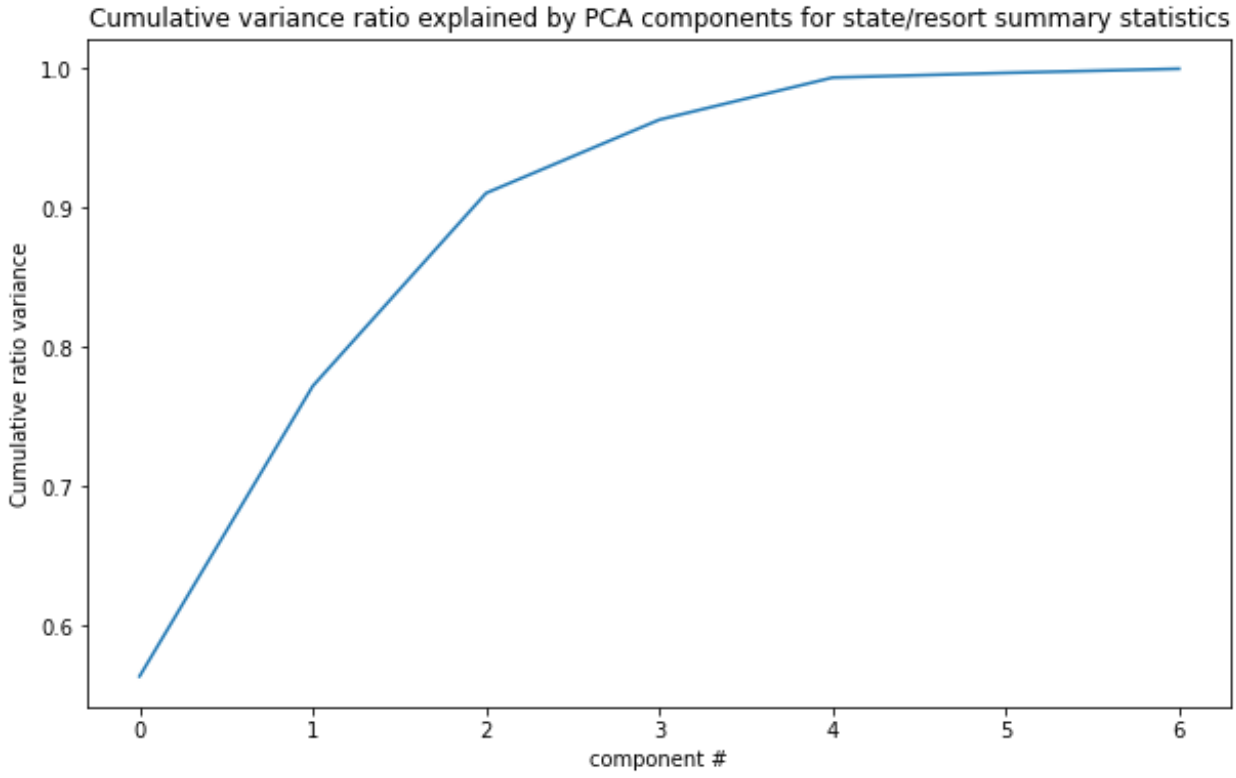
Project report

Data wrangling

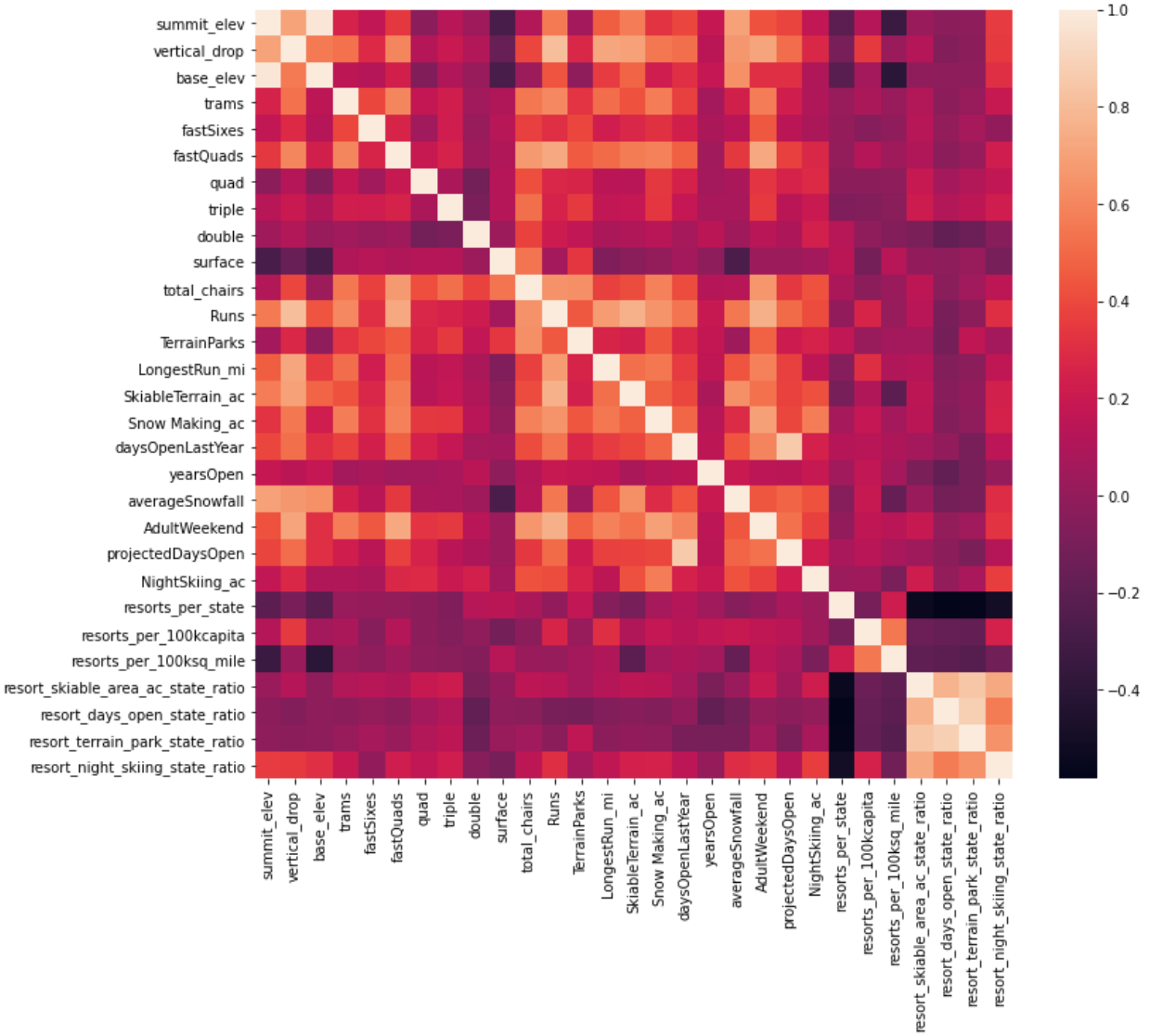
The first step was to clean and explore the dataset. The resort of interest is the Big Mountain ski resort and the goal is to increase the revenue. In the data wrangling phase I cleaned the missing values and got of rid of unnecessary values. I also looked at how the Big Mountain Ski resort compared with other resorts.

Exploratory Data Analysis

In the exploratory Data Analysis phase Principal Component Analysis is used to identify clusters of similar features and study the correlation to visualize the high dimensional data. As a result, the first two components seemed to account for over 75% variance and first four principal components account for 95% of variance.



We can also look for



correlation by a correlation map.

