Cherry Agarwal

Emily Buzzelli

* Description of Project Goals

The purpose of this project was to use data to determine the ideal type of restaurant the team could open in order to maximize customer ratings. The initial hope was to open a restaurant with features aimed at maximizing profits, but more data was found on customer ratings/restaurant features than could be found on profits, so it was decided to use customer ratings as a stand in for profits. The assumption being that more highly rated restaurants most likely do better monetarily. The team started with a dataset from Kaggle called “[Zomato-restaurants](#Zomato)”. This data set contained restaurant name/address information, details on the average cost/price range, whether the restaurant had online delivery or took reservations, the cuisine type and details on customer ratings (aggregate score & level) and the number of total votes used to calculate those ratings. The team set out to determine the following:

* + What area/locality should the restaurant be open in? – The goal was to find an area that was simultaneously a place where restaurants are on the more expensive side (in order to maximize the amount of money to be made) and also are highly rated (meaning people still like the restaurants in this area, even though the food is expensive).
  + Which cuisine is highly rated, but does not have many restaurants in the area/locality chosen? – The goal being to find a cuisine for which there was a lot of demand (people liked it a lot, but there were not too many restaurants serving it).
  + What extra features should be added to the restaurant to improve ratings? Would adding online delivery or allowing for table reservations improve customer ratings at the new restaurant?
* Exploratory Analysis

The team began by exploring what area of the world the majority of the data came from. It turned out that the vast majority of data related to restaurants in [India](#Restaurants_By_Country) (over 8000 observations in a data set that had around 9500 observations), so it was decided to focus on restaurants in India for making determinations. In addition, it was discovered that there were many [restaurants without ratings](#Restaurants_wo_Ratings). Since the purpose of this project was to discover the link between ratings and restaurant features, these observations were removed. Leaving them in would have falsely lowered rating scores for features that just happened to have more unrated restaurants.

From here, the team looked into what locality would be best to open a restaurant in. The ideal locality had highly rated restaurants and a high average cost for dining in the area – as this implies that people like restaurants in these areas and are willing to pay a premium for dining in these places. It turned out that of the restaurants in India, many of the rated restaurants in the data corresponded to restaurants in [New Delhi and its suburbs](#Restaurants_By_City). Therefore, the search was narrowed for a locality within the New Delhi & surrounding area. To find ideal locality, first the team found the most highly rated areas within New Delhi by taking the average rating and average cost among all restaurants in each sub-area. Next, sub-areas with fewer than 20 restaurants were removed (if areas with just a few restaurants had high ratings, this may skew the analysis). Based on the [results](#Localities_Cost_Rating), the team decided to target opening a restaurant in one of the top ten localities, but held off on deciding on which one until a determination of cuisine could be made. More details on this are below.

The second step was to determine what types of cuisines were well liked, but that represented relatively small number of restaurants in the New Delhi area. The cuisine data in the data set listed multiple cuisines related to each restaurant separated by commas. In order to determine the rating for each cuisine type individually, each cuisine was separated into a different column. Then these cuisines were counted for popularity (the number of times they appeared in any restaurant) and the [top ten cuisines](#Top_10_Cuisines) were selected to work with. The top ten were selected because even though the team wanted to choose a cuisine that did not have that many restaurants, we did not want to have to create a market for a new cuisine type. We wanted something that would be known to the population. Then the aggregate rating for each of the cuisine was plotted using a [box plot](#Rating_Top_10). Finally, the team decided to also look at [average price for two for each of the top ten cuisines](#Cost_for_two) – to determine which cuisine types demand a premium. From these results, the team decided to open an Italian/Continental restaurant because these cuisine types showed high ratings, but lower counts of restaurants in the New Delhi area and also showed a higher average cost for two. The final step before cementing the restaurant cuisine was to understand how these cuisine types performed in the localities chosen in the first step. A [table](#Cuisine_By_locality) was created showing the rating for these cuisines in the top localities – along with the number of restaurants in these areas with these cuisines. From this table it can be seen that Italian/Continental cuisines get high ratings, but have low percentage of representation in Epicuria Food Mall & Sector 29. Therefore, the decision to focus on an Italian/Continental restaurant in Epicuria Food Mall or Sector 29 was taken.

