GUIDED CAPSTONE PROJECT REPORT

HUNG CONG TRAN

1. Problem Identification

Problem: What is the best ticket price for Big Mountain resort to increase the revenue that can compensate for the additional operating costs of $1,540,000 within 6 months and gain at least doubled profit? Are there any opportunities to combine the best ticket price option with the operating cost cut with in 6 months to make the profit at least to be doubled?

Context: Big Mountain Resort is a ski resort located in Montana which serves 350,000 people ski and snowboard at Big Mountain each year. The resort has recently installed an additional chair lift to accommodate more visitors but it made the operating costs increased by $1,540,000. The resort would like to be advised on the best ticket price and combining it with the cost cut to make the profit to be at least doubled within 6 months.

Criteria for success: Get the best ticket price and combining it with the cost cut to gain at least double profit with in 6 months

Scope of solution space: Analyze the features (location, snow condition, facilities,...) and ticket prices of 330 same market share resorts (including Big Mountain Resort) in the US to decide the best ticket price.

Constraints within solution space: Lack information about the financial pictures of other resorts in the same market share. The time to evaluate the effectiveness of the new price and the cost cut may be long.

Stakeholders to provide key insight: CEO, Director of Operations: Jimmy Blackburn, Database Manager: Alesha Eisen.

Key data source: Data provided by the database manager that contains information from 330 same market share resorts (including Big Mountain Resort) in the US.

2. Data Wrangling

There are 330 rows in the original data which includes our own resort (Big Mountain resort) and there are 27 columns (Name, Region, state, summit_elev, vertical_drop, base_elev, trams, fastEight, fastSixes, fastQuads, quad, triple, double, surface, total_chairs, Runs, TerrainParks, LongestRun_mi, SkiableTerrai_ac, Snow Making_ac, daysOpenLastYear, yearsOpen, averageSnowfall, AdultWeekday, AdultWeekend, projectedDaysOpen, NightSkii_ac).

We modified the dataset as follows:

(1) Fix one incorrect value from the column SkiableTerrain_ac using information from the Internet.
(2) Drop the fastEight collumn since half of values in this column is missing and all but the others are zero values.
(3) Drop the row with yearsOpen column value is greater than or equal 1000 to eliminate the outliers.

(4) Rows with missing values on both AdultWeekend, AdultWeekday columns have been removed since it can not provide any information about ticket prices.
(5) Values in AdultWeekend column and AdultWeekday column are almost the same. The differentce mostly applies to the resort with weekeday prices less than 100. Weekday and weekend prices in Montana seemed equal. We drop the AdultWeekday column since it contains more missing values.
(6) Drop rows whose AdultWeekend values are missing.

We finally obtains a new data set with 277 rows and 25 columns. There are still some missing values. The target feature for the project is the AdultWeekend column.

3. Exploratory Data Analysis

Through the EDA above, we basically perform the following steps:

(1) We first explore data containing different relevant features (number of resorts, skiable_area, days_open, number of terrain_parks, number of nightskiing area, population, area) of 35 states. All features are numerical except the state names which are used as the index column.
(2) We replace some features (population, area) by other more useful features (resort density in terms of population and area).
(3) We standardize each feature and use PCA technique to reduce the number of features (from 7 to 2) while keeping "enough" information (77.2%) (i.e.variance (77.2%)).
(4) We plot data to see the relationship between the two new features and the ticket prices. However, we have go back to the original features (before using PCA) since they are more "interpretable". We observe that the resort density (in terms of population and area) are the ones contributing most the two PCA new features.
(5) We combine (left merge) the resort data to the state data and "localize" the resort features based on states resorts are located. We also create some new features by combining existing features and find the connections beween each resort features and price.

