Python Springboard Data Science Bootcamp

Data Science Guided Capstone

“Big Mountain” ski resort ticket price strategy and optimization regarding the existing facilities – report.

Contents

Figures

1. INTRODUCTION:

Every data analysis problem can be broken down into several subsequent stages, which are:

- Problem statement.
- Data wrangling.
- Exploratory data analysis (EDA).
- Model pre-processing and features engineering.
- Modeling.
- Documentation and communication.

In this last stage of documentation, communication and storytelling, we will explain and summarize all the previous stages as per the above.

2. PROBLEM STATEMENT:

This stage defines the objective of the analysis through some clear question(s) statement. It involves understanding the business context and identifying the success criteria. A well-defined problem statement guides the entire workflow, and it ensures alignment with the stakeholders' goals.

The actual problem at hand concerns a Montana ski resort called "Big Mountain" which applies its ticket pricing based on the other ski resort ticket pricing to which it adds some premium. Executives suspects the ski resort to under capitalize on its facilities and would like some data analysis answering the problem statement which is "Which facilities changes to consider cutting cost and to sustain or to increase revenue at the Big Mountain Resort?". The available data is focusing on 330 ski resorts, including "Big Mountain" resort.

3. DATA WRANGLING:

Data wrangling prepares raw data for analysis. The purpose is to transform messy data into a structured and reliable dataset suitable for analysis. This stage first includes cleaning errors, missing values approach, distinction between categorical and numerical features and identification of the target feature.

3.1. Categorical features:

The relationship between categorical values can be further studied as their individual importance.

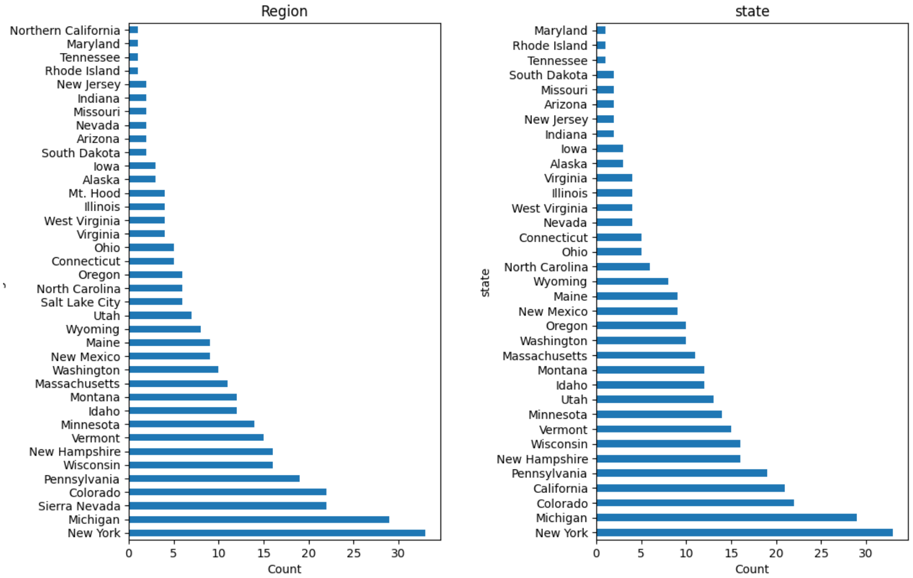


Figure 1: distribution of resorts per region and state.

Some exploration analysis of the ticket price allows us to check and ensure the same market share among the data as well as some difference between weekday and weekend ticket prices.

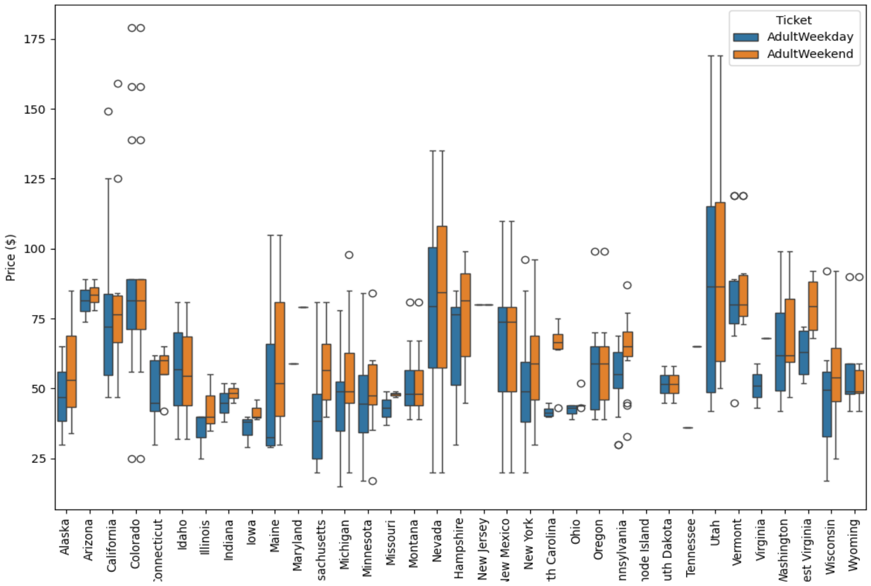


Figure 2: ticket prices variation.

3.2. Numerical features:

In this EDA stage, it is important to clearly visualize the distribution of this numerical features data. These visual numerical features distribution supports the identification of any outliers susceptible to hindering the features data. Are suspicious, the distribution which are not, at minimum skewed.

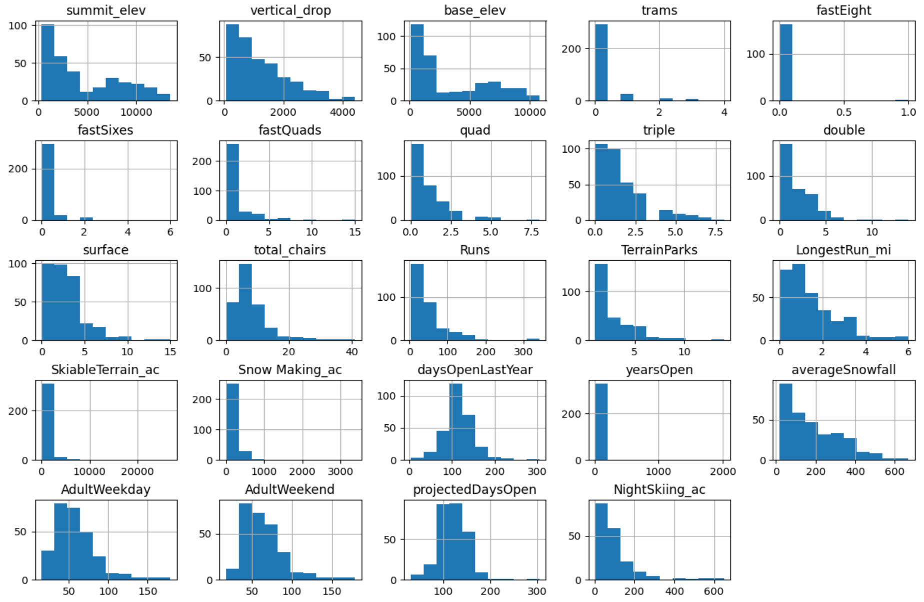


Figure 3: numerical features distribution before deletion of outliers.

