"Guided Capstone Project Report"
Springboard Data Science Career Track
Samantha Lee

Introduction:

Big Mountain Resort is a well-known ski resort in, Montana. The resort has various types of ski lines, provides spectacular views of the Flathead National Forest and Glacier National Park. It has 11 lifts, 2 T bars, and 1 magic carpet for inexperienced skiers. The lowest point is 4,464 feet, the highest point is 6,817 feet, and the longest run is 3.3 miles long. Big Mountain Resort has added a new chair lift to better distribute tourists across the mountain. This has resulted in a $1.54 million rise in their operational costs this season. Big Mountain Resort needs a new pricing plan based on data collected from several ski resorts around the country. In this report going to touch base on how we can develop a pricing strategy that can predict a competitive price for clients while appropriately reflecting the significance of Big Mountain Resort's facilities. This project's goal is to develop a pricing model for ski resort tickets in our market sector. Big Mountain feels it is not optimizing its returns in relation to its market position.

Name	Big Mountain Resort
Region	Montana
state	Montana
summit_elev	6817
vertical_drop	2353
base_elev	4464
trams	0
fastEight	0.0
fastSixes	0
fastQuads	3
quad	2
triple	6
double	0
surface	3
total_chairs	14
Runs	105.0
TerrainParks	4.0
LongestRun_mi	3.3
SkiableTerrain_ac	3000.0
Snow Making_ac	600.0
daysOpenLastYear	123.0
yearsOpen	72.0
averageSnowfall	333.0
AdultWeekday	81.0
AdultWeekend	81.0
projectedDaysOpen	123.0
NightSkiing_ac	600.0

Problem:

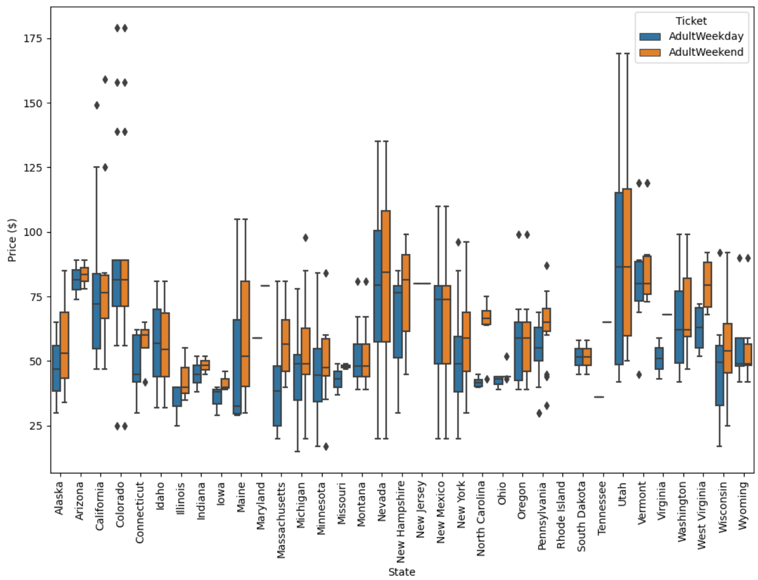
The dataset contains various significant statistics, such as the total vertical drop, the number of lift chairs, and AdultWeekend vs AdultWeekday rates; could it have been indeed beneficial to have a separate price for the weekend? As seen in the figure below, most states charged the same fee for both.

Four Step Strategy Plan:

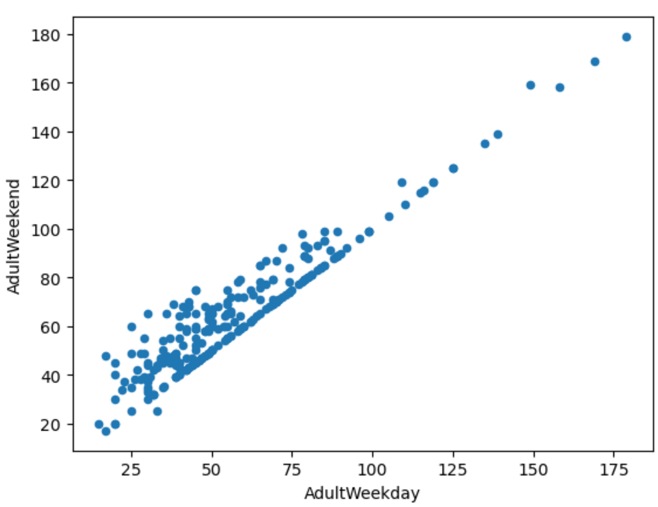
1. Data Wrangling
2. Exploratory Data Analysis,
3. Pre-Processing and training data
4. Modeling

