Big Mountain Resort

Capstone Project Report

Mehak Sethi

Overview

Big Mountain Resort is one of the oldest and most premier resorts in Montana, USA. It has a wide variety of offerings for its guests including adventure activities like Snowboarding, Skiing, Chair Rides etc.

Some facts about the resort:

- The resort was established in 1947 (72 yrs old).
- It has the most well lit area for Night Skiing (600 acres) which is the highest in Montana and its neighbouring states viz South Dakota, Idaho and Wyoming.
- It is in the top 3 resorts in Montana having the most number of Chair Lifts.
- It receives an average snowfall of 333 inches which is 3rd highest in Montana.
- It has the highest area in Montana covered by snow making machines adding to the snowfall it gets.
- It has 105 runs on the resort which is the third highest in Montana and 5th highest among the neighbouring states.
- The resort is 2nd in terms of triple seater ChairLifts in its peers in Montana.
- It was open for 123 days last year and 350,000 guests visited the resort last year.
- The resort has the most premierly priced tickets for both weekdays and weekends in Montana.

Approach Followed:

- Problem Identification (stated below)
    - Understanding the background and the problem faced by the resort which is further broken down into actionable tasks to achieve goals.
- Data Wrangling and Preprocessing
    - Data Ingestion, organisation, setting up of repositories, data definition and cleaning
- Exploratory Data Analysis
    - Data profiling, outliers analysis, understanding data relationships, identification and creation of features
- Preprocessing and Training Data Development
    - Creating dummies for model development, standardising the magnitude of features using scaling, splitting the data into train and test set.
- Modelling
    - Feeding the training data to the Linear Regression model, evaluating the validation metrics and developing new iterations to finalise the model.
- Documentation
    - Making it meaningful for everybody!

Problem Overview

Big Mountain Resort has incurred a capex of $1,540,000 recently for installing an additional chair lift in the resort so as to accommodate more and more guests. It wants to recover the capex while maintaining its 9.2% profit margin this year which makes for a revenue of atleast $1,696,035 to be achieved this year.

It is estimated that 350,000 guests visited the resort last year and we have assumed that the same number of guests shall visit this year too.

Goals

1. To achieve revenue of atleast $1,696,035 this year so that capex can be recovered with profit margins.
2. To understand factors contributing to the pricing so that they can be leveraged to maximise revenues.

Data Source:

The analysis has been done on the csv file shared by the DataBase Manager. The file has information about various resorts of the country, their ticket prices, their distinguishing features about their location, chairlifts, terrain, snowfall (including snowfall machines), etc.

Stepwise Analysis:

I. Data Wrangling

This crucial step involved loading the data for analysis, organising it into different objects and tables to facilitate further analysis after setting the repositories. Data Types were checked to make sure that all fields were read correctly, summary statistics generated, erroneous features like 'Unnamed: 0' we dropped as it was redundant. The next step here was to check for missing values which have been deleted/ imputed.

Data Information - Total observations 303, total variables 27 out of which float64 (13 variables), int64 (11 variables), object (3 variables).

Unique Values (as percentage of size of the data)

```
Name                 3.692480
Region               0.426487
state                0.392817
summit_elev          2.985410
vertical_drop        2.166105
base_elev            2.738496
trams                0.056117
fastEight            0.022447
fastSixes            0.078563
fastQuads            0.157127
quad                 0.078563
triple               0.101010
double               0.112233
surface              0.134680
total_chairs         0.347924
Runs                 1.234568
TerrainParks         0.123457
LongestRun_mi        0.471380
SkiableTerrain_ac    2.166105
Snow Making_ac       1.604938
daysOpenLastYear     1.245791
yearsOpen            0.808081
averageSnowfall      1.200898
AdultWeekday         0.931538
AdultWeekend         0.909091
projectedDaysOpen    1.010101
NightSkiing_ac       0.875421
dtype: float64
```