4. Pre-processing and Training Data Development

We splits the data into training set (70%) and test (30%). We code the evaluation metrics for regression problems (R2 scores, mean absolute value error, mean square absolute value error) and we also call sklearn library for these metrics. We also use Sklearn to call various models (dummy model, linear model, random forest regressor model) to fit the data, check their parameters, and evalute their predictions (on traing set, testing set, and using cross-validation). We work on data imputation, fill missing data, and scaling data. We use pipeline tool from sklearn to to combine various of traing steps, and use GridSearch tool to choose the best parameters for models. Finally we choose the random forest regressor model with best parameters choosen from Grid Search to move forward since it perform the best on the evaluation metric.

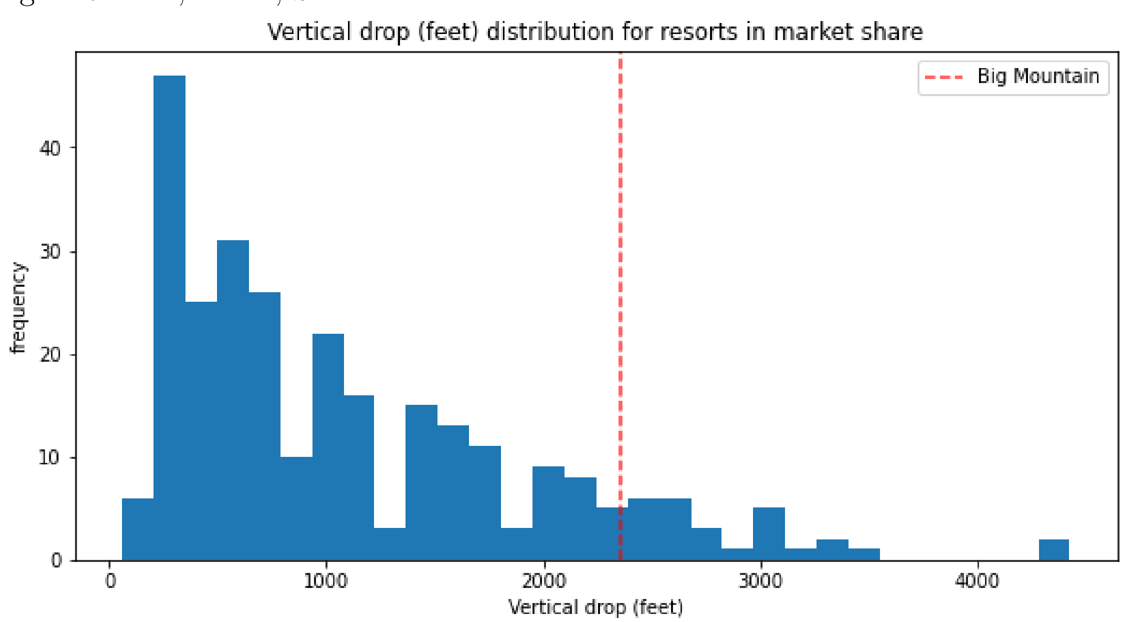
5. Modeling, Scenario modeling, pricing recommendation, and conclusion, and future scope of work