After outliers' deletion, the visual numerical features distribution is revised to assess the impact of these deletions:

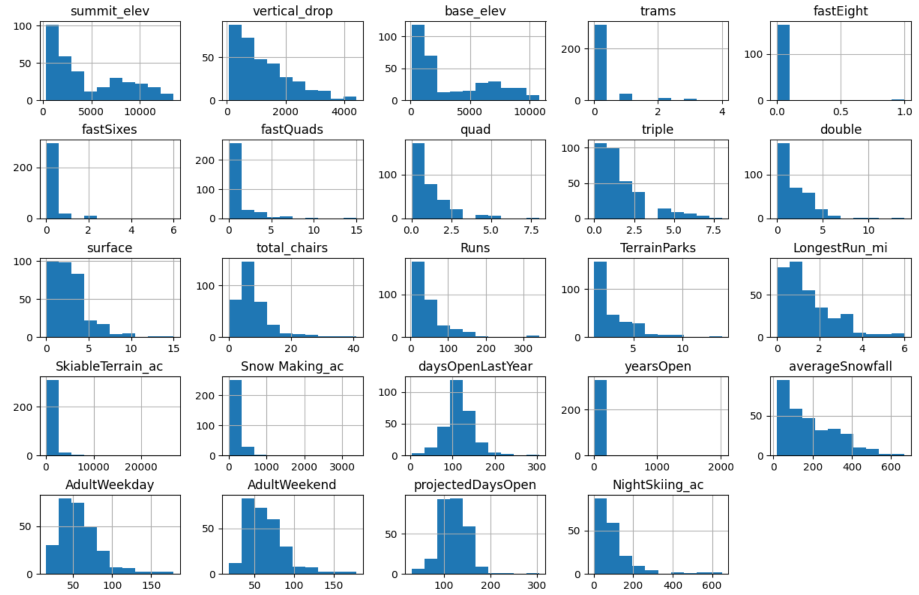


Figure 4: numerical features distribution after deletion of outliers.

Back to the categorical data, some additional data is concatenated to the isolated categorical dataset based on the state label and concerning its area and population.

Concerning, the target features identified as the weekday and weekend ticket prices, some relations between both can be analyzed as well as their null values for future imputation I the subsequent stage of preprocessing. There is a strong correlation between the weekday and the weekend price.

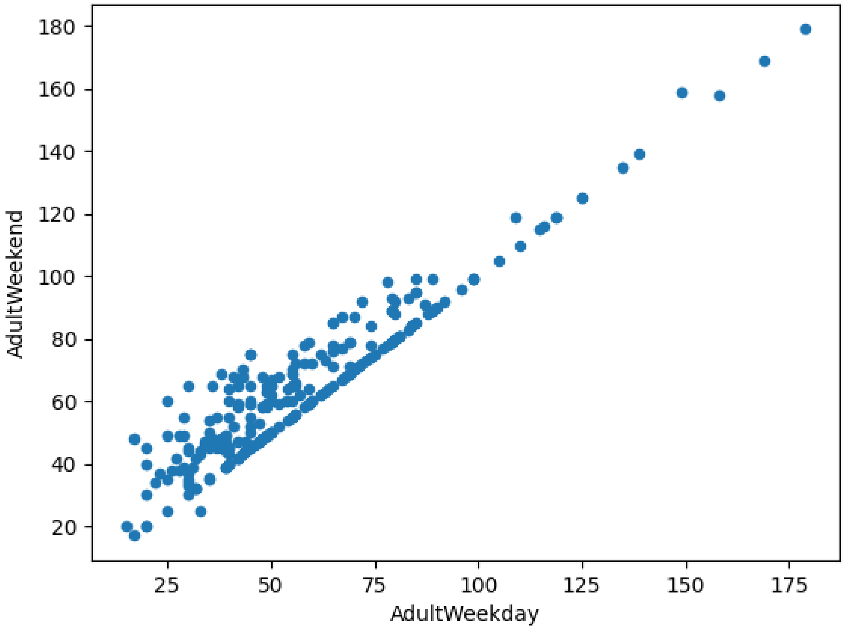


Figure 5: strong positive correlation between weekend and weekday ticket price.

Finally, the categorical dataset and the complete dataset are saved separately.

4. EXPLORATORY DATA ANALYSIS (EDA):

This stage illustrates the use of statistics and visualization to understand data distributions, patterns and relationships. This helps to uncover insights; anomalies and it guides features and model selection.

4.1. Exploration of the categorical data:

The objective of this exploration is to determine how to use this state data. The first exploration of the top states by order for each feature and the creation of state density features raises more questions.

An approach to disentangle the data relationships is to use the “principal component analysis” (PCA). This PCA first requires a scaling of the data by standardization which centers the mean on zero and with a standard deviation of one.

The fitting of the PCA object shows that the first two components, which explain 75% of the variance, are suitable for visualization.

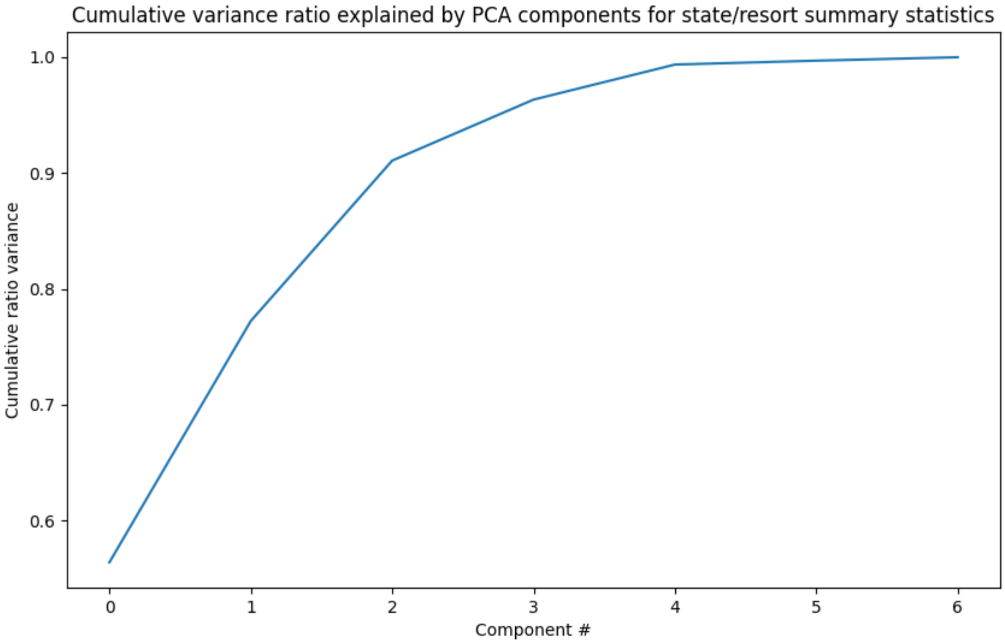


Figure 6: first two components gather 75% of the variance.