Next, the team did a bit of analysis to understand how online delivery, table booking and having more than one cuisine affected ratings. To do this, [box plots for each of these features](#Features_v_Rating) were created. By doing this, it could be seen that restaurants with online delivery and table bookings had slightly better ratings than those without these features on average. Based on these results, it was decided that the restaurant should have online delivery and table booking as features. Additionally, it could be seen that restaurants with more than one cuisine associated to them had better ratings. Therefore, the decision to go with a multi-cuisine restaurant in the previous step was confirmed in this step.

* Solution and Insights

The team ran several models in an attempt to validate the decisions from EDA and understand if further insights could be gleaned on which features were most important for high ratings. The following features were used in an attempt to predict ratings: has online delivery, latitude/longitude, table booking average cost, multicuisine or not, and cuisine. Several models were used to try to predict rating (Excellent/Very Good rating being the positive class, and all other ratings being the negative class), including: Naive Bayes, Linear Discriminant Analysis, K-nearest neighbors, Decision Tree and Logistic Regression. None of the models were able to outperform the baseline prediction (91% for the negative class) on test sets. However, logistic regression did show positive correlation between ratings and Italian/Continental cuisines, as well as table booking – which are all features to be incorporated into the team’s restaurant. In addition, while logistic regression showed a negative impact from online delivery = Yes, it was less of a negative impact than that assigned to online delivery = No. The fact that both yes and no for online delivery was associated with a negative impact on rating probably has to do with the fact that the likelihood of the negative class is high in general. That online delivery = Yes was less negative was taken as an indication that it was less associated with bad ratings than the converse – which confirmed what the team found in EDA. In general though, while the models (specifically logistic regression) seemed to provide some confidence for the decisions made coming out of EDA, they did not provide many additional insights to the team.

In summary, the team decided to open an Italian/Continental restaurant in one of two localities in the New Delhi metro area (Epicuria Food Mall or Sector 29) that has online delivery and takes table bookings. This combination of features was chosen, because the analysis indicated that this type of restaurant would bring higher customer ratings and also corresponded to high average cost for two. In addition, the cuisine types were well known amongst the population, but underrepresented when compared to the most popular cuisines – indicating restaurants of this type would be welcomed in the market. Given the high baseline prediction for the negative class (reviews less than Very Good) in the dataset, models did not seem to be very helpful for determining these features, so the team relied on charts and tables to come to these conclusions.

A few things could not be determined with the dataset, leaving the team wanting additional data to further research this question. To start, no financial information on the restaurant was included, so no analysis could be done on whether or not high ratings are actually tied to restaurants being profitable or not. While this is probably the case, the team would feel more confident in the results if some proof of this were able to be calculated. In addition, more details on the reviews given would have been helpful. For example, which reviewer gave what review would have helped the team figure out if certain cities/areas tended to have people who gave worse reviews on average than other places. Perhaps people living in some areas are just pickier than others – leading to lower reviews, but not necessarily a less popular restaurant. Finally, data from additional countries/areas would have been helpful and could have expanded the type of analysis that was able to be done on this data set. Not only could the team have researched ideal localities within a city in India, but could have determined the ideal spot in the world to put a restaurant in order to maximize ratings.

