U.S. Ski Resort Analytics - Big Mountain

Introduction & Summary

Much of a ski resort's success lies in its ability to correctly value its facilities and set an appropriate ticket price. If prices don't reflect the resort's facilities, it either risks falling short of operational costs and incurring revenue loss, or losing seasonal visitors unwilling to pay high prices for mediocre facilities. BIg Mountain Resort, one such resort in Montana, is struggling to balance ticket prices with operational costs particularly since adding a new chair lift. This is in part due to the fact that they charge a small premium on-top of the national average ticket price, while maintaining above-average facilities.

Given a vast set of ski-resort data, the challenge became to understand what features contributed most to ticket prices. The use of a correlation heat map as well as Principal Component Analysis helped reveal correlations within our high-dimensional dataset. The state a resort was in counted for little in determining ticket prices, while features such as the number of runs, capacity for snow making, and amount of fast quads available did contribute to ticket prices.

Two models were fitted to the dataset in order to predict ticket prices, Linear Regression and Random Forest, and their results compared. Between the two, the Random Forest Model produced a lower mean absolute error and exhibited less variability. When using this model to predict the ticket prices of Big Mountain based on its available facilities, the model predicted a

ticket price of $102.18, about $21 higher than their current price, with a mean absolute error of $10. We tested a number of proposed changes, such as closing down 10 of the least used runs, increasing the vertical drop, and adding snow making coverage, in order to see how they affect the ticket price prediction. Of the proposed options, adding a run, increasing vertical drop by 150 ft., and installing an additional chair lift contributed most positively to the ticket prices.

Problem Statement

How can Big Mountain Resort offset the $1,540,000 operating cost increase for this year's season by altering its pricing strategy or operating costs?

Data Wrangling

The provided ski resort data contained information about 330 ski resorts in the United States, with 26 features.

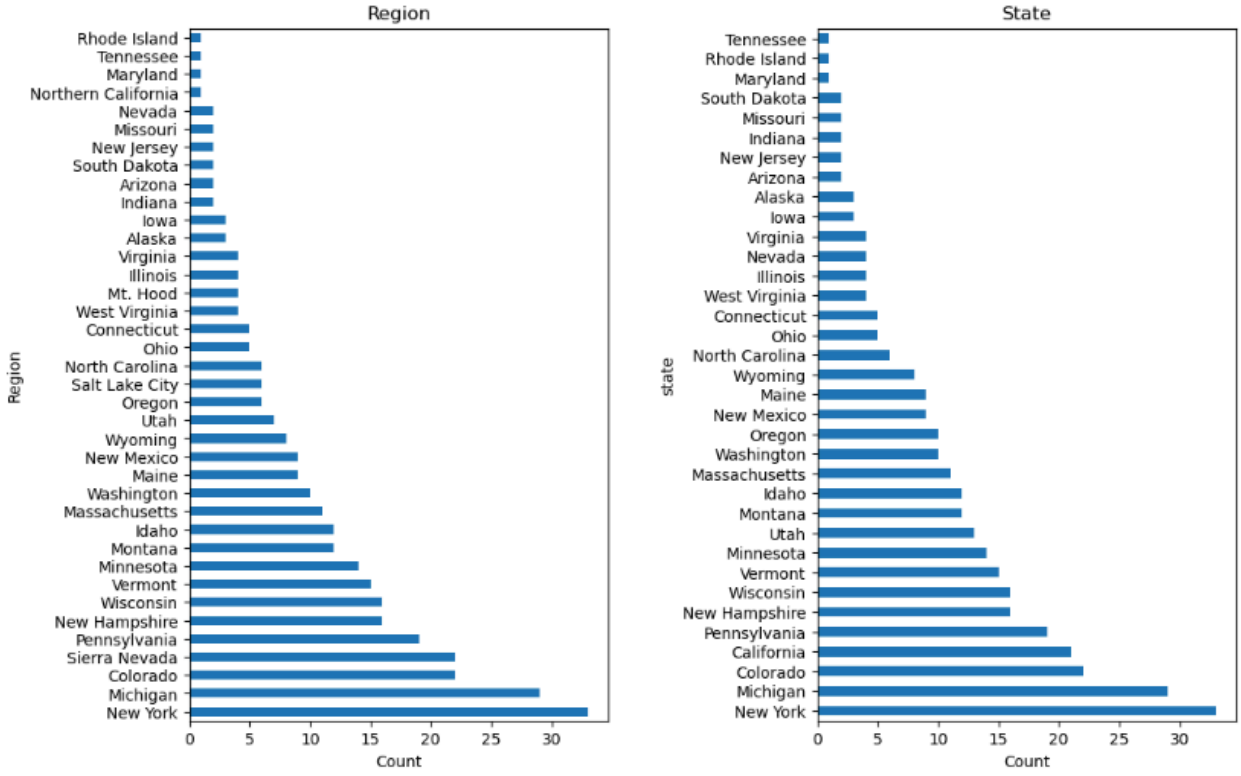
```
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Name               330 non-null    object
 1   Region             330 non-null    object
 2   state              330 non-null    object
 3   summit_elev        330 non-null    int64
 4   vertical_drop      330 non-null    int64
 5   base_elev          330 non-null    int64
 6   trams              330 non-null    int64
 7   fastEight          164 non-null    float64
 8   fastSixes          330 non-null    int64
 9   fastQuads          330 non-null    int64
 10  quad               330 non-null    int64
 11  triple             330 non-null    int64
 12  double             330 non-null    int64
 13  surface            330 non-null    int64
 14  total_chairs       330 non-null    int64
 15  Runs               326 non-null    float64
 16  TerrainParks       279 non-null    float64
 17  LongestRun_mi      325 non-null    float64
 18  SkiableTerrain_ac  327 non-null    float64
 19  Snow Making_ac     284 non-null    float64
 20  daysOpenLastYear   279 non-null    float64
 21  yearsOpen          329 non-null    float64
 22  averageSnowfall    316 non-null    float64
 23  AdultWeekday       276 non-null    float64
 24  AdultWeekend       279 non-null    float64
 25  projectedDaysOpen  283 non-null    float64
 26  NightSkiing_ac     187 non-null    float64
dtypes: float64(13), int64(11), object(3)
```