However, the scatter plot showing the relationship between those two components, once the PCA object data transformation made, does not show any identifiable pattern.

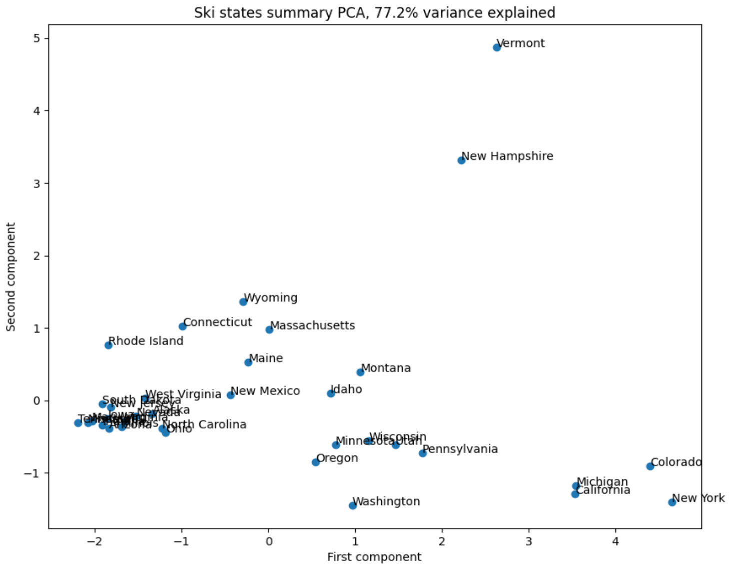


Figure 7: absence of clear pattern in the first two components relationship.

Besides, the superimposition of the average ticket price broken down per quartile does not seem to provide any additional direction for any pattern. This gives good reasons to treat all the states the same and to keep the state dataset.

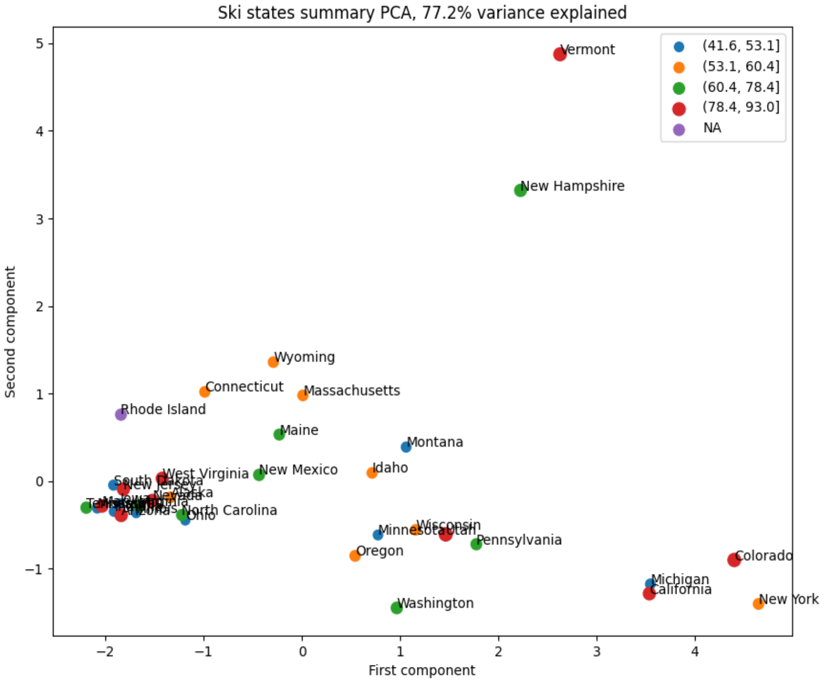


Figure 8: no additional pattern based on price breakdown per quartile.

4.2. Exploration of the numerical data:

Feature engineering supports the merging of the two datasets between the initial ski dataset and the saved state-related dataset, on the state label. Some interesting ratios between resorts and states are created and the state-related labels are dropped.

The non-numerical columns are dropped, and a correlation is set up followed by a Seaborn heatmap, which helps to better visualize any correlation.

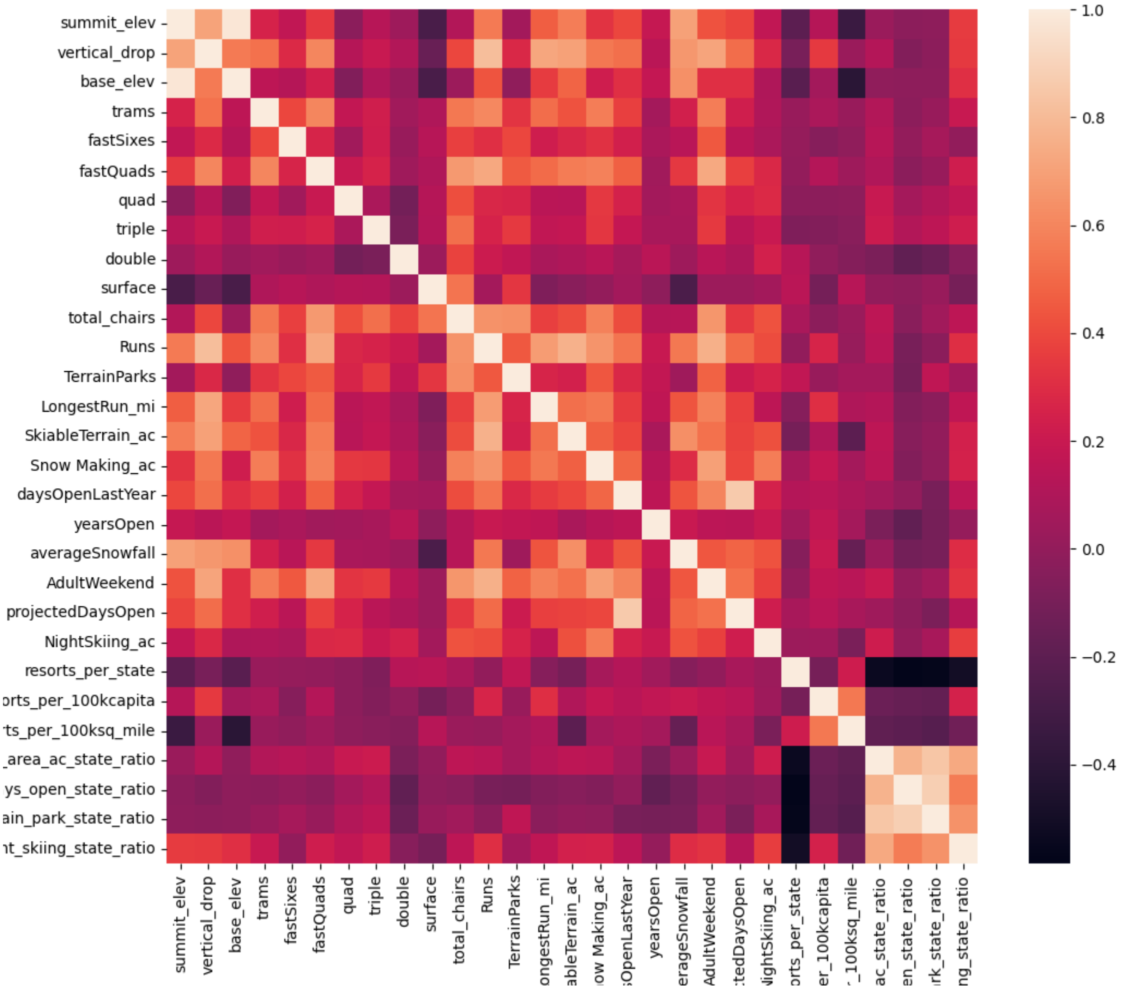


Figure 9: numerical features heatmap and correlations.

