Team Members: Aryan Jain, Mallory Sico

Milestone 2 Report

In this milestone, we draw conclusions about the best ways to determine which serving type, price range, cuisine type and location is best to open a new restaurant. The methods for determining this include confounder identification, hypothesis testing, distribution analysis and correlation analysis. All figures will be included in the appendix for reference. The success metric for being a "successful restaurant rating" is defined as the 90th percentile of ratings which is 4.2.

Confounder Analysis

Methodology

Looking at serving types of restaurants by count as shown in Figure 1, we see that Delivery and Dine-Out dominate as these are more broad categories when compared to others. Among the more specific categories, we see that Bangalore has a considerable amount of "Dessert" restaurants. It is important to note that a restaurant listed as "Cafe" would also fall under "Dine-Out" (and possibly "Delivery"). Additionally, we also have restaurants that are more generic that would just be listed as "Dine-Out". Figure 2 shows the ratings distributions of the serving types. Some important observations we can make from Figure 2 include the following:

1. Drinks&Nightlife and Pubs&Bars and Buffets are generally rated higher than other restaurant types.
2. Drinks&Nightlife and Pubs&Bars have very similar distributions which could be because both names are used almost interchangeably in most scenarios
3. Even though Desserts are a common restaurant type they do not fare as well with the population.

Between the 3 top rated restaurant types (Drinks&Nightlife and Pubs&Bars and Buffets), we would like to choose one. For this we look at any potential confounders that could be influencing the higher ratings and attempt to account for them. Since we assume that Pubs&Bars and Drinks&Nightlife refer to the same restaurants we will compare Buffets and Drinks&Nightlife using potential outcomes. The confounder that we want to account for is approx cost for 2.

We will have these 3 random variables for our analysis:

r: ratings	t: restaurant type (t=1: Buffets, t=0: Drinks&Nightlife)	c: cost (c=1: cost is $$$$, c=0: cost is lower than that)

Results

Looking at the unadjusted ATE, we calculate $\mu_{r|t}(1) - \mu_{r|t}(0)$ and find that the observed unadjusted ATE is -0.01. This tells us that Drinks&Nightlife restaurants are slightly more highly rated than Buffet restaurants. We will now factor in the cost parameter, to see if the higher rating is due to Drinks&Nightlife restaurants being more posh.

Adjusted ATE = $E[po_1] - E[po_0] = E[\mu_{po1|c}(c)] - E[\mu_{po0|c}(c)]$. The Adjusted ATE is 0.22. After adjusting for the confounder (cost for two), we see that buffets would be a better rated restaurant over all other categories. If we are setting up an expensive restaurant, we would go with Drinks & Nightlife. If going for a lower budget option, choose Buffet. This also shows us that our dataset had evidence of Simpson's Paradox, which has now been accounted for using potential outcomes.

Analysis

We found that the top 5 restaurants in Bangalore (in terms of ratings), were in the '$$$$' price category. This brings up the question of whether to consider a lower end price range.

Figure 3 shows the ratings distribution by price type. We see that getting to a very high rating is much harder for lower budget restaurants. This however does not mean that those restaurants would not be successful, but in our framework where rating is a proxy for success, it would not be in our interests to go with them. One could choose to open a restaurant with price type $$$ and have it be very successful (25% restaurants in that category are rated higher than the 90 percentile rating). Another important inference is that not all $$$$ priced restaurants are highly rated. Only 50% of those restaurants lie above the 90 percentile mark. This warrants the need for this project even more to try and find ideal configurations. With this information, we can refine our result to say that if you are opening a restaurant in the $$$$ price range, choose Drinks & Nightlife. If you are opening a restaurant in the $$$ price range, choose Buffets.

Cuisine Hunt

Methodology

Let us say that one does not want to open a restaurant with either Buffet or Drinks&Nightlife types. You seem to think that Buffets are not profitable or Drinks & Nightlife does not align with your philosophy of a restaurant, then what? We highlighted the fact before that the 5 serving types were just niches and a whole array of restaurants fell under 'Dining' and consequently 'Delivery'. Looking at both these aspects, we now look forward to answering the cuisine question. This along with knowledge about the location will help us appeal to a broader audience looking to open restaurants in Bangalore under a more generic Dining umbrella.

Figure 4 shows the ratings distribution of the top 20 cuisines in Bangalore. Before making a conclusion and analyzing the best rated 2-3 cuisines, we would first like to look at the distribution of cuisines across the 5 most popular locations in Bangalore (by count of restaurants). Those locations are BTM, Koramangala 5th Block, HSR, Indiranagar, and JP Nagar.

When grouping by cuisine category and location, the count of restaurants in each group severely decreases. Therefore, we are looking at only the top 10 to keep a significance level. Figures 5 - 9 show the ratings distributions and the counts of the 10 most popular cuisines in each of the 5 locations. Analyzing these distributions, we can draw the following inferences.

Results

For BTM, the best choice is Continental Cuisine. The top 2 highest rated restaurants in the area are North Indian. However, North Indian overall is not rated too special. Continental restaurants seem to be our best answer here as they are rated the best and also have much fewer restaurants than expected, so we can see the market is not saturated.

For Kormangla 5th Block, the best choice is North Indian Cuisine. We could argue that on the basis of ratings, there are multiple cuisines that do well: Continental, Cafes, Desserts, Italian and Indian. However, when we look at the Actual vs Expected graph, we see that 4 of the cuisines are extremely saturated and competitive. North Indian restaurants would be a good bet here (or Chinese, but this could be restaurant dependent as it has a high IQR).

For HSR, the best choice is Cafe. The 4 most popular cuisines do not perform great here, so it does not make sense to go with those. On the basis of having low variability, Cafes and Italian restaurants would be a good choice. However, similarly with BTM, Italian Restaurants have low variability, but also low number of restaurants.

For Indiranagar, the best choice is to explore outside the top 10 cuisines. Even though people here tend to vote generously, none of the top 10 cuisines stand out to the eye. This would also indicate that there are cuisines outside of this bracket that are doing very well in this location. We also see that the median rating has the slightest fall (when compared to other locations) when we include the top 20 cuisines, implying their likeliness of other cuisines. This also follows through knowledge of the area as it has a very big "young population".

For JP Nagar, the best choice is Chinese cuisine. No cuisine is a stand out performer, irrespective of high ratings. Even though North-Indian has the highest IQR, it still could give the best results if done right. However, Chinese would be a better pick as most of their parameters are the same, but Chinese restaurants have a higher median.

Overall, the best choice among each of these options would be a Continental Restaurant in BTM due to the high overall ratings and low competition. The worst choice (among the top 5 areas) would be JP Nagar due to its saturated market.