Preprocessing and Training

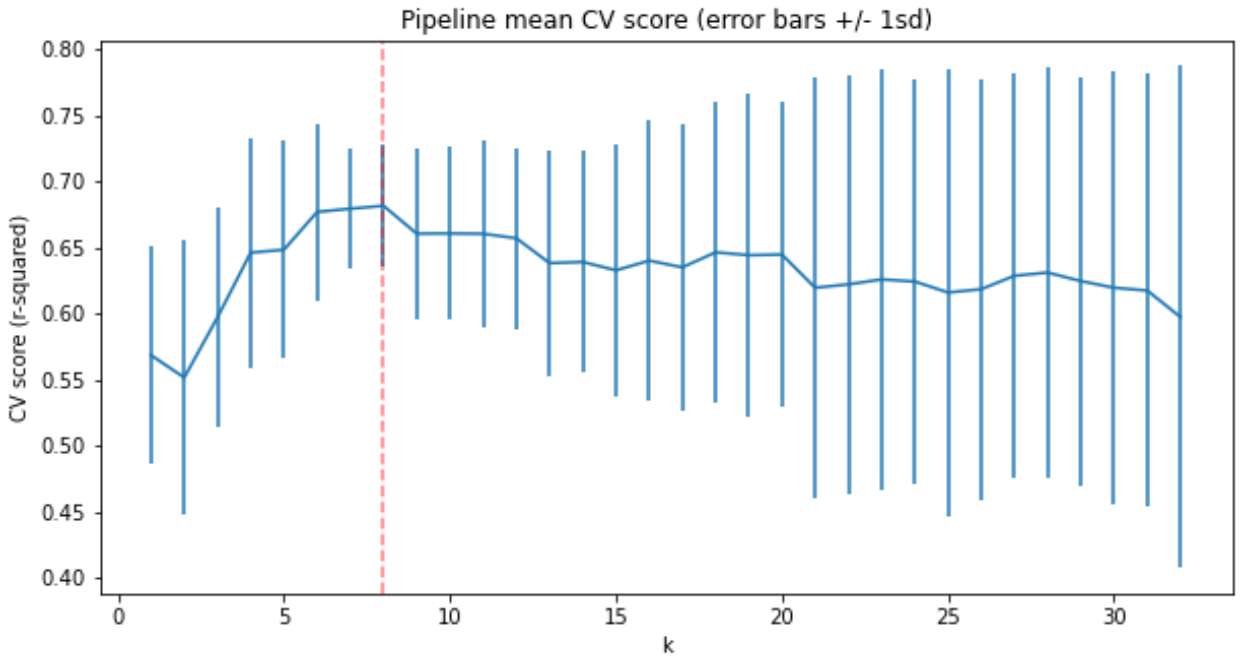
Start by taking the mean value as the predictor by using the dummy regressor. The results are compared by using R squared. The R-squared for the mean will be 0. The mean absolute error and the mean squared error for the training set are 19.136 and 581.436 respectively.

Next step is to impute missing values using the median and the R squared for the median training set is 0.72 i.e., the median explains 72% of the variance. The mean absolute error and the mean squared error for the training set are 9.41 and 161.73 respectively.