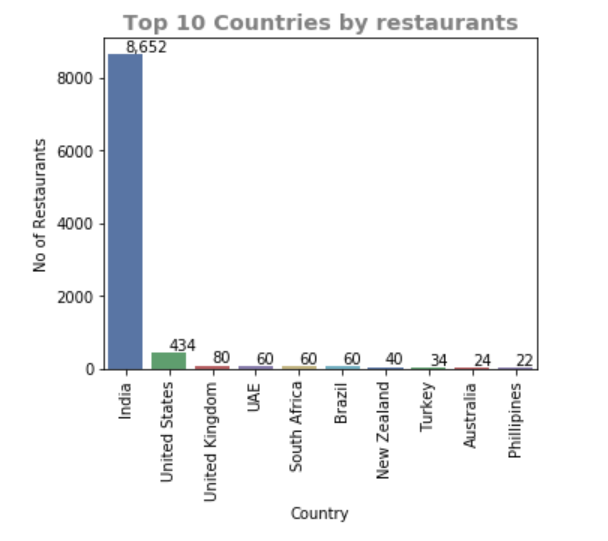
* Links and figures

**Links:**

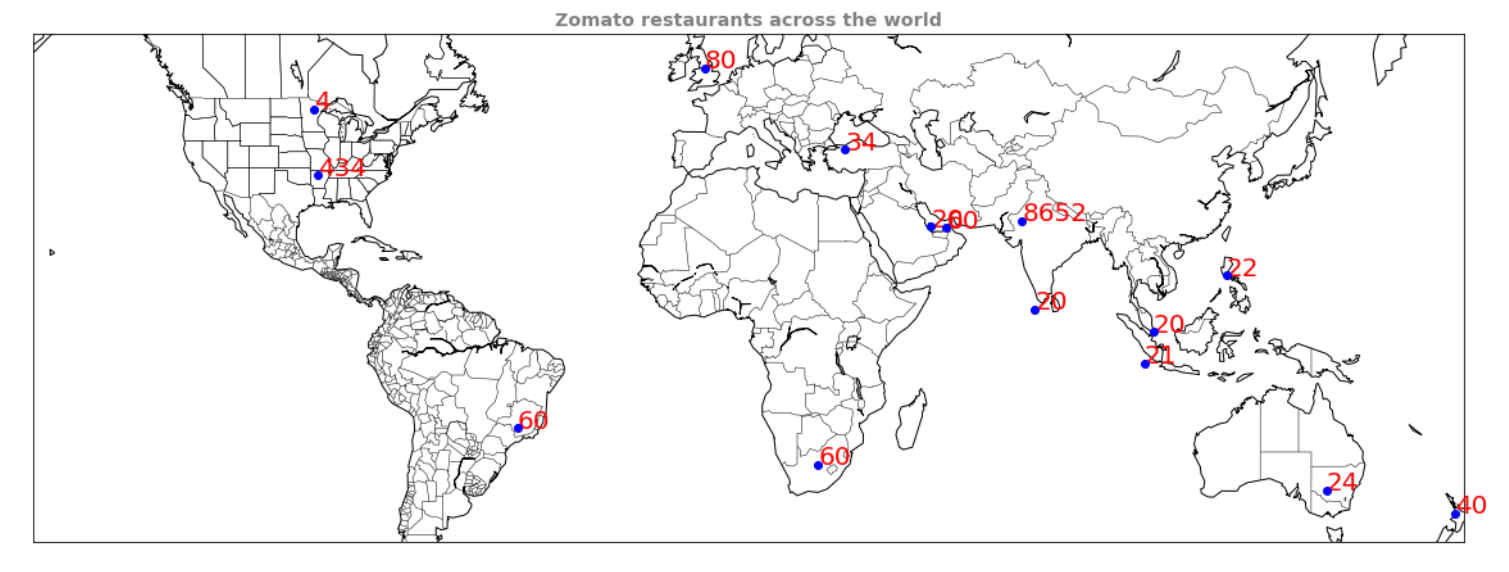
1. <https://www.kaggle.com/shrutimehta/zomato-restaurants-data>

**Figures:**

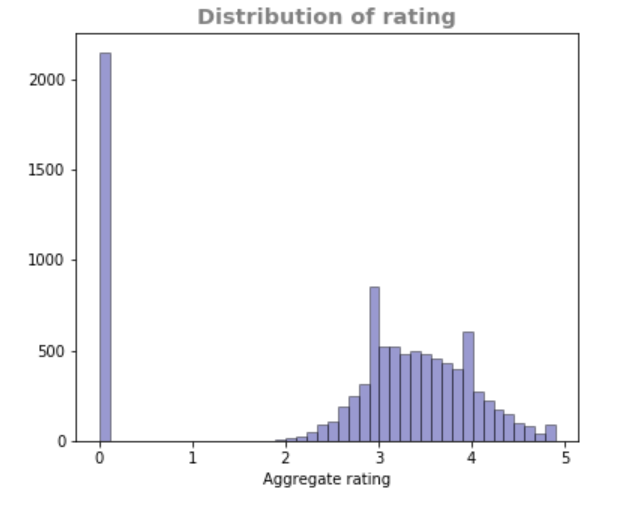
* + Figure 1
    - A - Restaurants by Country



