Recommender Systems Part 1

Hongying Tao

April 23, 2024

1 Introduction

In this analysis, we explore a dataset of restaurant ratings and reviews to uncover patterns and insights that can inform recommendation systems. We employ a multi-faceted approach including: exploratory data analysis for data cleaning and visualization, popularity matching to identify high-performing restaurants, content-based filtering for personalized recommendations, and natural language analysis using Jaccard distance to optimize our suggestions. We're committed to providing clear, accurate analytics to support granular restaurant recommendations.

2 Exploratory Data Analysis

2.1 Data Import and Cleaning

In the preliminary data cleaning phase, the following steps were taken:

- **Duplicate Rows:** Identify 8 duplicate rows and delete them to avoid deviations in the recommendation system.
- **Missing Values:**
  - For missing data in `MaritalStatus`, `HasChildren?`, `PreferredModeofTransport` and `Vegetarian`, I made to label these entries as `Unknown`. For missing data in `ReviewText`, I made to label these entries as `No Review`. This approach was taken to maintain the integrity of the dataset, allowing the recommendation system to acknowledge and process instances where user information is not disclosed
  - `AverageAmountSpent` missing values were imputed with the most common spending level per restaurant category to reflect typical customer behavior.
  - `Numeric Variables:` Missing numeric data were imputed using the median value of the respective `Cuisine` category to maintain a robust distribution and mitigate outlier effects.