Analysis

We looked at restaurants opened after this dataset was generated that suit our "ideal restaurant criterias" to find some interesting insights. We first look under the Continental Restaurants opened in BTM relatively recently. We find the ratings of these newly opened restaurants **(K-OS Game Bar: 4.5, Fogg Lounge: 4.4, Mudpipe Cafe: 4.2)**. Even though the ratings of the first 2 would depend on other factors (High Price for 2, New concept (i.e. K-OS is a darts bar)), we see a lower price restaurant (Mudpipe Cafe) did succeed as well.

Very few exclusively North Indian restaurants were opened in recent times in Kormangla in the $$ and $$$ price range. We were able to find 2 **(Kathpals: 4.3, Jalandhar Street: 4.0)**. We would say that this was a missed opportunity for potential investors.

We also look at recent Cafes opened up in HSR which are in the $$ and $$$ price range. On a general view, all of them seem to be rated above 3.9. Additionally we have a good chunk rated higher **(Hustle: 4.9, Burger Yard: 4.8, Cuppa Redefined: 4.5)** which shows that our analysis was on the right track. In fact it was a very big hit here as 2 of the city's highest rated restaurants are now cafes in HSR.

Since, we also looked at Buffets being a successful category, we look at recently opened buffets in the $$, $$$ section **(Kitchen De Buffet: 4.7, Maxims Buffet: 4.2, One Atria Cafe: 4.3, Zodiac: 4.0, The Buffet Table: 3.5)**. A lot more variability here with what most people would call a "hit and miss", signifying that there are other factors that need to be taken in consideration as well. This looks at Buffets in all locations which could account for variance.

Hypothesis Testing Time (Other Factors)

Methodology

We also have access to the phone numbers of the restaurant. Numbers starting with +91 (Country code of India) belong to personal mobile phones whereas numbers starting with 080 (Area code of Bangalore) belong to landlines. We hypothesize that restaurants that have personal phones listed have less organizational structure in place and would therefore suggest that they belong to a lower price range. As we know there is a correlation between rating and price range, it would suffice to say that they would be rated lower.

Designing the Hypothesis Test: Test Statistic (t) = median(personal_phone) - median(landline) | H_0 : t = 0 | H_1 : t > 0

Results and Analysis

Since we have flags (1 represents personal phone and 0 represents landline), we can set up a permutation test. Due to the number of values, it would not be computationally possible to do this for every permutation. Therefore, we apply a monte-carlo approach. The p value of this test is 0.0. Due to the low p-value, we can reject the null hypothesis. This would imply that restaurants with landline numbers listed tend to be rated higher than those with personal phone numbers listed. It could be argued that the phone number listed does not have any implicit impact on the ratings (irrespective of the statistically significant value), but we would use this knowledge to recommend a new restaurant have a good organizational structure and ideally a landline number listed.

Analyzing Correlation

Methodology

Figure 10 shows the correlation matrix between the serving types and ratings. Most insights offered by this correlation matrix have been found and discussed earlier. The one worth mentioning is the high correlation between Drinks&Nightlife vs Rating when compared to Buffets vs Rating. This goes to show how our earlier suggestion holds. From the correlation matrix, Figure 11, which shows the correlation matrix between ratings, votes, costs and cuisine count, we gain 2 insights.

Results

The first insight is that there is a high correlation between ratings and votes. The number of votes depends on the success of the restaurant as well as how long the restaurant has been running. However, this tells us that when starting a restaurant, we should incentivize voting. This could be in the form of schemes like "If you rate us on Zomato, you get a free dessert".

The second insight is that the number of cuisines offered is positively correlated with the rating. This was unexpected as offering fewer cuisines would make the restaurant more specialized which we would expect to be a good thing as opposed to restaurants that try to do too much and may generally be worse. To determine the confidence of this inference, the confidence interval for the correlation coefficient was calculated using a bootstrap analysis with Fisher’s transformation. Using a bootstrap sample size equal to the original dataset, we find that there is an extremely tight confidence interval over 1000 batches. The 95% confidence interval for the correlation coefficient is [0.2067 - 3.33E-16, 0.2067 + 3.33E-16]. Given that we have a sample size large enough to obtain such a small confidence interval, we can conclude that the correlation is reliable.

Analysis

The confidence interval implies that the correlation is statistically significant. Does that tell us that the more cuisines you offer, the better your rating? **NO**. We see that the degree of freedom here will be very high (n-2), so even a very slight change in correlation values would appear statistically significant. Additionally, we cannot conclude that this correlation implies causation of better ratings.

Next, we explored the reason why cuisine count has a positive correlation. We see that the median number of cuisines offered over all restaurants is 2. However, the median cuisines offered by Buffets and Drinks&Nightlife is 3. Since these are the best rated categories they effectively contribute to telling a story that cuisine counts and rates are positively correlated.

Milestone 3 Plan: How do we know if our ideal configurations are indeed ideal?

Through our analysis in the previous sections, we offered restaurant configuration options. We take an ML approach to build a rating predictor regression model using ratings as our success metric. We will use the current dataset as training and testing data to choose the right ML model and do the required hyperparameter tuning. We will then test our ideal restaurant configurations using the model to predict their ratings if they existed.

To set a baseline for the model performance, knowing that there is strong correlation between approx_cost and ratings, we built a simple linear regression model using approx_cost as the input variable. The accuracy of this model is the baseline accuracy moving forward. For this baseline model, the mean absolute error is 0.33.

As we had highlighted before, approx_cost is certainly an important parameter, but not all that is important. We saw that 50% of $$$$ restaurants don't outperform the 90 percentile mark as seen in Figure 3. In fact, 25% of those restaurants lie below the median rating in Bangalore. This tells us that there are certainly other factors that make the ideal restaurant. So this brings the question, what are these factors (other than those we already described)?

To answer this question, we will look at the reviews in the next phase to identify things that matter the most to a consumer in Bangalore (especially in the realms of our ideal restaurant). This could highlight things like service and taste to show us what we should invest in the most.

