Pre-processing and Training Data Summary

The first model of the average price data for an adult weekend ticket was to simply find the mean. The mean was $63.81. Splitting the test and training data with a 70-30 split, the mean absolute error of the test data was around $19. This means the model (guessing the mean) could be off by $19. Then, empty cells were imputed with the median, but this has little effect on the data compared to imputing the empty cells with the mean. A linear model with all of the cleaned data’s features seems to be a poor fit, so the model was improved with sklearn’s pipeline features.

Sklearn features a function called GridSearchCV that looks for the best number of features (k) to keep in the model. GridSearchCV returned a graph in 4.9.8 that shows a best k would be 8. Linear model coefficients show which features to choose. Some of the top features include: vertical\_drop, fastQuads, fastSixes, and averageSnowfall. NightSkiing\_ac is the number 13 feature on the list. I feel that I need to revise my hypothesis based on this information.

Sklearn also offers random forest modeling, which was cross-validated after the pipeline was coded. The best\_params\_ attribute returned 88 estimators, simpleimputer\_\_strategy = median, and standardscaler = None. This means the best model for random forest has 88 trees, imputed medians, and no scaling of the variables. The results of the random forest are different, as the top four features are: fastQuads, Runs, SnowMaking\_ac, and vertical\_drop. The repeat in the fastQuads in both the improved linear model and the sklearn forest model may indicate a new desired target.

The linear regression model yields a mean absolute error of $9.90 and a standard deviation of $1.56. The random forest yields a lower mean absolute error to the improved linear model by about $1 and a lower standard deviation (lower variance too!). Thus, the sklearn random forest model was chosen as the best model moving forward. Also, a learning\_curve function was called to check if more data for training would be necessary, but the optimal training set curve starts to flatten at around 40-50, which is significantly less than the 193 used. No additional data is needed for training.