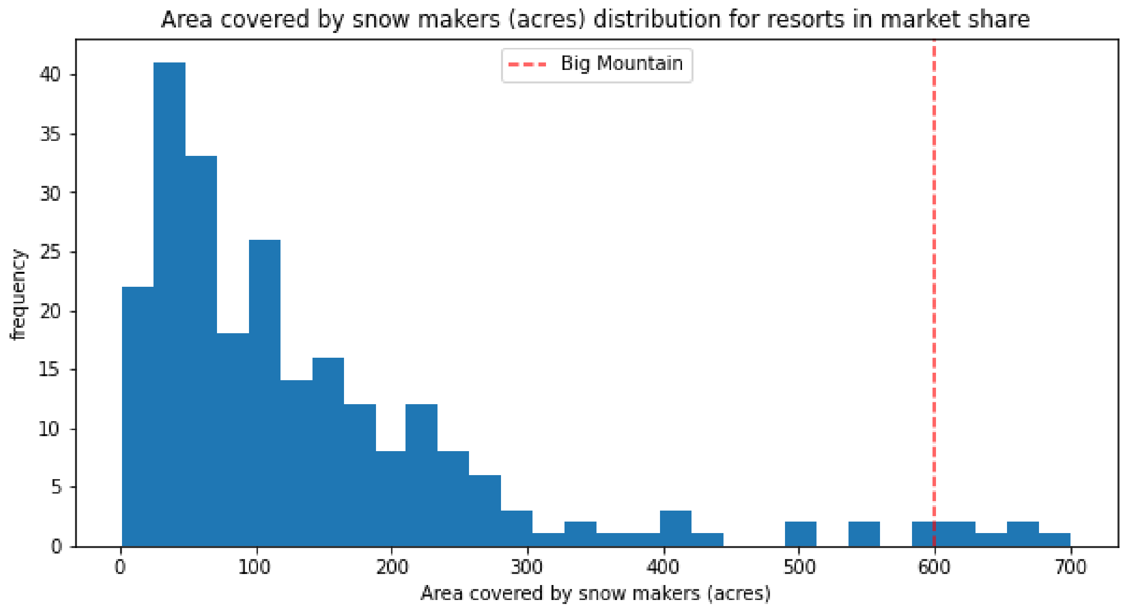
Big Mountain currently charge $81.00 for a weekend ticket and the model suggests the ticket should be increased to $95.87. We discuss the rank of each facility of Big Mountain among the corresponding facility of other resorts and convince the business leader than the price change makes sense. We discuss various scenarious of changing Big Mountain's facilities, use the model to predict ticket price changes, and use the assumptions on potential number of tickets sold to predict the revenue changes.

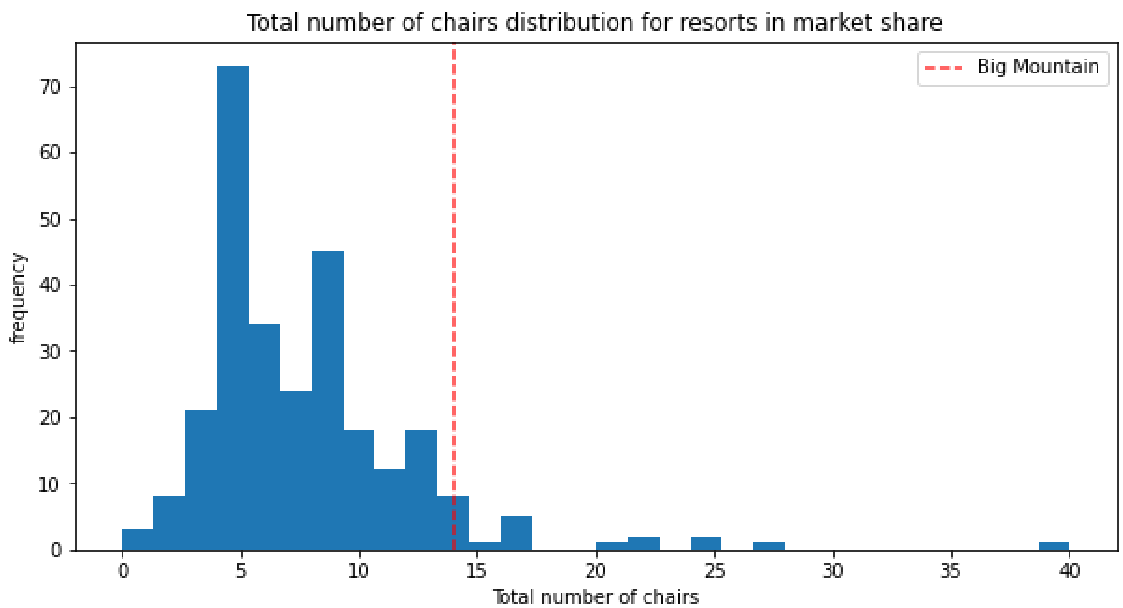
The data contains information of ticket prices and we can predict the revenue based on the assumption of expected number of tickets sold. However, the number of tickets sold may strongly depends on the change of ticket price as well as the change of facilities. So the potential future model should use additional data on the number of ticket sold and set the revenue as target instead of ticket prices. Moreover, we also need to integrate the model result with the operational cost to have a clearer financial picture.

