

# Guided Capstone Report

What opportunities exist for Big Mountain to recoup the increased operational cost of \$1.54 million over the next year through modification of pricing or optimization of importance of their facilities?

## Introduction:

Big Mountain recently installed a new chair lift to increase the distribution of visitors on the mountain. It is likely that the resort's premium pricing is not being maximized. The goal is to create a machine learning model to predict the price Big Mountain should charge depending their place in the market

## Data Wrangling:

The provided data contains 26 descriptive features of 328 unique ski resorts. Some of these features are name , region, state, summit elevation as well as adult weekday price and adult weekend price. The last two being potential target features.

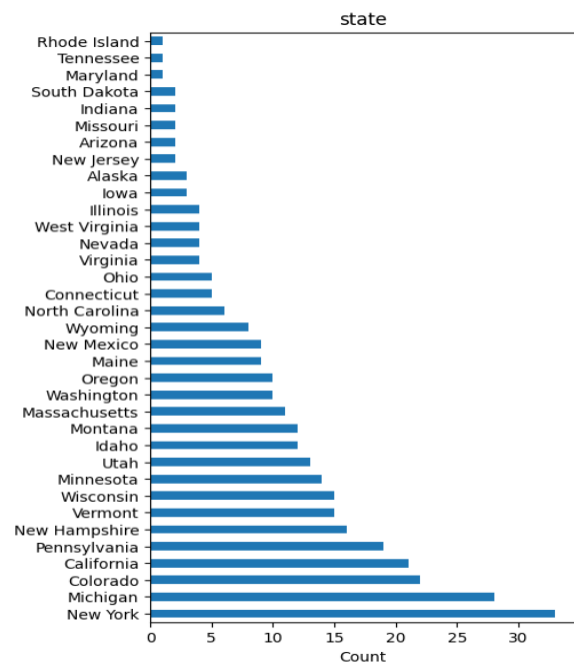
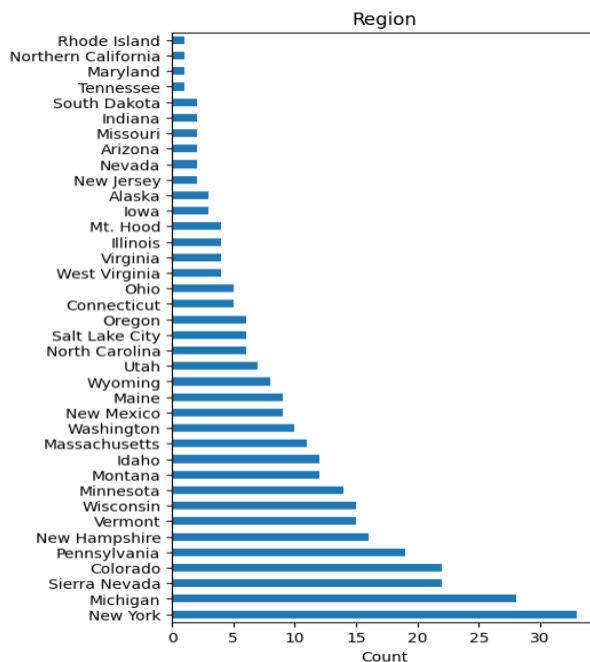
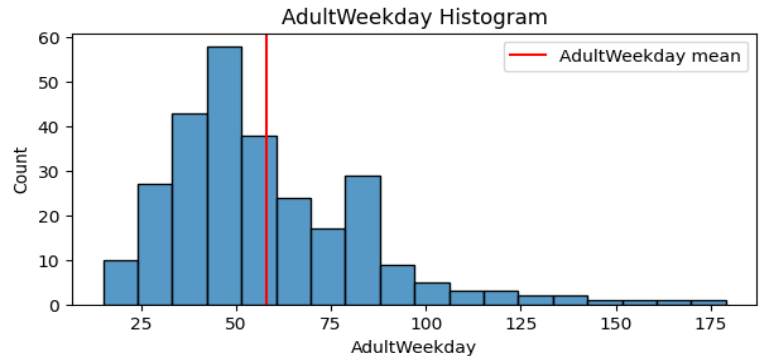
There were a few rows that did not include any price data and were therefore dropped. The fastEight feature contained heavily imbalanced classes and would not provide any useful insight so it was also dropped. The yearsOpen feature contained a value of 2019 that was most likely an input error that was corrected to match the remaining data of the column.

