Khang Thai, David Park, Jonathan Ho, David Favela

kNN will check the data surrounding the new data to determine what value it has on the line of best fit. Linear Regression will take in all the data and then find the line of best fit while kNN is adding more data overtime and changing the line. Generally kNN wants fewer predictors when trying to find the correlation and in our example of diamonds, this is shown because when we did regular kNN we only got a 94% correlation compared to Linear Regression's 98%. After scaling the data, kNN achieves a higher prediction compared to the linear regression. When we did the Decision Trees for Linear Regression, it did very little when trying to predict the correlation. The correlation for a Decision Tree showed us that it was better than the linear regression even by a small margin, it was able to achieve a higher correlation. When pruning the tree however, the correlation dropped below the linear regression correlation.

For Logistic Classification, our base accuracy was around 73.1%. When we did kNN classification, we were able to achieve about a 73.3% accuracy which was better than the base Logistic accuracy. kNN showed us that it was useful for our data even if the increase is small. When we used decision trees, we saw a decrease in the correlation between the accuracy and the accuracy from the base logistic. Based on these results, it seems that the rating mean was not able to distinguish a big enough difference between satisfied and dissatisfied based on the rating.

K-means clustering, hierarchical, and model-based clustering are all different types of clustering algorithms. K-means clustering finds k centers (centroids) within the data, and groups other observations based on nearness to the closest centroid. It first starts with a random assignment knowing how many clusters there are, then it loops and assigns each observation to its closest centroid and recalculates the centroids. Observation assignment and recalculation occurs after every iteration until convergence occurs.

Hierarchical clustering uses a calculation that measures distance between pairs of observations, and combines these clusters into a hierarchy. Unlike k-means clustering,

hierarchical clustering does not require knowing how many clusters there are. In hierarchical clustering, each observation is placed into its own cluster, distance between each cluster and every other cluster is calculated, and the two closest clusters are combined. Distance and combinations keep occurring until all clusters merge into one cluster.

The last clustering algorithm looked at is model-based clustering. Model-based clustering takes a statistical approach into clustering the data. It assumes that data can be derived from a combination of probability distribution to try and further improve the fit between data and a mathematical model. This type of clustering is almost an extension of k-means since it is similar, however, it adds a component of weight to be able to assign data points to clusters and generate new means. The two steps model-based clustering utilizes expectation and maximization. The expectation step assigns data points to clusters based on probability while the maximization steps estimates the parameters of the model.

PCA and LDA were supposed to be used to take out unnecessary data, however, we already took out most of our unnecessary data. In most cases, this would be useful because taking out data that is irrelevant will allow the prediction to be higher because some predictors might influence the prediction causing it to be less accurate. In our data, PCA and LDA gave less accuracy than logistic regression. This can be due to our custom reduction of the data beforehand, as PCA's and LDA's purpose is to reduce the dataset into just the most meaningful or significant data. PCA and LDA differ in that PCA is unsupervised, while LDA is supervised meaning PCA does not use labels. In addition, LDA has a pre-processing step that calculates the mean vectors from the labels since it is supervised.