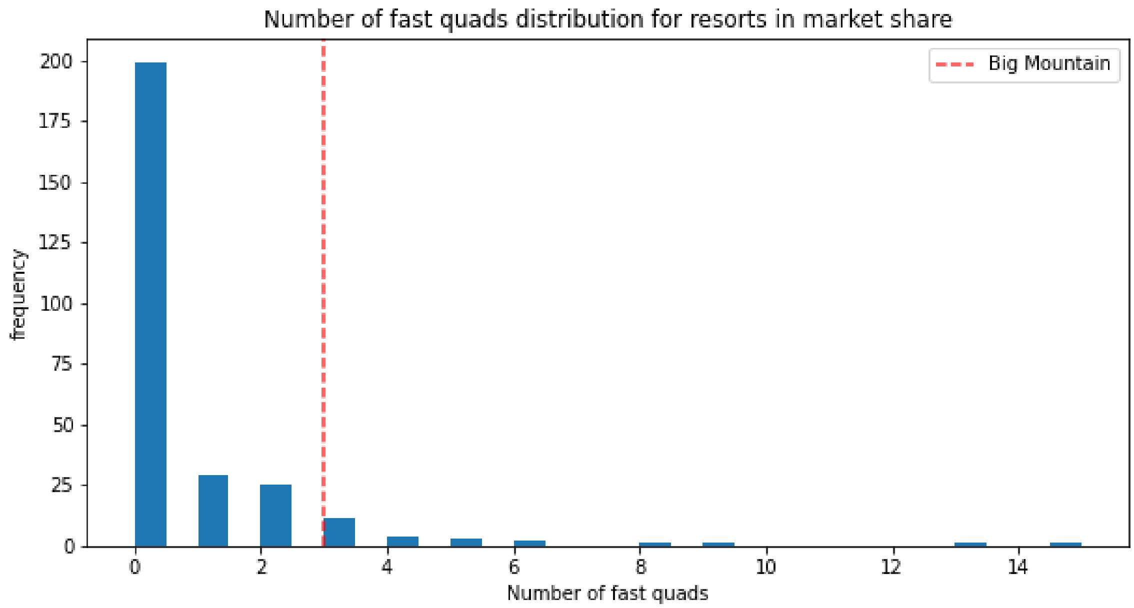
6. Big Mountain Resort In Market Context charts and plots