Data Wrangling:

This process focuses on gathering data, organizing it, and ensuring that it is adequately defined. In the 'data wrangling and cleaning' phase rather than exploratory data analysis, looking at feature distributions determining whether the values are reasonable and whether there are any clear outliers to examine. We're interested in whether distributions appear plausible or incorrect in this context.



Box plot shows similarity in weekend and weekday prices for the state of Montana (Big Mountian Ski Resort is located in Montana).



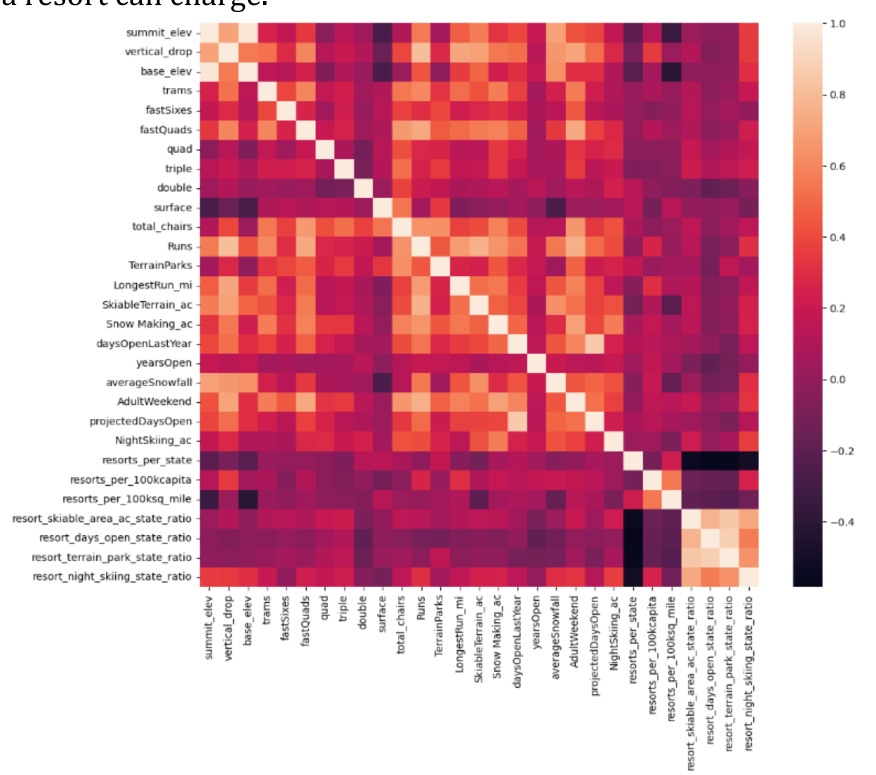
A handful of observations are possible. For starters, there is a significant divide between weekend and weekday costs. Weekend fares appear to be confined to resorts priced around $100.

The weekend and weekday numbers match for every item in Montana, the state where the resort is located, but AdultWeekend had numerous missing values. As a result, the AdultWeekend column was dropped. Apart from AdultWeekend, the fastEight column was dropped since half of its values were null and the other half were generally 0. Apart from those two major ones, there were a few minor columns that had to be dropped, as well as many missing data. Once this was completed, we were left with 277 of the original 330 rows.