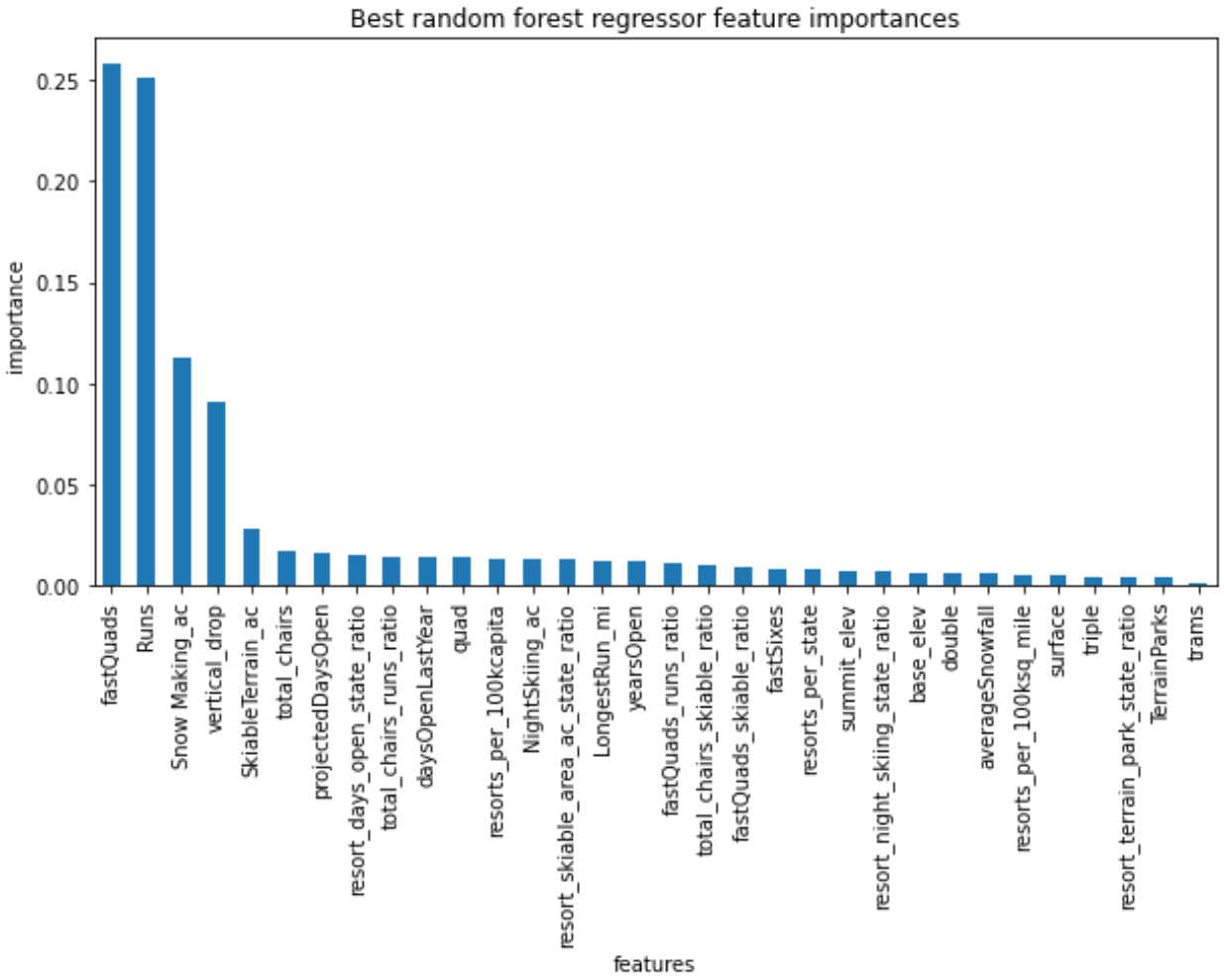
After that we impute the missing values using the mean and the R squared for the mean training set is 0.716 i.e., the median explains 71% of the variance. The mean absolute error and the mean squared error for the training set are 9.42 and 164.39 respectively.

We then apply the linear regression algorithm to the model and the r squared, mae and mse are 0.63, 10.48.

Using cross validation at this point to determine the number of features is a good idea and the best value of k comes out to be 8. Therefore, we use 8 features for prediction.



We then apply Random Forest and apply feature importance in random forest. The random forest model has a lower cross-validation mean absolute error by almost \$1. It also exhibits less variability. Verifying performance on the test set produces performance consistent with the cross-validation results.



We then see different scenarios the increase the revenue. To conclude the revenue will be highest if we add a run, increase the vertical drop by 150 ft, and install an additional chair lift. This increases the ticket price by 8.61$ and the total revenue by $15065471