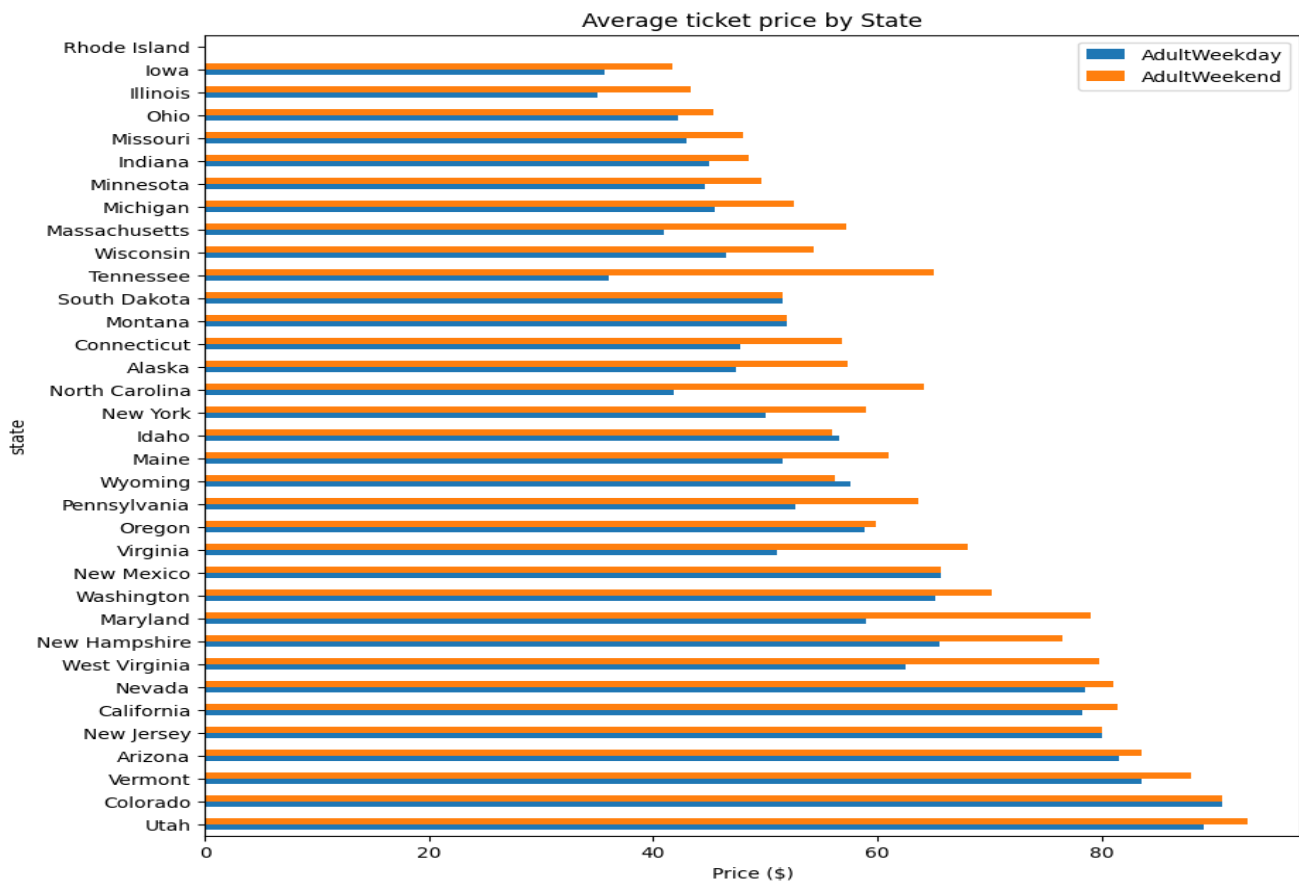
Data about state population was acquired from Wikipedia to create features such as resorts\_per\_state, resorts\_per\_100kcapita, resorts\_per\_100ksq\_mile, resort\_skiable\_area\_ac\_state\_ratio, resort\_days\_open\_state\_ratio, resort\_terrain\_park\_state\_ratio, and resort\_night\_skiing\_state\_ratio.

## Exploratory Data Analysis:

There were a few notable discoveries found while exploring each column. Firstly, the target feature of AdultWeekend is right-skewed with an average of 64.28.

