Zomato Project

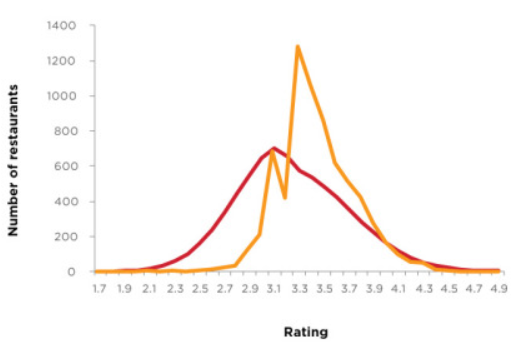
In Bengaluru, there are thousands of restaurants and the number keeps increasing every day. To restaurant owners, their priority job is to make their business survive from fierce competition and stay competitive. According to Zomato dataset, except for URL, name, and address of the restaurant, there are fifteen variables that affect the performance of the restaurant in daily operation. One of the variables is rate, whose range is between 0 and 5, and the higher rate means a better performance of the restaurant. The rate also can be interpreted that it represents customer satisfaction. The purpose of this project is to assist restaurant owners to make optimal decisions and get a higher rating on Zomato. At the same time, this project can give foodies a general idea of a restaurant before they step into it.

The project will be only based on Zomato dataset, so the rate is my response variable and the other fourteen variables are my predictors. The analysis is to find out which of them are my strong predictors that have a major influence on rate. After researching the dataset, cuisines, phone and city are excluded. The reason is I don’t believe a certain type of cuisine is better than others and will generate a high rate for the restaurant. Phone and city are also irrelated to the analysis.

The tools will be used in this analysis includes, but not be restricted to, R studio, JMP14, SQL, and Tableau. There is no limitation on methods or algorithms. The general steps of the analysis are Data Cleaning & Understanding, Data Preparation, Prescriptive Model Building, and Predictive Model Building.

Data Understanding, Cleaning & Preparation

My whole project is to analyze the performance of the top restaurants and find out what makes them outstanding. According to the blog of Zomato, they used to apply a weighted average of absolute scores to the restaurants rating, which is the orange line puts over 50% of the restaurants in the range of 3.2 to 3.5. Several years ago, they launched a new anti-bias algorithm to create a comparative distribution curve, which is the red line in the graph. It puts fewer than 30% of the restaurants will fall within the 3.2 – 3.5 range.



However, if we look at the top and bottom restaurants, they don't change much, which means no matter how Zomato calculates, the algorithm doesn't affect the top restaurants. To restaurants owners, I believe, all of them want their business at the top of the list. If I were them, I'd like to know why some restaurants have a higher rating than mine and what changes should I make to catch up with the top ones. This is the reason why I choose to focus on the performance of the top restaurants.

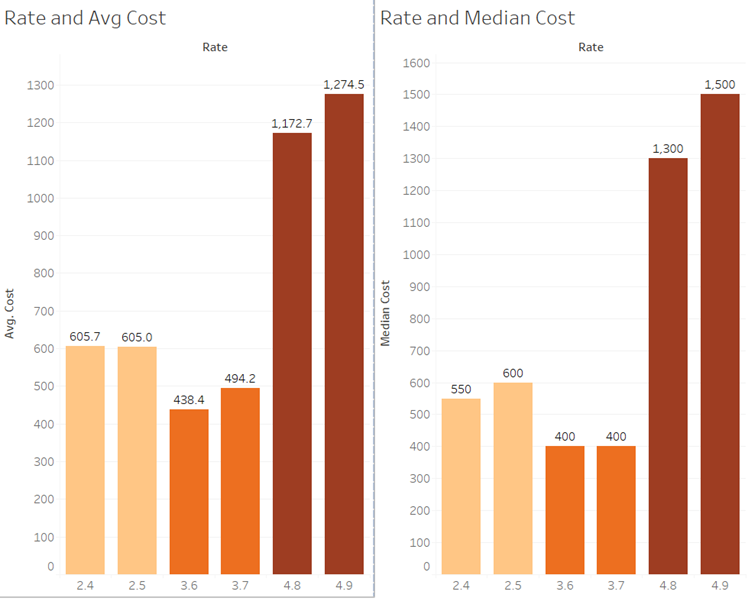
As we learned in Dr. Scott's course, data is categorized into four measurement scales: nominal data, ordinal data, interval data, and ratio data. When we look at the dataset, there are over 15 variables and only votes and costs are ratio data, which are worthy of attention. Reviews is another important variable but difficult analyzing with Excel, Tableau and Jmp 14. It is better to be analyzed with R. Other variables are information.