Among correlations are summit and base elevations, drop, the introduced ratio resort/state. Concerning the ticket price, a correlation exists with fastquads, runs, night skiing, number of chair lifts and snow making.

However, scatter plots can even better identify the relationship between ticket price and those numerical features. The scatter plots below show a strong correlation of ticket price with the vertical drop, fastquads, runs and total number of chair lifts.

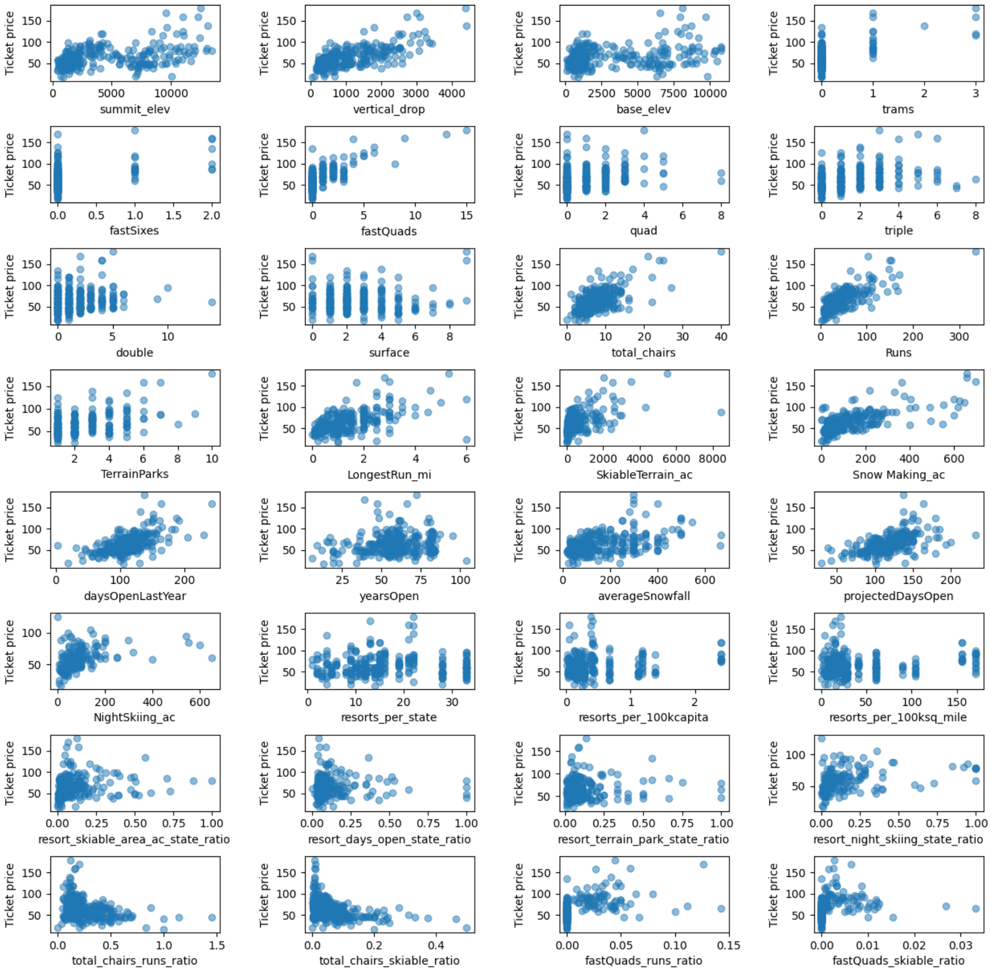


Figure 10: scatter plots to show correlation between ticket price and each numerical feature.

5. MODEL PRE-PROCESSING AND FEATURES ENGINEERING:

Usually, model preparation and training involve null values imputation for both train and test dataset, data scaling, training and fitting of the model, prediction of the model and assessment through performance metrics.

However, the concept of pipeline supports the integration of those individual steps into one pipeline object that can be trained and predicted from as a model. Cross-validation preserves the model from overfitting on the test dataset on the search of the ideal hyperparameters. Moreover, gridsearchCV can combine the pipeline and cross-validation in one command in the search for the ideal parameters.

5.1. Linear regression model:

Applied to a linear regression model, gridsearchCV identifies the number of features (8) needed for best performance as well as their names and their coefficient of importance can be computed.

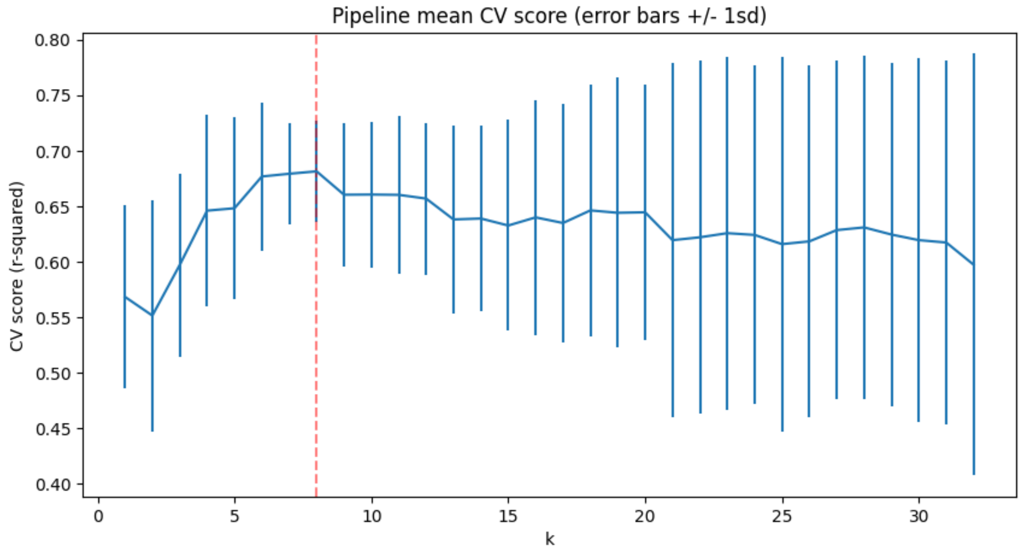
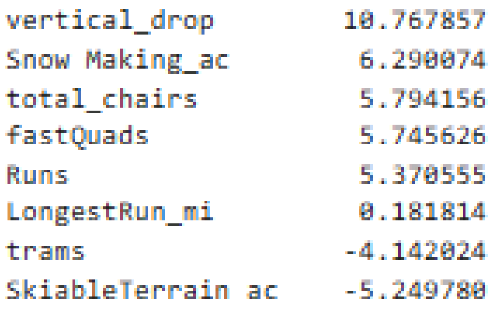


Figure 11: best number of numerical features to be analyzed for linear regression model.



```
vertical_drop          10.767857
Snow Making_ac          6.290074
total_chairs            5.794156
fastQuads               5.745626
Runs                    5.370555
LongestRun_mi           0.181814
trams                  -4.142024
SkiableTerrain ac      -5.249780
```

Figure 12: most important features and their importance coefficient.