Exploratory Data Analysis:

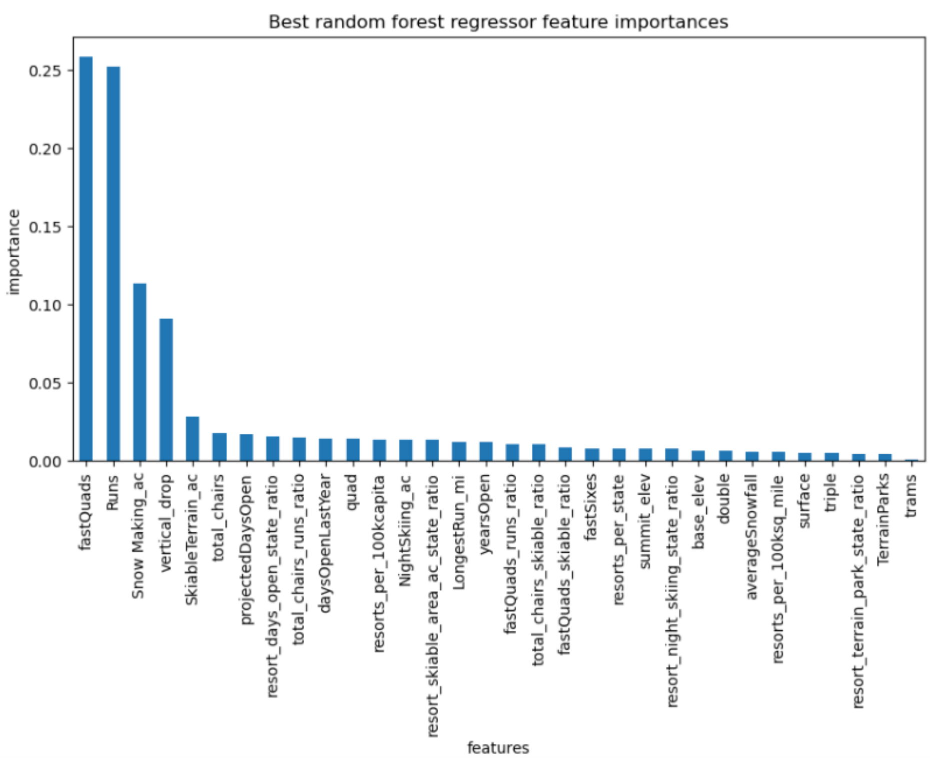
A correlation heatmap is a heatmap that depicts a two-dimensional correlation matrix between two discrete dimensions, with colored pixels representing data from a monochromatic scale. The first dimension's values display as rows in the table, while the second dimension's values appear as columns. The cell's hue is proportional to the number of measurements that correspond to the dimensional value. This makes correlation heatmaps great for data analysis since they show differences and variance in the same data while making patterns clearly accessible. A correlation heatmap, like a standard heatmap, is aided by a color bar to make data more legible and understandable.

When you look at your goal feature, the 'AdultWeekend' ticket price, you'll see several fair relationships. 'fastQuads,' 'Runs,' and 'Snow Making ac' stand out. The final one is very intriguing. Tourists appear to prefer more assured snow, which would incur expenditures in terms of snow production equipment, driving up prices and costs. 'Resort night skiing state ratio' appears to be the most associated with ticket price of the new features. If this is true, perhaps capturing a larger percentage of night skiing capacity is beneficial to the price a resort can charge.



Pre-Processing and training data:

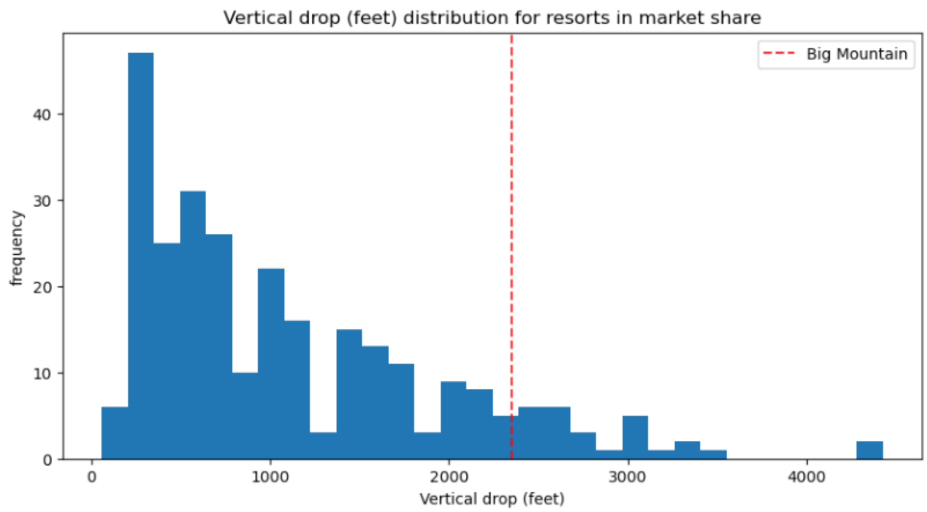
During pre-processing and training data, developed machine learning (ML) models with the Scikit-learn package. Utilizing train_test_split for the pre-processing phases to divide our data into training and test data sets. We also investigate whether values in our data set must be imputed and how our imputing method may impact the performance of our model. Evaluated two modeling approaches, linear regression vs. random forest regression. Linear Model used SelectKBest selects the k best features. The cross-validation for multiple values of k and use cross-validation to pick the value of k that gives the best performance. For regression RandomForestRegressorFeature is used and favorable in most cases. Built a best linear model and a best random forest model. The random forest model is more favorable 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.



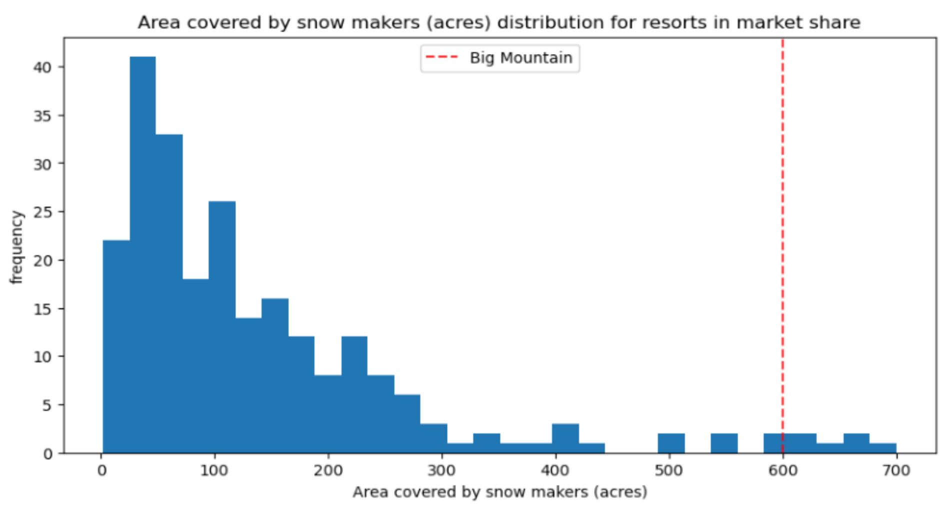
Modeling:

One would first believe that integrating Big Mountain in the model training will boost model performance in forecasting Big Mountain's ticket price. But here is where our business perspective comes into play. The inspiration for this entire undertaking stems from the realization that Big Mountain's pricing must be adjusted. To put it another way, we want to train a model to forecast Big Mountain's ticket price using data from all the other resorts. We don't want Big Mountain's present pricing to sway our decision. We want to compute pricing that is only dependent on its competitors.

