**Data Cleaning and Understanding – Part 1**

Keith Melton and Lucas de Paula

**Executive Summary and Preliminary Analysis**

Only four restaurant types return revenue over 1,600 Rupees per couple on average: fine dining, club, lounge and microbrewery. Fine dining averages 2,750 the highest revenue per two diners and averages over 21% more revenue than the second-place restaurant type club at an average of 2,171. Coming in at third, is lounge at 1,728 and is 37% less revenue per couple. The final restaurant type averaging over 1,600 is microbrewery at 1,633.

Fine dining is an easy choice over singular investments in clubs, lounges or microbreweries. The restaurant type club revenue makes it an attractive target. However, there are only six clubs in Bangalore which make the data a small sample size and coupled with our lack of club management experience make it less attractive option. The lounge option is attractive based on the revenue opportunity discovered in the data. The initial findings would be to combine lounge or bar with fine dining, for increased revenue potential, where dining space allows. Pursuing a microbrewery investment is not recommended due the high cost of equipment needed. The data shows that Bangalore fine dining is only ever combined with lounge, bar or microbrewery restaurant types.

Preliminary analysis was conducted on fifteen Bangalore micro-areas found to have a history of supporting fine dining or the affluence to sustain additional fine dining restaurants based on average location dining revenue.

*Initial Analysis*

A preliminary analysis indicates high revenue potential in select Bangalore markets making it an attractive opportunity for individual acquisitions or restaurant placement. Additionally, fine dining should be combined with a lounge or bar for increased profit potential, where space allows. The most popular cuisine options noted are American. Chinese, North Indian, Italian, Café, Biryani, Continental and South Indian and at least one of these options should be chosen in our establishments. The fine dining price median is 2,847 for two diners and is a realistic target for our restaurants and possibly higher if combined with a lounge or bar concept. Obviously, high rating is needed and preliminary analysis shows a 4.1 rating or higher is need to maintain the approximate median fine dining price point noted. Finally, booking shows a positive impact one revenue but online ordering show little effect.

More in-depth analysis and modeling will be conducted in the latter half of the project timeline. This additional analysis is needed to confirm these initial findings and/or improve the final recommendation. A final analysis with recommendations is due by August 21, 2022.

**Background**

Data cleaning, understanding and exploration are an important step of any analysis. The incomplete, incorrect, duplicate or otherwise inaccurate data must be corrected in the dataset. Understanding the data helps refine the overall analysis plan. Even with the provided dataset there are numerous paths one could take. Better understanding at this early juncture helps shape the path forward for the best and most focus results. Through data exploration, such as querying or visualization, an initial hypothesis is formed as found above in the Executive summary.

**Data Cleaning**

The data was cleansed after identifying several problems in the dataset by resolving corrupt, inaccurate or irrelevant data. Data cleaning took place in Excel and Tableau to enable data exploration and modeling. The following steps were taken:

1. Remove unneeded data
2. Removed duplicate data
3. Fix structural errors
4. Handle missing data
5. Filter out outliers
6. Grouped levels for Food List Type, Restaurant Type, Cuisine
7. Filtered and pre-processed Reviews List

First, 4,533 rows were removed from the dataset of 56,250 rows leaving 51,717 rows.

The issue appears to be a mismatch between the Phase 1 and Phase 2 data extraction.

Some corrupt characters were removed as well.

Several unneeded columns were eliminated such as:

1. URLs: the URL was used to extract the data, doesn’t add more value to the analysis
2. Location: address is very confusing and difficult to parse to something meaningful
3. phone (phone numbers): we calculated metrics such as number of phone lines out this variable

The dataset is still large with over 50,000 entries, and duplicate elements were investigated. They were found in scenarios such as a restaurant having two or more locations; or having different types of cuisines.

Here are the main transformations/treatments:

* Structural elements were addressed, for example, restaurant\_type and cuisine had data elements separated by commas. In restaurant\_type there were two types of restaurants noted and in the cuisine column up to 8 different cuisines were indicated. This problem was address in multiples ways. For the Tableau EDA was placed in separate column so no data cell contained more than one data element. In R the data similar restaurant types (ex., Bar, Pub, Lounge) and cuisine types (ex., Ice Cream, Mithi, Desserts) were combined to scale down data elements to model. We believe treating the data as so will help in the modelling phase by reducing the variability on the dataset.
* The rate column had a problematic syntax (e.g., 4.1/5); the /5 was removed and we converted it to a number in order to run mathematical operations on such as average.
* In the ratings column, “New” data elements were converted to “999”.
* To create a phone number R count code was used to clean the data (see figure below).