Region & State

One of the first questions which arose while investigating this data was what's the relationship between state and region? A cursory look at the Region column revealed that they are *mostly* states. A few, however, different from the State entry:

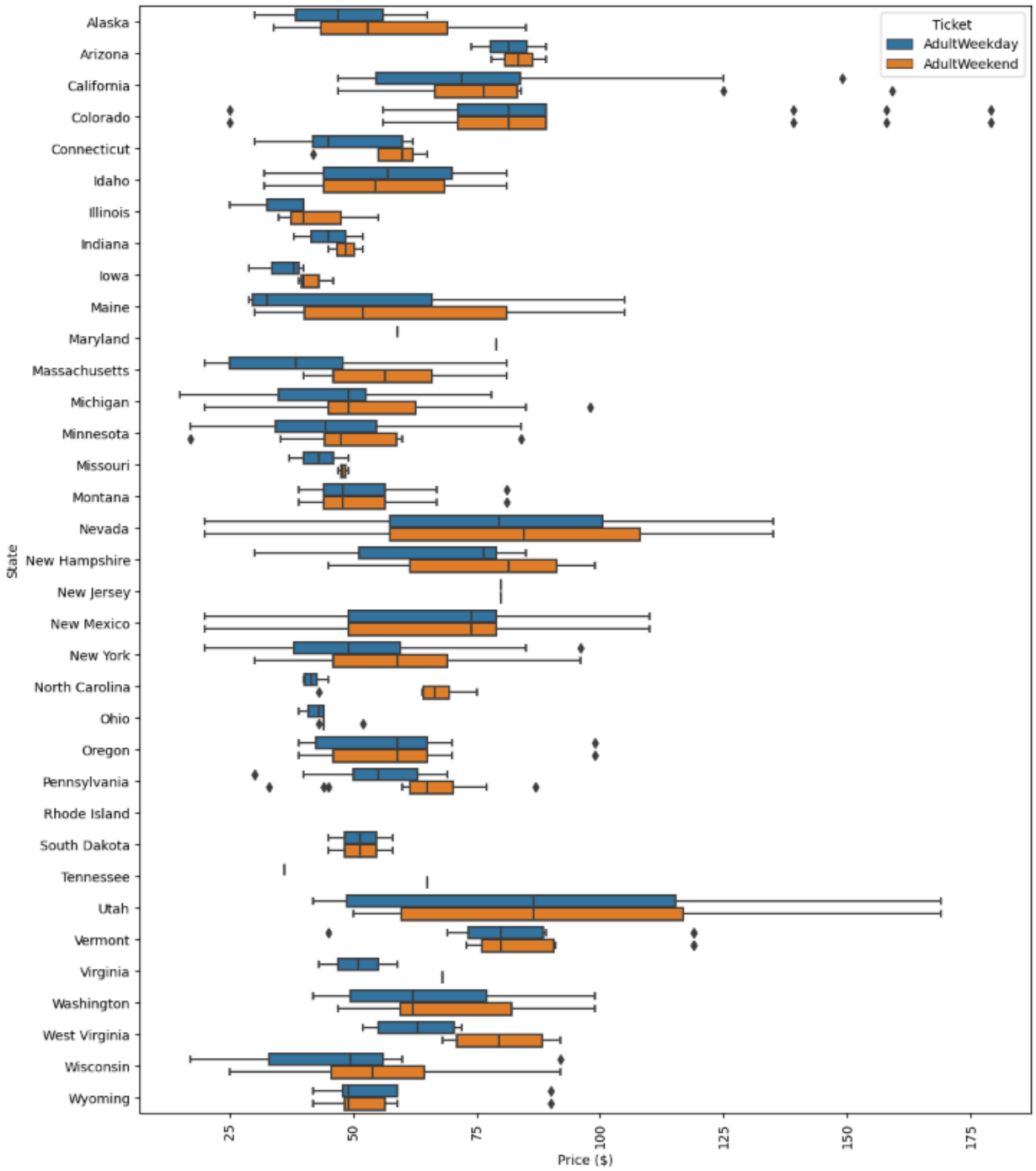
```
state       Region
California  Sierra Nevada          20
            Northern California     1
Nevada      Sierra Nevada           2
Oregon      Mt. Hood                4
Utah        Salt Lake City          6
Name: count, dtype: int64
```

Investigating the number of resorts per state and per region did not yield vastly different results. Therefore, resort region would not play an important role in determining ticket price.



