Big Mountain Resort Project Documentation

# Problem Statement

Big Mountain Resort wants to develop pricing and operational business strategy to capitalize on its facilities as much as possible to address the $1.54M increase in operational costs after installing a new chair lift this upcoming season.

# Data Wrangling

## Key processes and findings

The project used `ski\_resort\_data.csv` as a data source. Original table was in the wide format with 26 rows for different features of 330 resorts (columns) across the US, including the resort of our interest, Big Mountain Resort. The table was transposed to reflect data in tall (long) format. Data was evaluated for the presence of duplicate values among the Name-State-Region because there were resorts in different states with the same name, and some regions spanning across multiple states. No duplicates were found, all resort entries are unique.

Exploring the price distribution by state revealed that Montana's prices are somewhere in the middle range, giving our resort some potential leverage for the pricing, however industry experts should be involved to give the professional evaluation of the pricing for the particular resort.

## Issues

Some of the features raised concern for the validity and usability of data.

- Resorts with missing ticket prices: 3% missing either Weekday or Weekend ticket prices, 14% missing both and just over 82% having both prices.

- Area values `SkiableTerrain\_ac`, `Snow Making\_ac` were clustered down at the low end due to suspiciously high outliers.

- Some lift features like `fastEight`, `fastSixes` and `trams` were 0 for almost all resorts.

- For one of the resorts the `yearsOpen` was 2019, which suggests someone recorded the calendar year instead of the number of years.

## Remedial actions

- 14% of the rows have no price data. These rows are of no use because price is our target. Rows with no data for both prices dropped.

- Skiable area for Silverton Mountain, CO was recorded incorrectly: 26819 acres instead of 1819 acres, as verified via the Internet search. Data corrected.

- Heavenly Mountain Resort had a suspiciously high area in `Snow Making\_ac`, but it is not worth correcting it, because there is no ticket price data for this resort at all. Row dropped.

- Column `fastEight` had only one non-zero entry. Column dropped.

- The `youngest` resort in the data is 6 years old (as of 2019), so it's not optimal to introduce a new youngest 1 year old resort, so it feels best to drop this row. Row dropped.

## Summary of the resort features, adjusted to the state data

### Following statistics were calculated/added using public state data:

* Number of parks per state (`*state\_total\_terrain\_parks*`)
* Total skiable area per state, acres (`*state\_total\_skiable\_area\_ac*`)
* Total number of days resorts were open per state (`*state\_total\_days\_open*`)
* Total lit up area for night skiing per state, acres (`*state\_total\_nightskiing\_ac*`)
* Number of resorts for each state (`*resorts\_per\_state*`)
* State population and area (`*state\_population*` and `*state\_area\_sq\_miles*`)

The `state\_summary` DataFrame (35 rows and 8 columns) is exported as `*state\_summary.csv*` and saved in the `*../data*` folder.

### Ski resort data

There is no price difference between weekday and weekend tickets in Montana (our state of interest), also there are many resorts that have no premium for weekend tickets (Fig. 1). So, the `*AdultWeekend*` price column was chosen to be kept as it has less missing values than `*AdultWeekday*`, which was dropped.

The `*ski\_data*` DataFrame (277 rows and 25 columns) is exported as `*ski\_data\_cleaned.csv*` and saved in the `*../data*` folder.

