|  |
| --- |
| KPMG |
| Yelp Data Analysis |
| Timothy Yamaguchi |

The data was provided by the client in order to test the accuracy of Yelp’s ability to predict highly rated restaurants. The data itself comes from a data scrape from yelp.com and contains the information of restaurants in Arizona from January 2005 to December 2006. Because of the way Yelp categorizes different businesses, non-restaurants were included in this dataset, such as spas, bars, and clubs, due to the fact that they were tagged as a restaurant on yelp.com. Due to this oversight, it should be noted that additional steps should be taken in order to gain further accuracy in the ability to predict highly rated restaurants. Some of these steps were taken during the data preprocessing phase in Excel and using Python. The steps taken in Excel were: removing location variables such as GPS coordinates, the state the business was located in, the address, etc. the hours the business was open, the dates the reviews were created, general information about the user who made the review, and information about the type of restaurant such as if it was Italian food, American food, etc. In Python, dummy variables were added, Boolean variables were transformed from T/F to 1/0, and missing data points were assumed to be False or 0.

The initial summary of the analysis shows that there is a strong correlation between the dependent variable, the rating of the restaurant, and the remaining independent variables. The mean rating for all restaurants was a 3.74 and the data did not highlight anything out of the ordinary. For example, users who give higher ratings tended to increase the rating of the restaurant, users like restaurants that are nice but not too formal, the more reviews a restaurant receives the higher its rating will become, and the more reviews a restaurant reviews a restaurant receives and the higher the review of that restaurant is then the more like the next user will rate the restaurant high as well. To highlight the last two points, it turns into a chicken-vs-egg problem where does a better restaurant lead to better reviews and therefore more people go to the restaurant or do better reviews lead to more people going to the restaurant creating more reviews for a better restaurant.

A deeper look at the data was taken on specific variables chosen by the variables that had the greatest impact on the model. The additional regressions that were run were the users’ average rating of the restaurant, the dress code of the restaurant, the number of reviews for a restaurant, the price range for the restaurant, if the restaurant accepts credit cards, and a selection of variables chosen by the analyst. The users’ average rating regression was run to compare the impact of a user’s average rating they give businesses as a whole compared to the restaurant’s rating. The analysis shows that the higher a user’s average rating is the more likely they were to give a higher rated review for a restaurant. Therefore, one can conclude that a person is more likely to give a high rating to one restaurant if they give high ratings to other restaurants. The dress code regression examined how the dress code of the restaurant would affect its rating. The analysis showed that the dressier a restaurant is, the rating will be higher, however, at a certain point, a restaurant becomes too formal and the ratings go down. One can conclude that although people value nice restaurants, if they are too nice, ratings tend to go down because the expectations for those restaurants go up. The number of reviews regression was run to examine how the number of reviews a restaurant had correlated with the overall rating of that restaurant. The analysis shows that the greater number of reviews a restaurant has, the more likely it is to have a higher overall rating, which means the more reviews a restaurant has means more people have eaten there and if more people have eaten at the restaurant the rating should increase. The price range regression was run to examine how the price range of the restaurant affects its rating. The analysis shows that the more expensive the restaurant, the more likely it is to have a better rating. However, because this dataset is not sorted by types of restaurants, the analysis is incomplete to determine if the cost of dining at a specific restaurant affects its overall rating. For example, an expensive dive bar will rate very low while an expensive steak house may rate very high. The types of restaurants should be sorted at a later time for a full analysis. The credit card regression was run to examine the impact of a restaurant accepting credit cards and the effect on its rating. The analysis shows that if a restaurant accepts credit cards, the more likely it will have a higher rating. However, due to a low correlation, it should be easy to determine that accepting credit cards does not have a large effect on the rating of a restaurant. Rather, if a restaurant does not accept credit cards it will lower the rating. The personal choice regression was run to examine the impact of the variables that were important to the analysis when choosing a restaurant. The analysis shows that while the analyst places an importance on happy hour and having a TV, they are less favorable qualities of a restaurant, while accepting credit cards, serving just beer and wine, and not allowing smoking are more favorable qualities of a restaurant that lead to higher ratings.

The analysis and dataset does come without some caveats. There were problems with the data. There were many missing data points that had to be assumed when filled. There was unavoidable multicolinearity in the data due to the fact that places that were good for dancing are usually not good for children. There was also a duplicate column in the data that had the same exact name but different datapoints. And there was the fact that the dataset was not sorted by the type of restaurant rather, the information was run on all restaurants in the dataset. This included fast-food to casual bars to high-class restaurants.