2.2 Data Visualization

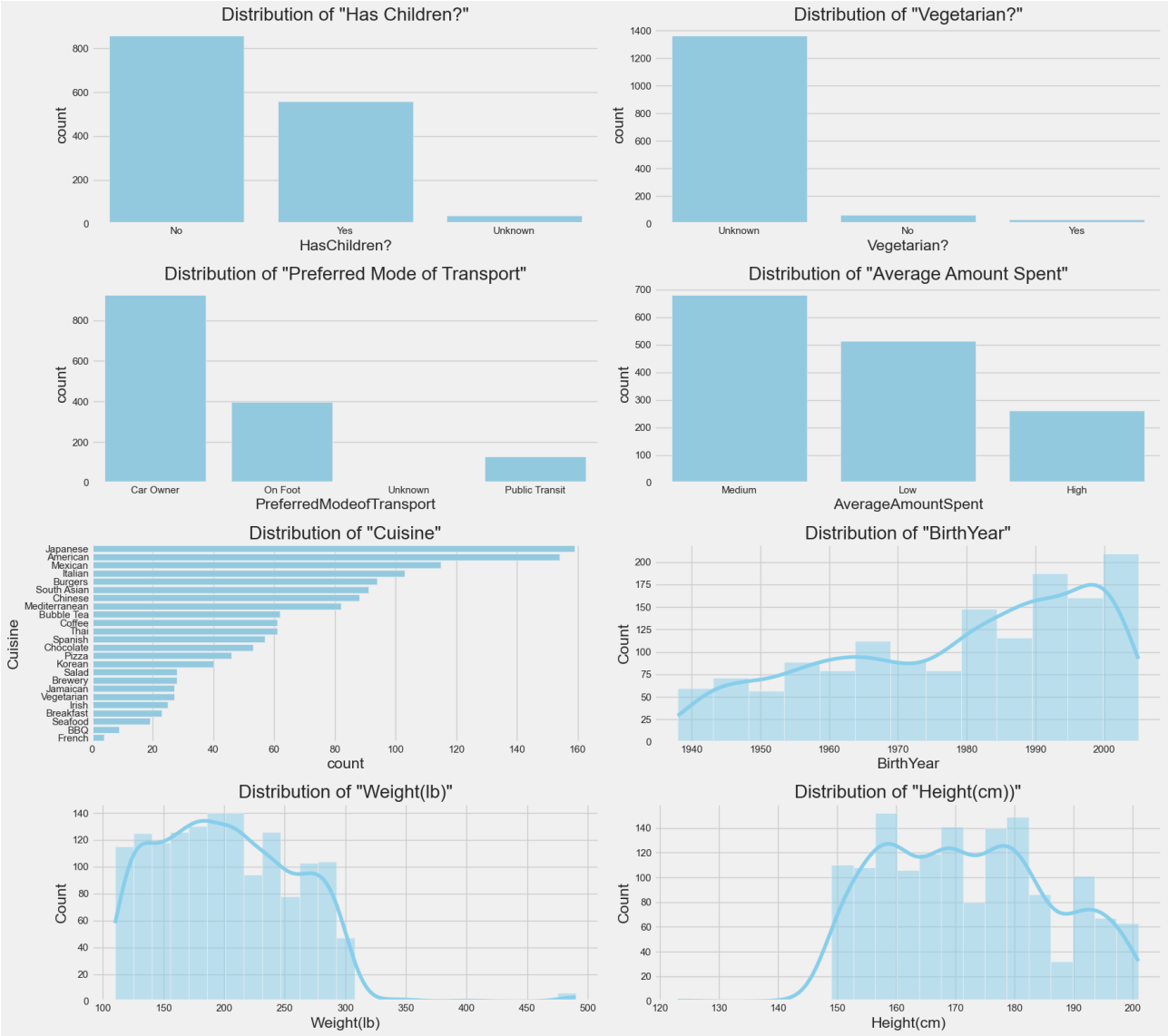

Figure 1: Distribution of Demographic Data

Analysis of demographic data shows that most reviewers prefer walking and generally spend a moderate amount on meals, suggesting that affordable restaurants within walking distance would be ideal for recommendations. Different birth years and a significantly younger population mean there is a trend for modern and socially vibrant dining venues. The popularity of Japanese food among various preferred cuisines demonstrates the demand for such options in the recommendation pool. Since many reviewers did not specify having children or vegetarian preferences, the system should prioritize versatility in the dining experience while still accommodating these considerations.

2.3 Clustering Analysis

Before performing clustering, I converted all the categorical demographic data to one-hot encoding.

2.3.1 K-Mean Clustering

- **Elbow Method:** I determined the optimal number of clusters by plotting the within-cluster sum of squares versus the number of clusters and determining the point at which SSE starts to plateau (the "elbow" point).

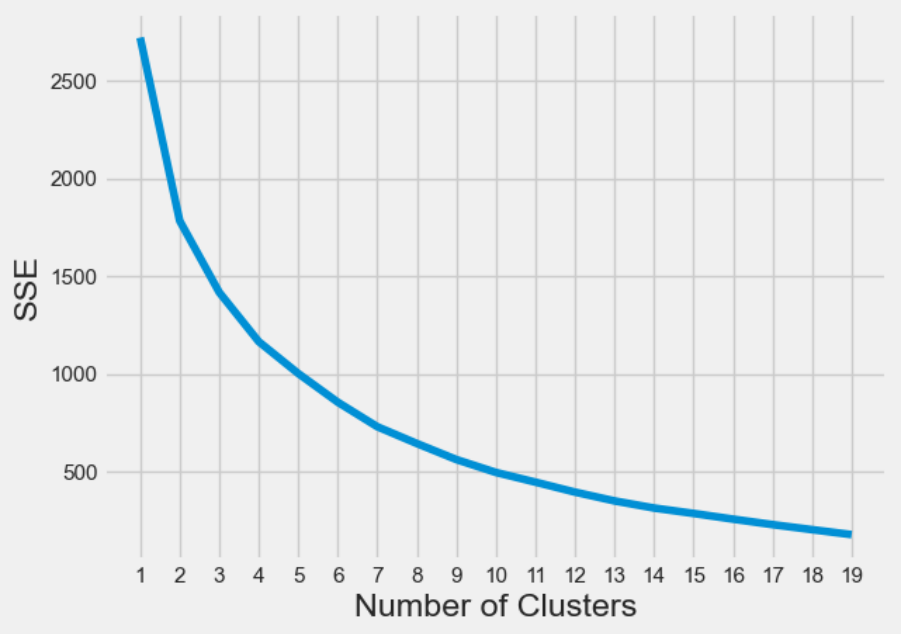

Figure 2: Elbow Plot

As shown in Figure 2, the line begins to level off at $6-8$, and I chose to naturally divide the data set into 6 different clusters.

- **Silhouette Score:** I chose to use silhouette scores to examine how similar an object is to its own cluster compared to other clusters, which calculated to be 0.56 for K-Means clustering. This moderate score indicates that the clusters are reasonably defined and separated from each other.

- **Features of Each Cluster:**

Cluster	Avg Rating	Review Count	Marital Status	Has Children?	Vegetarian?	Transport	NU Student?
0	3.92	233	Married	No	Unknown	Car Owner	No
1	3.86	224	Single	No	Unknown	On Foot	No
2	3.54	448	Married	Yes	Unknown	Car Owner	No
3	3.84	277	Single	No	Unknown	Car Owner	No
4	3.48	142	Married	Yes	Unknown	On Foot	No
5	3.90	132	Single	No	Unknown	On Foot	Yes

Table 1: Cluster Characteristics based on K-Means Analysis

From Table 1, after K-mean cluster analysis, most of the "`Vegetarian?`" in each cluster are "`Unknown`", which means that this variable is not a strong feature to distinguish different customer groups. Additionally, the average ratings for each group were close to medium and not significantly different.