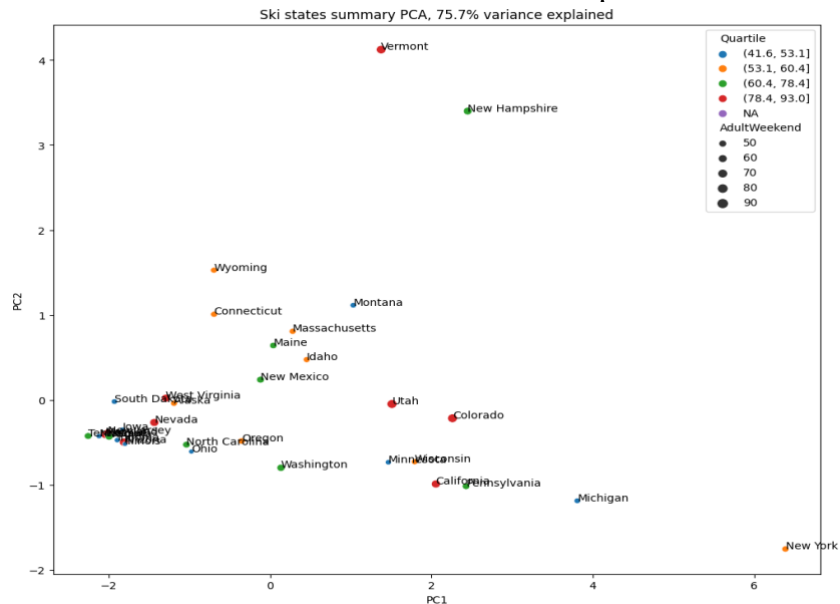
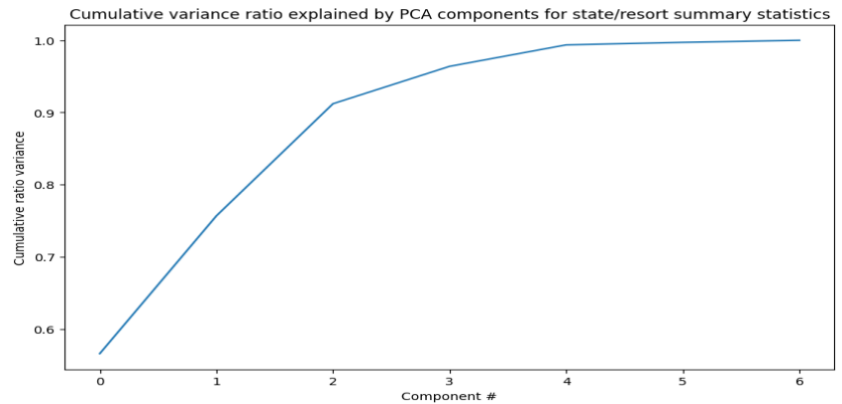
There was also a difference in the number of resorts per state compared to number of resorts per region. The New York, Michigan, Sierra Nevada and Colorado regions had the highest number of resorts while New York, Michigan, Colorado and California were the states with the most amount of resorts.