Distribution of Ticket Price by State

Since resort regions will be disregarded, we can investigate just the relationship between state and ticket price. The following boxplot shows both weekday and weekend ticket price distributions per state:



It is apparent that some states stand out in their ticket pricing. Utah and Vermont, for example, have the highest range in prices, while California, Colorado, and Utah have some very expensive ticket prices. This, however, isn't enough information to completely discount state information from our model.

State-wide Summary Statistics

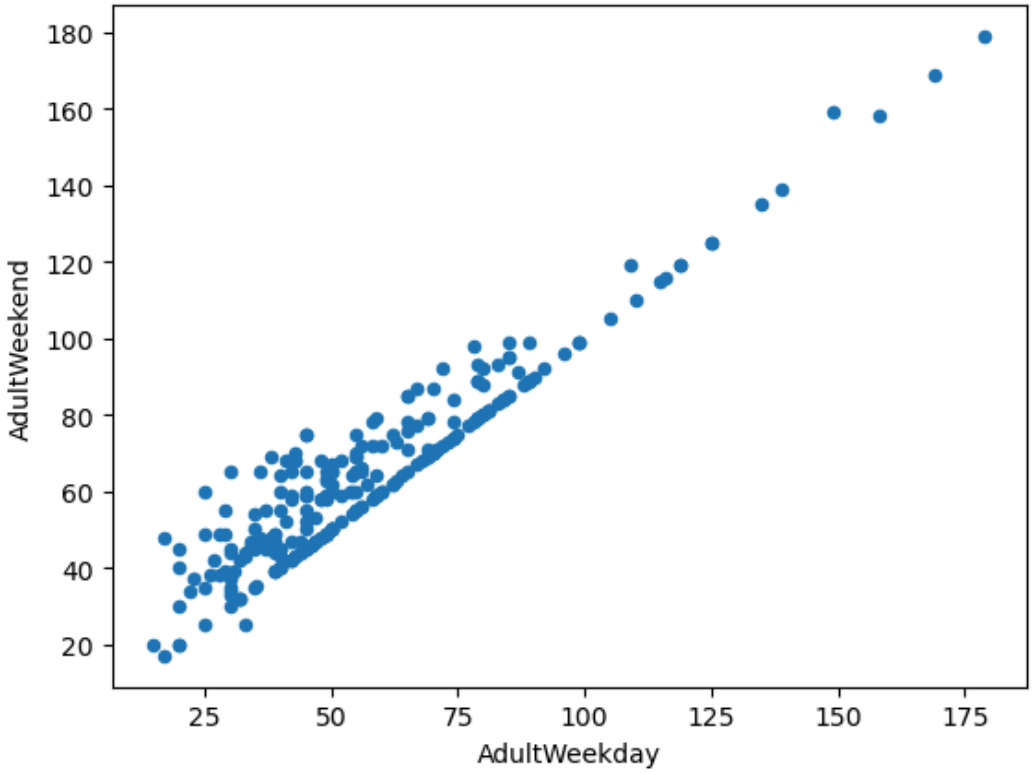
Since state-wide supply and demand of resort facilities may factor into pricing, it could be helpful to generate summary statistics for some features. The features that could drive demand include the number of terrain parks, skiable terrain, days open last year, and number of acres open for night skiing. The following are the first 4 rows of the summary statistics calculated for the entire dataset:

	state	resorts_per_state	state_total_skiable_area_ac	state_total_days_open	state_total_terrain_parks	state_total_nightskiing_ac
0	Alaska	3	2280.0	345.0	4.0	580.0
1	Arizona	2	1577.0	237.0	6.0	80.0
2	California	21	25948.0	2738.0	81.0	587.0
3	Colorado	22	43682.0	3258.0	74.0	428.0
4	Connecticut	5	358.0	353.0	10.0	256.0

Missing Values

Price

Calculating the relative frequency of missing values in ticket prices revealed that 14% of the rows were missing both weekend and weekday ticket values. These rows were dropped, resulting in about 3% of rows remaining with only one missing price value. Of this 3%, 4 values are missing from weekend prices, and 7 from weekday prices. Additionally, the relationship between weekday and weekend prices is quite linear:



In Montana, the prices are actually equal. Therefore, since weekend prices are missing fewer values, the weekday column was dropped and rows missing a weekend price were dropped.

fastEight

The column containing the number of fast-eight lifts is missing about 50% of its values. Investigating the remaining values reveals that all but one have a value of 0. Due to these factors, this column was also dropped.