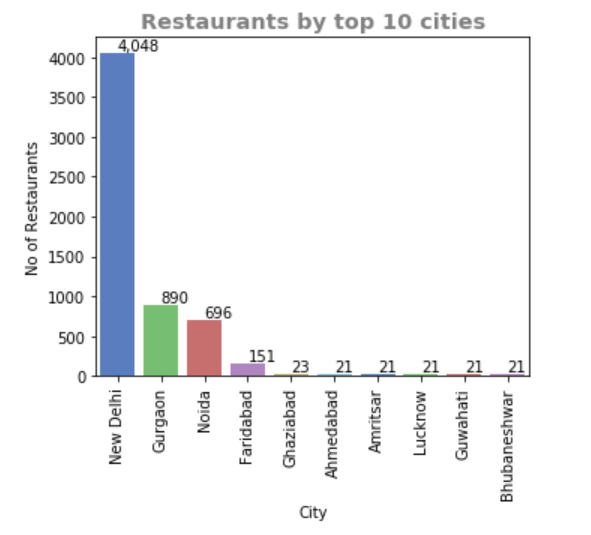
* + - B - Restaurants by Country



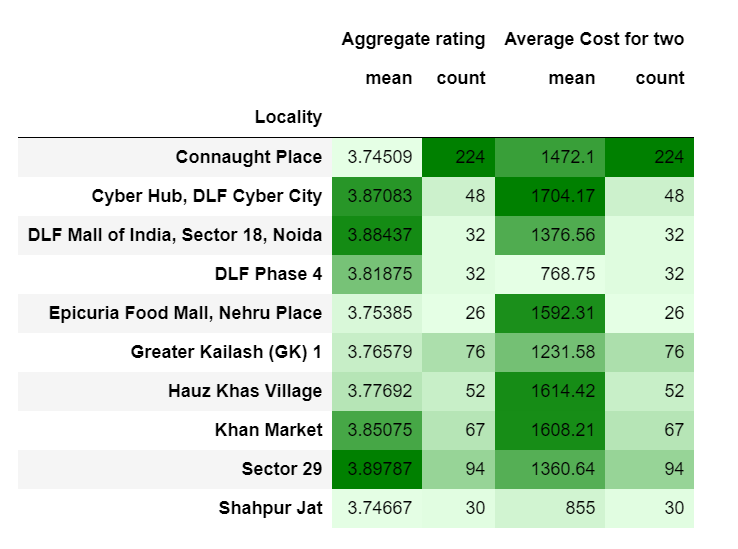
* + Figure 2 – Restaurants without Ratings