Rate is my response variable and votes, costs, and reviews are my predictor variables. To keep these columns, firstly I sorted the rate from High to low. I found many of the restaurants doesn't have a rate, so I only selected rows with a rate and copied them into a new Excel file. In the new Excel file, I replaced /5 with space, so I could sort the rate. Deleted all other columns to finish the data cleaning.

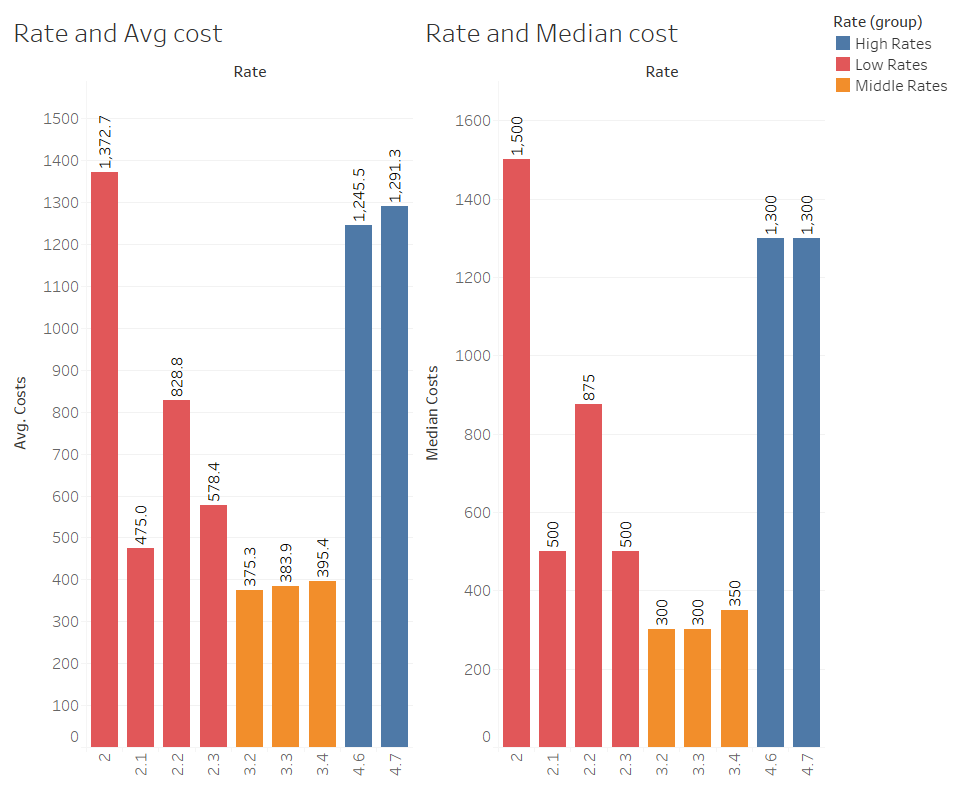
After the cleanness, the next step is to play the new dataset with Tableau. I created two groups of samples to make the analysis more unbiased. In group one, I picked a rate of 4.8 and 4.9 to represent a high rating, 3.6 and 3.7 for middle rating and 2.4 and 2.5 for a low rating. In group two, the rate of 4.6 and 4.7 represent the high rating, 3.2,3.3 and 3.4 for middle rating and 2.0, 2.1, 2.2, and 2.3 for the low rating. The re-sampling is also a procedure of cross-validation technique.

I made two types of bar charts. One is rates and averaged cost, and another is rates and median cost. Same as rates and votes. The reason is the dataset misses lots of data and Excel recognizes the missing data as zero. To avoid biased results, I brought the median into the analysis.