Team Members: Aryan Jain, Mallory Sico

Appendix

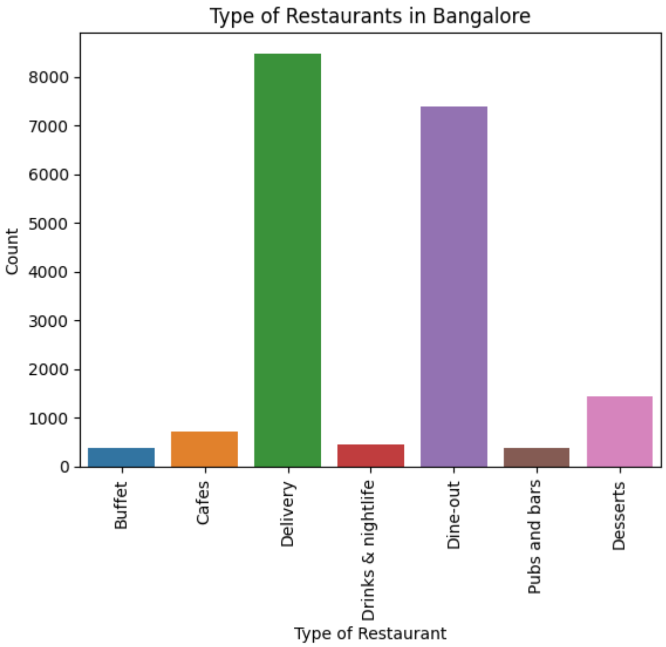

Figure 1: Serving Types of Restaurants in Bangalore

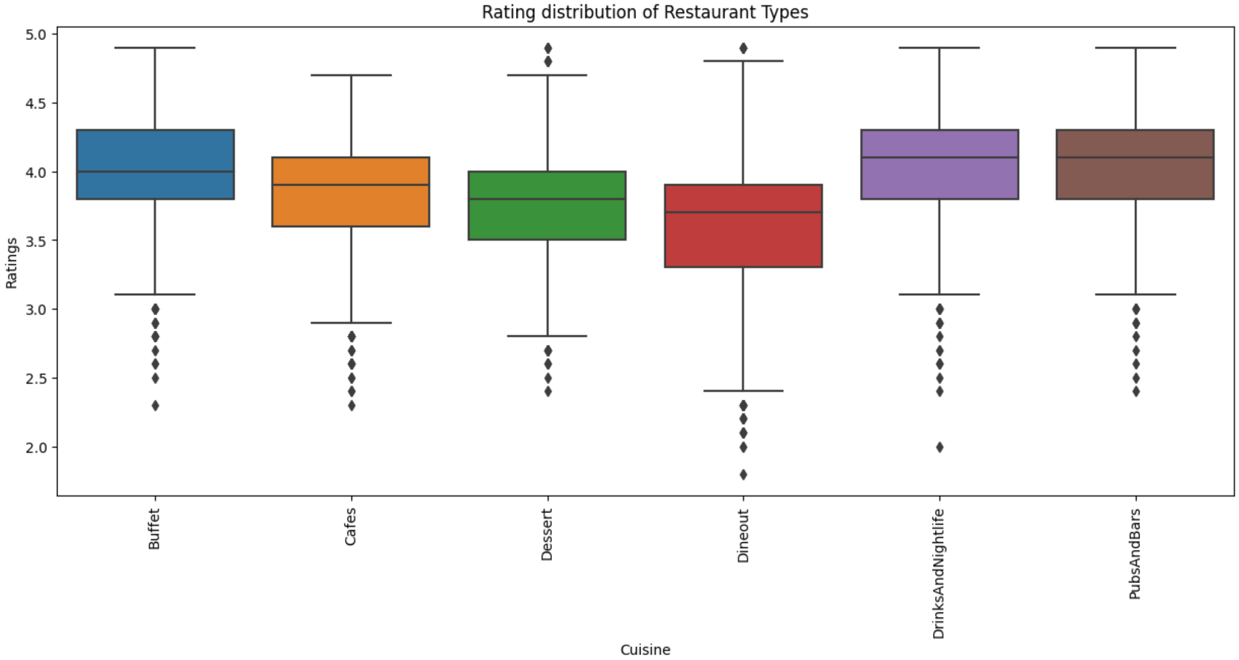

Figure 2: Ratings Distribution by Serving Types

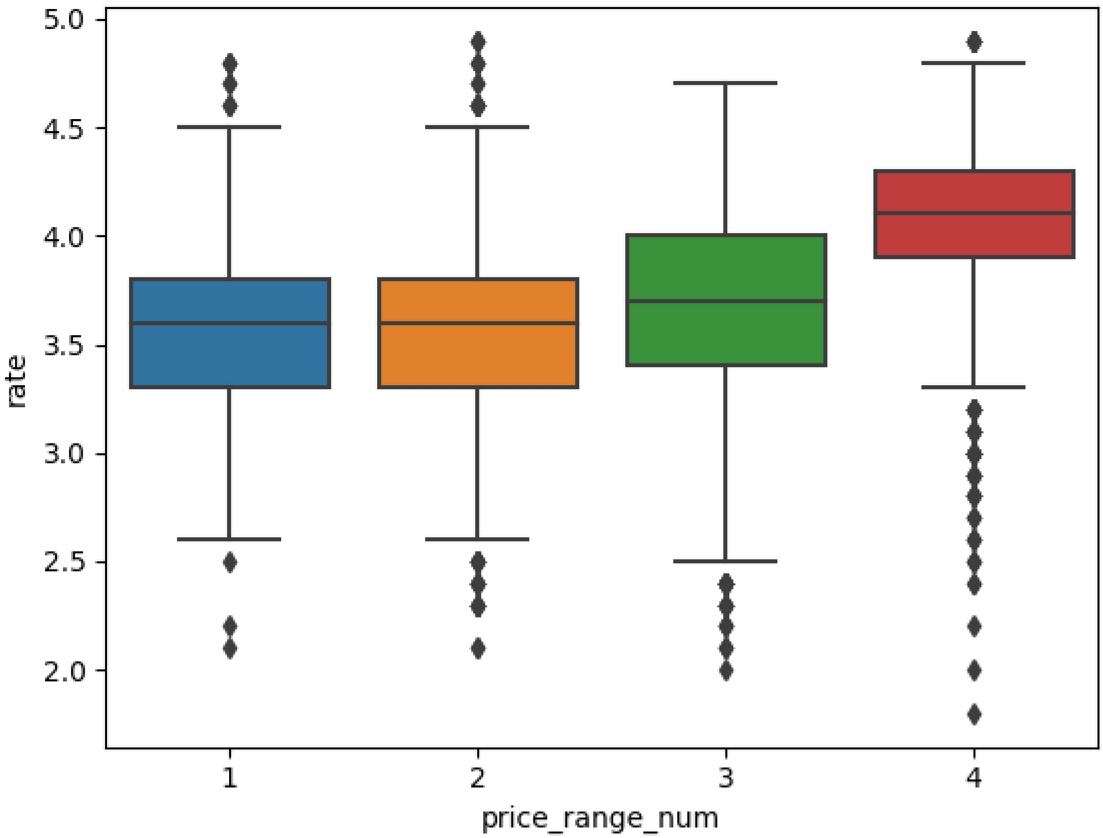

Figure 3: Ratings Distribution by Price Type

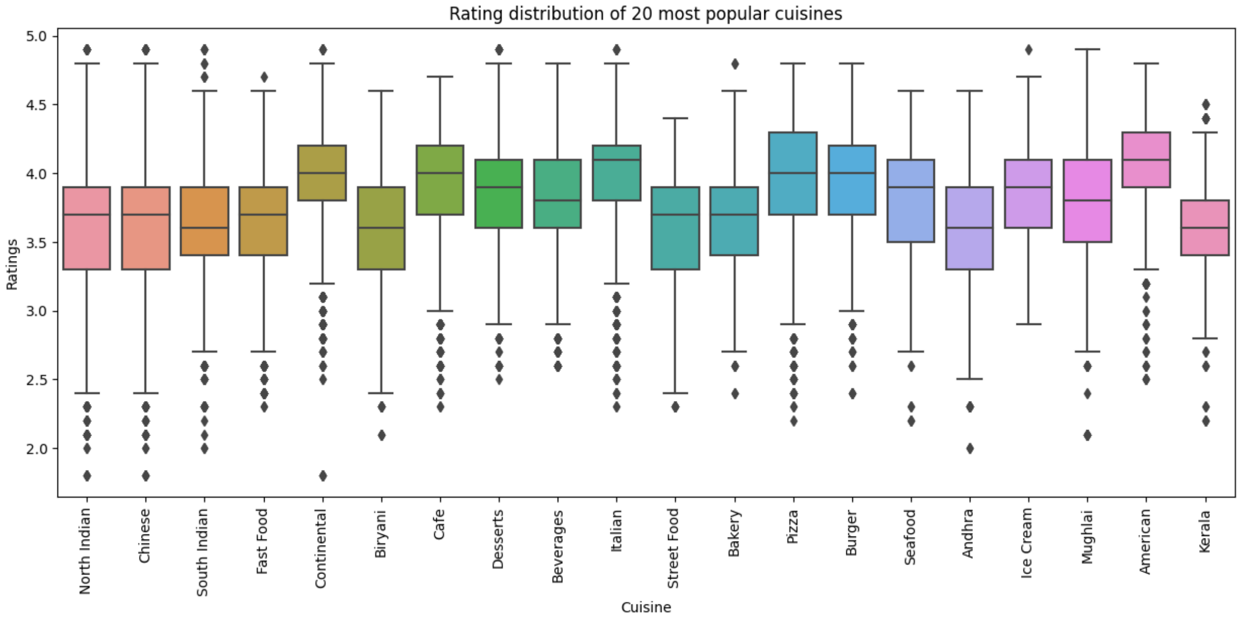

Figure 4: Ratings Distribution by Cuisine Type (Top 20)

