Analyzing Factors Affecting Customer Satisfaction with Top Non-American Ethnic Cuisines in the US

Authors: Han Sun (Claire), Bowie Liu | February 2023 | Team #30

1. Motivation

The motivation behind our analysis of Yelp restaurant data is to explore the factors influencing customer satisfaction and behavioral intention regarding four pivotal cuisines: Chinese, Japanese, Italian, and Mexican. These cuisines represent prominent non-American culinary traditions in the United States, deeply ingrained in the fabric of ethnic cuisine and widely embraced by consumers. As interest in diverse culinary experiences continues to grow, understanding the dynamics that drive preferences within these culinary categories becomes increasingly relevant.

Powell et al.'s research (2014) emphasizes the importance of considering how exposure to food advertising varies across demographic segments. Disparities in exposure could shape food preferences and consumption patterns, potentially impacting the popularity of specific cuisines among different racial/ethnic and income groups. However, research by Tomić et al. (2018) highlights the positive attitudes consumers exhibit toward ethnic food, reflecting a broader societal trend toward embracing culinary diversity. Similarly, the study by Boch et al. (2021) suggests that as aspects of ethnic culture assimilate into mainstream culture, the emphasis on the ethnic origins of these cuisines may diminish.

These research findings prompt questions about the consumption patterns of popular ethnic cuisines: are they uniformly favored across diverse demographic groups? Alternatively, could preferences be influenced by factors such as the concentration of specific ethnicities or the appeal of distinct flavors and culinary experiences?

Our objective is to unravel the intricate interplay between demographic characteristics and restaurant attributes that influence customer satisfaction with ethnic cuisines. Through comparative analysis of restaurant ratings for each of the four cuisines using Yelp data, we aim to discern the factors contributing to disparities in customer satisfaction (ratings) across these culinary categories. Ultimately, this analysis seeks to provide valuable insights to stakeholders in the food industry, fostering a deeper understanding of culinary diversity in the United States.

2. Data Sources

2.1. Yelp Data (Primary Data Source)

Yelp provides access to the data exclusively for academic or non-commercial purposes through its website. The data includes 5 datasets: “business”, “checkin”, “review”, “tip”, and “user”. We excluded the “checkin” and “user” datasets from our analysis as they are not relevant to our analysis objectives.

- The “business” dataset contains detailed information about businesses on Yelp, such as ratings/stars, location/zip code, review count, attributes, categories, etc.
- The “review” dataset contains every review for each business as well as the characteristics of each review.
- The “tip” dataset contains tips provided by users and complement ratings.

2.2. US Census Data (Secondary Data Source)

Accessing US Census Data involves a process that begins by visiting the US Census Bureau's website to obtain an API key. This key grants access to the database and the website's documentation outlines the available variables for retrieval.

For our project, we have selected nine demographic features to extract: median household income, population, household count, population with education, population with a bachelor's degree, median age, Hispanic Latino population, White population, and Asian population. Our primary focus is retrieving data for zip codes where restaurants from the Yelp dataset are located.

2.3. Summary of Data Sources

Table 1

Dataset	Business	Review	Tip	US Census
Source	yelp.com	yelp.com	yelp.com	census.gov
Format	JSON	JSON	JSON	JSON
Access Method	Download	Download	Download	API
Important Variables	'business_id', 'state', 'postal_code', 'stars', 'review_count', 'attributes', 'categories'	‘user_id’, 'business_id', 'stars', 'useful', 'funny', 'cool', 'text'	‘user_id’, 'business_id', 'text', 'compliment_count'	‘median_household_income’, ‘population’, ‘household_cnt’, ‘bachelors_degree’, ‘median_age’, ‘population_hispanic_latino’, ‘population_white’, ‘population_asian’
Number of Variables	14	9	5	9
Number of Records	150,346	6,990,280	908,915	803
Size	113.4MB	4.98GB	172.2MB	42KB

3. Data Manipulation Methods

3.1. Collect and Transform Data

- We began by reading the Yelp source data in JSON format and then **converted** it into pandas data frames.
- We collected demographic data by creating a function to **interact with the US Census API** using the “requests” package. We retrieved the data in JSON format and subsequently transformed it into a pandas data frame.
- Some columns in Yelp data, such as “attributes”, contain nested dictionaries with inconsistent keys across records, making them unsuitable for direct analysis. Therefore, we **flattened** these columns, merged the flattened data back into the original dataset, and subsequently dropped the original columns. This process resulted in the creation of multiple new columns and introduced missing values.

3.2. Subset and Drop Data

- We subsetted the Yelp data to include only restaurants located in the U.S. to create the “**restaurant**” table.
- We further subsetted the data to focus on 4 cuisines: Chinese, Japanese, Mexican, and Italian.
- To ensure that each record uniquely represents a single cuisine, we introduced a “MECE_check” column that aggregates the total count of cuisines associated with each restaurant. Records displaying a value other than 1 in this column are excluded, thus ensuring a mutually exclusive and collectively exhaustive (**MECE**) categorization of cuisine types within our dataset.
- We **dropped** columns that are not relevant to our analysis, such as “name” and ”address”, as well as columns that contain a single value and thus do not provide any useful information.

- Each time we drop data from a dataset, we **create a new dataset**. This practice is commonly recommended to preserve the integrity of the original data and ensure traceability and reproducibility throughout the analysis process.