# Exploratory Data Analysis

### Density of Resorts

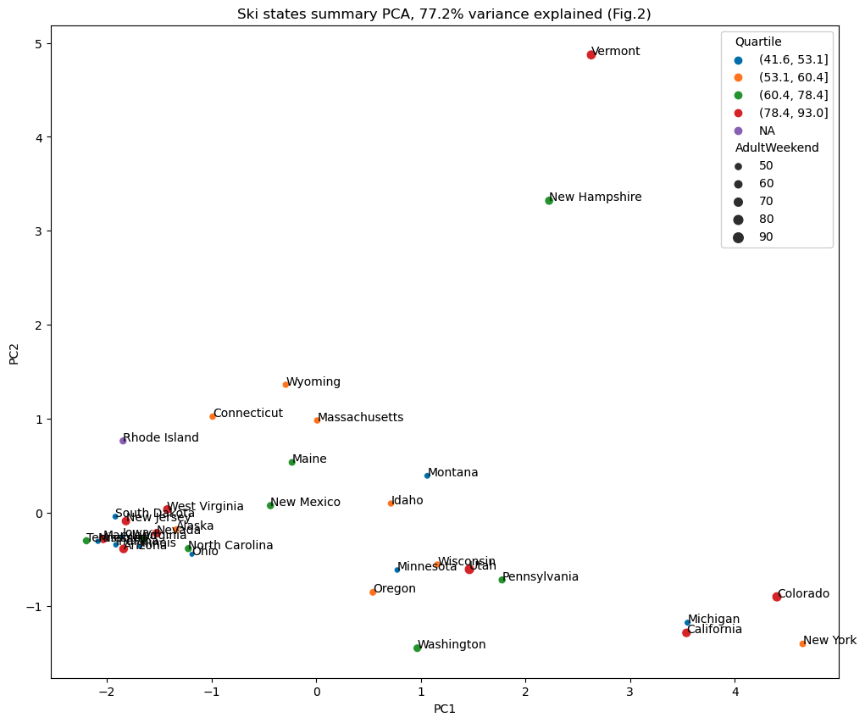
During the EDA the columns `state\_population` and `state\_area\_sq\_miles` of `state\_summary` table were converted to values reflecting density of resorts relative to the state population and size: `resorts\_per\_100kcapita` and `resorts\_per\_100ksq\_mile`.

Revealed top states by resort density:

* Per 100k capita: Vermont, Wyoming, New Hampshire, **Montana**, Idaho
* Per 100k square miles: New Hampshire, Vermont, Massachusetts, Connecticut, Rhode Island

### Principal Component Analysis

During the principal component analysis (PCA), relationships between states and ticket prices were analyzed. Plotting correlation between First and Second principal components, along with dot size-/color-coding the price quartile/value showed no obvious pattern. Multiple states were clustered in the lower-left quadrant of the plot, but there were price quartiles/values of different ranges among them, and other points overall (Fig. 2).



However, there were "stand-out" states of Vermont and New Hampshire, which had particularly large values of `*resorts\_per\_100ksq\_mile*` in absolute terms, and these put them more than 3 standard deviations from the mean. Vermont also has a notably large value for `*resorts\_per\_100kcapita*`.

### Conclusion on how to treat state data

As there was no obvious pricing pattern shown depending on the state's features, we decided to **treat all states equally** and **build a pricing model that considers all states together**. However some potentially relevant state data was captured, which may be useful for our business case.

### Feature engineering

Following “state resort competition” features were added after merging state summary features and ski resort data:

* ratio of resort skiable area to total state skiable area
* ratio of resort days open to total state days open
* ratio of resort terrain park count to total state terrain park count
* ratio of resort’s night skiing area to total state night skiing area

#### Feature correlation

Exploiting the feature correlation heatmap (Fig. 3), following conclusions were made:

* All four “state resort competition” features are negatively correlated with the number of resorts per state, due to lowering the share of any resort with the increase of the total number of resorts per state.
* Positive correlation between the ratio of night skiing area with the number of resorts per capita, which shows that when resorts are more densely located with population, more night skiing is provided.
* Ticket price is reasonably correlated with the number of fast quads, number of runs and area covered by snow-making equipment.
* The ratio of resort’s night skiing area to total state night skiing area is the most correlated with ticket price among the “state resort competition” features. Which is most likely because people want to be able to ski after sunset which may be particularly early during Winter.
* The number of runs and total number of chairs were converted to the ratio of number of chairs or fast quads to the number of runs and and area of skiable terrain respectively: `*total\_chairs\_runs\_ratio*`, `*total\_chairs\_skiable\_ratio*`, `*fastQuads\_runs\_ratio*` and `*fastQuads\_skiable\_ratio*`.

