This was a challenging assignment. I spent a great deal of time in the data preparation and exploratory data analysis phases. Part of that was due to the sheer sizes of the datasets we were working with. I also spent a lot of time trying to find the best way to downsize the train dataset. Initially, I just took the first 200,000 rows as my sample. As I thought about it more, I realized that this approach might not be the best one. Depending on how the train.csv data were organized, by simply taking the first 200,000 rows, I could be excluding data further down in the set that might be important. Therefore, I changed my approach to taking a random sample of 1% of the data. This represented a more complete downsizing approach that still provided a smaller test sample to work with.

I also used some graph analysis to help me decide on the best modeling approaches to use. A histogram of the hotel_cluster variable indicated that there was a good distribution across all 100 clusters. Because of this type of distribution, I didn't expect to get great prediction accuracy since most modeling methods perform best when data follows a normal distribution.

My second graph looked at the correlation matrix for all of the variables. I was looking for any linear dependencies between variables. The heat map, however, showed that there were no strong correlations between variables. What this suggested is that a linear-type of prediction model might not be the best method for this type of data.

When it finally came time for algorithm selection, I ended up choosing the following methods: KNN, logistic regression, random forest, and naïve Bayes.

KNN: I chose this as my first method because it was one I was familiar with. KNN is a non-parametric model, so I thought it would do well for our non-normally distributed data. The intent was to teach the model to predict the correct hotel cluster based on what the number of nearest neighbors. The accuracy of this method was 25.6%. This was not too bad for our first attempt. One advantage of this method was that it was fairly quick to run. We could likely improve our accuracy by adjusting our k-value. I just selected the default value of 5.

Logistic Regression: As mentioned earlier, our correlation matrix suggested that a linear prediction model might not be the best choice for our data. While logistic regression is considered a linear method, it performs better than a linear model when the dependent variable is categorical. In our case, the category is essentially either a yes or a no as to whether a specific hotel will belong to a certain cluster. This method took a very long time (over an hour) to calculate, but the end result showed an accuracy of about 30.4 %. A fair improvement over the KNN model. If we wanted to try and improve the accuracy, we could try changing the multi_class setting as well as adjust the number of folds used in the cross validation.

Random Forest: This is also a non-parametric classification method and I assumed the accuracy value would be close to what I saw with the KNN model. Because this type of method is prone to overfitting, I made sure to include the cross validation. This method calculated quickly and returned an accuracy score of 25.0 %, which was very close to the KNN value. There are a number of parameters associated with the random forest method that could be adjusted in attempts to improve accuracy, but I only looked at one iteration.

Naïve Bayes: I specifically chose to use this method because my boss at work is really interested in Bayesian methods and I am trying to learn more about them. I had also heard that this method is extremely quick to evaluate and so I was interested to see if that was indeed the case. It turns out that this method was by far the quickest to return an answer. Unfortunately, it also performed the worst of the four methods by giving an accuracy of only 10.3 %. I think I would be able to improve the accuracy by adjusting the number of folds used in the cross validation. Other than that, however, there doesn’t seem to be many other parameters that can be adjusted. This suggests that the naïve Bayes method is probably not ideal for this kind of data.