Text

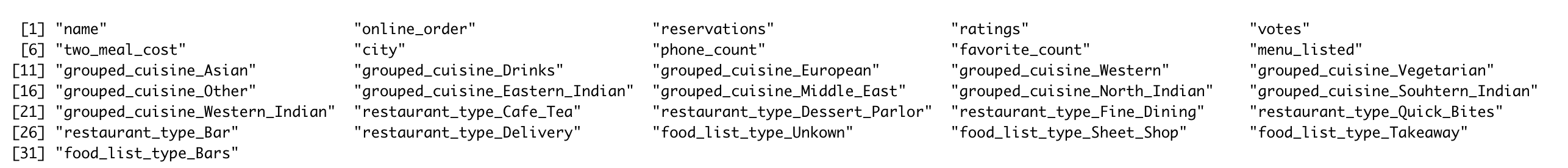
Description automatically generated

* Booking and online data were changed to numeric to prep for modeling.
* In the fourth step some missing data corrected in the first step when over 5,000 rows were removed. However, some missing data was rectified further in Tableau when conducting data exploration and visualization steps (especially null values).
* For text analytics, we had treated the reviews\_list variable as one string. This means we had to remove repetitive text like “RATED: 3.0 “ \n”, and many other special characters. We also removed missing reviews\_list such as “[]”.
* We grouped levels for variables such as cuisines. We grouped them by regions: north, south India, and macro regions Europe, Asia, etc. The same process was used for variable restaurant\_type and food\_list\_type. See below the code used.

Text, letter

Description automatically generated

* After we had grouped categories, we had to create dummy variables for these columns, and this allowed us to aggregate the data in order to have grouped by name and location of the restaurant. The output data columns looks like this:



* Outliers were corrected in Tableau when needed.
* For modelling purposes, we split the data into train and test, where 70% of the observations goes to 70% and 30% goes to test.

**Data Exploration**

*Data Understanding*

The data understanding process began the when scoping the project. Some hypothesis began being informed and when investing the dataset preliminary questions were produced. The data was assessed and elements of interests where noted such as costs, ratings, restaurant types and cuisines. Possible solutions to business problems are noted (ex. Determine which restaurant types are most popular per location and have the highest approx\_cost to guide recommendations). It is important to understand the high-level data elements so a data dictionary is provided here:

* **URL**: Zomato URL for the restaurants
* **address**: address of the restaurant
* **name**: name of the restaurant
* **online\_order**: whether restaurant accepts online orders
* **book\_table**: whether restaurant provides option for booking table
* **rate**: restaurants rating by customers
* **votes**: number ratings votes
* **phone**: contact details of the restaurant
* **location**: area where restaurant is positioned; smaller area
* **rest\_type**: type of restaurants
* **dish\_liked**: specific dishes customers liked
* **cuisines**: type cuisines offered by the restaurant
* **approx\_cost**(for two people) : average cost for two customers
* **review\_list**: reviews of the restaurant on Zomato
* **menu\_item**: menu items available in the restaurant
* **listed\_in(type)**: type of the restaurant
* **listed\_in(city)**: locality of the restaurant

*Data Visualization*

Data visualization and exploration is a crucial step is solving business problems presented in this exercise. Exploring the data further helps make sure the right questions are asked and allows for further confidence when forming a hypothesis. Showing data in a graphical way provides a path to see and understand trends, outliers and patterns in data not possible or at least more easily discernible. High-level findings are found below:

1. Exploration revealed fine dining offered the highest revenue opportunity with club, lounge and microbrewery with over 1,600 average costs as well.



1. Location analysis yielded two data element opportunities to explore, the location and listed\_in(city) data. The location data was more granular and allowed for more focused or targeted analysis for fine dining opportunities relative to the listed\_in(city) data.



1. The top 30 micro-areas were chosen from the list below for further exploration and analysis.



1. The 15 locations with the highest costs were investigated. The exploratory analysis is represented on the below dashboard showing 8 of the 15 locations. All 15 sites showed a history of fine dining establishments or were identified as potential locations for additional higher cost restaurants due to high average revenue per customer in that location. Even without a fine dining establishment the area shows potential for a desire for higher quality and the ability to pay for it.



1. Potential high revenue cuisines are shown below. The cuisines highlighted in blue are seen 10 or more times in the data for fine dining cuisines. The cuisines highlighted in gray represents a minimum of 5 instances in the data. The most popular cuisines are American. Chinese, North Indian, Italian, Café, Biryani, Continental and South Indian. The three highest grossing cuisines are American, Chinese and North Indian.



1. The optimal fine dining experience price range for two is between 3,098 to 2,308 Rupees with a median cost of 2,847. The price range data is seen below in the boxplot. A correlation between low ratings and high prices is noted. Potentially customers have higher and possibly unreasonable expectations if the price is too high. Ratings of 3.4 and 3.5 had average costs of 4,000 and 3,467 Rupees respectively.



1. The optimum price sweet-spot appears justified in the ratings starting at a 4.1 rating. Price increases are seen starting at 4.1. For ratings of 4.0 an average price of 2,300 is noted but goes to 2,826 at a 4.1 rating. The optimum price range is sustained from 4.1 to 4.9. There were no perfect 5.0 fine dining ratings in the data.