Descriptive Statistics

	count	mean	std	min	25%	50%	75%	max
summit_elev	330.0	4591.818182	3735.535934	315.0	1403.75	3127.5	7806.00	13487.0
vertical_drop	330.0	1215.427273	947.864557	60.0	461.25	964.5	1800.00	4425.0
base_elev	330.0	3374.000000	3117.121621	70.0	869.00	1561.5	6325.25	10800.0
trams	330.0	0.172727	0.559946	0.0	0.00	0.0	0.00	4.0
fastEight	164.0	0.006098	0.078087	0.0	0.00	0.0	0.00	1.0
fastSixes	330.0	0.184848	0.651685	0.0	0.00	0.0	0.00	6.0
fastQuads	330.0	1.018182	2.198294	0.0	0.00	0.0	1.00	15.0
quad	330.0	0.933333	1.312245	0.0	0.00	0.0	1.00	8.0
triple	330.0	1.500000	1.619130	0.0	0.00	1.0	2.00	8.0
double	330.0	1.833333	1.815028	0.0	1.00	1.0	3.00	14.0
surface	330.0	2.621212	2.059636	0.0	1.00	2.0	3.00	15.0
total_chairs	330.0	8.266667	5.798683	0.0	5.00	7.0	10.00	41.0
Runs	326.0	48.214724	46.364077	3.0	19.00	33.0	60.00	341.0
TerrainParks	279.0	2.820789	2.008113	1.0	1.00	2.0	4.00	14.0
LongestRun_mi	325.0	1.433231	1.156171	0.0	0.50	1.0	2.00	6.0
SkiableTerrain_ac	327.0	739.801223	1816.167441	8.0	85.00	200.0	690.00	26819.0
Snow Making_ac	284.0	174.873239	261.336125	2.0	50.00	100.0	200.50	3379.0
daysOpenLastYear	279.0	115.103943	35.063251	3.0	97.00	114.0	135.00	305.0
yearsOpen	329.0	63.656535	109.429928	6.0	50.00	58.0	69.00	2019.0
averageSnowfall	316.0	185.316456	136.356842	18.0	69.00	150.0	300.00	669.0
AdultWeekday	276.0	57.916957	26.140126	15.0	40.00	50.0	71.00	179.0
AdultWeekend	279.0	64.166810	24.554584	17.0	47.00	60.0	77.50	179.0
projectedDaysOpen	283.0	120.053004	31.045963	30.0	100.00	120.0	139.50	305.0
NightSkiing_ac	187.0	100.395722	105.169620	2.0	40.00	72.0	114.00	650.0

Missing Value Percentages and Imputations:

	percent
fastEight	0.503030
NightSkiing_ac	0.433333
AdultWeekday	0.163636
AdultWeekend	0.154545
daysOpenLastYear	0.154545
TerrainParks	0.154545
projectedDaysOpen	0.142424
Snow Making_ac	0.139394
averageSnowfall	0.042424
LongestRun_mi	0.015152
Runs	0.012121
SkiableTerrain_ac	0.009091
yearsOpen	0.003030

(**fastEight** → Missing values imputed with zero as a difference of total_chairs and fastEight was taken to check the actual status of missing values in this variable, accordingly it was decided to impute zero.

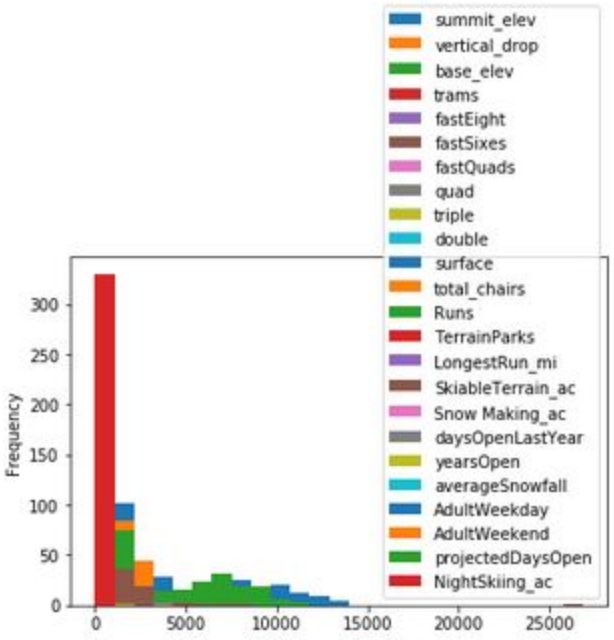
AdultWeekday → Missing values imputed with mean, this could not be filled with zero as this contains the most vital feature i.e. Price.

Runs, TerrainParks, LongestRun_mi, SkiiableTerrain_ac, SnowMaking_ac, daysOpenLastYear, yearsOpen, averageSnowfall, AdultWeekend, projectedDaysOpen, NightSkiing_ac → all were imputed with mean. Imputation with zero was not a good idea as these are key numeric attributes.)