3.3. Check Missingness

- There is a significant amount of missing data in the Yelp data, particularly after flattening some columns, as mentioned in section 3.1. To address this, we first calculated the percentage of missing values for each column representing restaurant attributes. Then, we decided to **drop** columns with **more than 50% missing** values.
- For columns containing key information for our analysis, such as “zip code”, we simply **removed** all records containing missing values in these columns to ensure complete and accurate data in essential areas.
- For columns that are less critical to the analysis, we utilized our experience and exercised judgment in determining the replacement strategy for missing values, opting for methods such as **replacing** them with 0, the mean value, the mode value, or encoding them as 99.
- Several iterations were required to address missing values, as subsequent data manipulation steps introduced new instances of missingness. For example, when merging the “restaurant” table with the “review” table, missing values were introduced because some restaurants had no reviews. In such cases, we simply **replaced** the missing values with 0.

3.4. Check Errors and Outliers

- We examined the data distribution by utilizing functions like “describe” to identify data errors or outliers. For instance, we discovered zero or negative values for population, median household income, and the percentage of people with bachelor’s degrees in the demographic data. To guarantee the accuracy and meaningfulness of our subsequent analysis, we **filtered out** these erroneous data points.

3.5. Aggregate Data

- The Yelp “review” and “tip” datasets are at the user level, while the “restaurant” table is at the business level. Since each restaurant can have multiple user reviews and tips, and our analysis focuses on restaurants instead of users, we **aggregated** the “review” and “tip” data to the business level to maintain consistency and avoid duplications when merging these datasets.

3.6. Merge Data

- Each time we merge different tables, we **create a new dataset** instead of overwriting the existing dataset. It's a common **best practice** that helps preserve the original data and allows for traceability and reproducibility in the analysis process.
- When joining the "restaurant" table with demographic data, we chose **inner join** to exclude any records with missing zip codes or demographic information, ensuring no missing data in key fields. This process led to the creation of a new dataset, named “master1”.
- Next, we used the “master1” table to **left join** aggregated review data to create “master2”.
- Following that, we performed a **left join** with aggregated tip data to create “master3”.

3.7. Engineer Features

- The original Yelp “review” and “tip” datasets contain text data, which cannot be directly used or aggregated. Therefore, we employed the sentiment lexicon “SentimentIntensityAnalyzer” to assess the sentiment of words in the text and **generate sentiment scores**. These scores range from -1 to 1, where:
    - 1 indicates a positive review,
    - 0 indicates a neutral review, and
    - -1 indicates a negative review.
- The “restaurant” table contains lots of categorical variables, each with multiple levels. To extract the main information from these variables, we grouped their levels and utilized **One-hot Encoding** to convert categorical attributes into binary numerical values.
- Given that our demographic data is organized at the zip code level, it naturally exhibits variations in population density and area among different zip codes. To address this variability, we **normalized** specific demographic

features to make comparisons more meaningful. For example, we calculated metrics such as "population_perRestaurant" and "household_perRestaurant" to normalize population and household data relative to the number of restaurants in each zip code.

- We further enriched our dataset by converting absolute values into relative ratios, thereby enhancing their utility. For instance, we **calculated percentages** such as "bachelors_pcnt", "education_pcnt", "hispanic_latino_pcnt", "white_pcnt", and "asian_pcnt" to reveal information relative to the total population or specific demographic groups.
- Finally, we **extracted** cuisine information in the "master3" table and subsetted the data to focus on the four cuisines of interest, creating 'master4' for subsequent analysis.

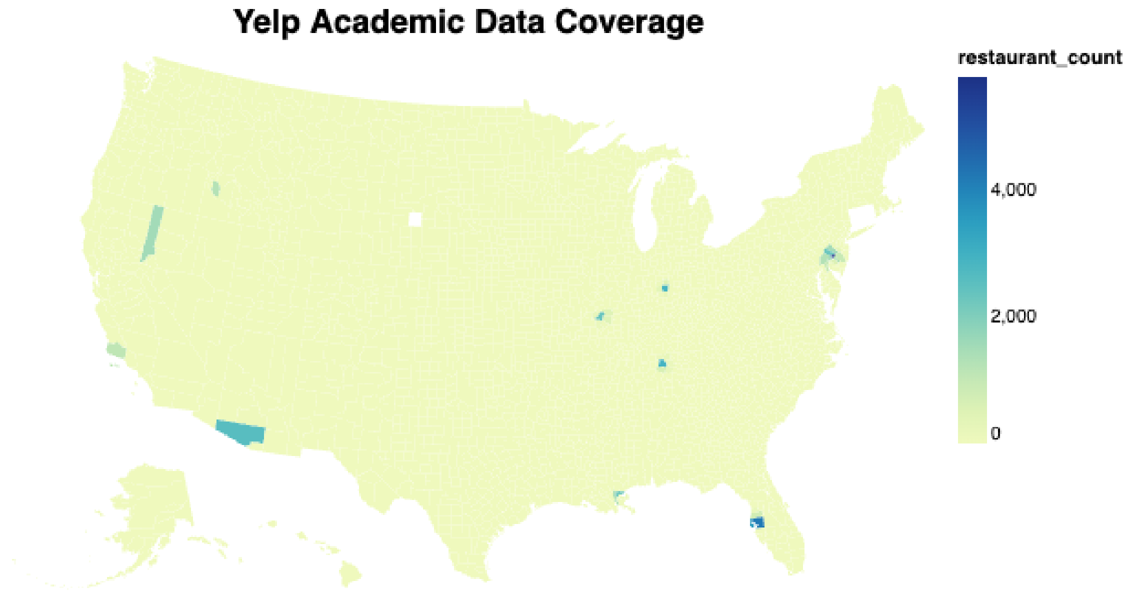
4. Analysis

4.1. Check Yelp Data Coverage and Potential Biases

We discovered that the academic dataset shared by Yelp is only a small portion of the real-world data. This subset covers merely 16 states, with 3 of them having only 1 restaurant represented.