Relationships between these features and ticket price show exclusive vs. mass market resort effect. With more chairs there are more potential visitors making lower prices possible, where with less chairs, prices must be kept high. However, there is no number of visitors per year in the data we were provided.

Also, the number of fast quads may limit the ticket price, but if the resort is covering the wide skiing area, having a small number of fast quads may be beneficial to ticket price.

# Modeling

## Baseline

To get an initial baseline idea of the performance, `*sklearn*`'s `*DummyRegressor*` was used with strategy `*mean*`. Getting a mean price of 63.81108808.

Training set `*y\_tr\_pred*` was then created with just this mean value repeated over and over for the length of `*y\_train*`. Calculating r-squared with these `*y\_train*` and `*y\_tr\_pred*` training sets give us 0, as expected. Then when applied to `y\_test` and `y\_te\_pred` test sets it returns `-0.0031235200417913944`, which also makes sense. Since the training set was already 0.0, and the test set is expected to perform worse than the training set, the r-squared should be also worse, and it became negative.

Calculating Mean Absolute Error of both training and test sets returned ~$17 and ~$19 respectively, will be our reference value when assessing our model's performance. Same for the Mean Squared Error: ~614 for training and ~581 for test sets.

## 

## Choosing a model based on performance

### Linear Regression Model

|  |
| --- |
| **Pipeline**  SimpleImputer  StandardScaler  LinearRegression |

The pipeline for the Linear Regression Model was created with different strategies (mean and median) to fill missing values. All features were scaled to zero mean.

Following r-squared scores were returned:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Trained model | | Test model | |
| Median | Mean | Median | Mean |
| R-squared score | 0.8177988516 | 0.8170154094 | 0.7209725843 | 0.71638147170 |
| Mean Absolute Error | 8.5478503018 | 8.5368840407 | 9.4070201186 | 9.41637562579 |
| Mean Squared Error | 111.89581254 | 112.37695055 | 161.73156451 | 164.392693095 |

These results are not very different from each other, so we will use median.

Then the pipeline was redefined to include the feature selection step:

|  |
| --- |
| **Pipeline**  SimpleImputer  StandardScaler  SelectKBest  LinearRegression |

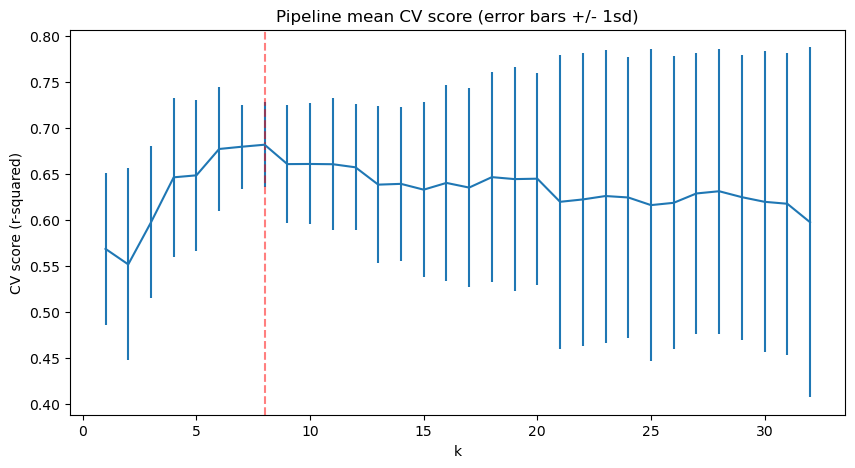
`SelectKBest()` with `f\_regression` as a score function.

With default k=10 this new change performed poorly:

|  |  |  |
| --- | --- | --- |
|  | Trained model | Test model |
| Median | Median |
| R-squared score | 0.7674914326052744 | 0.6259877354190833 |
| Mean Absolute Error | 9.501495079727484 | 11.201830190332057 |

Performance was validated with cross-validation (5-fold) with mean CV score of 0.6327128053007867 (SD = 0.09502487849877675).

Grid parameters were assessed using the `*GridSearchCV*` to find the best k value, which turned out to be `*k = 8*`.