5.2. Random forest model:

GridsearchCV is also applied to another model pipeline based on random forest. The values for following parameters are searched: number of trees, the scaler, the imputation. The result gives the features importance.

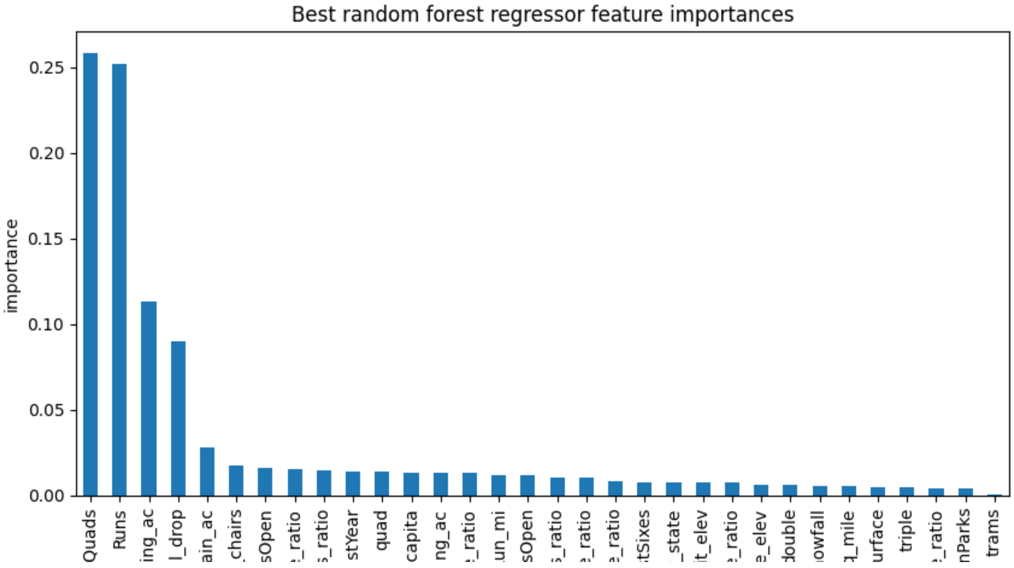


Figure 13: most important features and their importance.

5.3. Model selection:

Comparison between the linear regression model and the random forest model based on the cross-validation mean absolute error and its standard deviation. The conclusion is that the random forest model has a lower cross-validation mean absolute error and less variability.

The mean absolute error represents the average of the absolute differences between the predicted values and the actual values.

$$MAE = \frac{1}{n}\sum_{i}^{n}|y_i - \hat{y}|$$

Figure 14: mean absolute error (MAE).

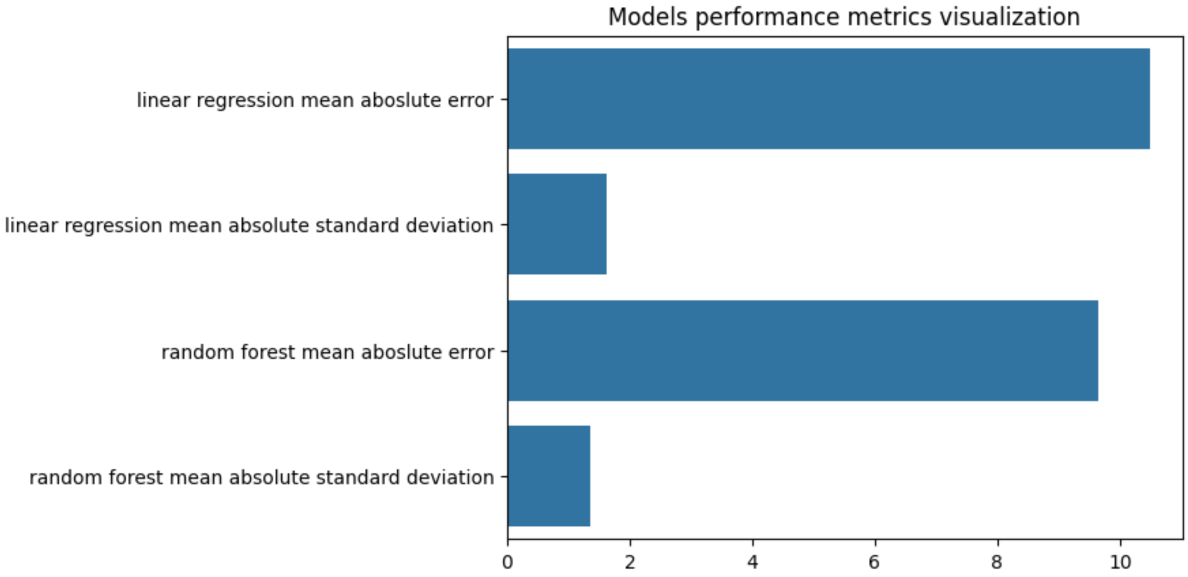


Figure 15: models performance comparison.

In terms of data quantity, the learning curve below shows that there is enough data in the initial dataset as the it is already levelling oof with a sample size of 40-50.

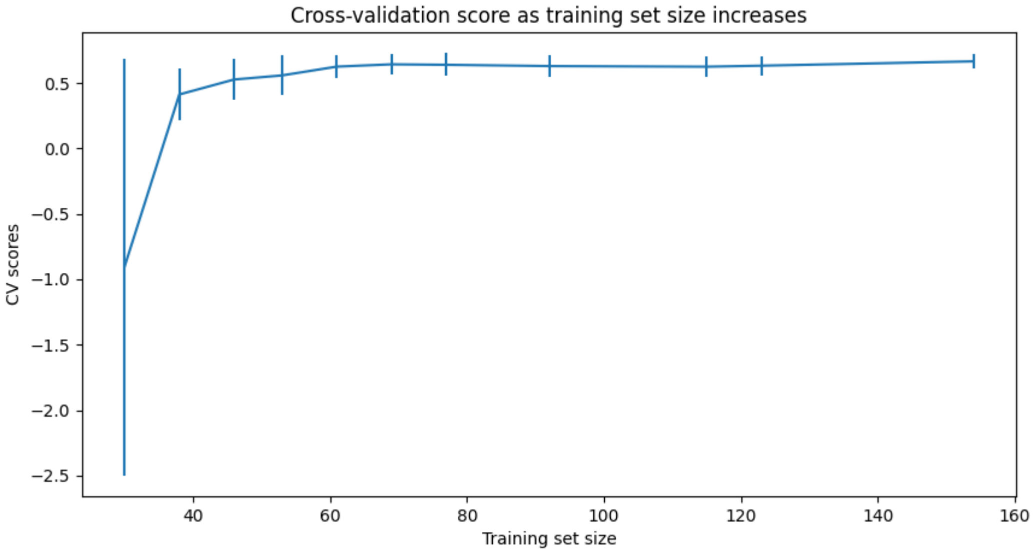


Figure 16: dataset and learning curve.