1. Restaurants with booking options showed higher revenues while online orders did not show a positive price impact.



1. What the data doesn’t tell us:
   1. There doesn’t seem to exist a big difference on average rating by Restaurant Type (grouped)

Chart, box and whisker chart

Description automatically generated

* 1. There doesn’t seem to exist a big difference on Vote Average by Restaurant Type

Chart, box and whisker chart

Description automatically generated

1. What the text data tells us
   1. We ran three analysis:
      1. simple word count by restaurant location; and the result, for the most part indicated Chicken, Service, Ambiance, Taste.
         1. However, some locations won’t mention the word as much as others. For example, ambiance isn’t as mentioned at Marathahalli as it is at Koramangala 4th block.

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

* + 1. We also calculated the same metric, but this time using metric called TF/IDF, where instead of looking at how many times a word appears in one location, we calculate how many times does that word appear in many other reviews about other locations, this way we get an importance across locations and reviews.
       1. This time, unfortunately we get more Hindi words, so it is difficult to interpret the results, but we can see that JNC is Jyoti Nivas Pre-University College, which is close to 4th Block. DYU is an art cafe, Boho is a bar close by, same goes to Atithi.

Chart, bar chart

Description automatically generated

* + - 1. Kakaji, ITPL are places close to Marathahalli. A kulhar is a traditional handle-less clay cup from South Asia that is typically unpainted and unglazed, and meant to be disposable.
      2. Chart, bar chart

         Description automatically generated
    1. Lastly, we ran a topic discovery algorithm to analyze 4 different topics with the top 10 words in each of them.
       1. For this case, we ran in the full dataset, and we didn’t break it by location. We didn’t find any major insight, but Chicken is clearly present in all of them.
       2. Topic 1 – mentions staff, service, ambience
       3. Topic 2 – mentions time, friends, pizza, veg
       4. Topic 3 – mentions beer, music, experience
       5. Topic 4 – mentions quality, service, taste

Chart, bar chart

Description automatically generated

**Sampling and Cross Validation – Part 2**

There are many sampling options and cross-validation approaches and no perfect approach. With any method taken one must be on guard for overfitting and randomly splitting the data helps avoid the overfitting problem.

The true test of a model is how it performs on data not seen. Options include:

1. waiting for new data or
2. simulating the existence of unseen data.

Since receiving new data is not possible the choice is easy. Simulating new data is used here by withholding some data to test the model. The downside of this approach is giving up some of the data which could help improve our model. However, the consensus in the data science community is cross validation and splitting is best practice despite this downside. With over the high number of observations found in the Zomato dataset the concern is further minimized.

A simple sample or splitting approach chosen here of 70/30 (training data to test data ratio); however, another common simple approach option is 80/20 when using a training and test sets approach. Alternatively, some choose an approach with a training, validation and test of 60/20/20, or 70/15/15.

Instead of using a training, validation, and test approach a more powerful but more computationally intensive method is called k-fold, which is explained below. With any approach taken, one must be on guard for overfitting. The concept is based on leaving a portion of the data out of the estimation process. The training data is directly used to estimate the model. It builds and evaluates the model on the training set. The test dataset is treated completely independent of model’s development and is used to further define the predication model’s success. The model’s performance against the test dataset provides an unbiased result of the model’s success. The goal is to choose the model that demonstrates the best predictive results to predict the loan risk for a specific medical profile.

The process we have followed so far for data preparation and splitting is the following.

A picture containing text, screenshot, outdoor, blue

Description automatically generated

For the 70/30 split, we used the following code. It creates a random selection on the dataset up to the number of observations generated. Then, we select those rows and all columns to split it into 70% for the training dataset and 30% for the testing.

Graphical user interface, text, application

Description automatically generated

An example of a 70 training to 30 test sampling approach in R.

For the initial splitting of the data, we first set the seed to provide reproducibility. Then, we used the initial\_split function from R. This function takes the previously prepared dataset, the proportion of observations that goes to training, and a stratification variable (optional). We partitioned the dataset stratifying by the response variable ratings. We also removed the variable name as it is almost unique and doesn’t aggregate to the training purposes.

Another technique employed on the Zomato dataset is k-fold cross-validation. In k-fold, k is a number. The data is split into k “fold” or partitions. It is common for k to equal 3,5 or 10 although there is no set rule. However, if the data is particularly large, a model complicated or computing resources limited it may be wise to choose a lower k value. Other techniques are often added to k-Fold to average the findings leading to greater confidences.

Our goal with k-fold cross-validation is to train a model on the next phase that is more generalized (meaning it can make predictions while reducing bias and sensibility to variability on the distributions of the input data).

Here is how the process of k-fold works as per [scikit-learn](https://scikit-learn.org/stable/modules/cross_validation.html) website. First, we split the dataset into k-folds. In this image we can see we chose 5 splits. Each fold will contain the samples of the training dataset, used to train the model and it will leave one-fold for the testing. Finally, we evaluate the best split on the test dataset.

Table

Description automatically generated with medium confidence

After setting the splitting to randomly into training and test data and setting up k-fold, the cleaned and transformed Zomato data set is ready for predictive modeling.