yearsOpen

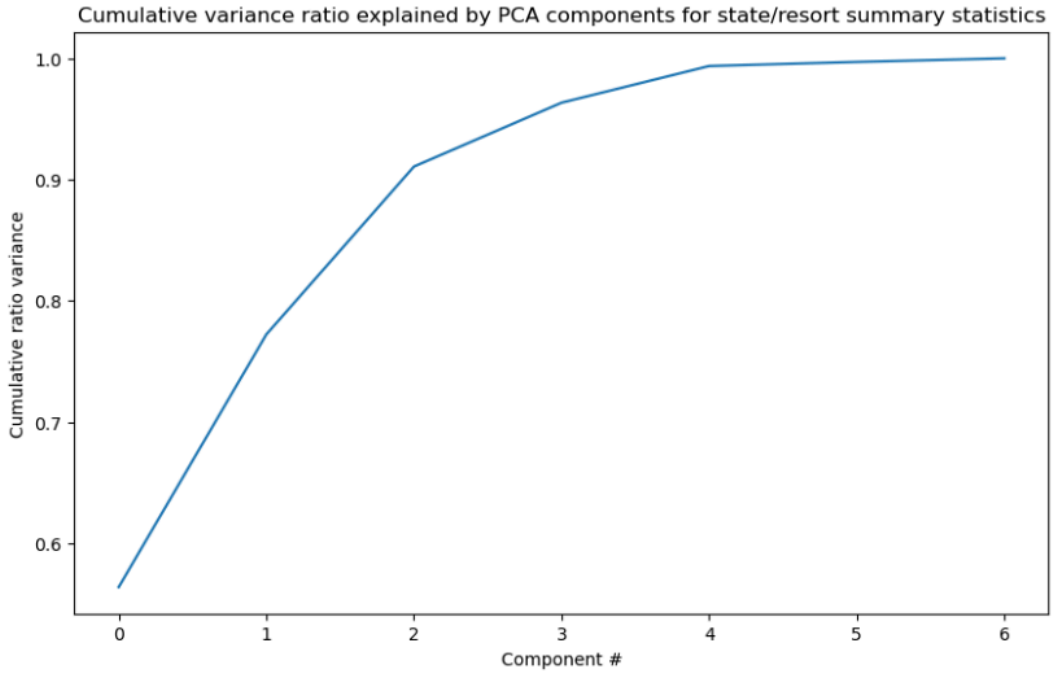
One resort contained the number 2019 for years open. Since this is impossible, and was skewing the distribution of that column, this row was dropped.

Exploratory Analysis

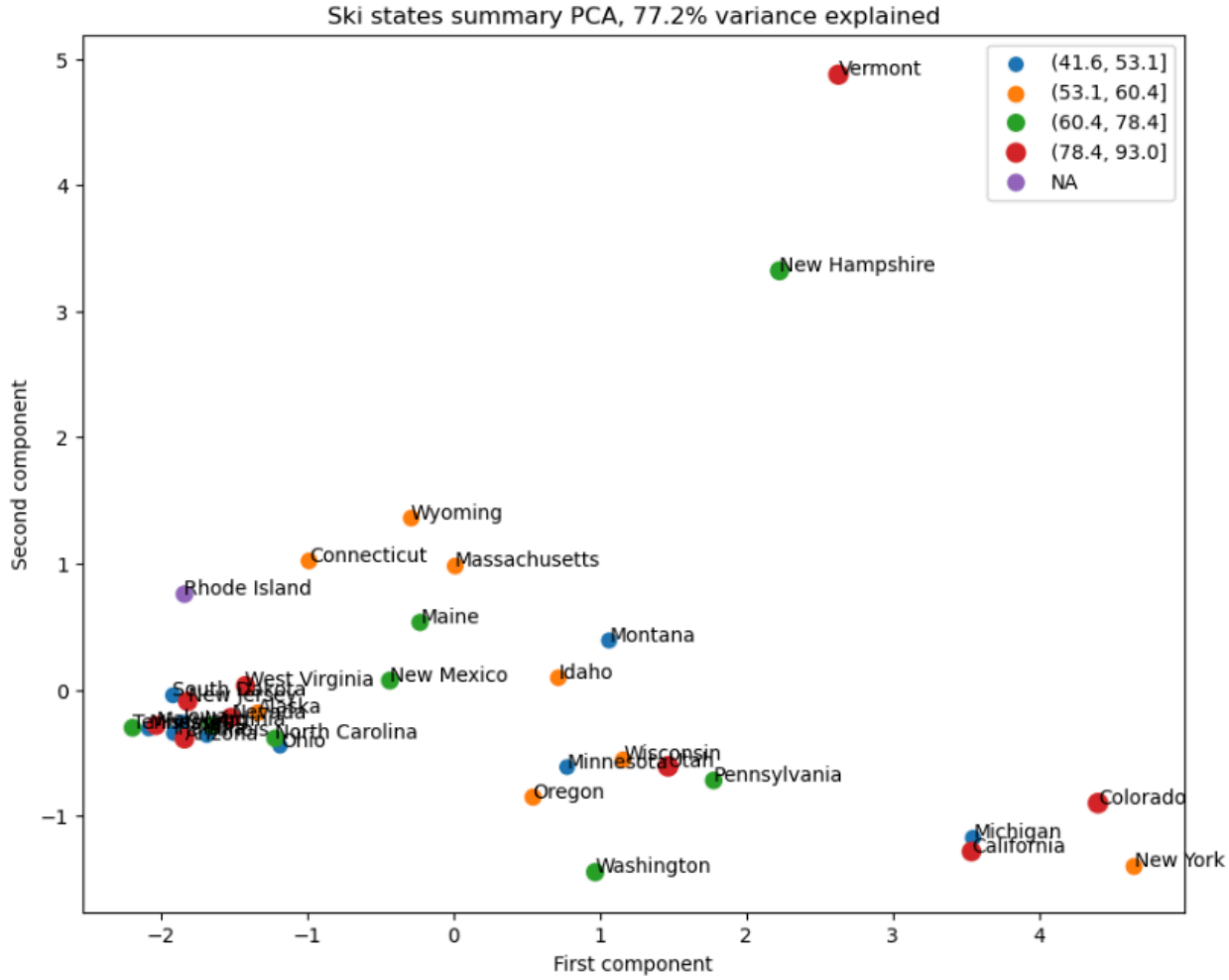
With the addition of features, the high-dimensionality of our data became more and more complicated to unravel. Therefore, Principal Component Analysis was implemented in order to disentangle these complex interrelationships.