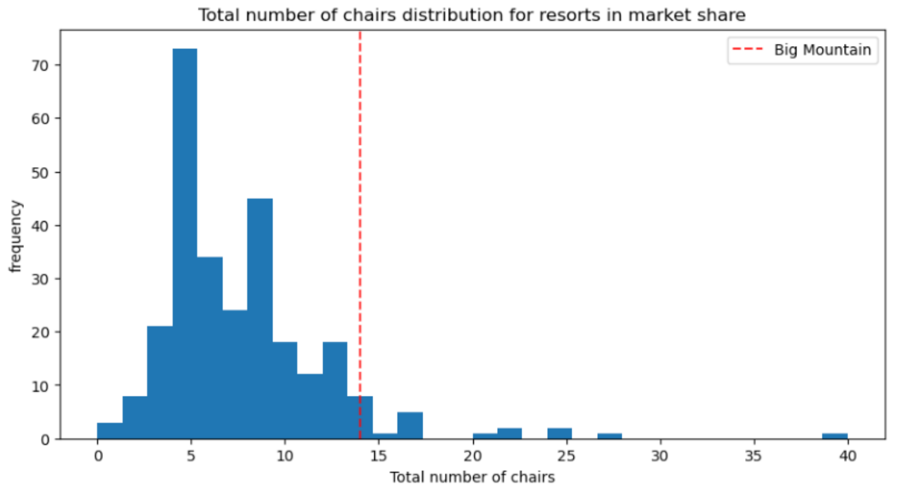
Custom functions are frequently handy for effectively viewing data. The function below takes a feature name as an argument and shows a histogram of that feature's values. It then indicates Big Mountain's position in the distribution by drawing a vertical line through Big Mountain's value using matplotlib's axvline function. It also cleans up any missing values and provides informative labels and a title.
Chose the top components and a regression approach. Can use them to build a model that provides us with a data-driven ticket pricing. To make the model as realistic as feasible, increased the number of components from six to eight by include total_chairs, LongestRun_mi, trams, and vertical_drop. To figure out a reasonable price, I looked at how Big Mountain Resort (shown by the dashed red line) scored in these criteria.



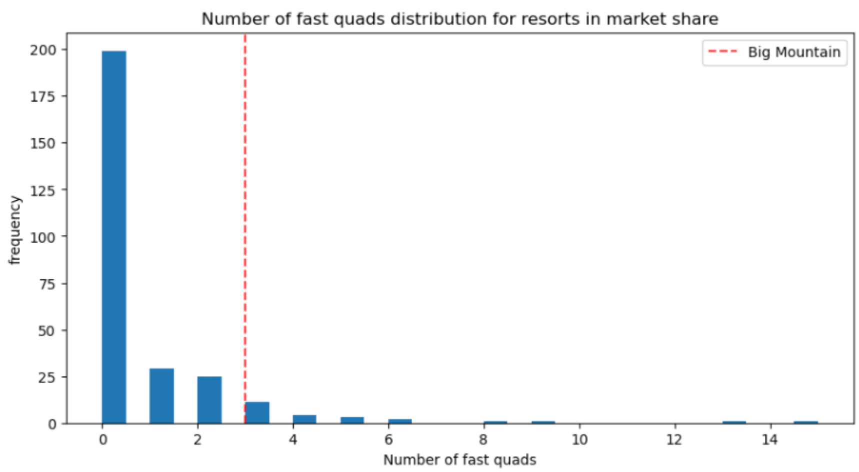
- Big Mountain is performing well in terms of vertical drop, however, there are several other resorts with a larger drop.



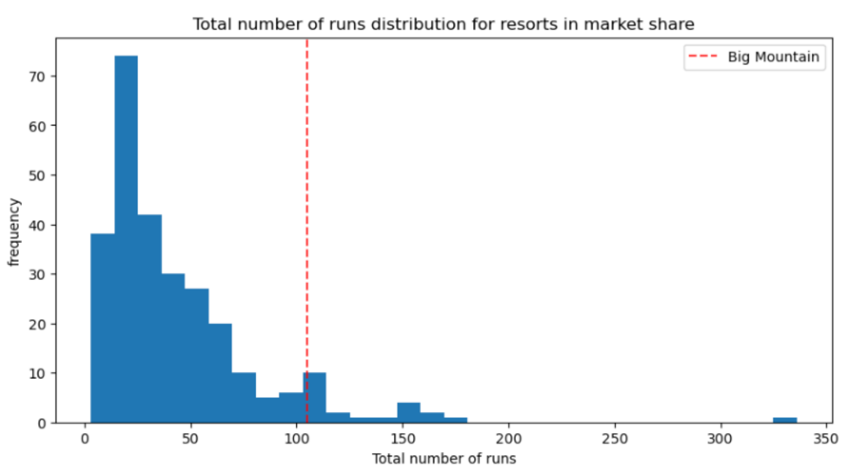
- Big Mountain is towards the top of the snowmaking league rankings.