Team Members: Aryan Jain, Mallory Sico

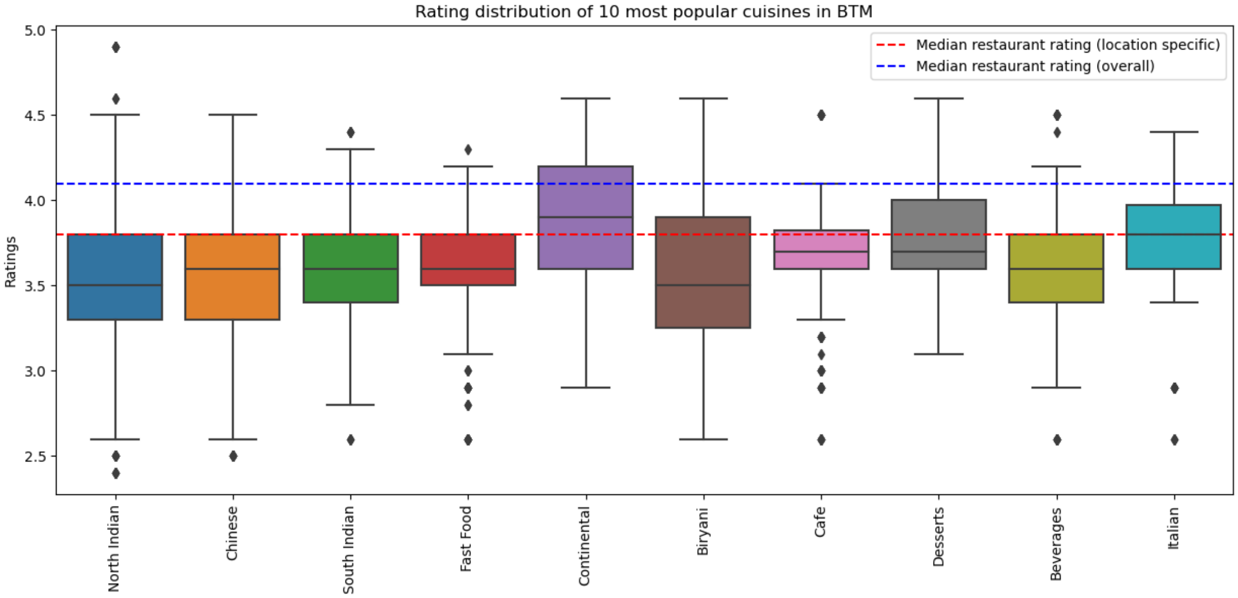

Figure 5a: Rating Distribution of 10 Most Popular Cuisines in BTM

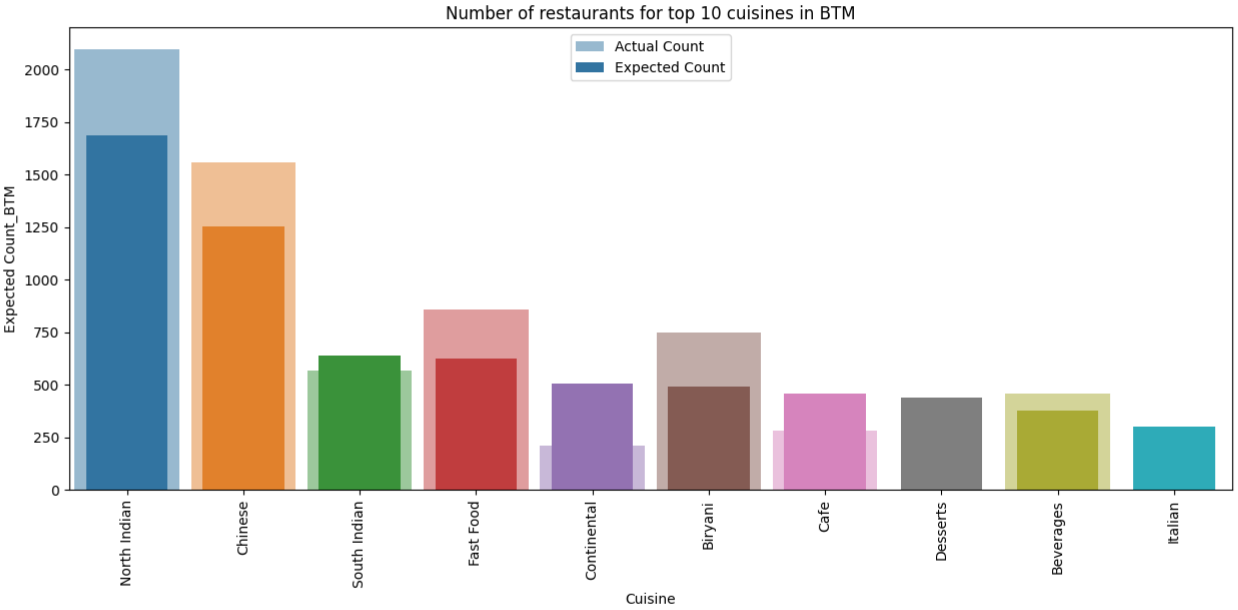

Figure 5b: Number of Restaurants for 10 Most Popular Cuisines in BTM