As illustrated in Figure 1, the Yelp data is evidently not randomly sampled at the zip code level, whereas our secondary data source - demographics - is structured at the zip code level. Therefore, our analysis is likely biased due to the data limitations, and our findings may not be applicable to the entire country.

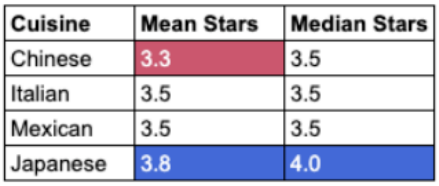
Figure 1



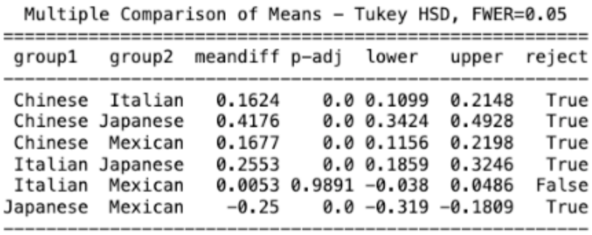
4.2. Compare Restaurant Ratings Across Cuisines

As shown in Table 2 and Figure 2 below, Japanese cuisine generally receives higher ratings compared to the other three cuisines, with the highest mean and median stars among all categories. On the contrary, Chinese cuisine has the lowest average ratings among the four, with a statistically significant difference observed.

Table 2



Cuisine	Mean Stars	Median Stars
Chinese	3.3	3.5
Italian	3.5	3.5
Mexican	3.5	3.5
Japanese	3.8	4.0



```
  Multiple Comparison of Means - Tukey HSD, FWER=0.05
=======================================================
 group1   group2  meandiff p-adj  lower   upper  reject
-------------------------------------------------------
 Chinese  Italian   0.1624    0.0 0.1099  0.2148   True
 Chinese Japanese   0.4176    0.0 0.3424  0.4928   True
 Chinese  Mexican   0.1677    0.0 0.1156  0.2198   True
 Italian Japanese   0.2553    0.0 0.1859  0.3246   True
 Italian  Mexican   0.0053 0.9891 -0.038  0.0486  False
Japanese  Mexican    -0.25    0.0 -0.319 -0.1809   True
-------------------------------------------------------
```

Figure 2

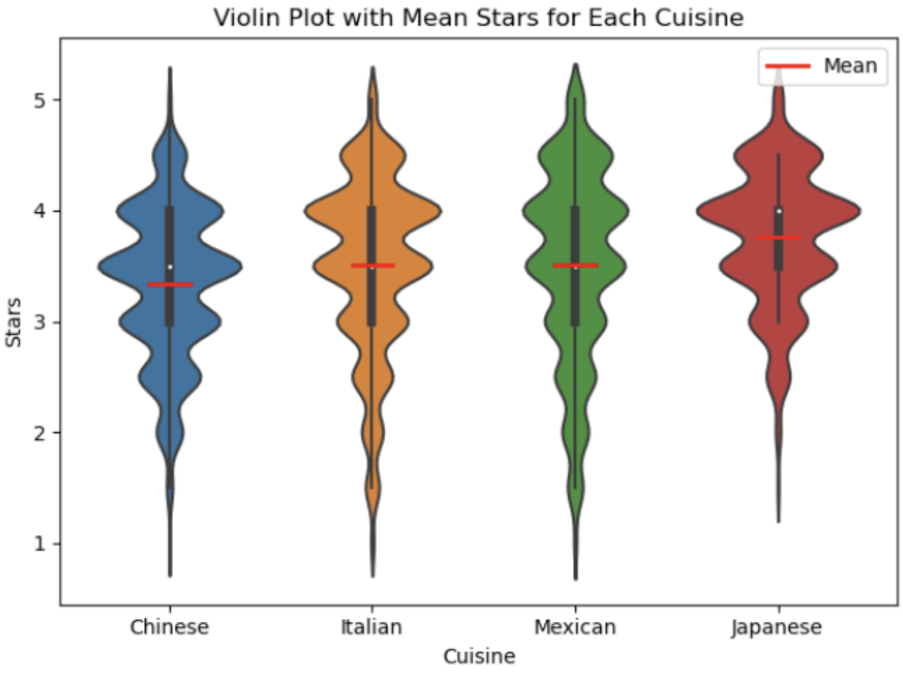
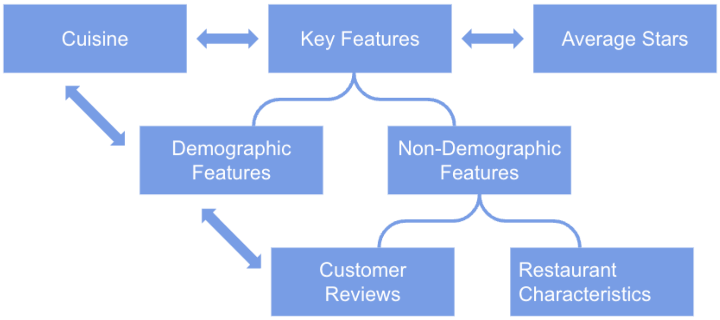


Figure 3



To understand the reasons behind the high customer satisfaction of Japanese restaurants and the lower ratings typically received by Chinese restaurants, despite both serving Asian cuisine, we investigated the factors influencing restaurant ratings across the four types of cuisine. Figure 3 provides an overview of our approach.

