorical variable represented by the number of $. For cuisine type, we will analyse and create a similar distribution plot.

1. Results (To Do: complete after performing the statistical analysis)

To Do: We will put above plots here and analyse their relationships

1. Discussion (To Do: complete after performing the statistical analysis)
2. Conclusion (To Do: complete after performing the statistical analysis)

References

Chiara Farronato and Georgios Zervas. 2022. [CONSUMER REVIEWS AND REGULATION: EVIDENCE FROM NYC RESTAURANTS](https://www.nber.org/papers/w29715). *NBER WORKING PAPER SERIES*, 29715.

Hodges et al. 2022. [Exploring Foodborne Illness and Restaurant Cleanliness Reporting in Customer-Generated Online Reviews Using Business Analytics](https://pubmed.ncbi.nlm.nih.gov/24848215/). *Journal of Environmental Health* (Volume 85), 3: 16-22.

Janet Fleetwood. 2019. [Scores on doors: restaurant hygiene ratings and public health policy.](https://pubmed.ncbi.nlm.nih.gov/31383956/) *Journal of Public Health Policy*, 40: 410–422.

Meek et al. 2021. [A big data exploration of the informational and normative influences on the helpfulness of online restaurant reviews](https://www.sciencedirect.com/science/article/pii/S0148296320308328). *Journal of Business Research*, 125: 354-367.

Siering. 2021. [Leveraging online review platforms to support public policy: Predicting restaurant health violations based on online reviews](https://www.sciencedirect.com/science/article/pii/S0167923620302293). *Decision Support Systems*, 143: 113474.

Tiina Läikkö-Roto and Mari Nevas. 2014. [Restaurant business operators’ knowledge of food hygiene and their attitudes toward official food control affect the hygiene in their restaurants](https://www.sciencedirect.com/science/article/pii/S0956713514001121). *Food Control*, 43: 65-73.