Team Members: Aryan Jain, Mallory Sico

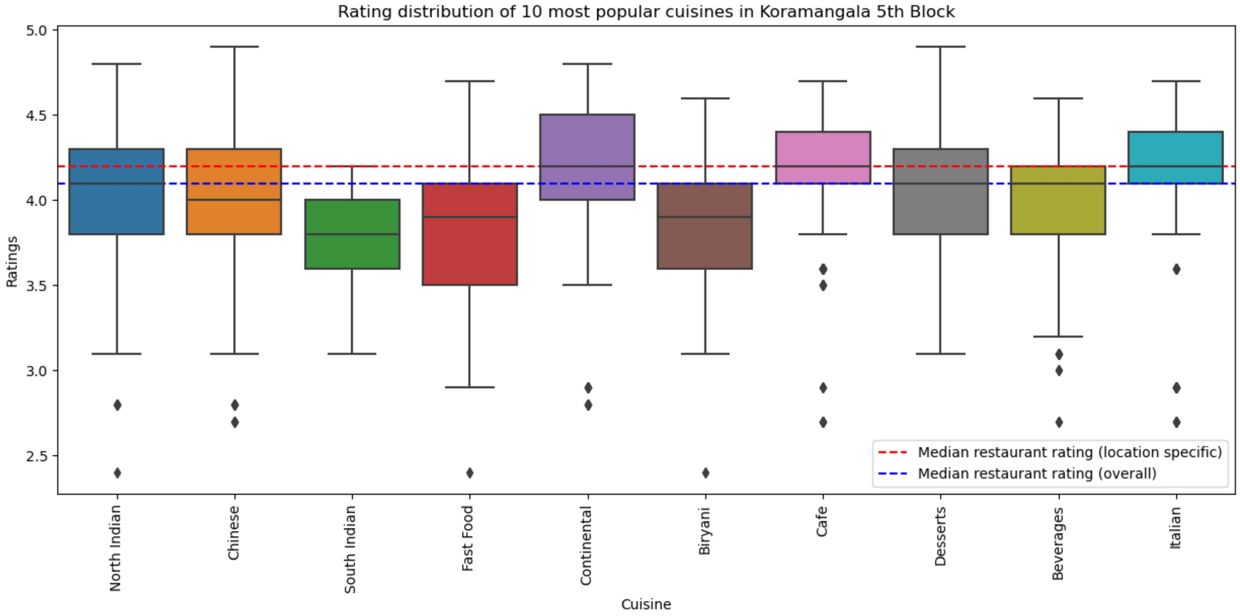

Figure 6a: Rating Distribution of 10 Most Popular Cuisines in Koramangala 5th Block

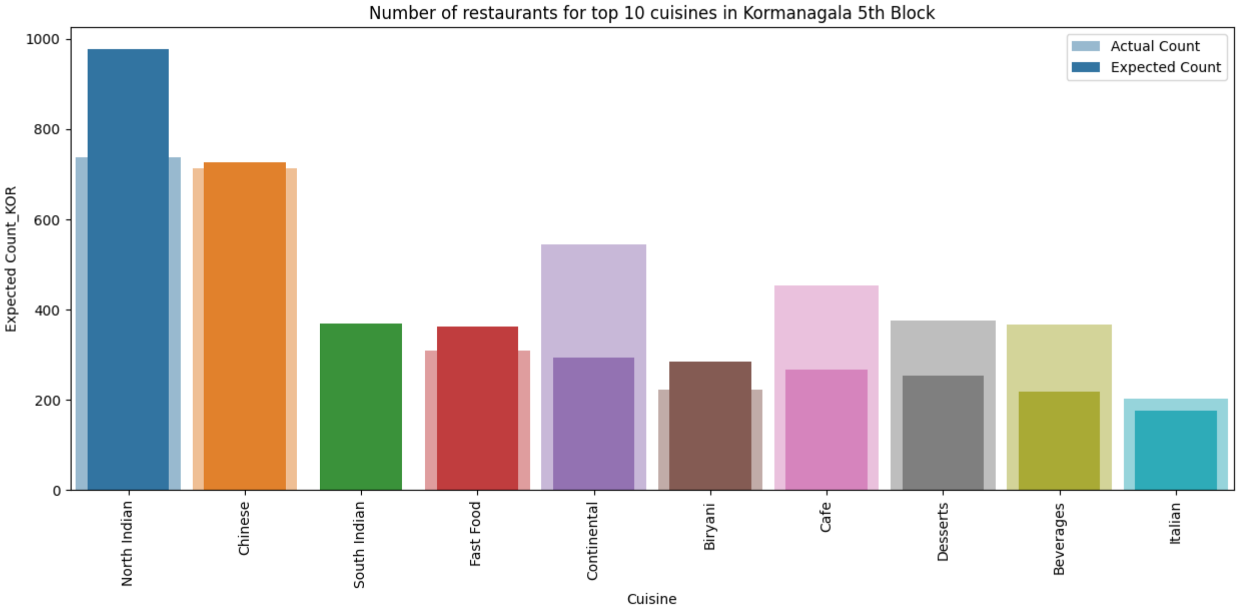

Figure 6b: Number of Restaurants for 10 Most Popular Cuisines in Koramangala 5th Block

