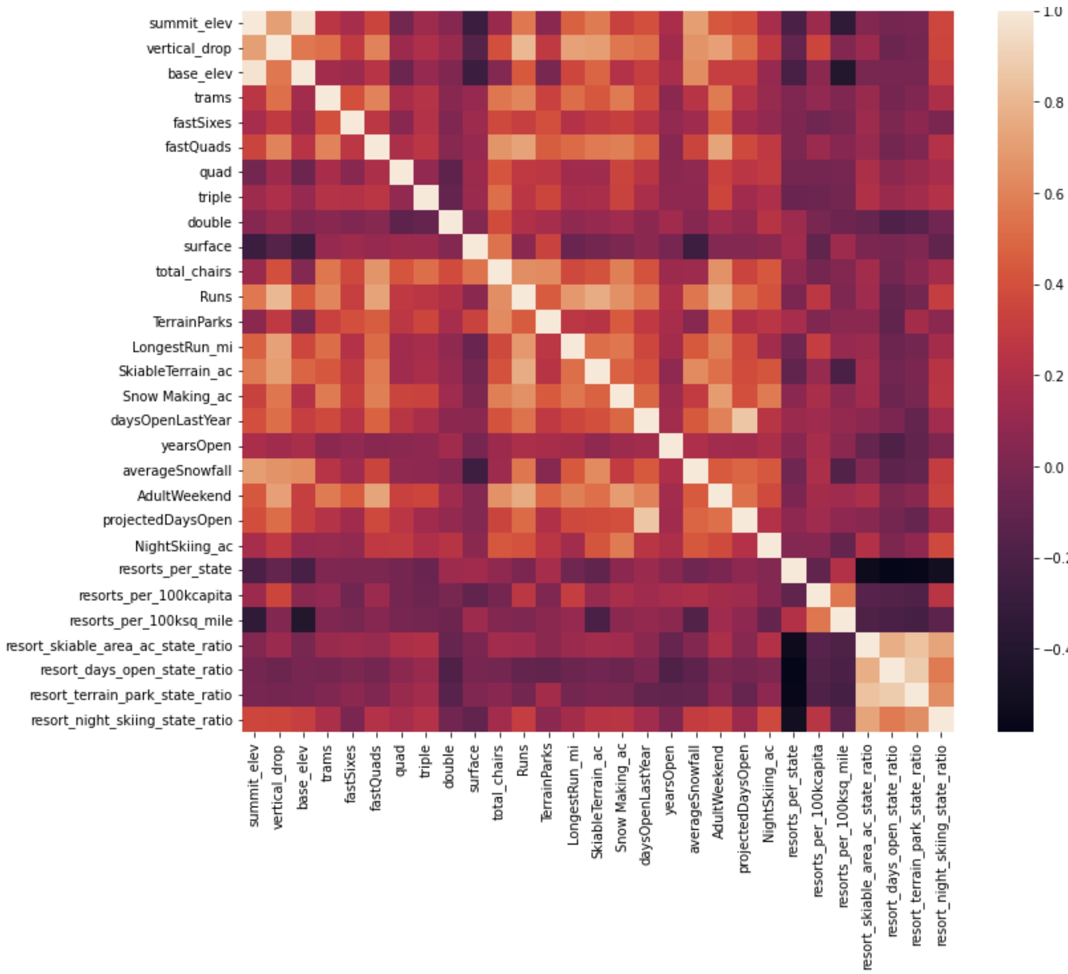
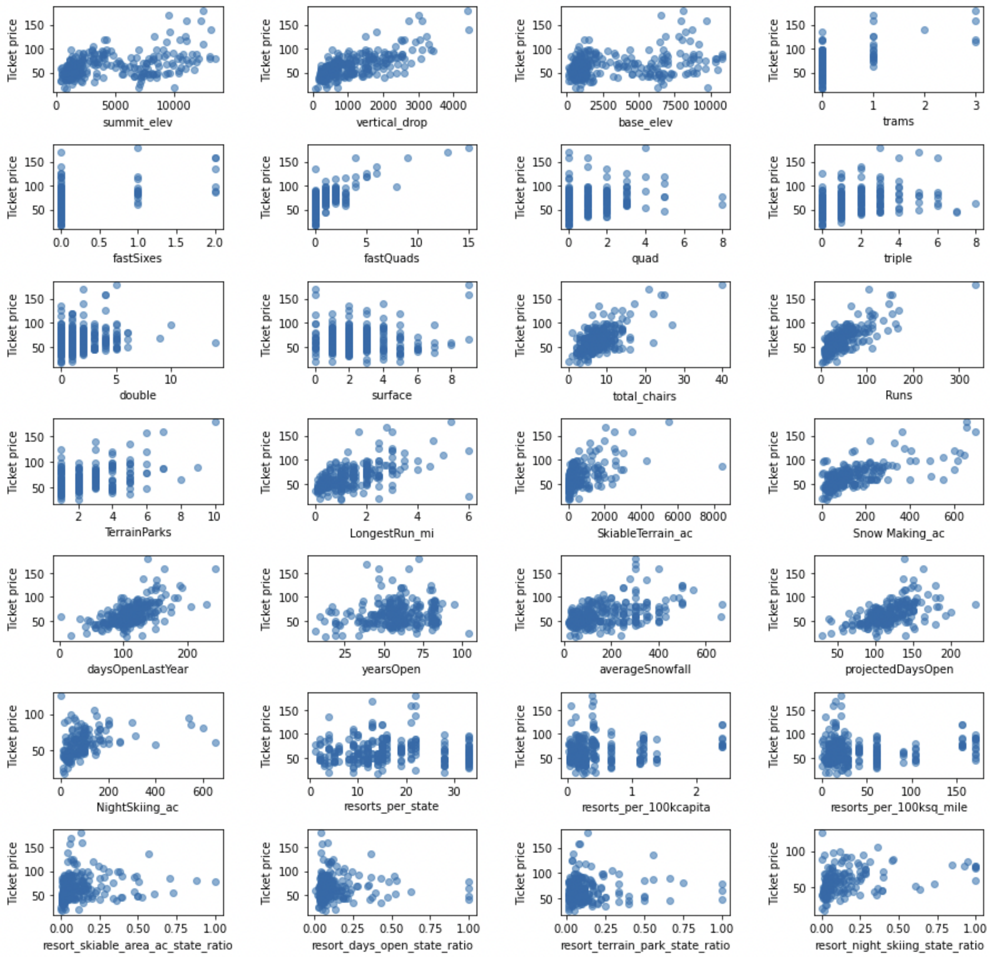
Guided Capstone Project Report