6. MODELING :

The modeling stage applies the selected algorithm to learn patterns from data and to make prediction (or classification).

6.1. First ticket price prediction:

The "Big Mountain" resort modelled price is $95.87, while the actual price is $81.00.

6.2. Market context:

For better understanding and to wisely select the facilities to consider for better management, several plots show the situation of the "big Mountain" resort regarding the most market impacting features. Concerning the ticket price, "Big Mountain" is at the maximum in Montana and on the high range considering all the states. Concerning the facilities features, as per above, Big Mountain resort is on the high range for all of them except for Trams. This motivates the optimization of the existing facilities in terms of ticket price making.

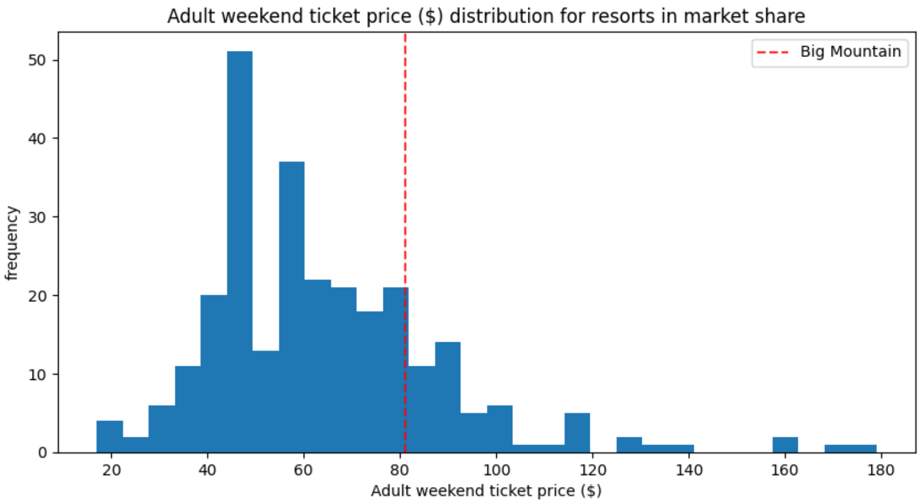


Figure 17: Big Mountain states ticket price.

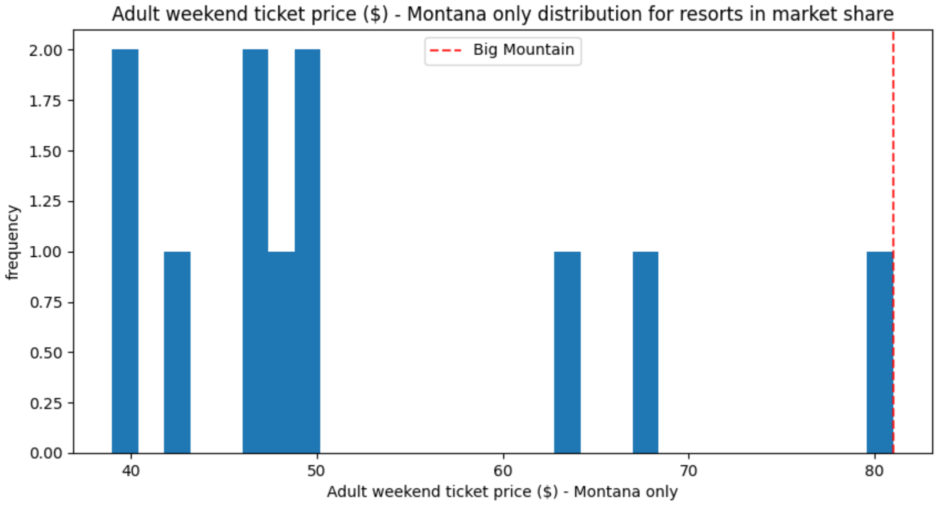


Figure 18: Big Mountain and Montana ticket price.

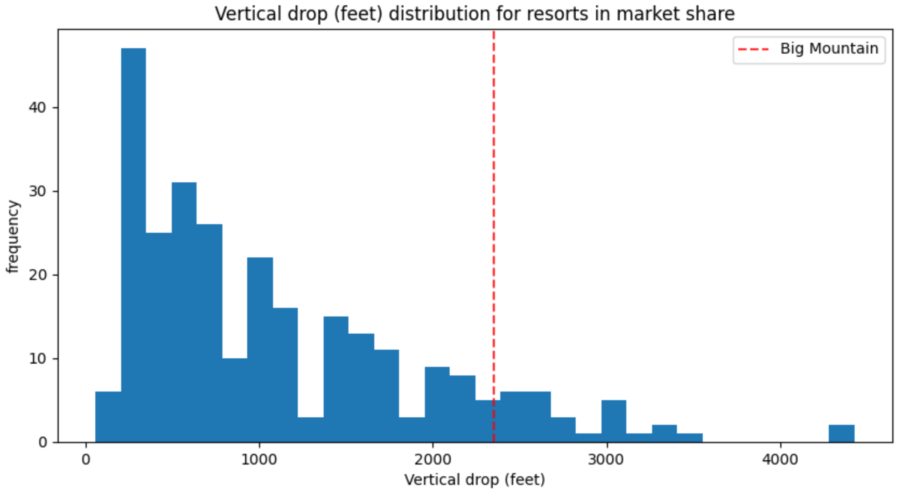


Figure 19: Big Mountain and vertical drop.

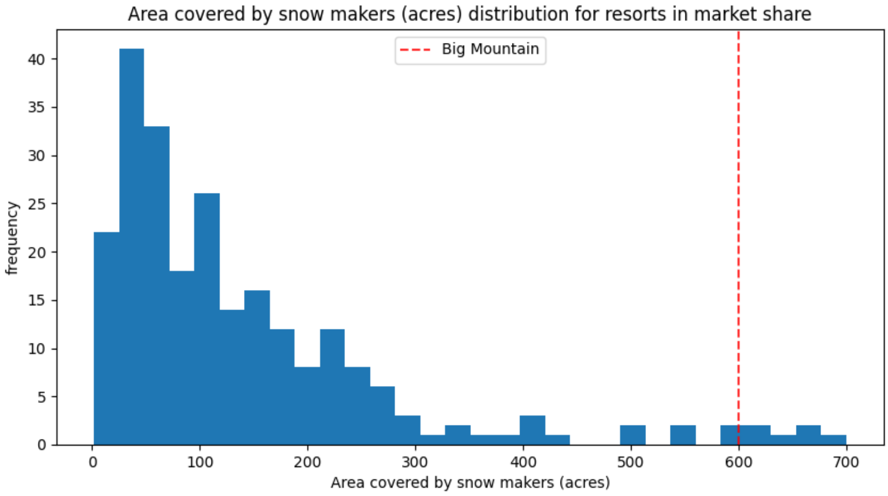


Figure 20: Big Mountain and snow making.