Model coefficients were evaluated, which showed similar results with what we were seeing during EDA: most impactful features are:

* Vertical drop (linear model coefficient: 10.767857)
* Area covered by snow making machines (linear model coefficient: 6.290074)
* Total number of chairs and fast quads (linear model coefficient: 5.794156)
* Number of runs (linear model coefficient: 5.745626)

Interestingly, the skiable area has a negative impact on the price (linear model coefficient: -5.249780), but this may be since larger areas can accommodate more people, thus making lower prices possible. However, we cannot tell that for sure because we lack information about visitors per resort.

### 

### Random Forest Model

The Random Forest Model was used (`*RandomForestRegressor(random\_state=47)*`) along with `*GridSearchCV`* for hyper-parameter search, with and without feature scaling and with mean or median imputing strategies.

|  |
| --- |
| **> GridSearchCV**  **> Pipeline**  SimpleImputer  StandardScaler  RandomForestRegressor |

Results are:

- Median seems to be better imputing strategy than mean.

- Scaling features do not help, so it should not be used.

With mentioned changes we've got much better results with a better mean CV score of 0.6976499257875012 (SD = 0.07095814462058088), compared to default CV results.

The dominant top four features (in common with linear model):

* Total number of chairs and fast quads
* Number of runs
* Area covered by snow making machines
* Vertical Drop

### Final Model Selection

Comparing the performance of the Linear Regression Model and Random Forest Model, the decision to use **Random Forest Model**, as it has a lower cross-validation mean absolute error by almost $1, than LM, while also having less variability.

# Data Quantity Assessment

Data quantity was assessed and determined to be sufficient as the model performance levels off by around sample size of 40-50.

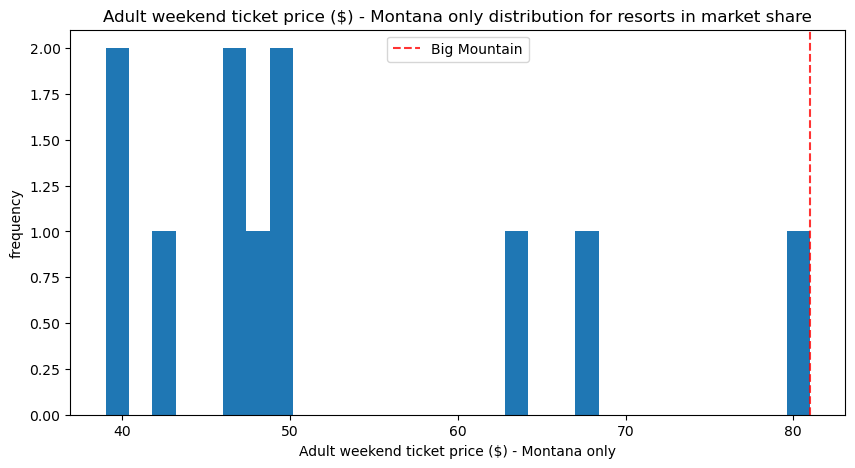
# 

# Big Mountain’s key feature overview

## Ticket Price

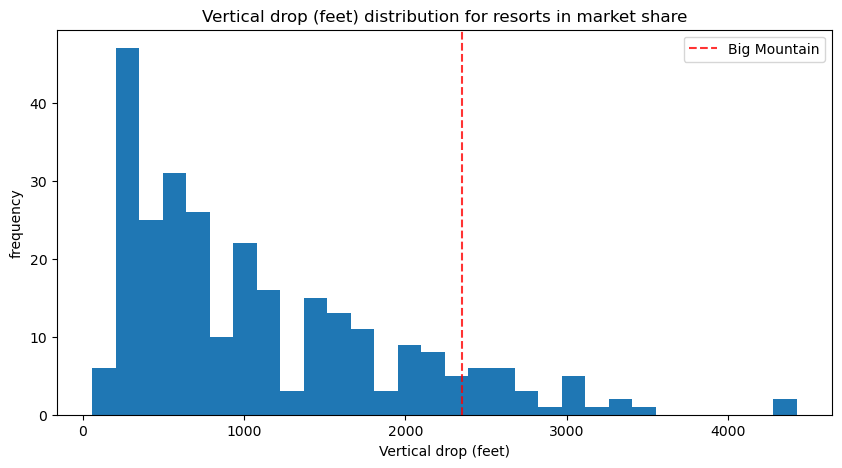
The distributions below show that Big Mountain’s prices sit somewhere at the higher half of the overall prices among all states, however its prices are the highest among all other resorts in the state of Montana.