The First Group



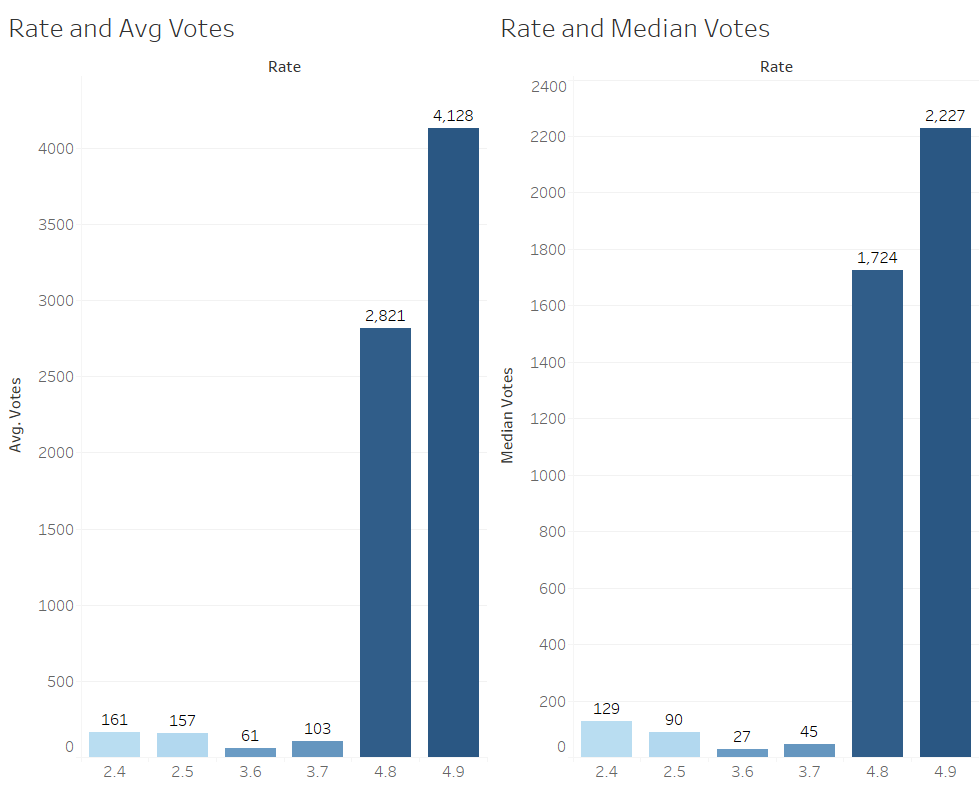
The Second Group



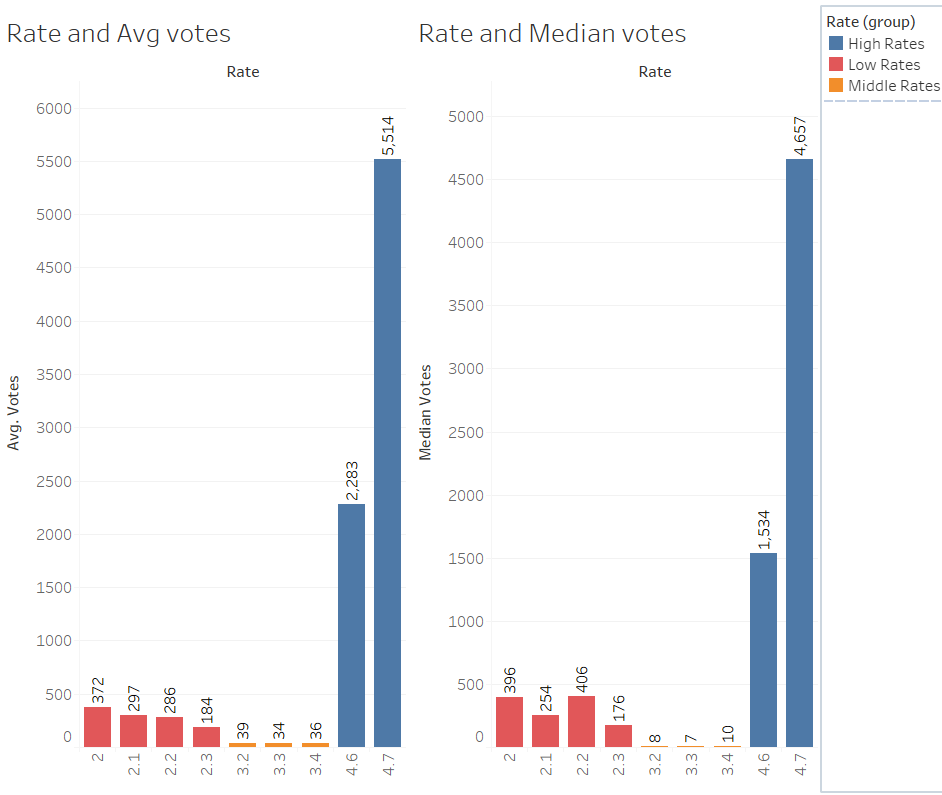
The result I think can give some information. Restaurants with high and low rating where foodies spend more than ones with a middle rating. Foodies spend much more money on high-rates restaurants. It is not difficult to understand this situation because most fine dining restaurants like Michelin starred restaurants, they are very expensive but offering excellent service and food. Foodies wouldn't withhold a good rate. One interesting fact is that foodies cost more in low-rates restaurants than medium ones. The only reasonable explanation is expensive but awful food or service offered by low-rates restaurants.