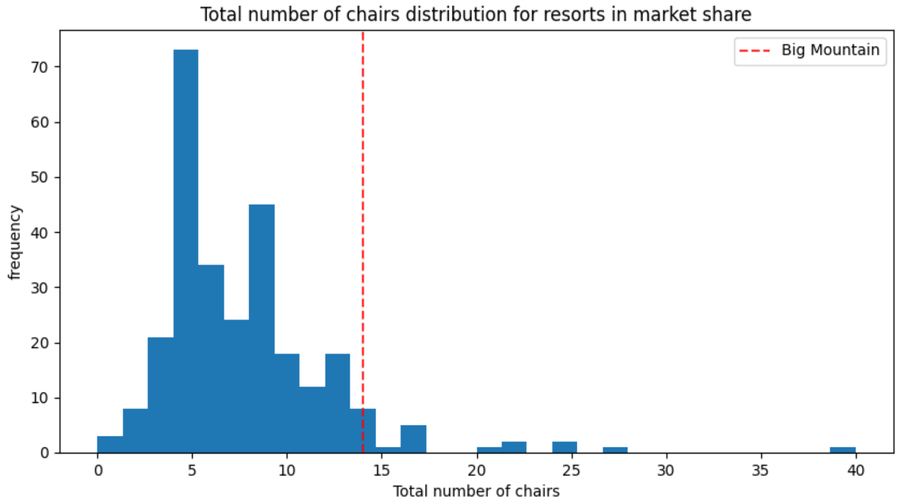


Figure 21: Big Mountain and total number of chairs.

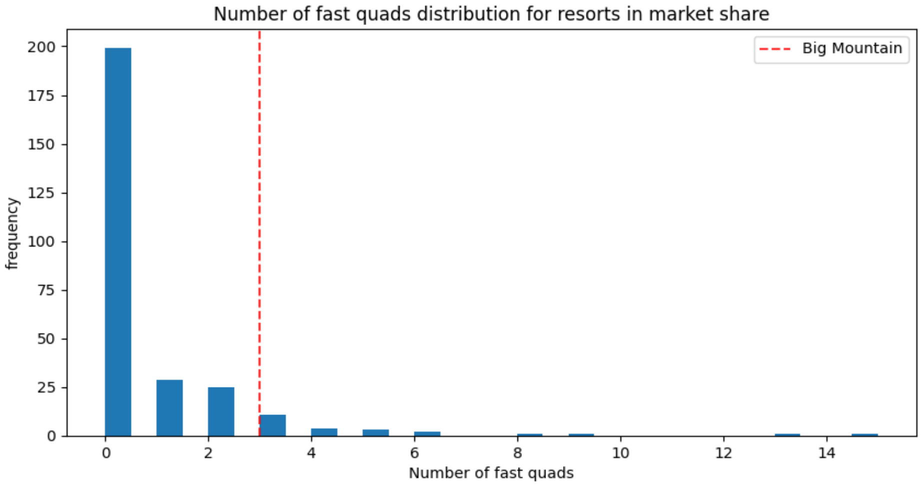


Figure 22: Big Mountain and number of fast quads.

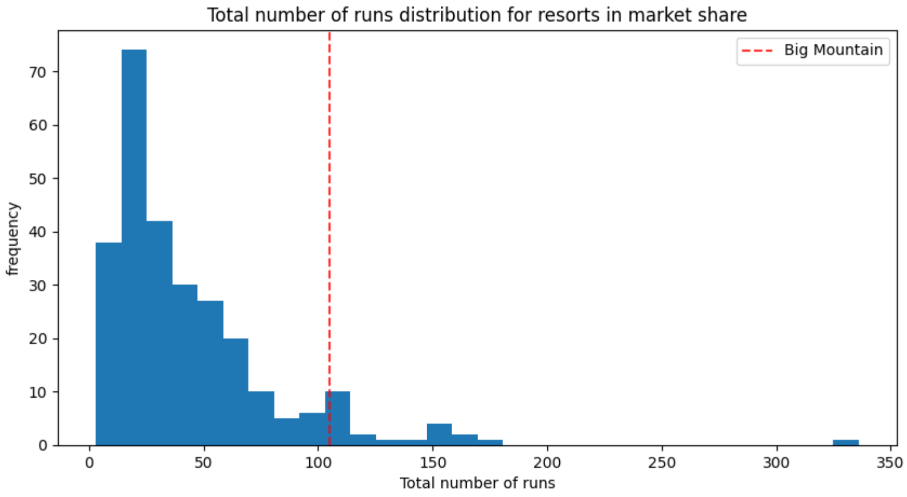


Figure 23: Big Mountain and number of runs.

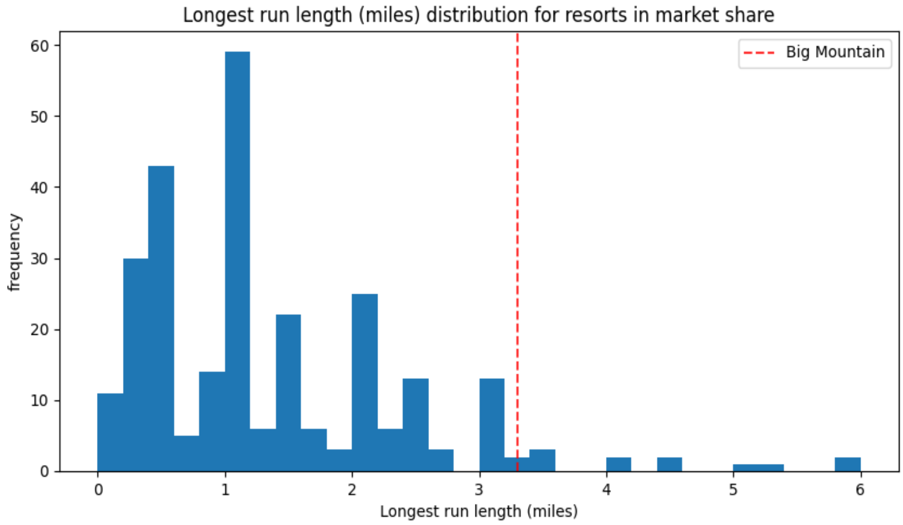


Figure 24: Big Mountain and run length.