## Vertical Drop

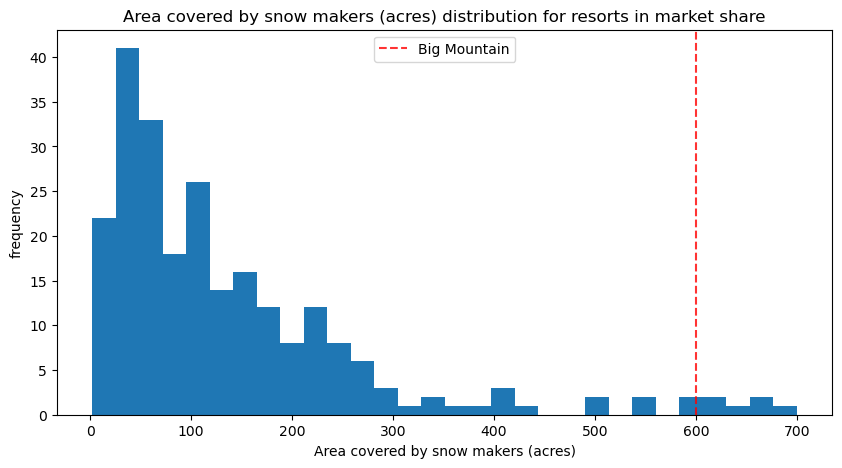
Big Mountain does well for the vertical drop overall. This is promising, as it is one of the most impactful factors on the ticket price. However, there are still quite a few resorts with a greater drop.



## 

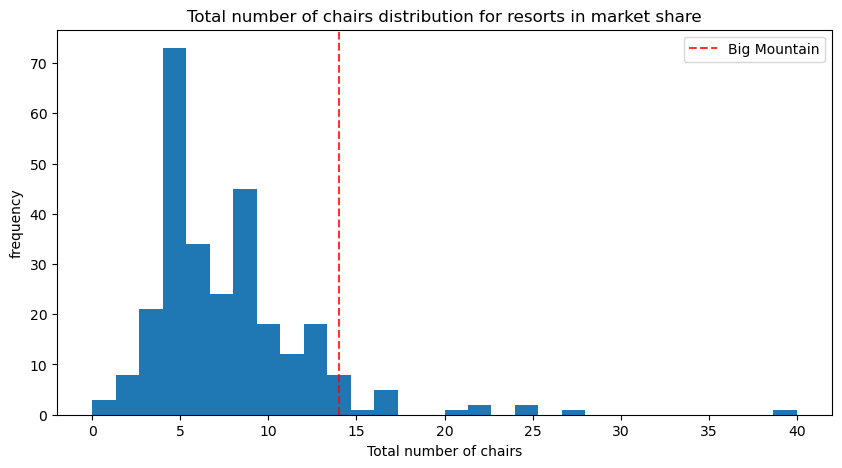
## Snowmaking Area

Big Mountain is very high up the table of the area covered by snowmaking machines.