Same as the relationship between rates and votes. Customers dine in popular restaurants. The more votes restaurants get, the more customers they have. But if the restaurants offer awful food and service, they also will get a lot of votes, mostly complaints.

The First Group



The Second Group

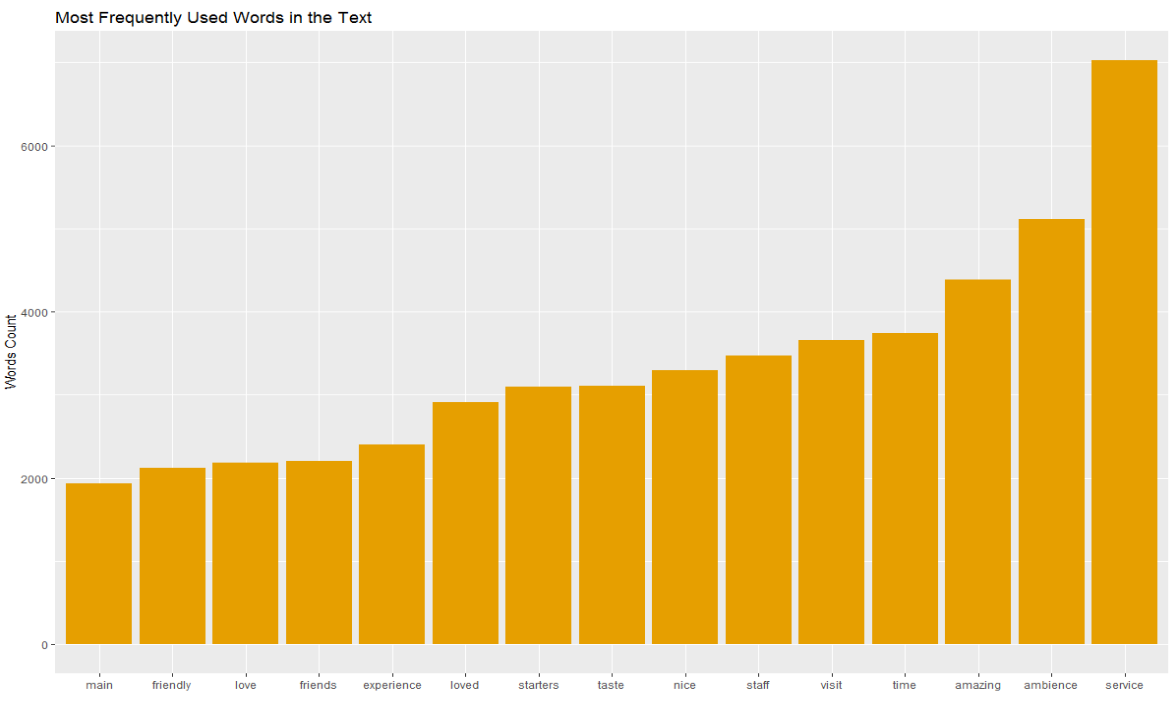


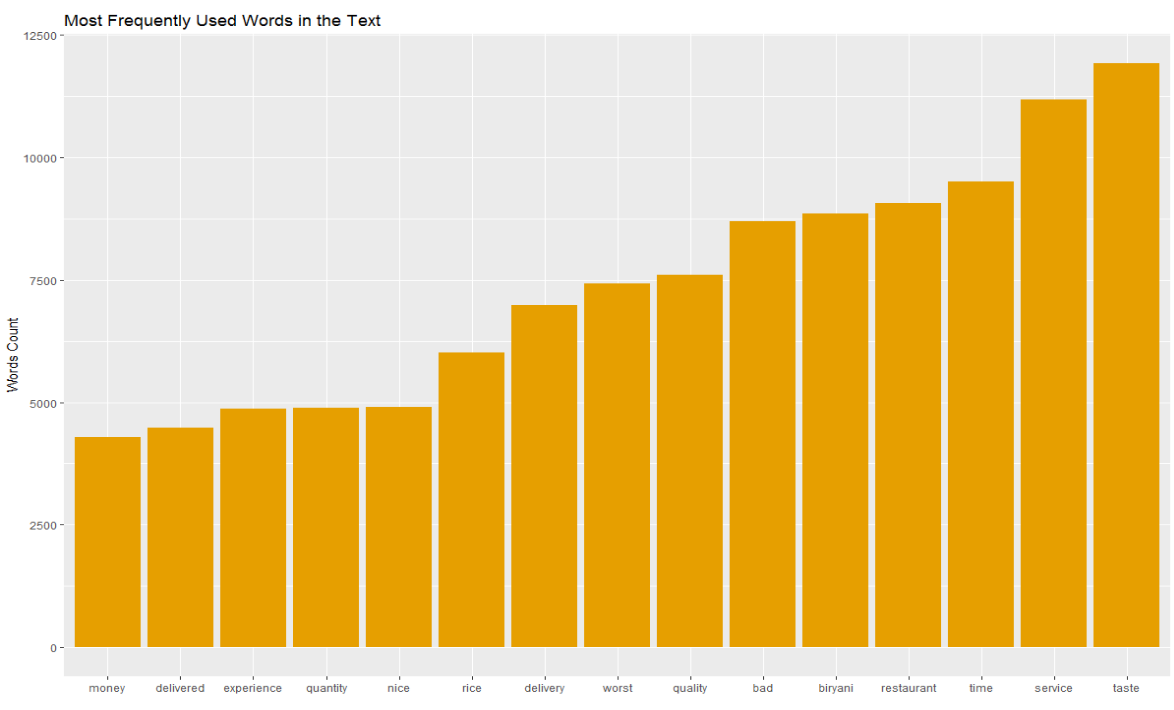
Prescriptive Model Building

The next part is to show foodies' attitudes and opinions toward restaurants, especially the top and low-rating ones. In my analysis, I categorize all rating over 4.7 as the top restaurants and keep all their reviews in an Excel worksheet named High.cvs and reviews of restaurants with a rating of 2.5 or lower, are kept in a worksheet named Low.cvs. In this way, I'm able to do text mining with RStudio. Hopefully, I can take a peek at the core of the reviews and learn why the top restaurants stay high-rating and why the bottom restaurants are not doing well.

The first step is pre-processing. I start the analysis by getting my libraries ready. Then read in the data from my High.csv file into R. There are about 400 restaurants in this file and over 2200 restaurants in Low.csv file. Next, break the text into tokens