II. Exploratory Data Analysis

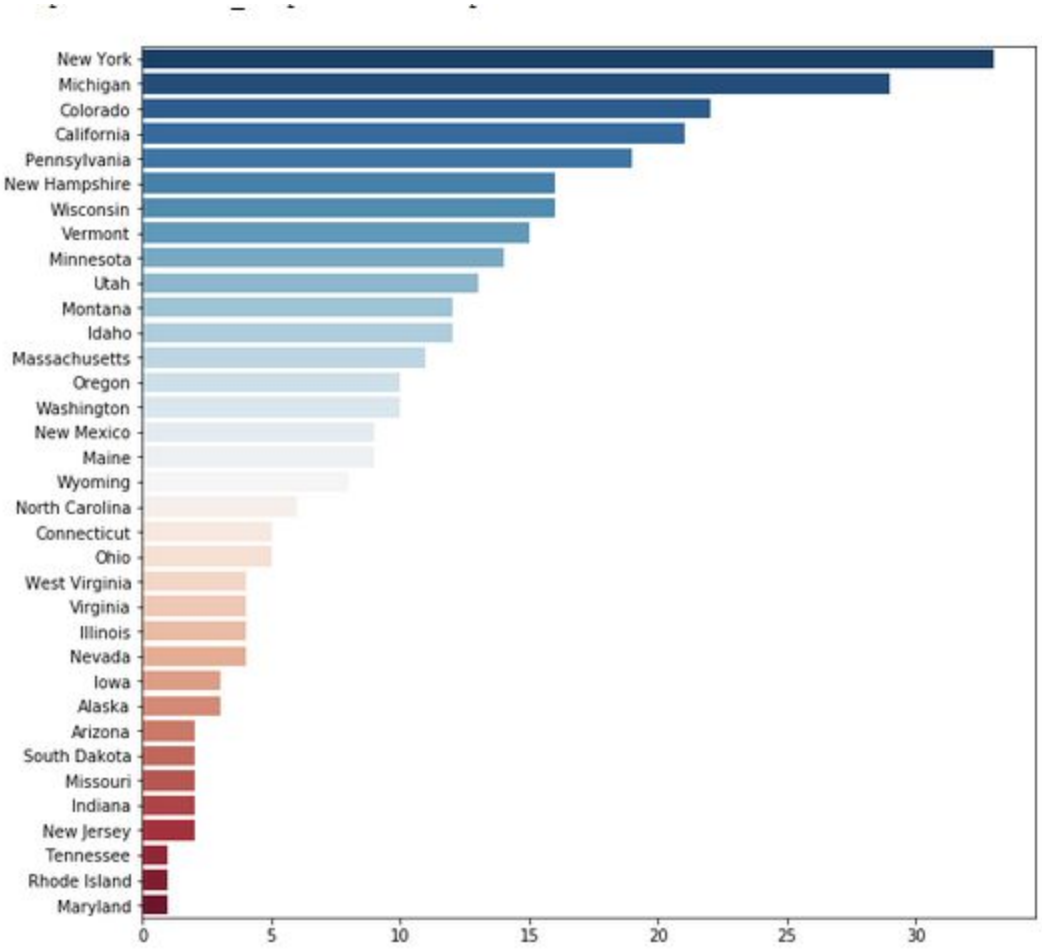
Further to the Data Wrangling step, the following analysis was done to understand the data and to ensure that it is error free before it was fed to the Linear Regression Model.