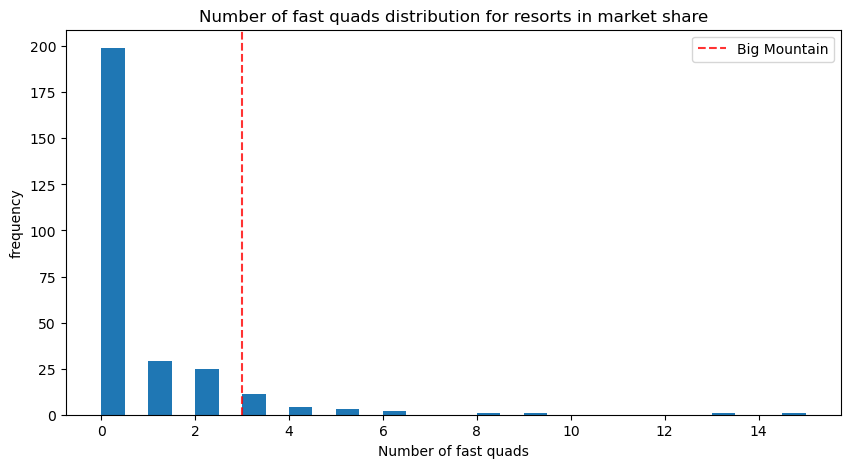
## Total Number of Chairs

Big Mountain is amongst the resorts with the highest number of total chairs, resorts with more appear to be outliers.

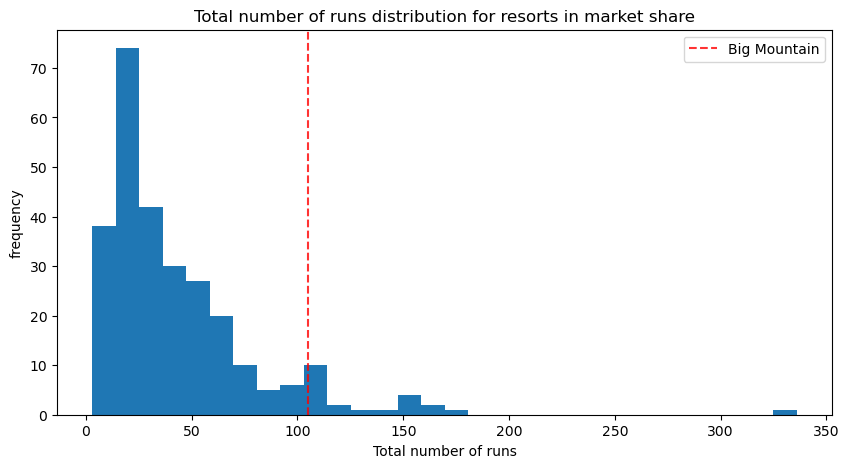


## Fast Quads

Most resorts have no fast quads. Big Mountain has three, which puts it high up that league table. There are some values that are much higher, but they are very rare.

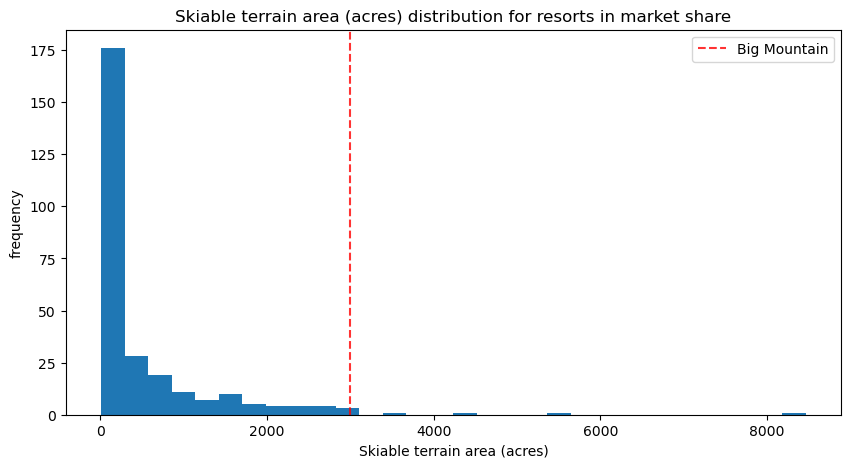


## Number of Runs

Big Mountain compares well for the number of runs. There are some resorts with more, but not many.

## Skiable Area

Big Mountain is amongst the resorts with the largest amount of skiable terrain.