and transform it into a tidy data structure. Since there are so many undesirable words that can muddy the analysis like food names and drinks, I need to remove them, and stopwords, punctuations, numbers, whitespace and special characters. After the cleanness, the tidy data frames look better. They are ready to get further steps.

To see what the top words in both files are, I count the top words and visualize them.

The Top Restaurants

The Bottom Restaurants

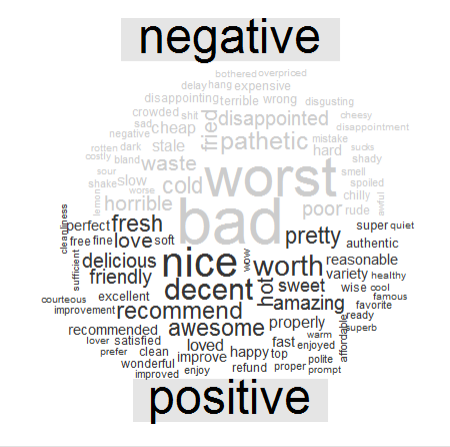
Both datasets talk about service and taste a lot, which is making perfect sense in restaurant reviews. The visualization also gives us some general ideas about the performance of the two categories of restaurants. The top ones have lots of pleasant words like nice, love and friendly, but the bottom ones only have one pleasant word, nice, and have many negative words like bad and worst. However, the top words analysis doesn't tell us a lot and it could be inaccurate. Think about what if a word like bad that is preceded by not? The meaning of the text will get changed.

Before we fix the potential mistake, let's work on the sentiments analysis. This allows us to see what the top positive and negative words are.