3 Popularity Matching

3.1 Restaurant Ratings and Reviews

In the popularity matching analysis, several key metrics were evaluated to understand the landscape of restaurant reviews.

- The overall **average rating** across all restaurants is approximately **3.74**.
- The **highest average rating** were *Evanston Games & Cafe*, *Fonda Cantina1* , *La Primary* and *Letour*, out of **5.0**.
- The **median number of reviews** per restaurant stands at **22.5**, which provides a measure of central tendency that is less affected by outliers and extreme values than the mean.
- *Campagnola* stands out as the **restaurant with the most reviews**, totaling **48**.

3.2 Sample Recommendation Engine

The recommendation system recommends the top three restaurants based on the type of cuisine specified by the user based on the restaurant's average rating and the number of reviews received.

Table 2: Top 3 Restaurant Recommendations by Cuisine

Type	RestaurantName	Average_Rating	Count_Review
Chinese	Joy Yee Noodle	4.290323	31
Chinese	Peppercorns Kitchen	3.545455	33
Chinese	Lao Sze Chuan	3.291667	24
Mexican	Fonda Cantina	5.000000	6
Mexican	La Principal	5.000000	1
Mexican	Zentli	4.764706	17
Spanish	Tapas Barcelona	4.206897	29
Spanish	5411 Empanadas	3.750000	28
Coffee	Evanston Games & Cafe	5.000000	1
Coffee	Philz Coffee	4.600000	15
Coffee	Brothers K Coffeehouse	4.533333	15

Table 4 recommends the most popular restaurants based on four specific categories: Chinese, Mexican, Spanish and coffee.