PCA Analysis

We continued investigating the relationship between state and ticket price through PCA analysis. A graph depicting the cumulative sum of the explained variance ratio of each principal component reveals that the first two components account for over 75% of the variance in our data:



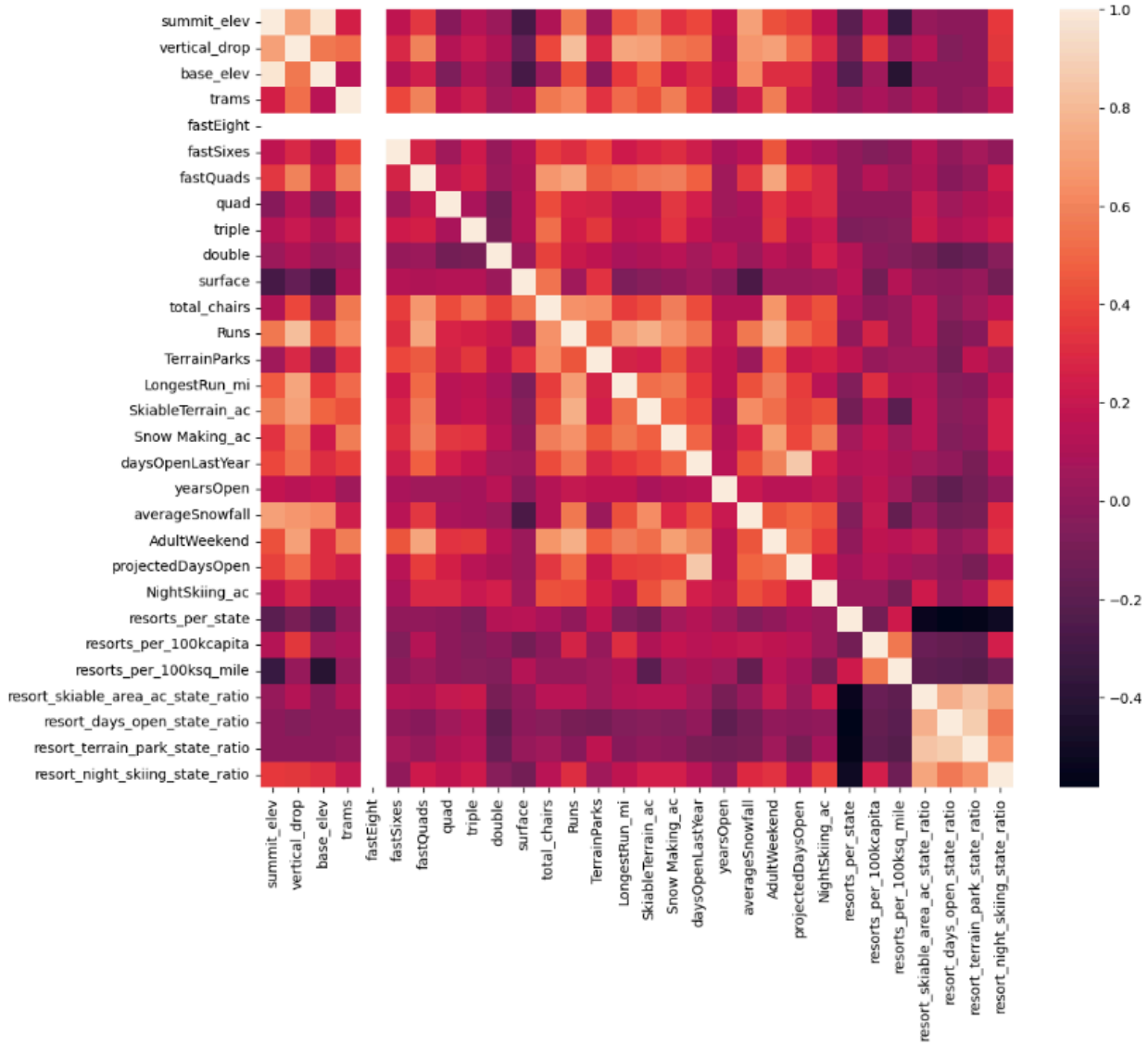
In order to utilize this to gain a better understanding of mean ticket price by state, PC1 and PC2 were added to a DataFrame with states as indices, and mean AdultWeekend prices as another column. In order to assist with visualizing differences in mean ticket price, they were separated into 4 quartiles. These set the colors of the points in the following scatterplot:



Unfortunately, there is no obvious pattern linking state to average ticket price. With the principal components, however, we were able to deduce that resorts per 100k sq miles and per 100k capita counted for a lot. This further reinforces our decision to handle all states equally.