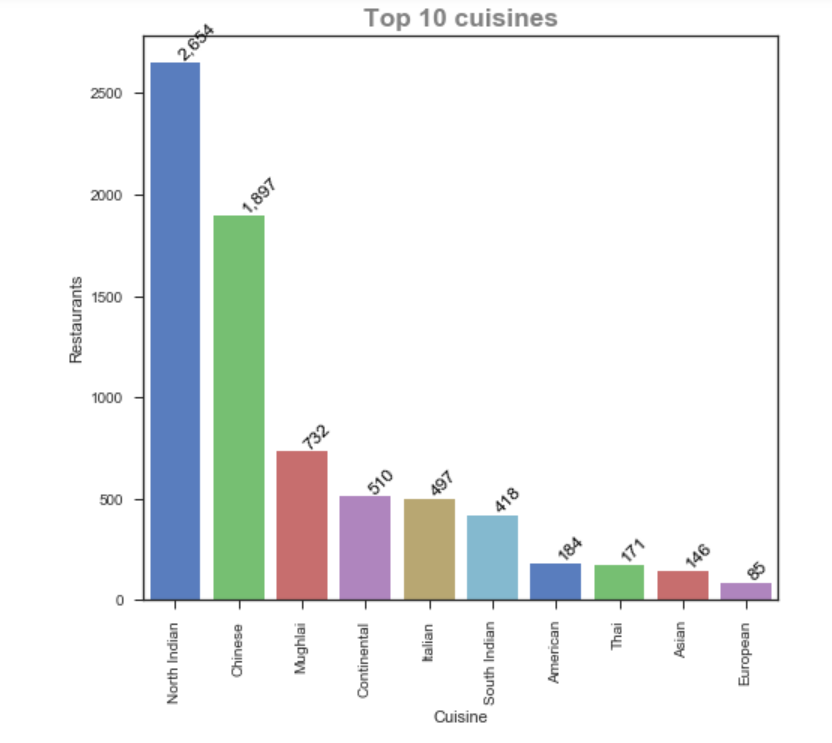
* + Figure 3 – Restaurants by City



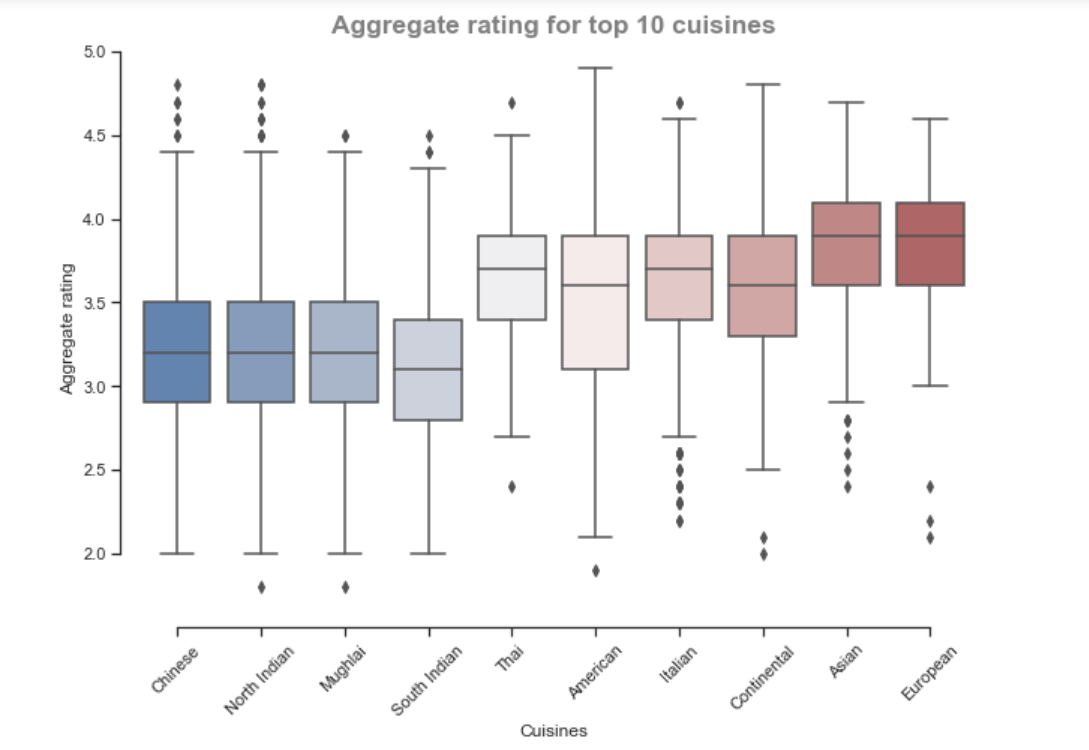
* + Figure 4 – New Delhi Metro Area Localities (Ratings & Cost)



