Write a summary of the work in this notebook. Capture the fact that you gained a baseline idea of performance by simply taking the average price and how well that did. Then highlight that you built a linear model and the features that found. Comment on the estimate of its performance from cross-validation and whether its performance on the test split was consistent with this estimate. Also highlight that a random forest regressor was tried, what preprocessing steps were found to be best, and again what its estimated performance via cross-validation was and whether its performance on the test set was consistent with that. State which model you have decided to use going forwards and why. This summary should provide a quick overview for someone wanting to know quickly why the given model was chosen for the next part of the business problem to help guide important business decisions.

Did an initial baseline modeling of the ski resort data by taking the mean of the ticket price to see if it can be used as predictor for price tickets. This was implemented using sklearn’s DummyRegressor model. With the use of MAE provided that on average using average ticket price as a predictor would result on guesses being off by ~$19. The next step was to then impute missing values that were discovered during EDA. Multiple strategies were applied for imputing the these missing values such as using the median, mean. Both strategies provided similar results but ultimately brought down the MAE to ~$8.54. This downward trend in error metrics is good, however, this may be due to overfitting of the model given the number of features used. The next step is hyperparameter tuning of the LinearRegression model utilizing ‘SelectKBest’ to picks the top features to be used when generating a model for predicting ticket pricing. To maximize on the optimal number of features utilized for the best results GridSearchCV was utilized to provide cross validation on the data set used to train the LinearRegression model along with a list of grid params providing range of values to provide for number of features to test against. This optimal number of features that resulted from this test was 8 features. Another model utilized for the predicting ticket price was RandomForest. RandomForest when fine-tuned compared to LinearRegression provided better accuracy and ultimately was chosen as the model to be used for predicting ticket pricing.