# Winning Model and Scenario Modeling

The chosen Random Forest model was refitted on all available data excluding Big Mountain Resort.

The final Mean Absolute Error of the model is 10.39492491059665 (SD = 1.470438795387226), this is consistent with previous results.

Big Mountain Resort modeled price is $95.87, actual price is $81.00. Even with the expected mean absolute error of $10.39, this suggests there is room for an increase. However, this price was predicted assuming other resorts are pricing their tickets accurately according to the market. According to the results, our resort is underpricing, but there is always a chance that other resorts are over-/underpricing as well. It may also appear that there are some other features (like operating costs) that we are lacking for the more precise analysis and prediction.

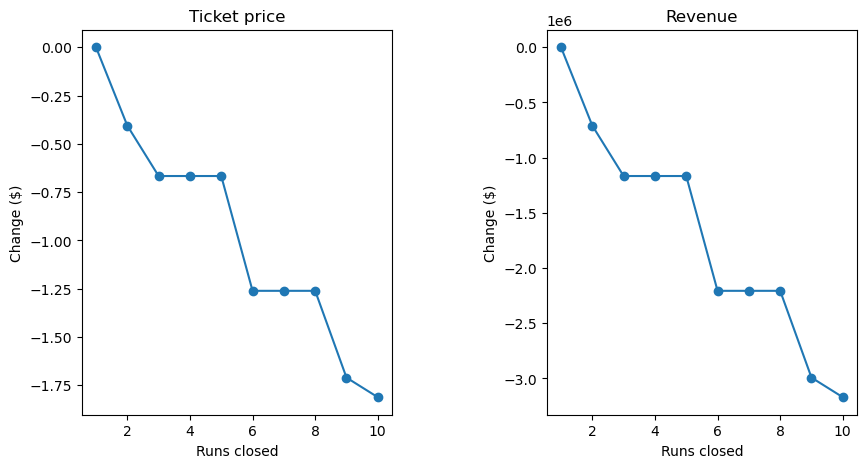
We still do not have the data for the visitors per year for each resort, but we were provided with this data for our resort. The expected number of visitors this season will be 350,000, with an average stay of 5 days.

Having this data, we can predict how the resort's revenue will react to changes of certain features that are currently anticipated by the business management.

## Scenario 1

- Permanently closing to 10 of the least used runs.

The model says closing one run makes no difference. Closing 2 and 3 successively reduces support for ticket price and so revenue. If Big Mountain closes 3 runs, it seems they may as well close 4 or 5 as there's no further loss in ticket price. Increasing the closures down to 6 or more leads to a large drop.



## Scenario 2

- Increase the 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, without additional snow making coverage.

This scenario increases support for ticket price by $1.99.

Over the season, this could be expected to amount to $3,474,638.

## Scenario 3

- Same as Scenario 2 but adding two acres of snow making cover.

This change of snow-making area makes no difference!

This scenario increases support for ticket price by $1.99 (same as Scenario 2)

Over the season, this could be expected to amount to $3,474,638 (same as Scenario 2)

## Scenario 4

- Increase the longest run by 0.2 mile to boast 3.5 miles length, requiring an additional snow making coverage of 4 acres.

This scenario also provides absolutely no difference to the revenue amount or ticket pricing support.

# Pricing Recommendations and Conclusion

Current ticket pricing of $81 definitely has some potential leverage for an increase. After modeling different scenarios, we concluded that **Scenario 2** might be the most optimal among others, as it doesn't lower the ticket price and doesn't involve additional snow making equipment. Scenario 2 increases support for ticket price by $1.99, which translates to the revenue of $3,474,638 over the season. However, operational and installation costs for the new chair lift should be considered. We might need additional data regarding this budget to give a more solid answer.

Additionally, if the resort wants to close any number of runs to lower operational costs, the less impactful on the ticket price would be closure of just one run. Closing 2 and 3 runs successfully reduces support for ticket price and revenue. If Big Mountain closes 3 runs, it seems they may as well close 4 or 5 as there's no further loss in ticket price. Increasing the closures down to 6 or more leads to a large drop.