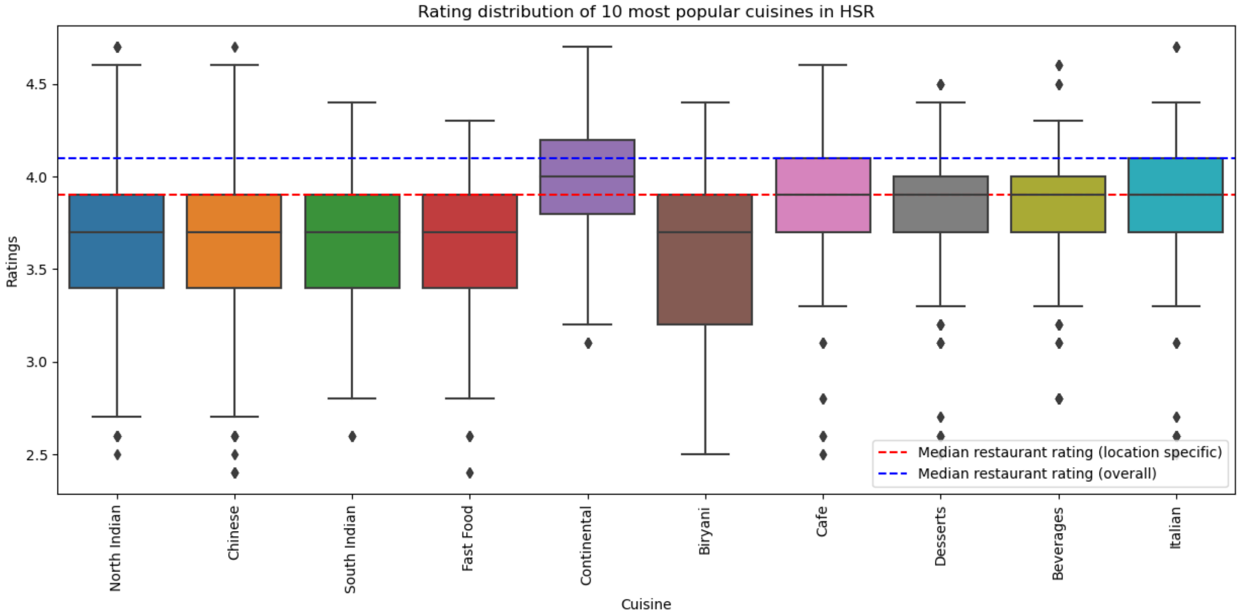

Figure 7a: Rating Distribution of 10 Most Popular Cuisines in HSR

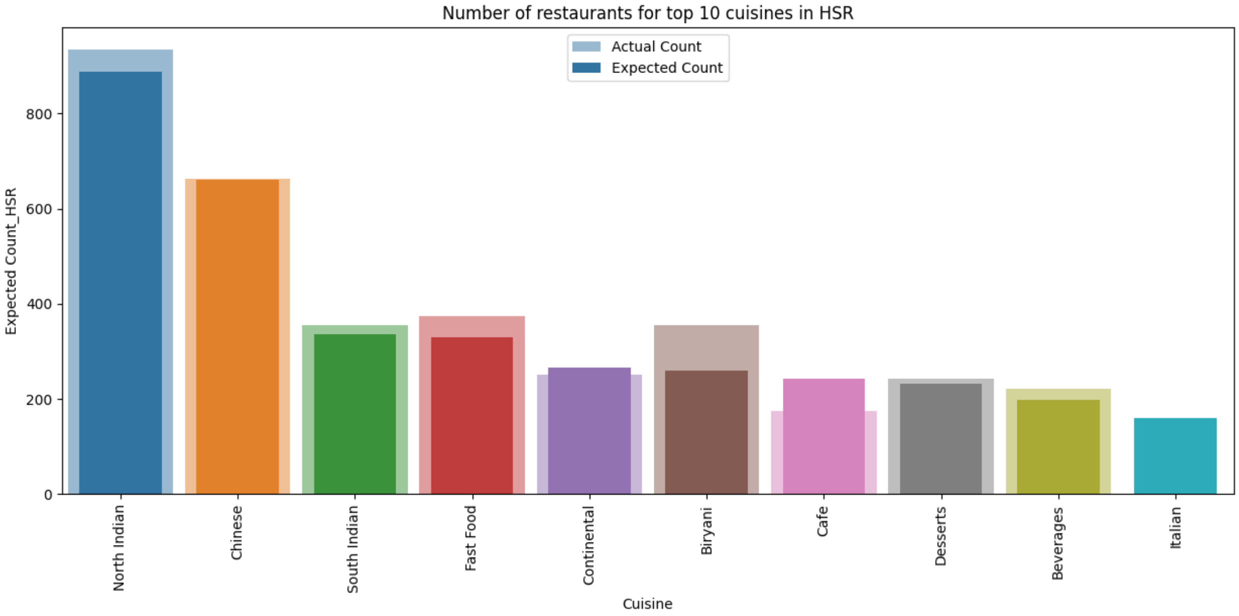

Figure 7b: Number of Restaurants for 10 Most Popular Cuisines in HSR

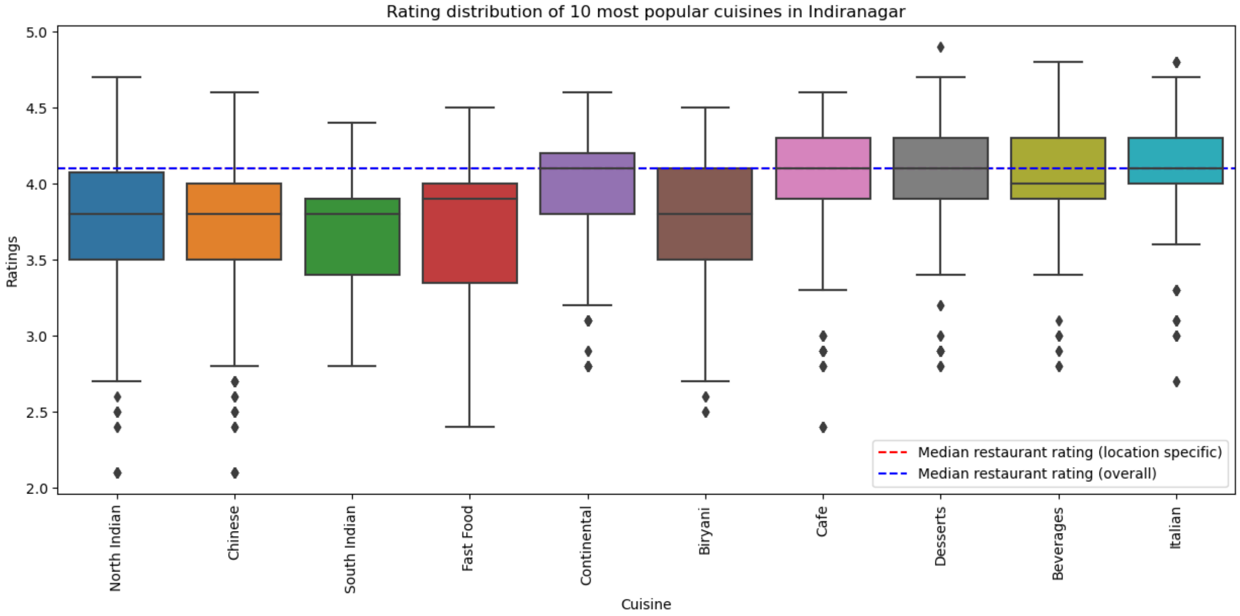

Figure 8a: Rating Distribution of 10 Most Popular Cuisines in Indiranagar