Correlations

In order to plot a heat map of the various correlations between our features, we created “state resort competition” features. These put various features, such as skiable area, days open, and terrain park count, in the context of its state. Having no more use for us, the state label was dropped, and the following heatmap created:



While this does provide numerous interesting insight, we are only concerned with those relating to ticket prices. Looking at those, it looks like the number of fast quads, snow making coverage, and number of runs correlate most with price. We reinforced these with scatter plots of the correlations.

Pre-Processing and Training Data

Thanks to the previous exploratory analysis, we have a broad idea of what to expect when modeling our data, and have done a fair amount of cleaning up our data into something usable.

Baseline

Before beginning to apply regression models to predict ticket prices, we set a baseline by using the mean ticket prices to predict a new ticket price for Big Mountain. Using the average produced an R-squared of -0.003, and mean absolute error of $19.

Pipeline

Imputation

Two options we considered for imputing missing values were the mean and median. To ensure that we are choosing the most accurate method, we applied a LinearRegression model after using both the mean and median to impute results. The resulting R-squared, mean absolute error, and mean squared error were not significantly different. Therefore, we'll stick with using the median to impute values.

Feature Selection

In order to lessen the odds of our models overfitting, we ensured we are using a judicious subset of features by using sklearn's SelectKBest model. We used the f_regression scoring function to select features.

Cross-Validation

Cross-validation was added to our pipeline through sklearn's cross_validate function. We implemented five-fold cross validation when testing our models.

Hyperparameter Tuning

Since our feature selection step in our pipeline, SelectKBest, needs a number of features to use, we can use GridSearchCV to cross-validate while also searching for the best number of features to use.

Regression Models

We compared results when using two different regression models, Linear Regression and Random Forest, to see which would be the best to use to predict ticket prices.

	Metric	Average Mean Absolute Error	Mean Absolute Error Deviation	Best Mean Absolute Error
Model				
Linear Regression		10.50	1.62	11.79