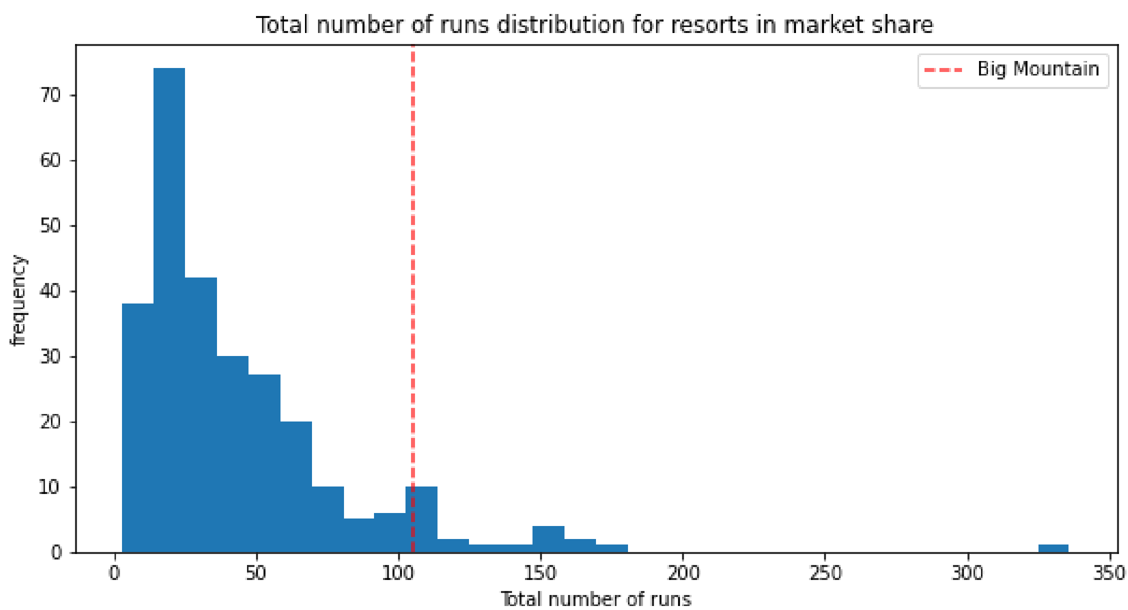
Features that came up as important in the modeling (not just our final, random forest model) included: vertical_drop, Snow Making_ac, total_chairs, fastQuads, Runs, LongestRun_mi, trams, SkiableTerrain_ac