The Top Restaurants



The Bottom Restaurants



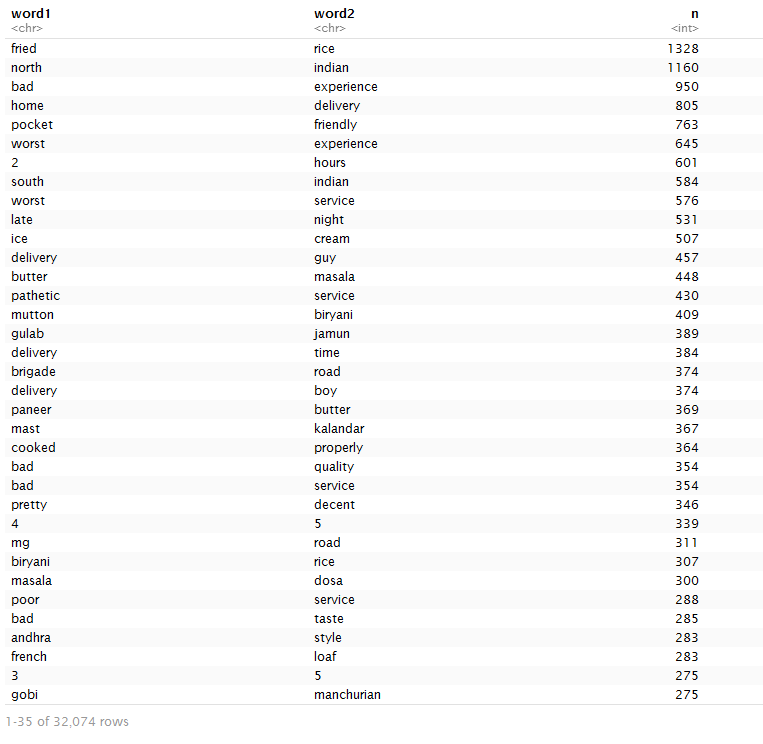
It's clear that the top restaurants have more positive words than negative words and the bottom restaurants have much more negative words in our datasets. Considering after the pre-processing the High.cvs has 425 thousand tokens and the Low.cvs has 860 thousand tokens. Bottom restaurants have a way higher percentage of negative words than the top restaurants.

Now, it's time to fix the potential problem brought by words counting. The method I choose is tokenizing by bigram. It tokenizes text datasets by pairs of adjacent words rather than by individual ones. With count function, we can see the most frequently used pair of words in datasets.