4.3. Identify Key Factors Influencing Restaurant Ratings

We categorized features into numerical and categorical, which facilitates the application of different methods in exploring the relationship between restaurant ratings (i.e. **stars**) and numerical versus categorical features.

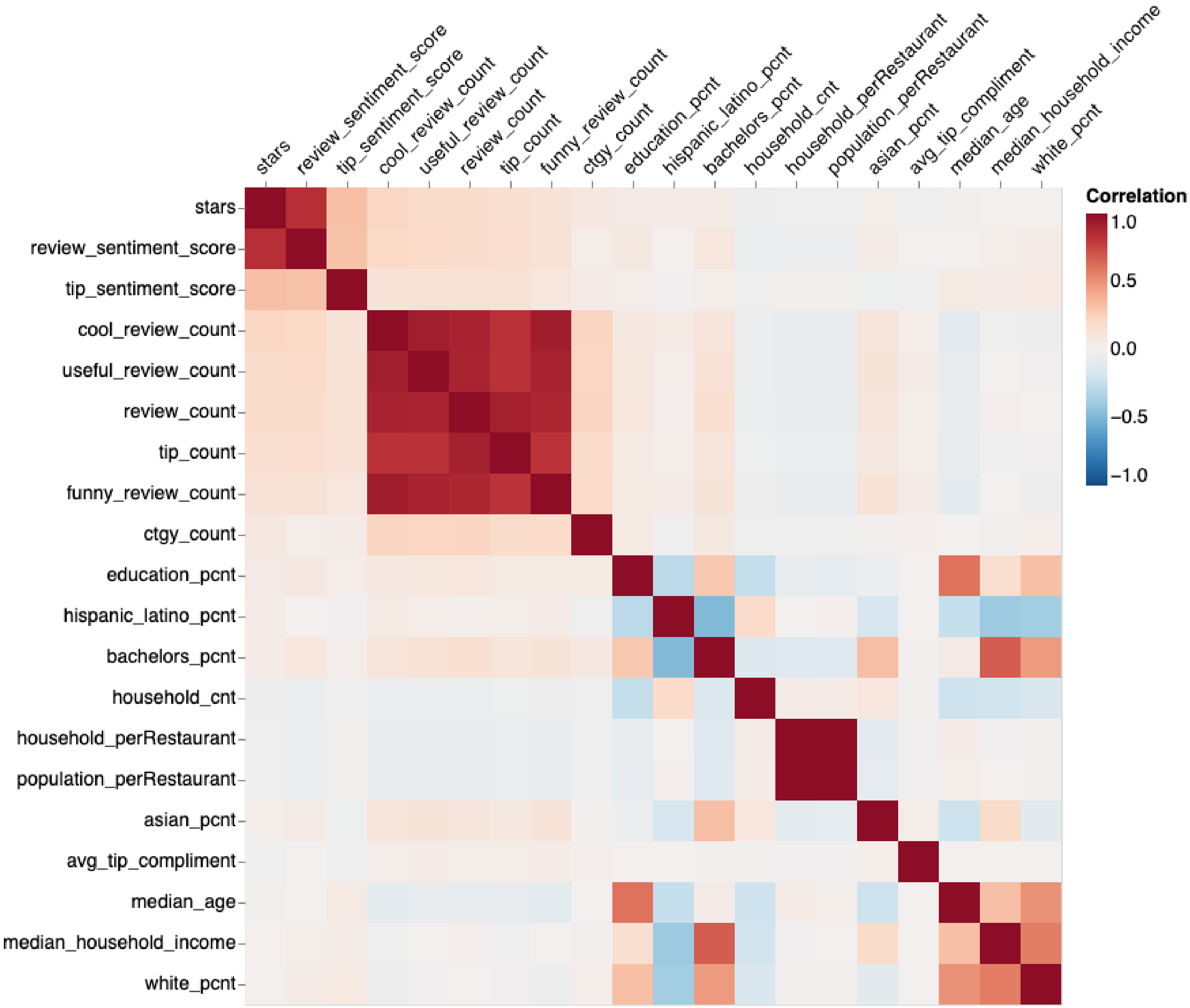
4.3.1. Identify Key Numerical Features

After examining the correlation matrix and heatmap (Figure 4), we observed strong correlations between restaurant ratings and review sentiment, tip sentiment, and related features. In contrast, demographic factors such as education, ethnicity, age, income, population density, etc., showed almost no correlation (correlation coefficient <0.05) with restaurant ratings.

Additionally, we found strong correlations between certain features. For example, the useful review count and cool review count displayed a high positive correlation, and similarly, the review count and cool review count also exhibited a strong positive correlation.

It is not surprising that restaurant ratings correlate strongly with review sentiment, tip sentiment, review count, and associated features. After all, the rating serves as a reflection of the overall customer sentiment and the restaurant's popularity. However, the lack of correlation with demographic factors was unexpected.

Figure 4



After aggregating data at the zip code level and revisiting the correlation matrix, we discovered stronger correlations between certain demographic features and restaurant ratings/reviews. This is particularly notable for the feature “percentage of White population”, as shown in Table 3 below. This could be attributed to the increased variability in demographic features when aggregated at the zip code level.