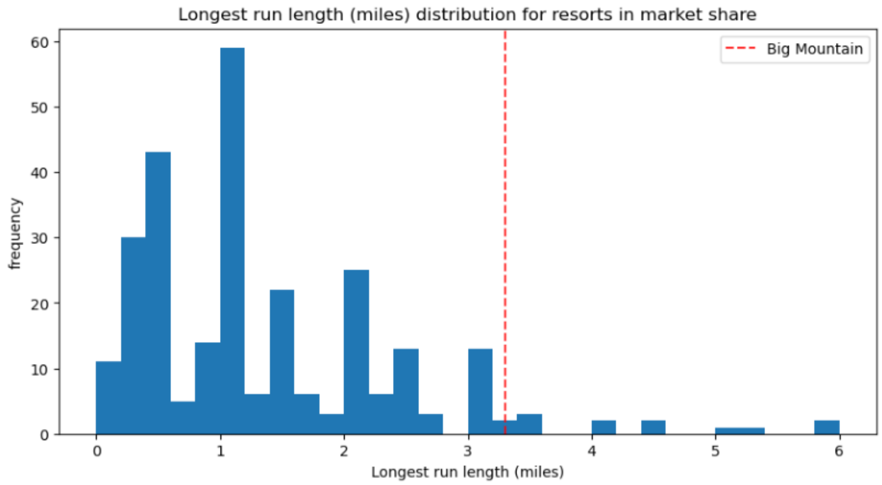
- Big Mountain has one of the most total chairs, with resorts with more appearing to be outliers.



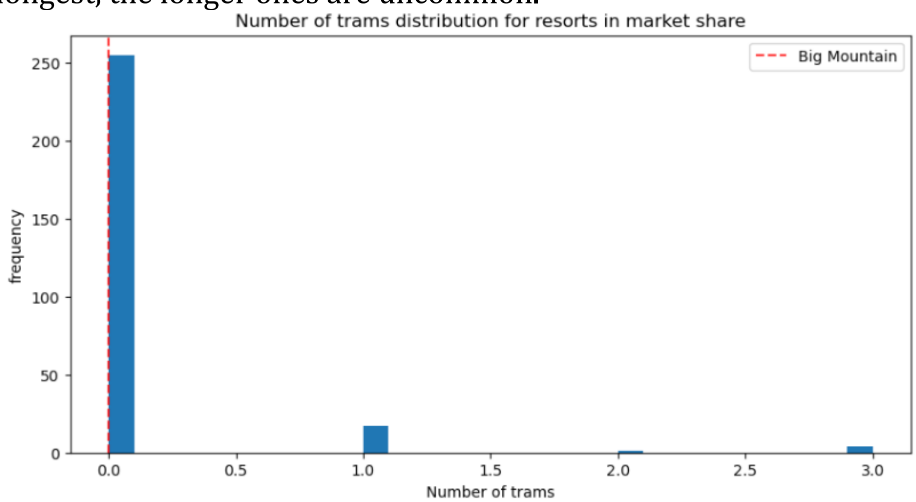
- Most ski resorts do not have fast quads. Big Mountain has three, putting it toward the top. There are certain values that are far higher, but they are not common.



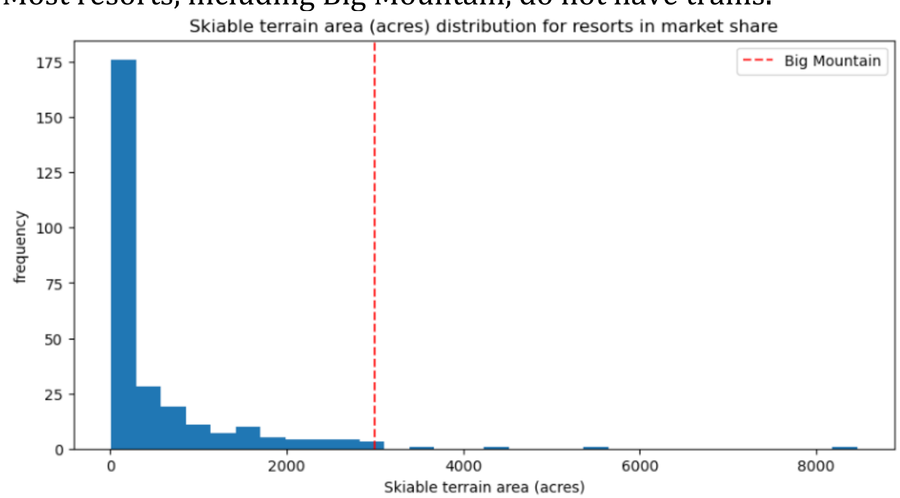
- In terms of the number of runs. Big Mountain compares favorably. Some resorts have more, but not many.