The Top Restaurants

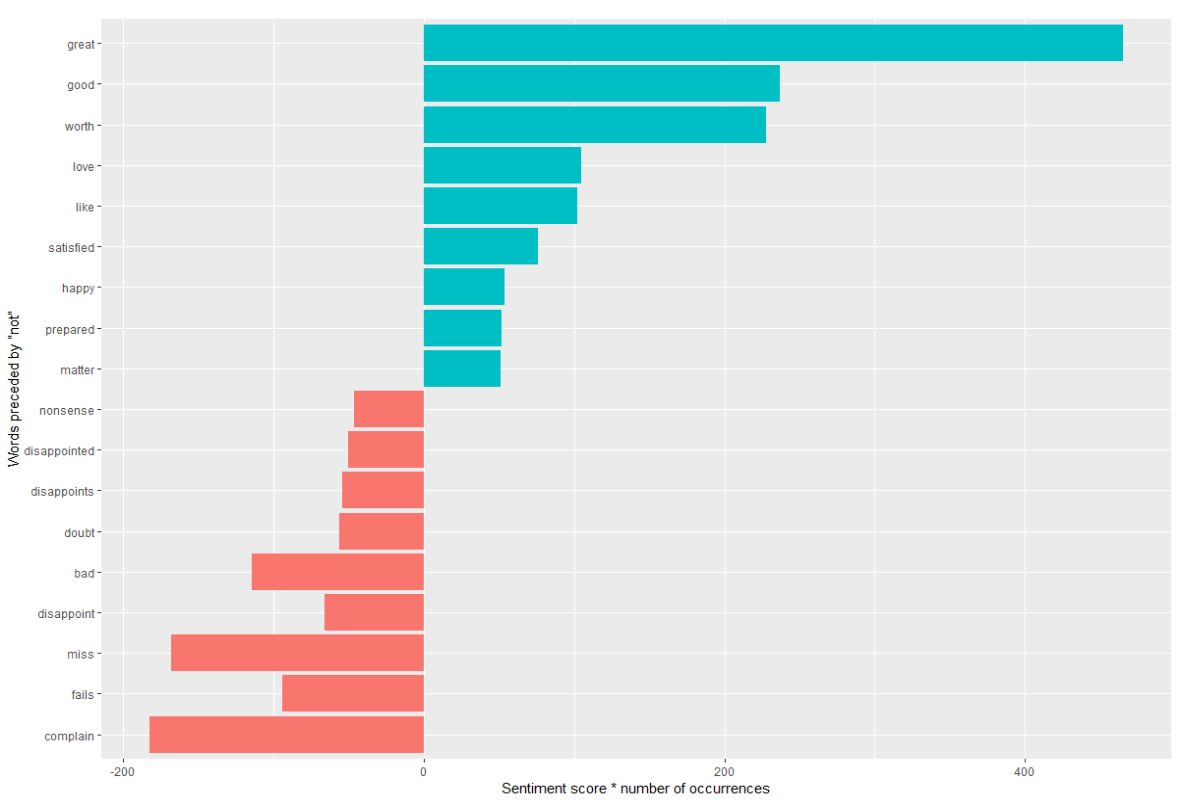


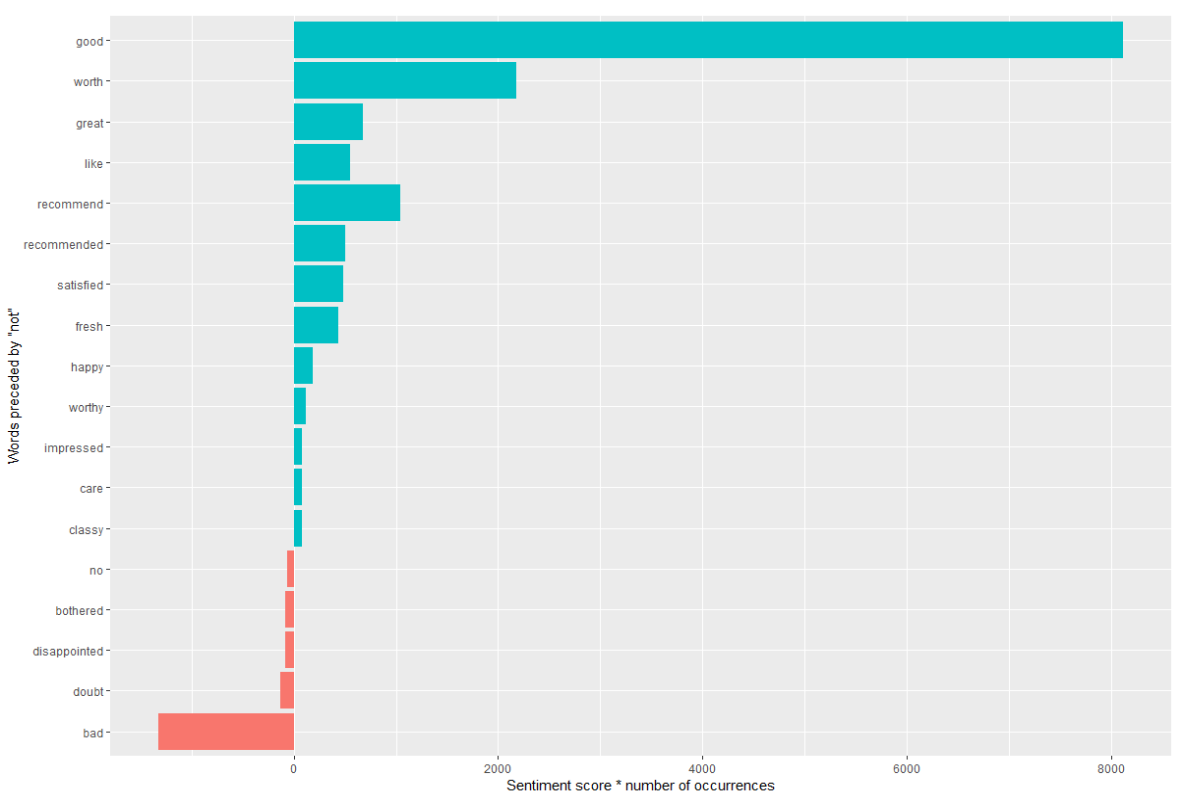
The Bottom Restaurants



We can see the top pairs of the High dataset, like 5 + service, amazing + service, highly + recommend and friendly + staff. These are all the pleasant things that foodies love. On the other side, Low dataset has many unpleasant pairs, such as bad, worst or pathetic + experience, poor + service. It really shows why foodies like the top restaurants and why they don't like the bottom restaurants.