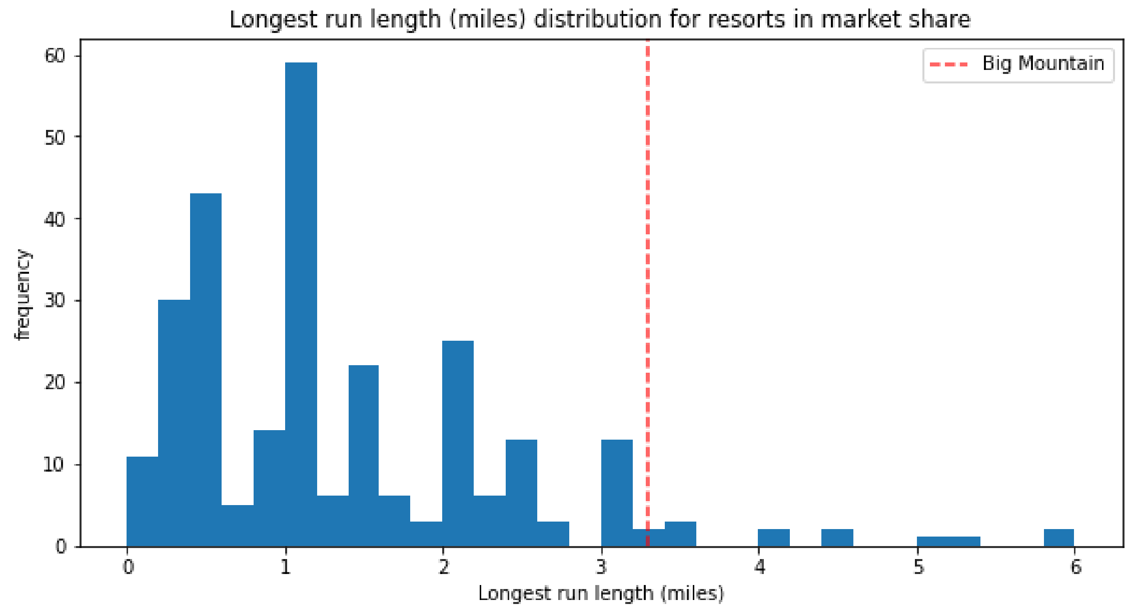


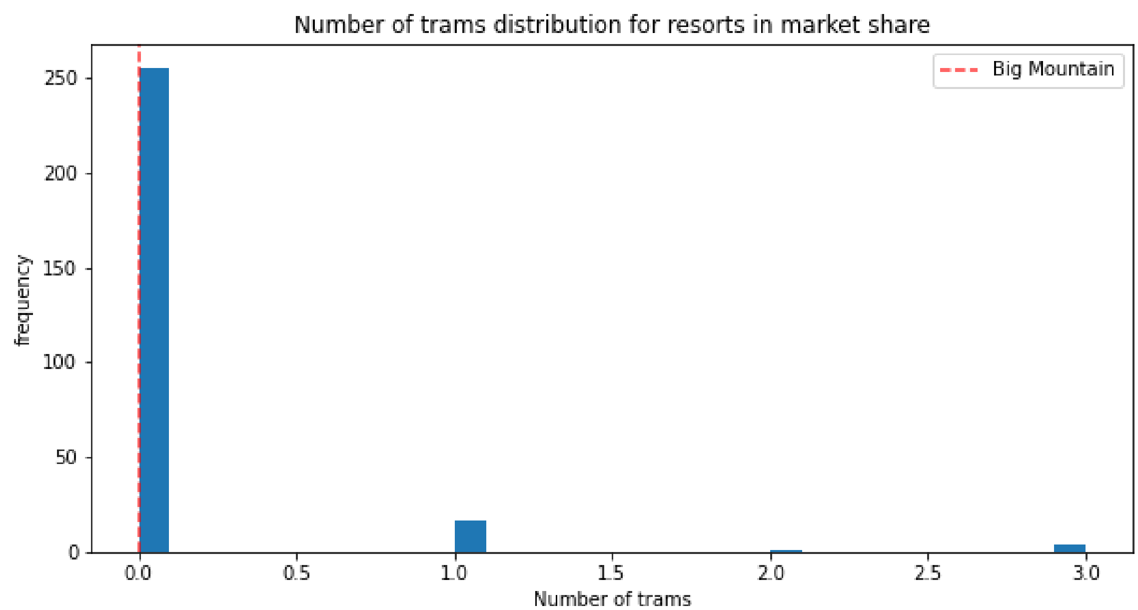

Area covered by snow makers (acres) distribution for resorts in market share
Big Mountain
frequency
0
5
10
15
20
25
30
35
40
0
100
200
300
400
500
600
700
Area covered by snow makers (acres)