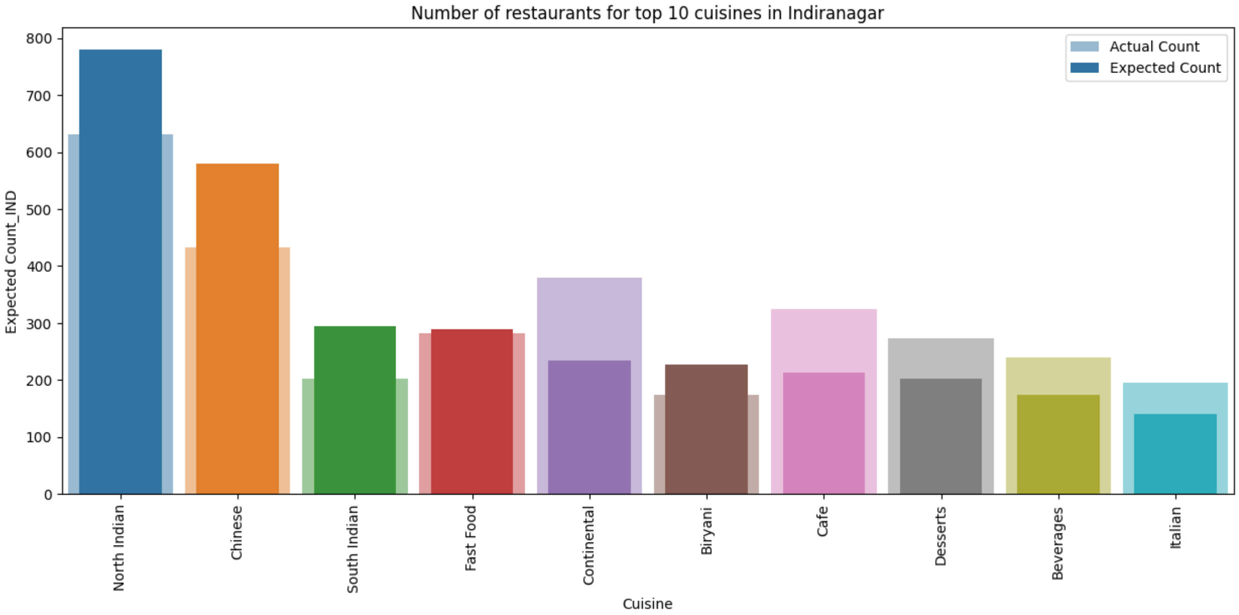

Figure 8b: Number of Restaurants for 10 Most Popular Cuisines in Indiranagar

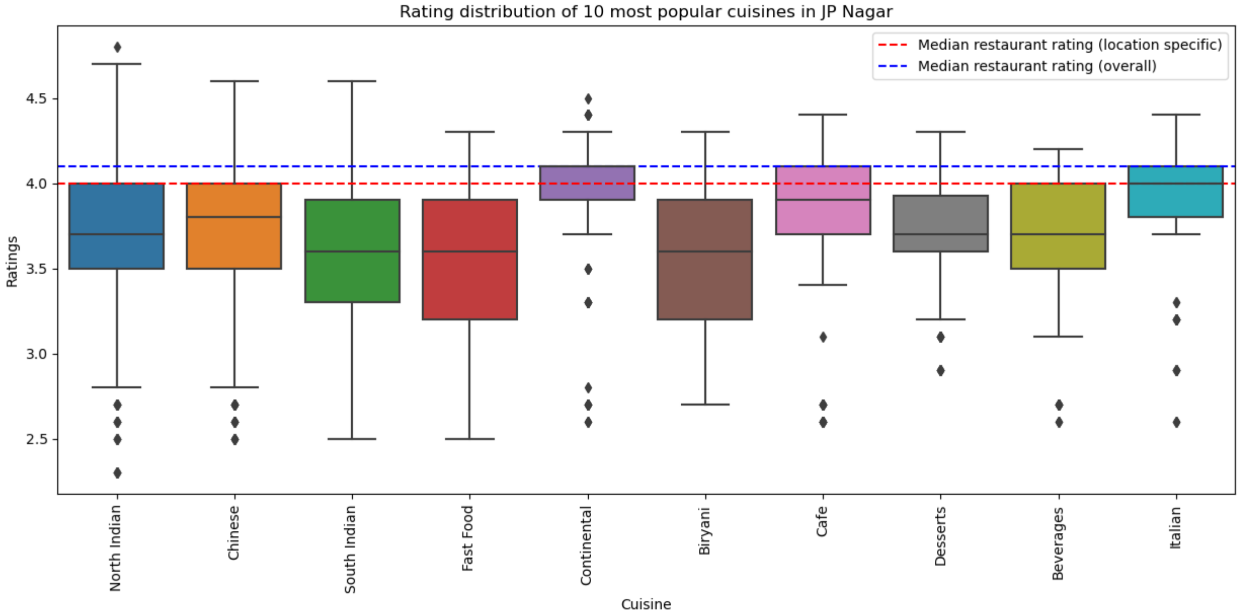

Figure 9a: Rating Distribution of 10 Most Popular Cuisines in JP Nagar

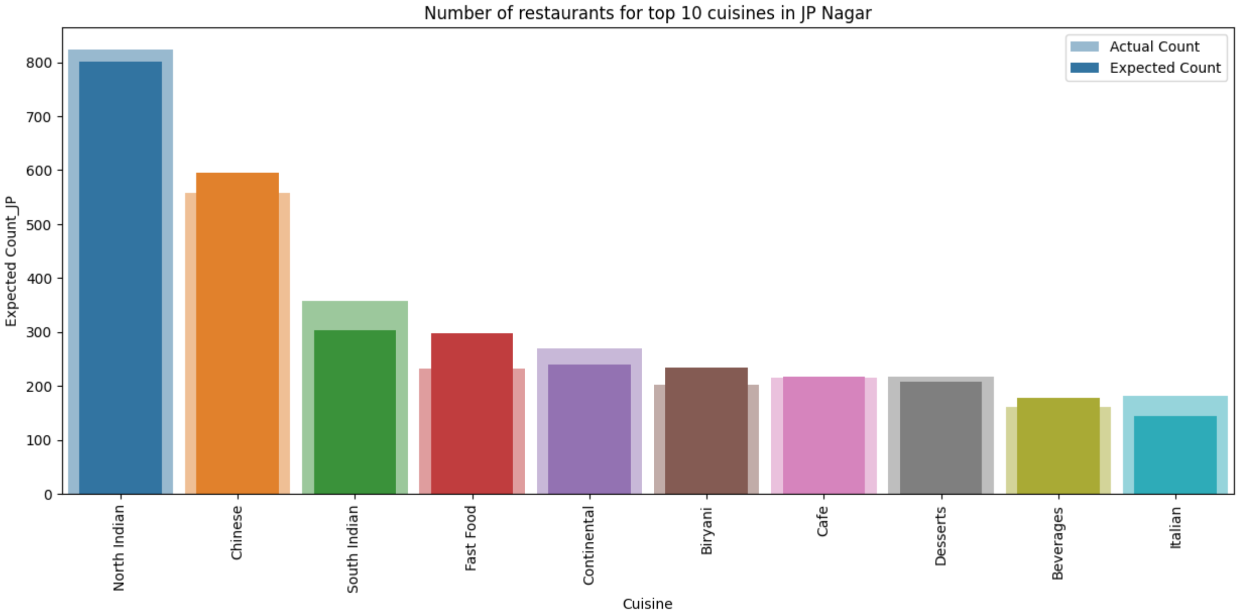

Figure 9b: Number of Restaurants for 10 Most Popular Cuisines in JP Nagar