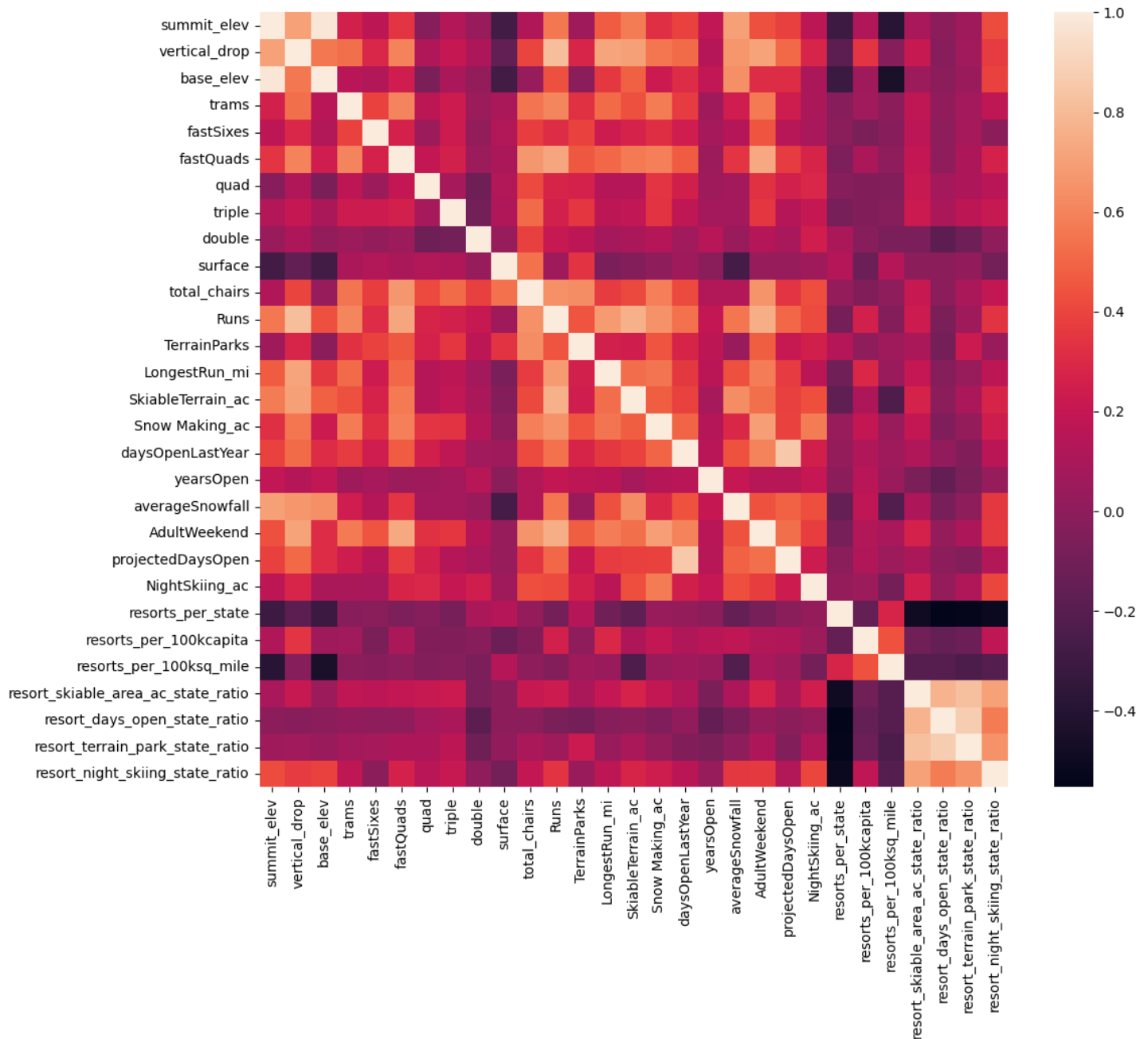


When examining the ticket prices per state, Utah, Colorado, Vermont and Arizona had the highest prices. But no clear pattern was found when examining the prices per state.

Principal component analysis was used to find patterns within the data. It was notable that the first two components account for over 75% of the variance while the first four component account for over 95% of the variance. Vermont, New Hampshire and New York appear to be outliers but there appears to be no clear pattern with the distribution of states with the first two components.



There were also a few features that were closely correlated. Notably, Resort\_night\_skiing\_state\_ratio and AdultWeekend as well as fast\_quads, total\_chairs, Snow Making\_ac.



## Preprocessing and Training:

A 70/30 train/test split was conducted on the data. Since there was still a bit of missing data, it was imputed with the median since some of the features had discrete values rather than continuous values. Finally, the data was scaled around 0 using a standard scaler algorithm.

## Modeling and Evaluation:

The data was used to build two models: a linear regression model and a random forest regressor model. Both models were created using five-fold cross-validation to find optimal hyperparameters. The best linear regression model had an average cross-validation score of 0.633 with a standard deviation of 0.095 while the random forest regressor had an average cross-validation score of 0.713 with a standard deviation of 0.0701. This indicates that the random forest regressor is better at accurately predicting the price. Its most important features are Runs and fastQuads with Snow Making\_ac and vertical\_drop also being important.

A learning curve was used to determine if collecting more data would be required to find better results. This shows that the cross-validation scores level off by around 40-50 samples, indicating that there is sufficient data.

### Results:

There were multiple scenarios to examine. The first scenario involved closing up to at least 10 of the least used runs. The closure of a single run makes no difference while closing 2 and 3 runs reduces the support for an increase in ticket price. Closing 4 and five runs produces the same loss in revenue as closing 3 runs and closing six or more runs leads to a large drop in ticket price and revenue.

The second scenario involves adding a run, increasing the vertical drop and installing a new chairlift. This scenario supports the increase in ticket price by \$1.61 and over the season, this price increase is expected to amount to \$2,815,217.

Scenario 3 involves adding a run, increasing the vertical drop by 150 feet, installing a new chairlift, and adding 2 acres of snow making. This scenario gave the same results as the second scenario.

The fourth and final scenario involves increasing the longest run by 0.2 miles and guaranteeing snow coverage by adding 4 acres of snow making capability. This scenario resulted in no difference.

### Conclusion:

Comparing the four different scenarios, the second scenario yields the best results. It will allow for an increase in price that will cover the seasonal operational cost of adding a new chairlift.