Table 3

restaurant level correlation coefficients

	stars	review_sentiment_score	review_count
review_count	0.18	0.18	1.00
white_pcnt	0.00	0.04	0.00
hispanic_latino_pcnt	0.04	0.00	0.02
median_household_income	0.00	0.03	0.01
median_age	0.00	0.00	-0.06
asian_pcnt	0.02	0.05	0.10

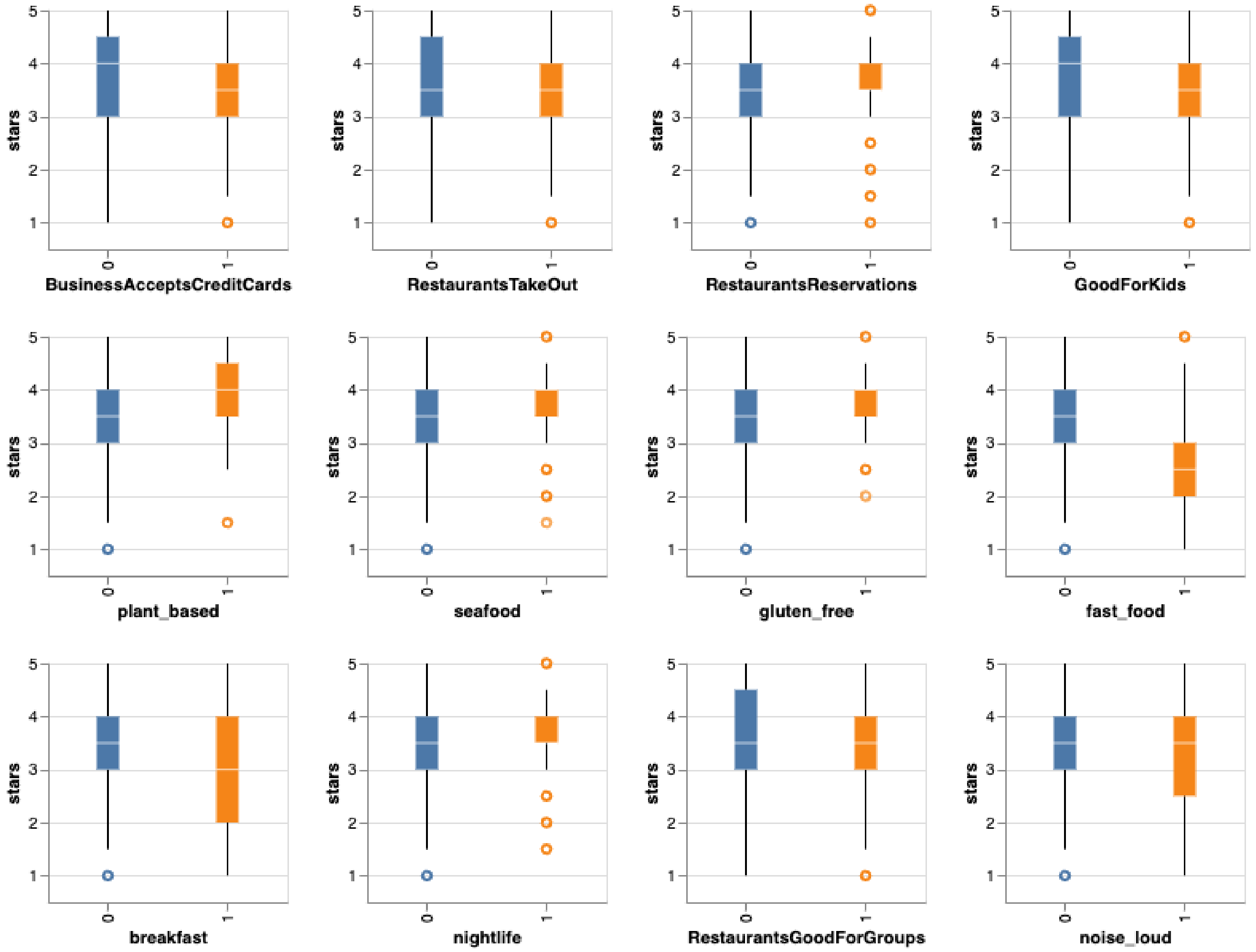
zip code level correlation coefficients

	stars	review_sentiment_score	review_count
review_count	0.23	0.30	1.00
white_pcnt	0.16	0.22	0.07
hispanic_latino_pcnt	0.09	0.02	0.05
median_household_income	0.08	0.15	0.11
median_age	0.07	0.12	-0.01
asian_pcnt	0.02	0.08	0.25

4.3.2. Identify Key Binary Categorical Features

By examining the differences in means and boxplots, we found 12 binary features, as shown in Figure 5, that significantly influence restaurant ratings. For instance, fast-food restaurants tend to receive lower ratings, while restaurants offering plant-based or seafood dishes tend to receive higher ratings.