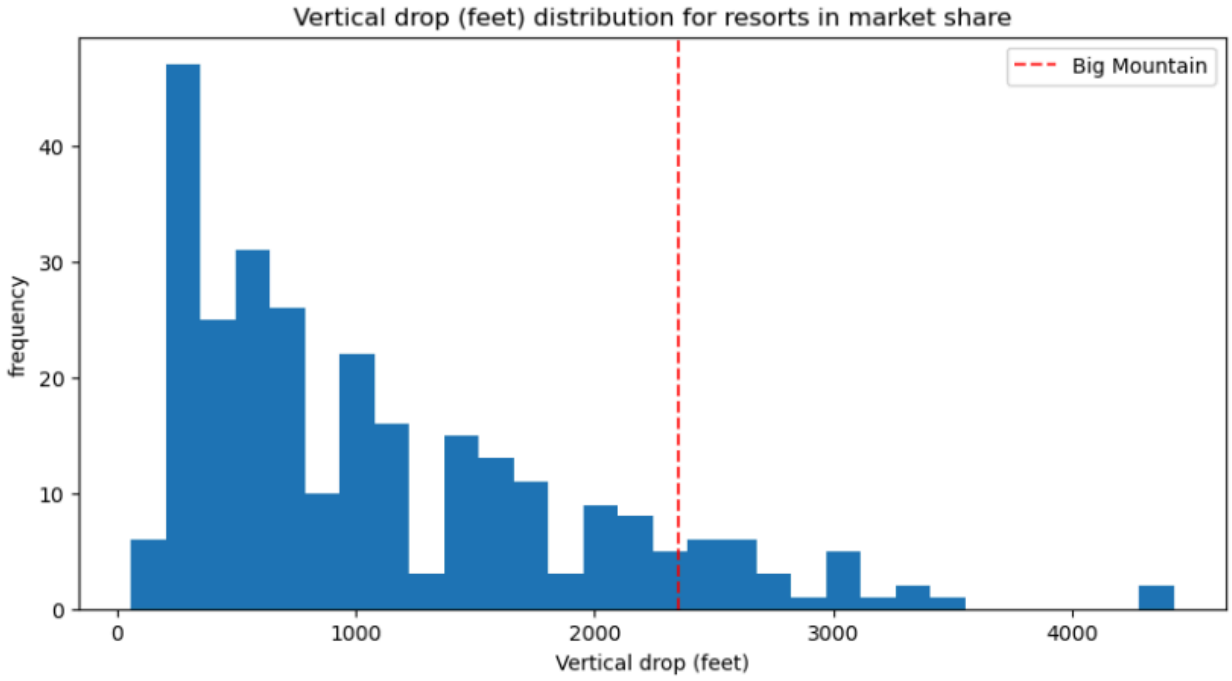
Random Forest		9.66	1.26	9.48

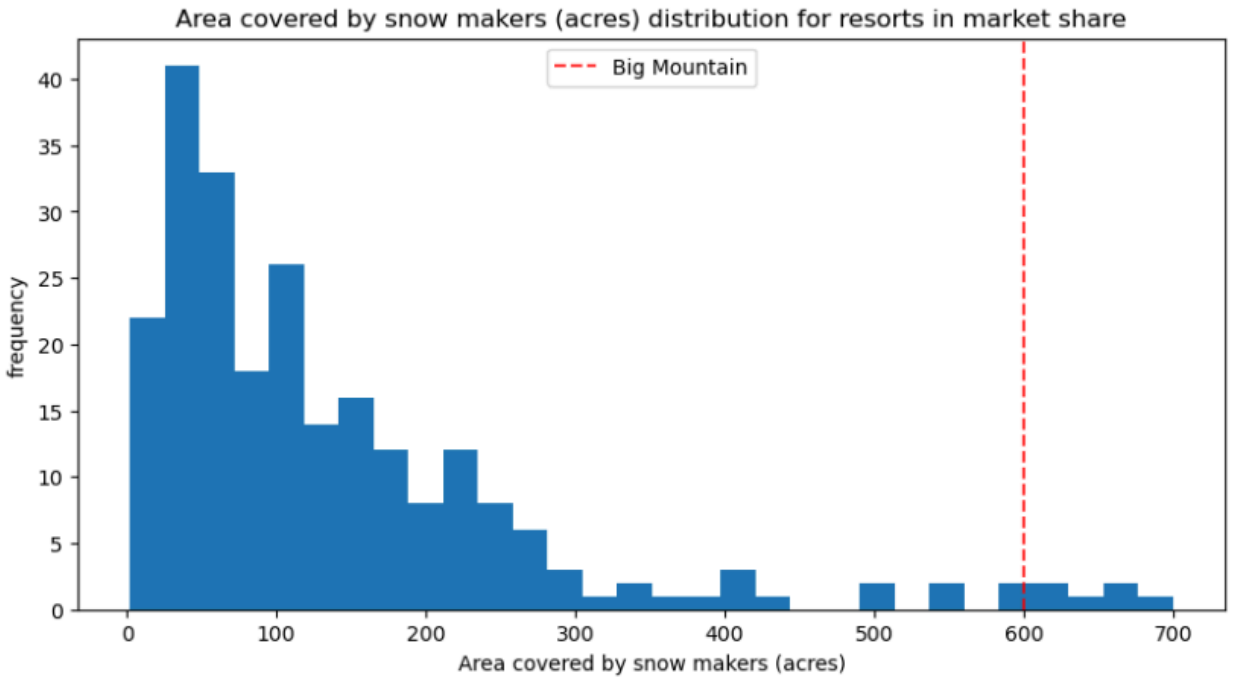
According to these metrics, the random forest model has a lower cross-validation mean absolute error by almost $1, and exhibits less variability. We'll go ahead and use that model to model ski resort ticket prices.