Big Mountain Resort, a ski resort located in Montana offers spectacular views of Glacier National Park and Flathead National Forest, with access to 105 trails. Every year about 350,000 people ski or snowboard at Big Mountain, it can accommodate skiers and riders of all levels and abilities with the service of 11 lifts, 2 T-bars, and 1 magic carpet. The longest run Hellfire is 3.3 miles in length, with a base elevation of 4,464ft, a summit of 6,817ft, and a vertical drop of 2,353ft. Big Mountain Resort has recently installed an additional lift chair to help increase the distribution of visitors, however, the chair increases their operating cost by $1,5400,000 this season. In response to the cost change, the resort's pricing team has decided to charge a premium price for the tickets. The price increase criteria will be analyzed from collected data of other resorts across the country to maximize the profit margin.

The dataset received from Alesha, the Database Manager, presents data for 330 ski resorts in different cities and states, including Big Mountain Resort. The data provides information such as weekday ticket prices, weekend prices, vertical_drop, runs, snow-making covered areas and etc that could be used to analyze and predict ticket prices in the future steps. Although Big Mountain Resort does not have any missing key values, there are missing data in other resorts that could potentially affect the analysis. FastEight was the first column removed due to over 50% of the values being absent and the rest being 0. AdultWeekday column was also dropped due to having more missing values than AudultWeekend, and the price for both are the same in Montana, where Big Mountain Resort is located. Rows that have a missing value for both AdultWeekend and AdultWeekdays are dropped. There are two other issues found, first is that Silverton Mountain resort's skiable was unrealistically bit, so it has been updated with correct information; the second is that one resort has the year's open value of 2019, which is also unrealistic, the row has been dropped. After all the processes above, the data now has 277 rows left.

The ski data and state summary data were explored and resort density per capita and area were calculated. Then Principale component analysis (PCA) was performed to find linear combinations of the original features that are uncorrelated with one another and order them by the amount of variance. The categorical features were names of the resort, its region, and state, the numerical features are numbers of chairs, vertical_drop height, number of runs and etc. Based on seaborn heatmap of correlations, fastQuad, Runs and Snow Making_ac were observed to have correlations with the ticket price; based on scatterplots of price with other numeric features, the vertical drop has a strong positive correlation with the ticket price. These features that were observed should be used for Big Mountain Resort ticket price modeling.





The first model is a baseline performance comparator for any subsequent model, a 70/30 train/test split was first performed on the ski data. Three different metrics are used: R-squared, mean absolute error, and mean squared error. The R-squared on the test set is - 0.00312, it is expected to be negative since performance on the test set generally is slightly worse than on the training set (in this case 0). The mean absolute error is higher on the test set than on the training set, but the test set's mean squared error performs better than the training set. The initial model started with imputing missing feature values, where both median and mean methods were used. The imputation was applied to both train and test splits and the data were scaled in order for them to be on a consistent scale. Those data were used to train the linear regression model to make predictions. The R-squared performance was around 0.8 for both median and mean impute on the training set and around 0.7 for both test sets. MAE and MSE were similar for both the mean and median filling methods for the training set and test set. This means that the result would be similar no matter whether using mean or median to fill missing values. In order to verify if the model was overfitting, the SeletKBest function was used to select k best features and the score function will be f_regression. After using the cross-validation technique to estimate model performance, GridSearchCV was used for hyperparameter finding that the best k is 8. After displaying these eight features, we can observe that vertical drop has the biggest positive value, and the area covered by snow-making equipment is also strongly positive but trams and skiable terrain are negatively associated with the ticket price. A random forest model was performed to test if it fits in this case, cross-validation will use the default setting so we can investigate different hyperparameters. In this model, the dominant features are fastQuads, runs, the area covered by snow-making machines, and vertical drops. By comparing its performance with the linear regression model, the random forest model has lower cross-validation mean absolute error by almost $1 and exhibits less variability so it would be the model I have decided to use going forward.

Big Mountain currently charges USD81,00, the modeling suggests a ticket price of USD95.87 that could be supported in the marketplace by Big Mountain's facilities. There is room to increase the ticket value even considering the expected mean absolute error of USD10.39. The resort has intentions to either cut costs or increase revenues, so four scenarios were created: 1. Cutting cost by closing down 10 least used runs, which by closing down one makes no difference but closing more than two would affect the ticket price therefore decreasing the revenue; 2. Increase revenue using vertical drop by adding a run to a point 150 feet lower down but requiring the installation of an additional chair lift to bring skiers back up, this scenario increases the ticket price by USD1.99; 3. Same as a scenario but adding 2 acres of snowmaking, the ticket price increase remains USD1.99, which means adding 2 acres of snowmaking isn't enough to affect revenue; 4. Increase revenue by increasing the longest run by .2 miles and guaranteeing its snow coverage by adding 4 acres of snowmaking, the result shows no difference in revenue for this scenario