Figure 5



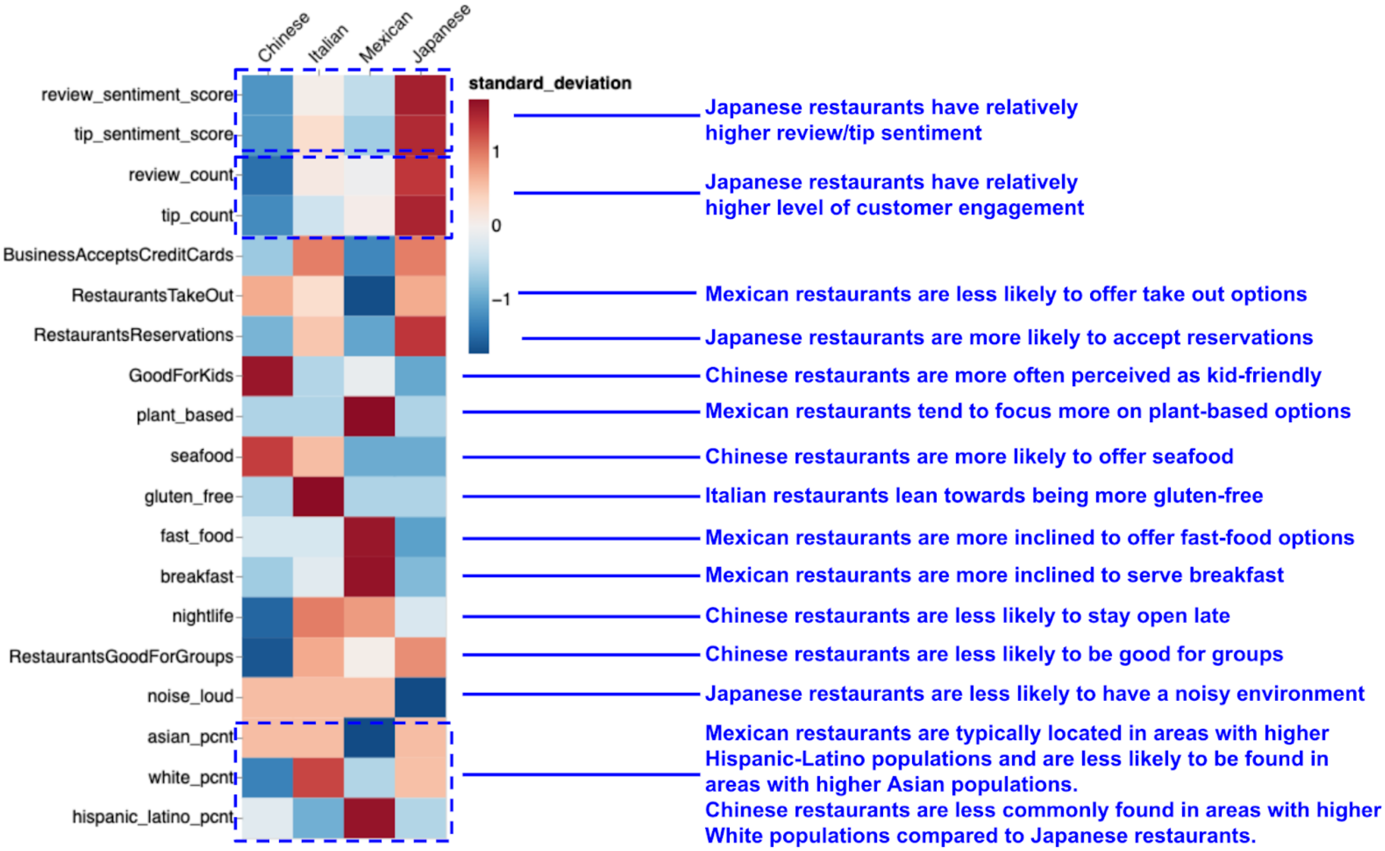
4.4. Analyze Disparities in Key Features Among the 4 Different Cuisines

We aggregated the key features identified in sections 4.3 by cuisine, standardized the data for each feature, and visualized the results in a heatmap, as depicted in Figure 6 below. This allows us to easily observe the discrepancies in key features across different cuisines. The greater difference in shading indicates more significant disparities.

For instance, Mexican cuisine tends to focus more on plant-based options, Italian cuisine leans towards being more gluten-free, Chinese cuisine is often perceived as kid-friendly and offers seafood dishes, and Japanese cuisine is less likely to have a noisy environment.

Additionally, we can observe that Mexican restaurants are typically located in areas with higher Hispanic-Latino populations and are less likely to be found in areas with higher Asian populations. On the other hand, both Chinese and Japanese restaurants are prevalent in communities with higher Asian populations, but Chinese restaurants are less commonly found in areas with higher White populations compared to Japanese restaurants.

Figure 6



Japanese restaurants have relatively higher review/tip sentiment

Japanese restaurants have relatively higher level of customer engagement

Mexican restaurants are less likely to offer take out options

Japanese restaurants are more likely to accept reservations

Chinese restaurants are more often perceived as kid-friendly

Mexican restaurants tend to focus more on plant-based options

Chinese restaurants are more likely to offer seafood

Italian restaurants lean towards being more gluten-free

Mexican restaurants are more inclined to offer fast-food options

Mexican restaurants are more inclined to serve breakfast

Chinese restaurants are less likely to stay open late

Chinese restaurants are less likely to be good for groups

Japanese restaurants are less likely to have a noisy environment

Mexican restaurants are typically located in areas with higher Hispanic-Latino populations and are less likely to be found in areas with higher Asian populations.
Chinese restaurants are less commonly found in areas with higher White populations compared to Japanese restaurants.

