Hotel Recommendations Project | Lara Clasen

For this assignment we want to predict which hotel cluster or group that a user will select based upon their search characteristics. After cleaning up the data a bit, there were ultimately multiple algorithms that could have been applied to the data, but I used the K-Nearest Neighbors algorithm for this assignment. This is an algorithm that we have learned about quite a bit lately and is becoming more familiar to me. Our data set is a multi-class set without any inherent pattern. Because of the randomness of real-world data, it requires quite a bit of honing and curation in order to apply this data to an algorithm and produce insights. I feel that Multinomial Logistic Regression would have provided the most accurate results because of the data structure, but my program had a difficult completing that as a task.

To begin we can see that nearly all of our fields are numeric, which is great. I first reduced the total size of the data set that I would be addressing since it is a large set. I used 1% of the total rows, to result in 241,179 rows, and read that amount of data in. An initial visualization of the data gives us a good look at the clusters and their variation. Manipulating a few of our fields helped to be able to more clearly observe the data. For example, the check-in and check-out date columns were adjusted so that only the year was utilized. Once we have adjusted the applicable features and removed features that are not valuable for our analysis, we can start to take a look at how things are correlated. Specifically, we want to see if there are any variables that correlate strongly with our hotel\_cluster variable. Since none of our variables appear to correlate in a linear fashion with hotel\_cluster, we can determine that linear regression is not ideal for our analysis. When we run the K-Nearest Neighbors classifier on our data, we get a result of 0.267. This gives us a relatively weak classifier for our problem, so we would likely want to run a Multinomial Logistic Regression algorithm on the same data in order to obtain the most accurate prediction results.