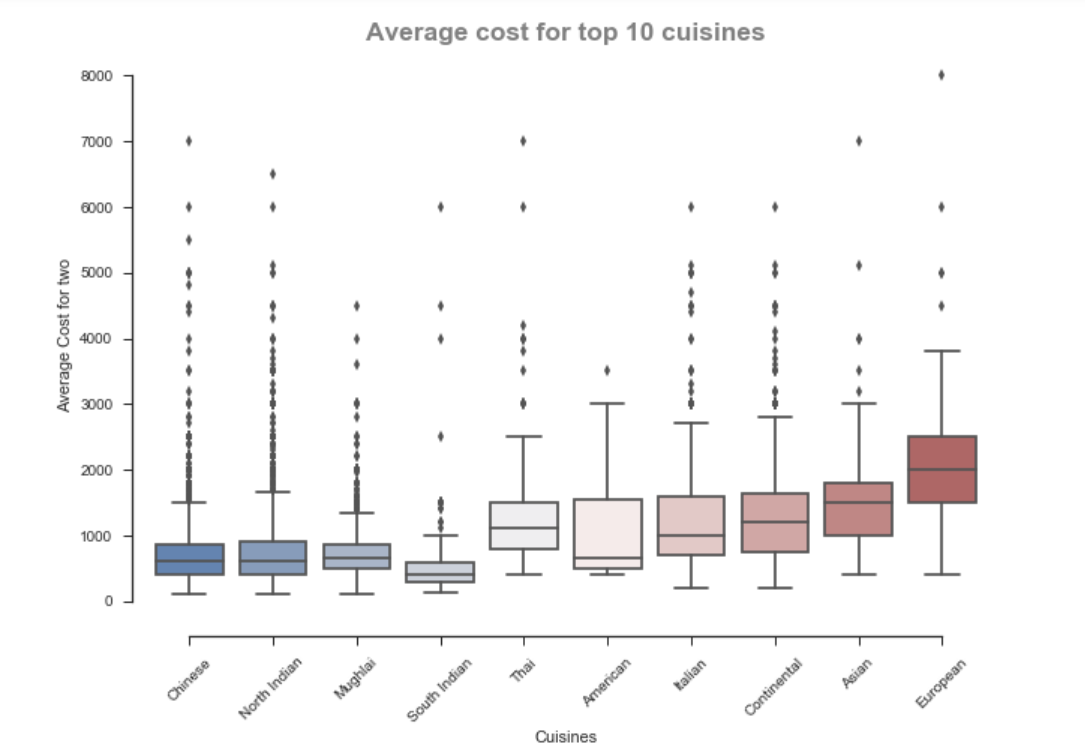
* + Figure 5 – Top 10 Cuisines



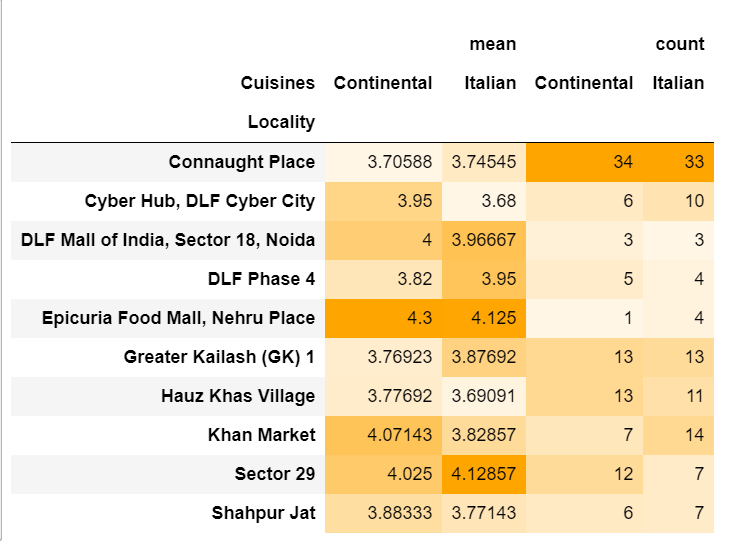
* + Figure 6 – Rating for Top Cuisines



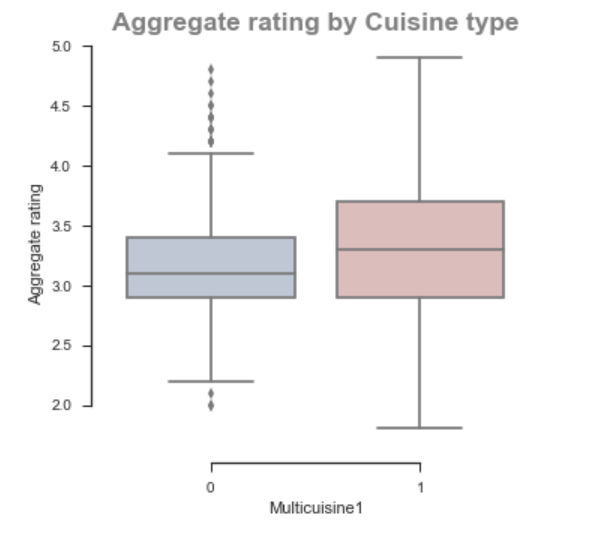
* + Figure 7 – Average Cost for Two (Top Cuisines)



* + Figure 8 – Italian/Continental in localities



* + Figure 9
    - A – Number of Cuisines & Ratings



* + - B – Online Delivery & Ratings



* + - C – Table Booking & Ratings