3.3 Shrinkage Recommendation Engine

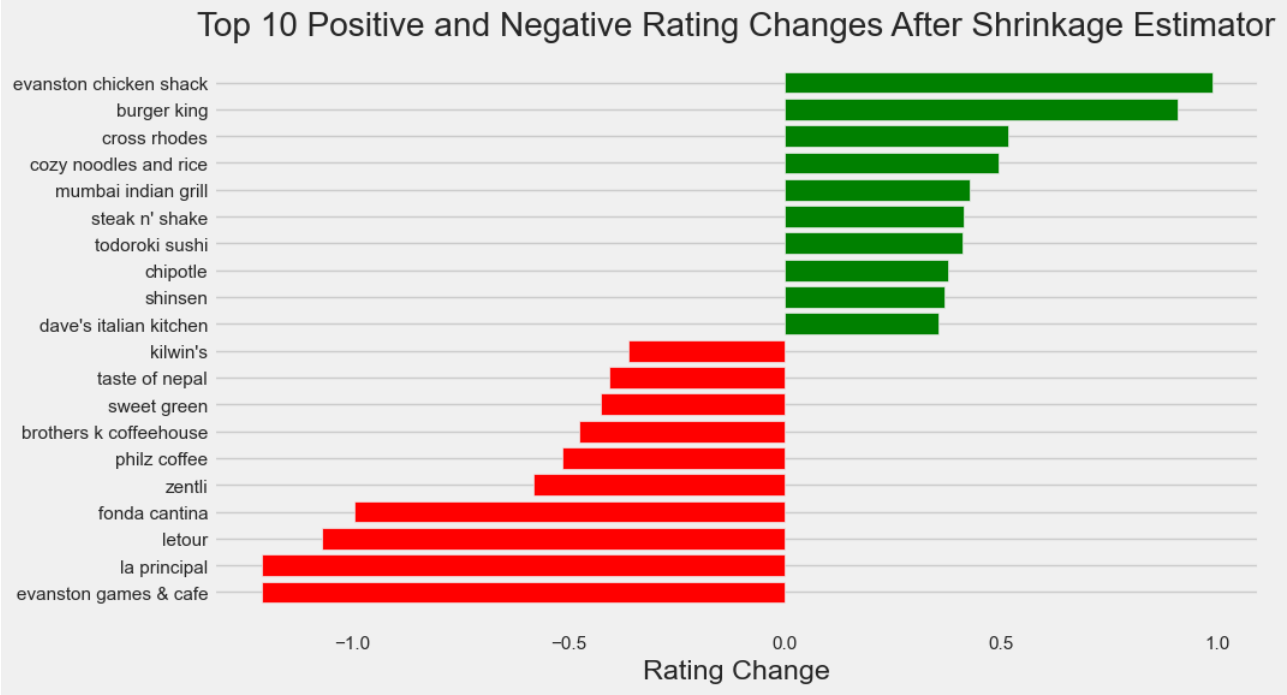

Figure 3: Top 10 Positive and Negative Rating Changes After Shrinkage Estimator

The shrinkage estimator combines a restaurant's average rating with the overall average rating across all restaurants, weighted by the number of reviews for each restaurant relative to the average number of reviews. The purpose of this method is to stabilize the ratings of restaurants with fewer reviews by "shrinking" their average ratings down to an overall average. If a restaurant has few reviews, its average rating will be more volatile and potentially less reliable. By using a shrinkage estimator, its adjusted ratings are pulled toward the global mean, thereby reducing the impact of possible outlier reviews. Likewise for restaurants with many reviews, the impact of the shrinkage estimator will be smaller because their own average ratings have greater weight.
Thus, after using a shrinkage estimator, the restaurants that benefited most were *Evanston Chicken Shack* and *Burger King*, and the restaurants that were most negatively affected were *Evanston Games & Cafe* and *La Principal*.

4 Content-Based Filtering

4.1 Distance Measures

- If the Euclidean distance of two restaurants is short and the cosine similarity is high, they may be similar in multiple dimensions and have eigenvalues that are close in size and direction. If the Euclidean distance is short but the cosine similarity is not very high, it may mean that they are similar in some features but different in other features. If the Euclidean distance is long and the cosine similarity is high, they may be similar in the direction of the features but different in the absolute magnitude of the feature values. If both are low, it may mean that the restaurants are not similar along multiple dimensions.

- For "Peppercorns Kitchen" and "Epic Burger", the Euclidean distance is 1.436, and their cosine similarity is 0.5284. The large Euclidean distance between them indicates that they have large differences in the feature space, while the moderate cosine similarity implies that these restaurants may be similar in some dimensions, although they differ in the size of the features.

- For "Peppercorns Kitchen" and "Lao Sze Chuan", the Euclidean distance is 0.315, and their cosine similarity is 0.9808. The small Euclidean distance and very high cosine similarity between them show that the two restaurants are very similar in multiple dimensions, not only in the absolute magnitude of the feature values, but also in the direction of the feature combinations. This suggests that they may be targeting very similar market segments and customer groups.

4.2 Recommendation System Based on Euclidean Distance

Ranking	**RestaurantName**	**Euclidean_Distances**
1	Clarkes Off campus	0.930041
2	Hecky's BBQ	1.235294
3	Edzo's Burger Shop	1.490472
4	Pâtisserie Coralie	1.502059
5	Philz Coffee	1.578572
6	Evanston Chicken Shack	1.641670
7	Le Peep	1.656719
8	Fridas	1.664978
9	Prairie Moon	1.695856
10	Mumbai Indian Grill	2.009934