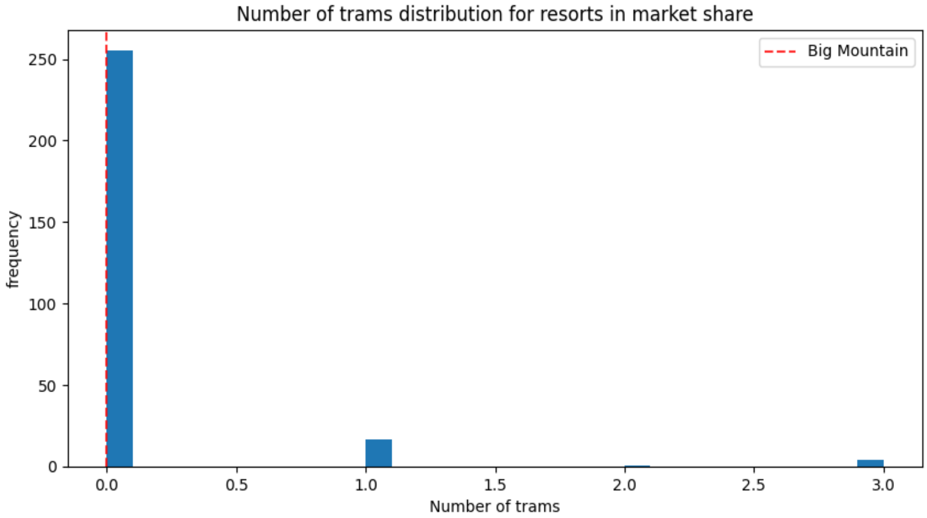


Figure 25: Big Mountain and number of trams.

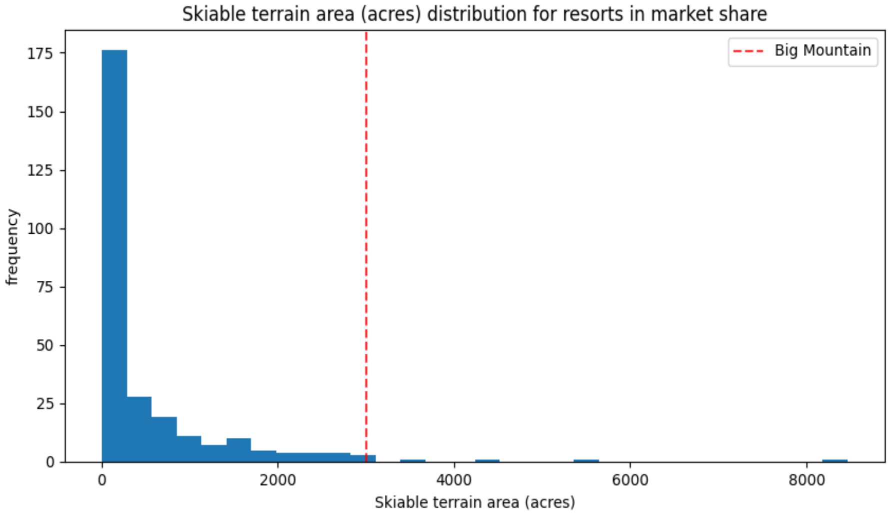


Figure 26: Big Mountain and skiable terrain.

6.3. Facilities management modeling scenarios:

As a kind reminder, the purpose of this study and research was to instruct some potential operations cost cuts and their potential to sustain or increase the resort revenue. 4 scenarios have been proposed by Big Mountain executives to capitalize on the resort facilities:

- **Closing at least 10 runs:**

Closing 1 run does not impact the ticket price. However, closing more than 1 run will negatively impact the ticket price, hence the revenue.

As a suggestion, perhaps another scenario could be to close 1 run as well as the potential related chair lift as well as the potential related snow making if any. This could help decrease the operations cost while keeping the same ticket price, which could increase the revenue.

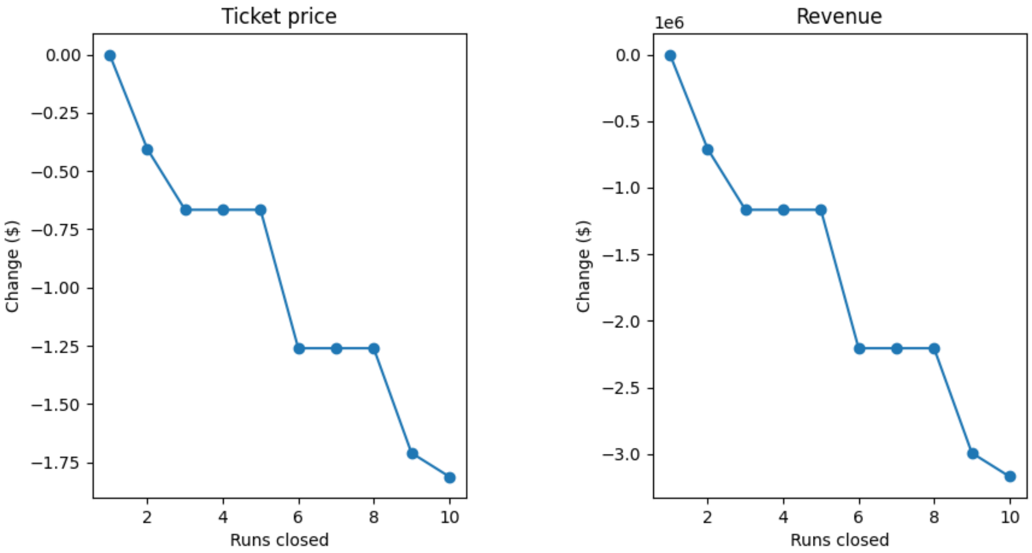


Figure 27: impact of number of runs closing on ticket price and revenue.

- **Increase vertical drop by 150 m + 1 chair lift:**

This scenario supports ticket price and revenue increase.

"This scenario increases support for ticket price by $1.99 Over the season, this could be expected to amount to $3474638".

As a suggestion, the additional cost of the new chair lift could be wise if it is associated with this increase in the vertical drop. If the new chair lift is not associated with this increase in drop, we can also simply check if the revenue benefit from this scenario covers its cost.

- **Increase vertical drop by 150 m + 1 chair lift + 2 acres of snow making:**

No difference in ticket price and revenue.

- **Increase the longest run by 0.2 miles + 4 acres of snow making:**

No difference in ticket price and revenue.

7. DOCUMENTATION AND COMMUNICATION:

This actual stage (present document) presents findings, methods, and conclusions clearly to the resort stakeholders. It ensures transparency and supports informed decision-making.

8. RECOMMENDATIONS AND FUTURE SCOPE OF WORK:

There might be some missing data in the initial dataset, such as:

- Actual number of customers per year.
- The actual number of customers stays per year.
- Opening dates of the different facilities.
- Staff wages.
- Equipment maintenance.
- Energy costs.
- Capital depreciation.
- Etc.

All this data could help to build a model less reliant on other resort price strategies and to build up Big Mountain prices based on its facilities and cost management. Other ski resort data could still be kept for market benchmarking.

To support the Big Mountain resort Executive team in their search of the best facilities management and ticket price strategy, the model might be accompanied by some intuitive and user-friendly graphical interface to easily capture the different scenario proposals and to return tables and charts of its outcome for better understanding.

****