4.5. Principal Component Analysis

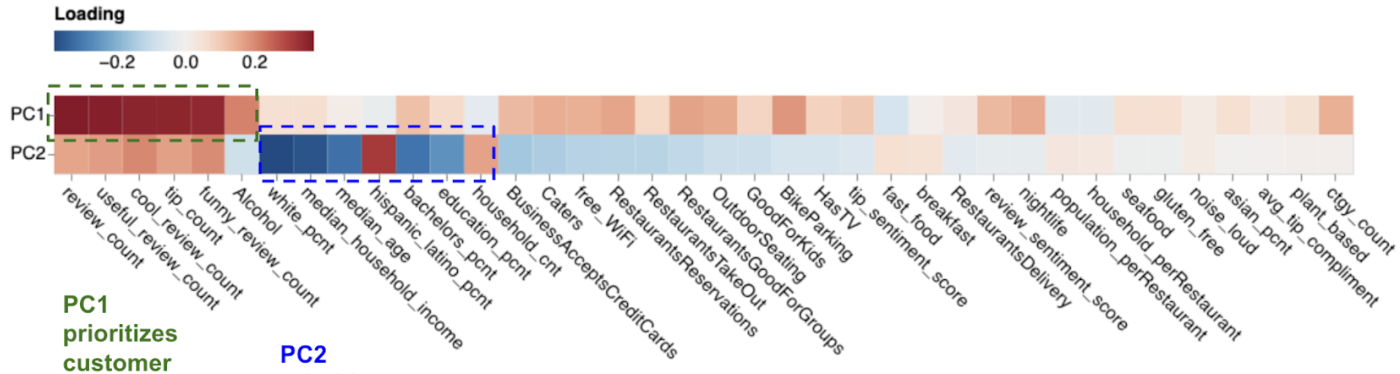
We conducted a Principal Component Analysis (PCA) on the key variables identified in section 4.3. This statistical technique enabled us to identify underlying patterns, reduce noise, and extract the most influential features that contribute to explaining restaurant ratings. By transforming the high-dimensional data into a lower-dimensional space, PCA helped us visualize the structure of the data and identify the principal components that explain the majority of the variance. This process facilitated a clearer understanding of the factors driving customer satisfaction and allowed us to focus on the most relevant features in our analysis.

PCA transforms the original variables into a new set of variables called principal components. Each principal component is a linear combination of the original variables, where the coefficients of this combination are known as loadings. These loadings represent the contributions of each original variable to the principal component.

The loadings are crucial because they indicate how strongly each original variable influences the principal component. High loadings (positive or negative) suggest a strong influence, while low loadings suggest a weak influence or no influence at all. By examining the loadings, we can interpret the structure and relationships within the data and understand which variables contribute the most to each principal component.

The heatmap presented in Figure 7 below illustrates the loadings for the first two principal components (PC1 and PC2). It is evident that PC1 emphasizes customer interactions, such as reviews and tips, while PC2 emphasizes demographic features.

Figure 7



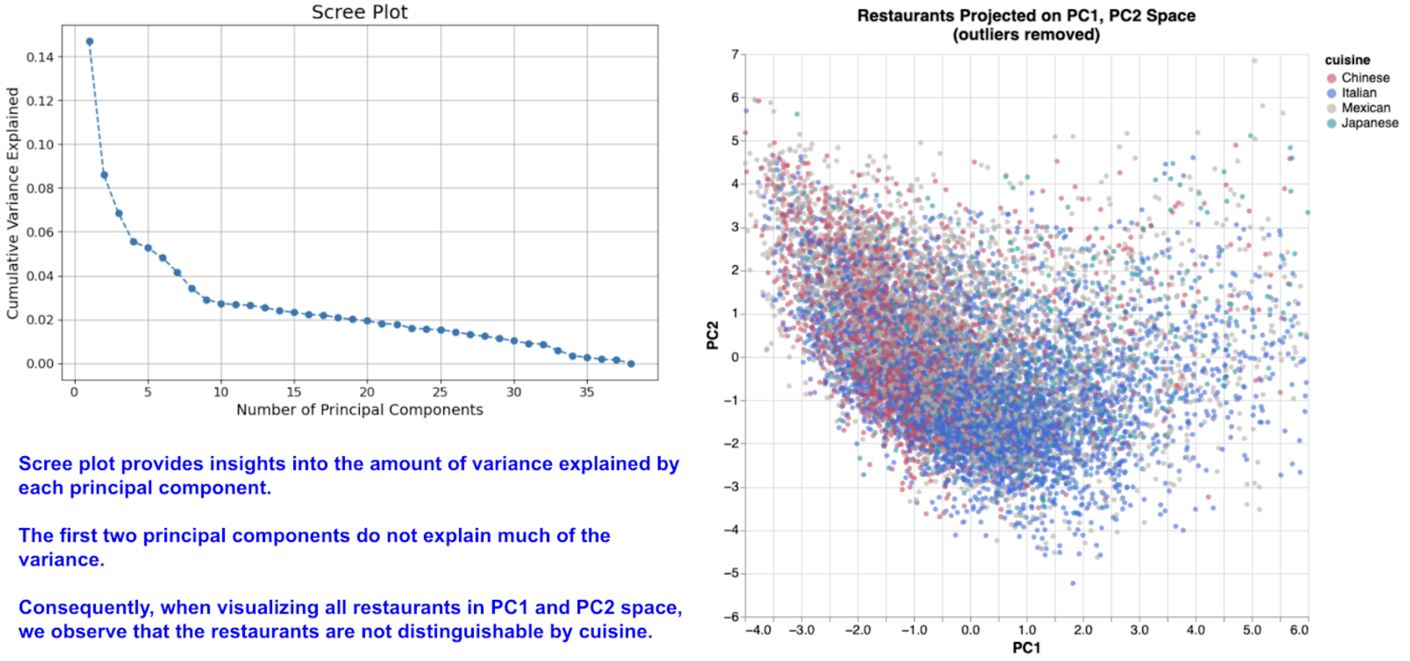
PC1 prioritizes customer interactions, such as reviews and tips.

PC2 prioritizes demographics features.

The Scree plot displays the eigenvalues of each principal component in PCA. Essentially, it provides insights into the amount of variance explained by each component and thus helps decide how many components to retain based on the amount of variance they explain.

By examining the Scree plot in Figure 8 below, it becomes apparent that the first two principal components do not explain much of the variance. Consequently, when visualizing all restaurants in PC1 and PC2 space as shown in Figure 8, we observe that the restaurants are not distinguishable by cuisine.