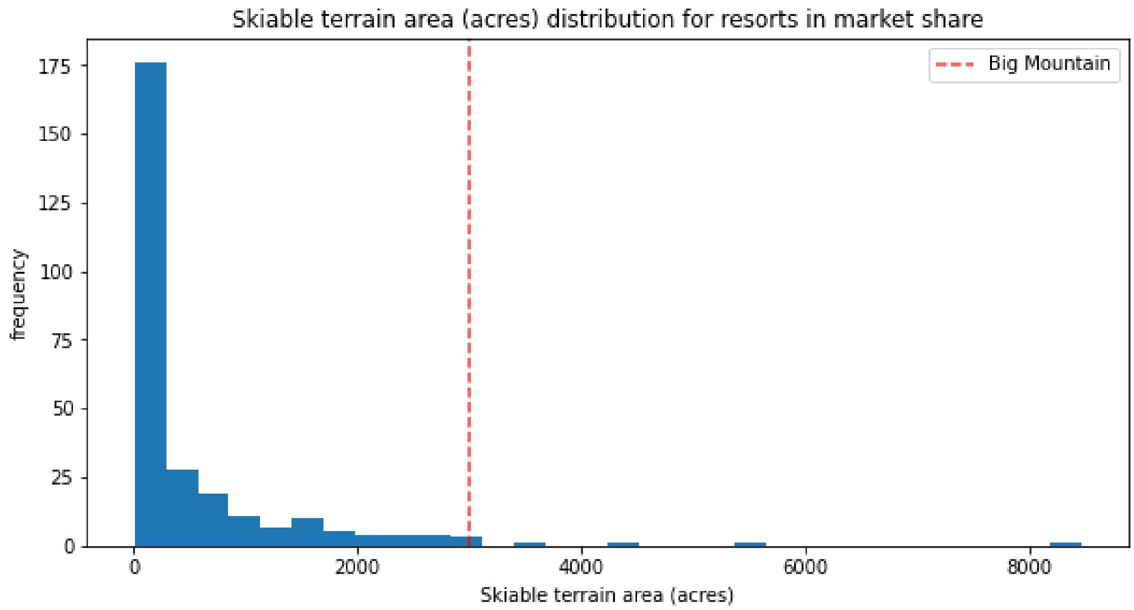

Total number of chairs distribution for resorts in market share
Big Mountain
frequency
0
10
20
30
40
50
60
70
0
5
10
15
20
25
30
35
40
Total number of chairs

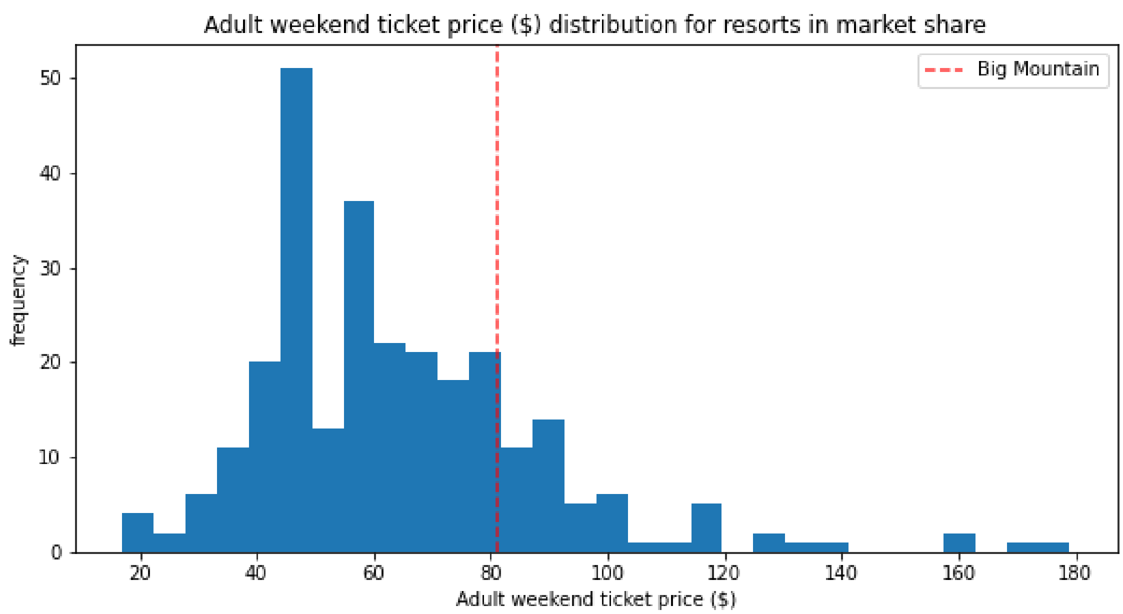

Number of fast quads distribution for resorts in market share
Big Mountain
frequency
Number of fast quads
0
25
50
75
100
125
150
175
200
0
2
4
6
8
10
12
14


Total number of runs distribution for resorts in market share
Big Mountain
frequency
Total number of runs
0
10
20
30
40
50
60
70
0
50
100
150
200
250
300
350


Longest run length (miles) distribution for resorts in market share
Big Mountain
frequency
0
10
20
30
40
50
60
0
1
2
3
4
5
6
Longest run length (miles)


Number of trams distribution for resorts in market share
Big Mountain
frequency
0
50
100
150
200
250
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Number of trams


Skiable terrain area (acres) distribution for resorts in market share
Big Mountain
frequency
175
150
125
100
75
50
25
0
0
2000
4000
6000
8000
Skiable terrain area (acres)


Adult weekend ticket price ($) distribution for resorts in market share
Big Mountain
frequency
50
40
30
20
10
0
20
40
60
80
100
120
140
160
180
Adult weekend ticket price ($)