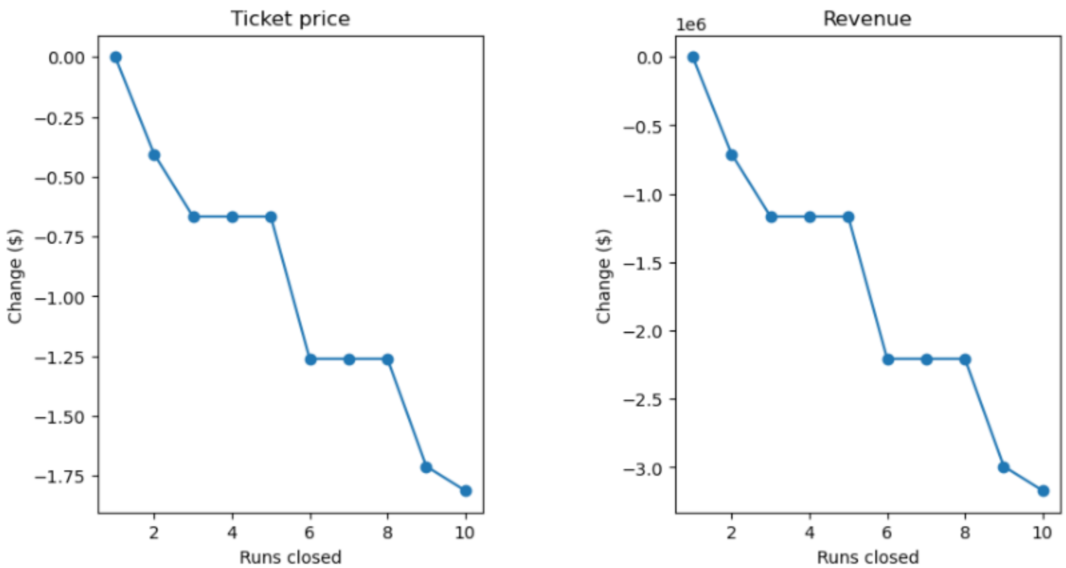
• One of the longest runs is on Big Mountain. Although it is just slightly longer than the longest, the longer ones are uncommon.



• Most resorts, including Big Mountain, do not have trams.



• Big Mountain has one of the most skiable terrains of any resort.



- According to the model, shutting one run makes no effect. Closing 2 and 3 limits support for ticket prices and hence income. If Big Mountain shuts three runs, it appears they may as well close four or five because there is no additional loss in ticket price. Increasing the closures to 6 or more results in a significant reduction.

Suggested Predicted Outcome:

There are two ways to profit: either foresee a price increase/decrease or save money by removing features: The best model was used, and it was discovered that the model price for Big Mountain Resort is 95.87 USD, whereas the real price is 81.00 USD. Even with the predicted mean absolute error of 10.39 USD, this indicates that there is space for growth. A 5 USD increase might reach 7.5M Dollars per year (350K visitors per 5 days stay on average). Several cost-cutting changes were proposed. Adding a run, increasing the vertical drop by 150 feet, and adding a new chair lift; a $1.99 USD increase that might total 3.5 million USD per year. (Adding 2 acres of snow on top of this does not appear to have an incremental effect.)

Conclusion:

The price prediction algorithm predicts an increase of 81 dollars and 1.99 dollars, for a total of 83 dollars per ticket price. To sustain the price rise, the facility would need to be further improved. The resort is already at the top of the market and is priced reasonably in contrast to other resorts with comparable characteristics. Because Big Mountain has a geographical advantage in the premium market, improving the amenities to charge a higher price while attracting more guests will increase the resort's competitive edge over a pro-market cycle.