Figure 8



Scree plot provides insights into the amount of variance explained by each principal component.

The first two principal components do not explain much of the variance.

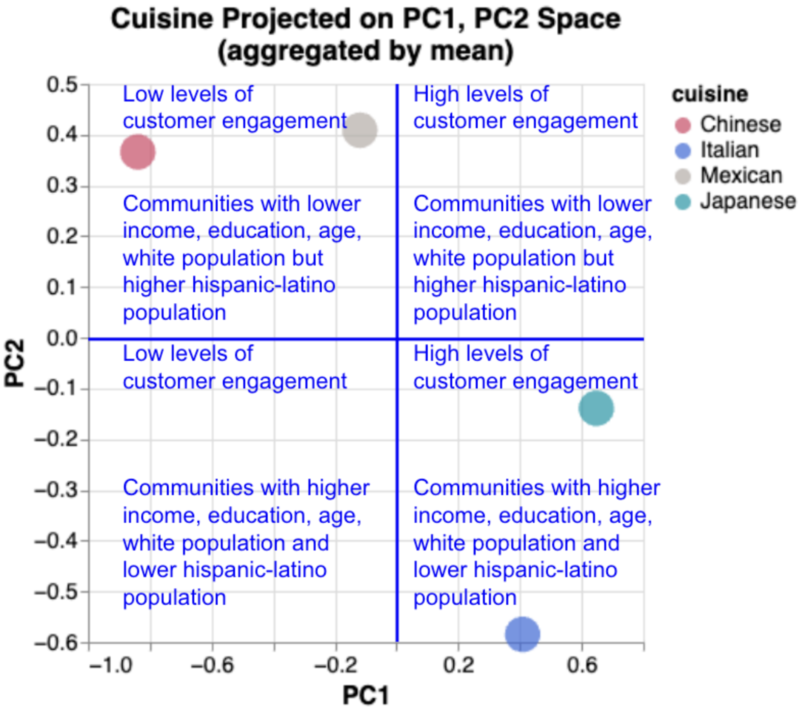
Consequently, when visualizing all restaurants in PC1 and PC2 space, we observe that the restaurants are not distinguishable by cuisine.

However, if we aggregate restaurants to the cuisine level by computing the average of all features, and then visualize the cuisines in PC1 and PC2 space while considering the loadings in Figure 7, we can better understand the main differences between cuisines.

As illustrated in Figure 9 below, positive PC2 values suggest that Chinese and Mexican cuisines are predominantly situated in neighborhoods characterized by lower income, education levels, average age, and a higher Hispanic-Latino population. Additionally, negative PC1 values indicate that both cuisines tend to display lower levels of customer engagement, with Chinese cuisine showing a more pronounced low level.

Conversely, negative PC2 values suggest that both Italian and Japanese cuisines are typically found in communities with higher income, education, average age, and a higher White population, but a lower Hispanic-Latino population, with Italian cuisine being more extreme in this regard. Additionally, positive PC1 values indicate that they both demonstrate higher levels of customer engagement, although Japanese cuisine tends to exhibit even higher levels.

Figure 9



4.6. Conclusions

By synthesizing findings from section 4.3. through section 4.5., we can gain deeper insights into the question posed in section 4.2 regarding why Japanese restaurants generally receive higher ratings. It is plausible that Japanese restaurants are more favored in White communities than Chinese and Mexican restaurants, where residents are more inclined to offer positive feedback.

However, when comparing Japanese with Italian cuisine, the picture becomes less clear. This discrepancy could potentially be influenced by unaccounted-for customer behaviors within the dataset, such as the tendency for individuals of Asian descent to contribute more reviews and assign higher ratings to Asian cuisine compared to Western cuisine. These nuanced insights underscore the intricate interplay of factors influencing restaurant ratings, highlighting the necessity for expanded data sources and more robust analytical approaches to fully elucidate these dynamics.

5. Statement of Work

Han Sun (Claire)	Bowie Liu
● Co-drafted proposal ● Data cleaning and manipulation ● Feature engineering ● Exploratory data analysis ● Comprehensive analysis and visualizations ● Drafted the ‘Motivation’ and ‘Analysis’ sections in the final report ● Revised the entire report	● Co-drafted proposal ● Data collection through API ● Exploratory data analysis ● Calculated correlations ● Drafted ‘Data Sources’, ‘Data Manipulation’ sections in the final report

6. References

6.1. Data Sources

https://www.yelp.com/dataset
https://www.census.gov/
https://simplemaps.com/data/us-zips

6.2. Articles

Tomić, Marina, et al. "Consumers' attitudes towards ethnic food consumption." *Journal of Central European Agriculture* 19.2 (2018): 349-367. https://hrcak.srce.hr/201705

Boch, Anna, Tomás Jiménez, and Katharina Roesler. "Mainstream flavor: Ethnic cuisine and assimilation in the United States." *Social Currents* 8.1 (2021): 64-85. https://journals.sagepub.com/doi/full/10.1177/2329496520948169

Powell, Lisa M., Roy Wada, and Shiriki K. Kumanyika. "Racial/ethnic and income disparities in child and adolescent exposure to food and beverage television ads across the US media markets." Health & place 29 (2014): 124-131. https://www.sciencedirect.com/science/article/abs/pii/S1353829214000926

Lim, Hoon, "Understanding American customer perceptions on Japanese food and services in the U.S"
(2010). UNLV Theses, Dissertations, Professional Papers, and Capstones. 654.
http://dx.doi.org/10.34917/1757765