The last step is to see the most frequent words that are preceded by negation words (not, no, never and without). We can identify words which contribute to the most in the "wrong" direction.

The Top Restaurants

The Bottom Restaurants

In High dataset, about 440 great and 220 good are negated. About 190 complain is negated. These are the inaccurate information that shows in our analysis.

In Low dataset, the wrong direction is much worse. Over 8000 good is negated and only 1000 bad goes the wrong way.

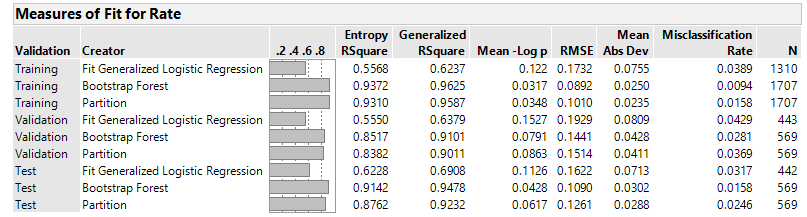
In conclusion, the analysis shows foodies have higher positive opinions and attitudes toward to the top restaurants and they are very disappointed at the bottom restaurants. It sounds no surprise, but text mining is a fun and different way to digging out detailed information from the dataset.

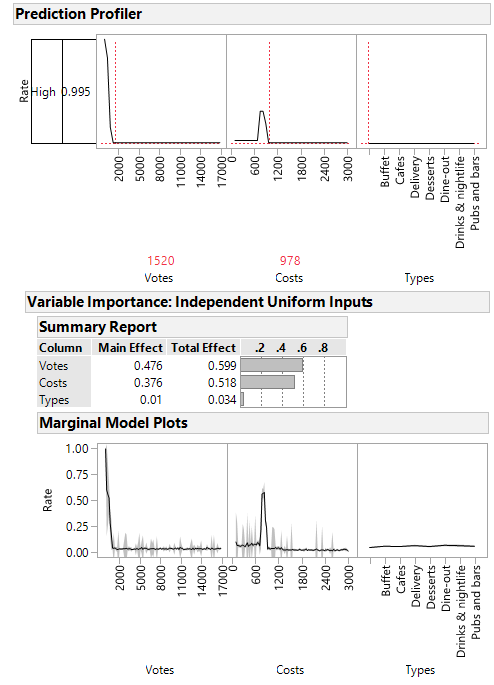
Predictive Model Building

The dataset is the same as the one in the previous analysis. The only difference is I include the types of restaurants, so there are three predictors: votes, costs, and types. My response variable is binary, high or low. In machine learning, many methods have been proposed and the list is growing. The common goal of all these methods is to select the most informative variables to obtain useful predictions. What method I apply to this project are linear (penalized logistic regressions) and nonlinear (random forest and decision trees). The central approach of the methods is cross-validation. It is built on the concept of leaving a part of data out if the estimation process. As the model starts growing, the prediction success of the model is investigated on data that is held out. Estimates are obtained, when the model prediction stops improving.

The first step is to use 6/2/2 principle to split the data into three parts: training, validation, and test. Training is used to build the model. Validation works to determine when the model building should stop, and test is to see how the model's performance in new observations. The decision rule is to choose the model that performs best on the test.

The second step is to build the model with three methods: penalized logistic regressions, random forest, and decision trees. Then, compare the results.



Jmp uses different names for statistical learning methods. Random forest is called bootstrap forest. In this chart, it shows us bootstrap performs the best among the three. It has the biggest RSquare and smallest Misclassification Rate on test data, so random forest is my champion method.

Re-run Jmp with random forest. The model predicts that if a restaurant has over 1500 votes and 900 costs, it has a 99.5% probability to get a high rating on Zomato. Votes and costs have high total effects on the response variable, and the relationships with rates are nonlinear. The variable of types has no influence on the rating of restaurants.

All in all, this project gives both restaurant management a general idea of how to achieve a high rating based on votes, costs, and reviews. Also, Zomato users would figure out a rating range of a new restaurant if they have some information about votes and costs. For restaurant management, getting a high rating could give the restaurant a top position on Zomato list. To realize this goal, based on my analysis, service, food, drinks, and costs are aspects to work on. In a word, try to please the foodies and get them to come back. It sounds cliché but it is the truth that the dataset tries to tell us.