Team Members: Aryan Jain, Mallory Sico

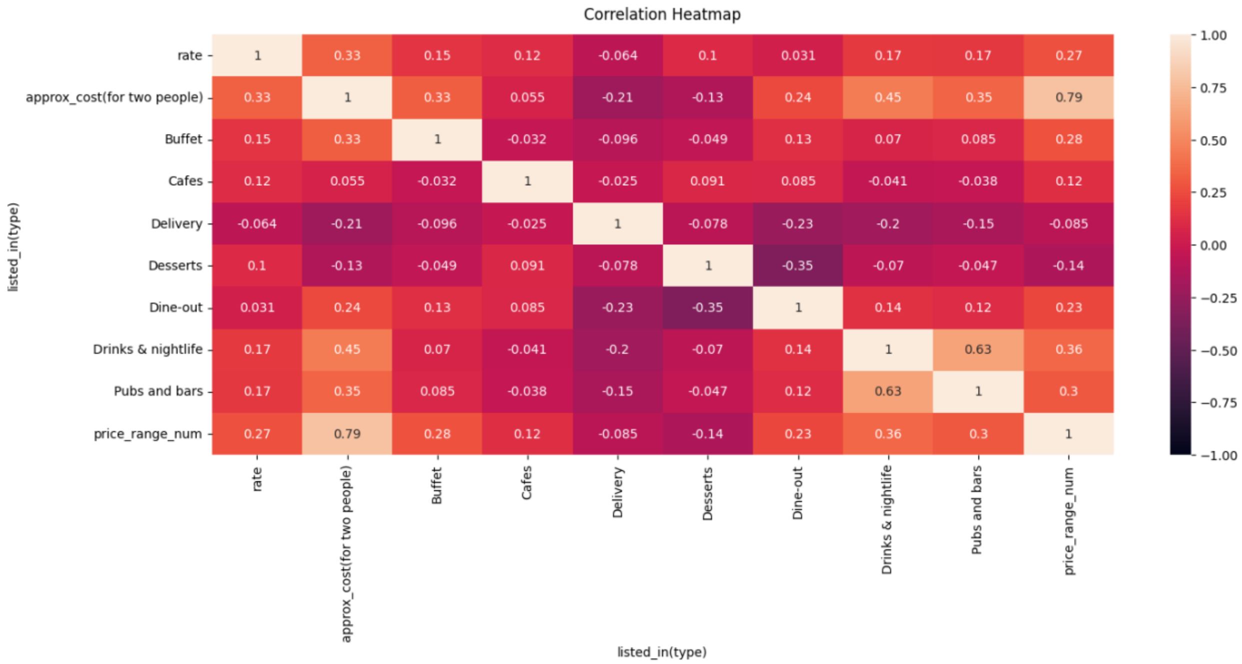

Figure 10: Correlation Matrix for Ratings, Serving Types and Price Range

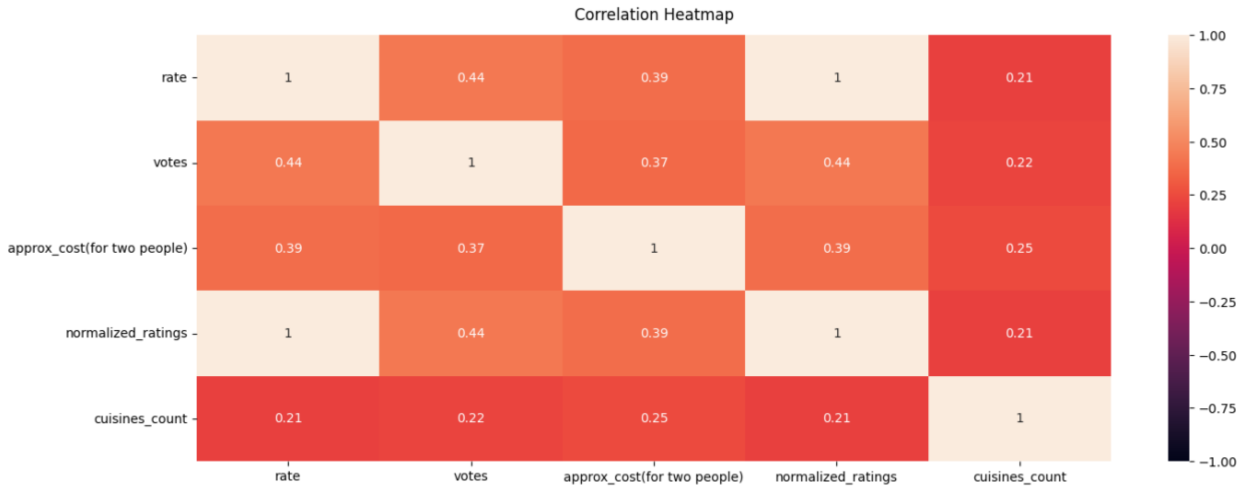

Figure 11: Correlation Matrix between Rates, Votes, Costs, Normalized Ratings and Cuisine Count