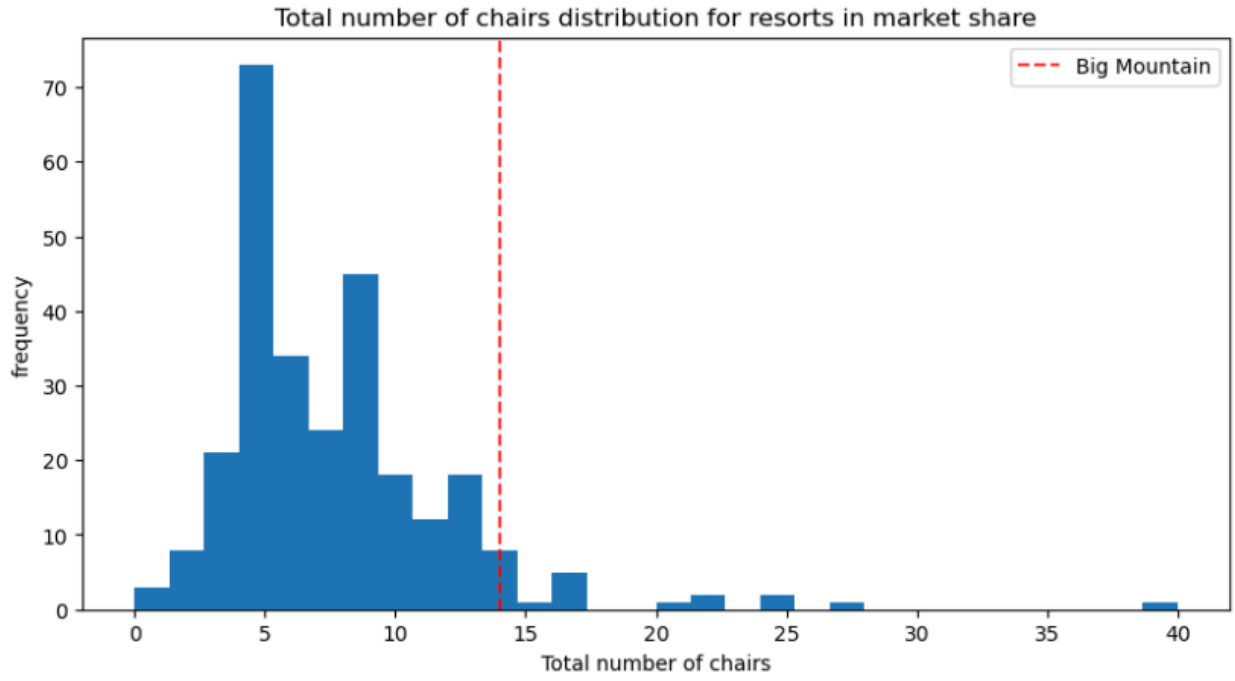
Modeling

Expected Big Mountain Ticket Price

Based on Big Mountain's available facilities, our Random Forest Model predicted a ticket price of $102.18, $21.18 dollars more than the actual ticket price. This does have an expected mean absolute error of $10.24, which still leaves plenty of room for increase even when it's factored in. When comparing Big Mountain's facilities with those of the other resorts, it is clear that they are above average. Here are a few comparisons:







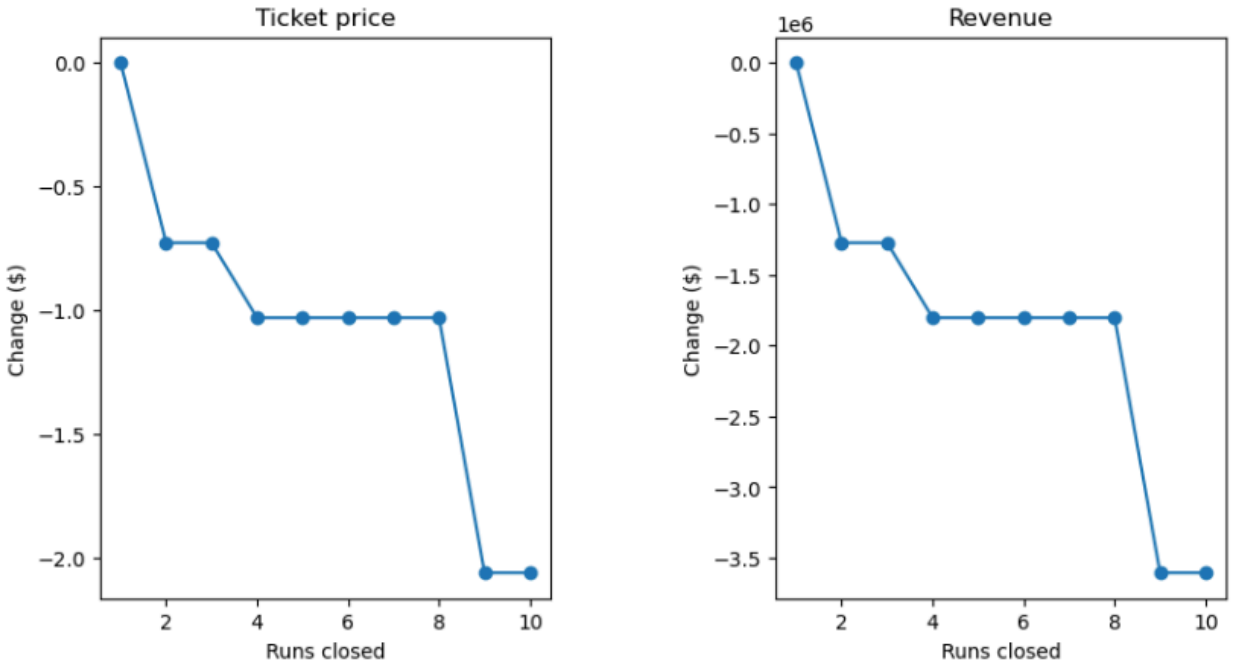
Similar comparisons of Longer run, Number of runs, Skiable terrain, and fast quads reinforce that Big Mountain has significantly above average facilities.

Modeling Scenarios

A few potential scenarios were reviewed by Big Mountain to either cut costs or increase revenue. These were assessed using our model in order to estimate the effects on ticket price.

Closing Down Runs

One scenario was to close down up to 10 of the least used runs. The following charts reflect the change in ticket price as runs are closed:



Clearly closing down runs will reduce support for higher ticket prices.

Adding a run, increasing vertical drop, and installing additional chair lift

In this scenario, one run is added, the length of the longest run is increased by 150 , and another chair is added to bring skiers back up from the new run. This scenario increased support for ticket price by $1.58, resulting in an overall revenue increase of $2,757,576 dollars.

Repeating previous scenario and also adding 2 acres of snow making

In this scenario, the resort is still implementing the previous changes while also increasing snow making capability to 2 more acres. This scenario also increases ticket support by $1.58. Increasing the amount of snow making capability did not have any effect on the price at all, while probably also adding alot of overhead.

Conclusion

Although only three possible scenarios were suggested by Big Mountain, numerous others can be modeled using our model in order to either cut further costs or capture an even higher ticket price. The bottom line is that Big Mountain does have above average facilities and can support higher ticket prices, as predicted by our model. Further data can also be collected to further enhance our predictions or find justification for Big Mountain keeping ticket prices relatively the same.