Table 3: Willie Jacobsen's Favorite Top 10 Most Similar Restaurants

4.3 Compare Euclidean Distance and Cosine Distance

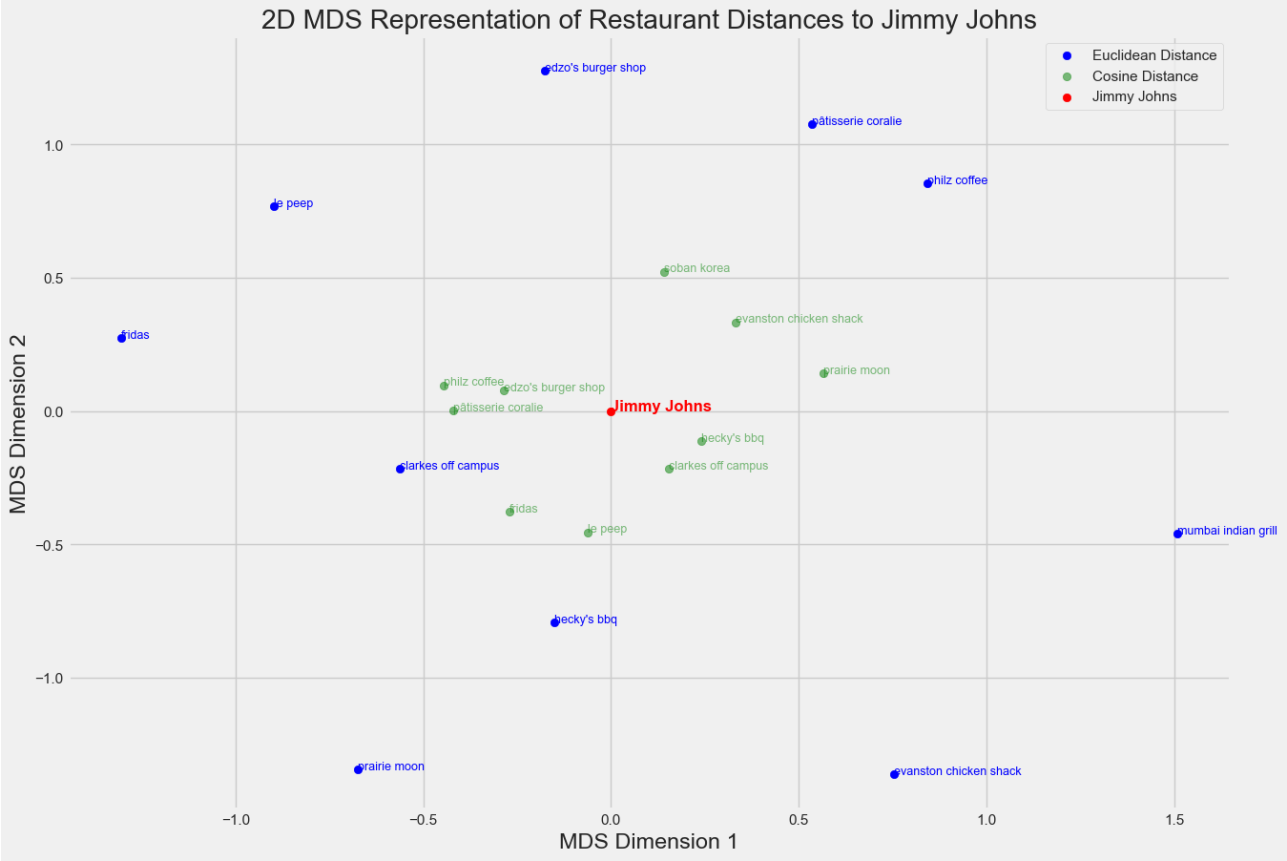

Figure 4: Willie Jacobsen's Top 10 Restaurant Recommendations

I used euclidean distances and cosine distances to make 10 restaurant recommendations based on Willie Jacobsen's favorite restaurant, Jimmy Johns. In order to compare which method recommended the restaurant that is most similar to Jimmy Johns, I used Multidimensional Scaling to draw these restaurants in the second place. Dimension up to the Jimmy Johns view. It is easy to see from the image that the restaurants recommended by the cosine distances method seem to be more similar to Jimmy Johns and may be preferred by Willie Jacobsen.

5 Natural Language Analysis – Version A

Table 4: Jaccard Distance Between Restaurant Pairs

Restaurant Pair	Jaccard Similarity
Burger King - Edzo's Burger Shop	0.190476
Burger King - Oceanique	0.080000
Lao Sze Chuan - Kabul House	0.187500

I generated a column in the original data frame called Augmented Description, which appends the dish type to the end of each restaurant's short description. This may be done to enhance the descriptive information available for each restaurant.

In the table above, Burger King and Oceanique are the least similar, and Burger King and Edzo's Burger Shop are the most similar.

5.1 Recommendation System Based on Jaccard Distance

Recommendations for Calvin Smith

Restaurant Name	Jaccard Similarity
Alcove	0.230769
Shangri-La Evanston	0.208333
Oceanique	0.185185
Union Pizzeria	0.173913
5411 Empanadas	0.160000

Recommendations for Solomon M

Restaurant Name	Jaccard Similarity
Tealicious	0.190476
Evanston Games & Cafe	0.187500
Tomo Japanese Street Food	0.181818
Fonda Cantina	0.176471
Kuni's Japanese Restaurant	0.160000

6 Conclusion

By combining different analytical methods, this report presents a multi-angle view to understand and predict customer preferences. Exploratory data analysis reveals the underlying characteristics of the dataset, while popularity matching and content-based filtering techniques enable us to provide users with customized restaurant recommendations. The application of natural language processing further improves the accuracy of recommendations and provides valuable insights for future catering business decisions and customer experience optimization.