Visualisations

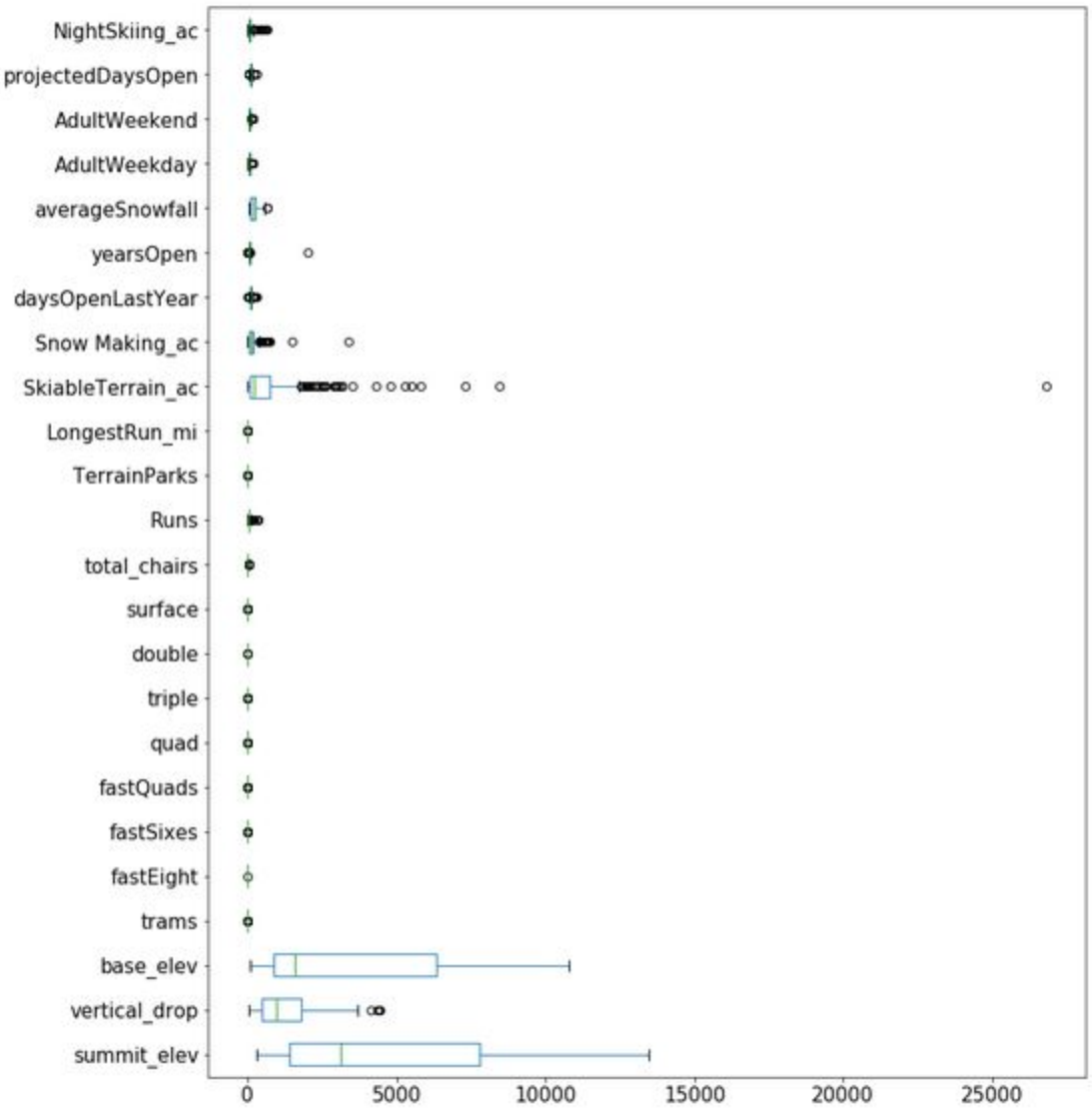


NightSkiing_ac has the biggest spike in the data frame.

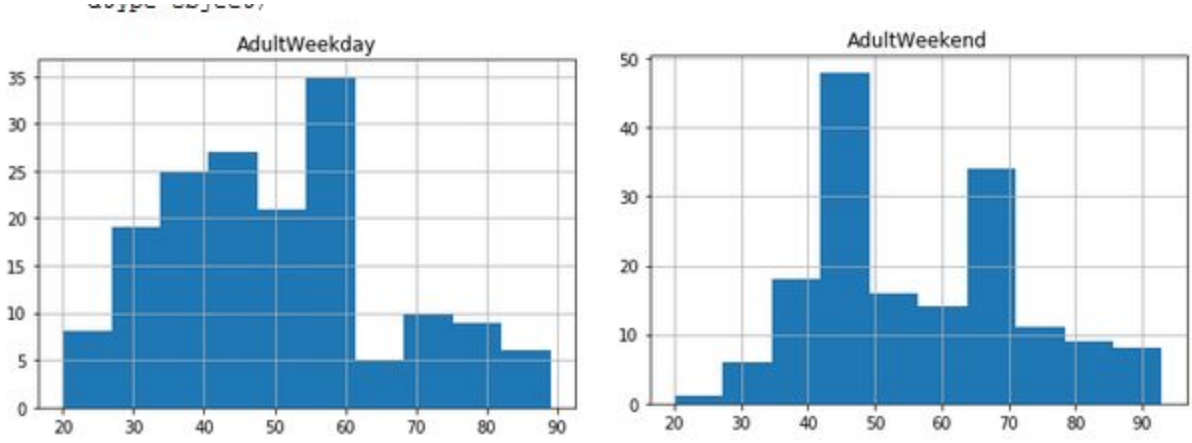
Region and **State** had the same values (as in the below chart) and were highly collinear so Region was dropped. **New York** has the highest number of resorts as per the data. **Montana** is in the middle of the distribution (11th spot). **Maryland** has the least number of resorts.



Boxplot analysis was done to identify the presence of outliers in the data. However, these were retained as removing them was resulting in removal of Big Mountain Resort row which was defeating the purpose of the analysis. And these potential outliers did not contain any impossible/ redundant values, so retaining them was necessary.

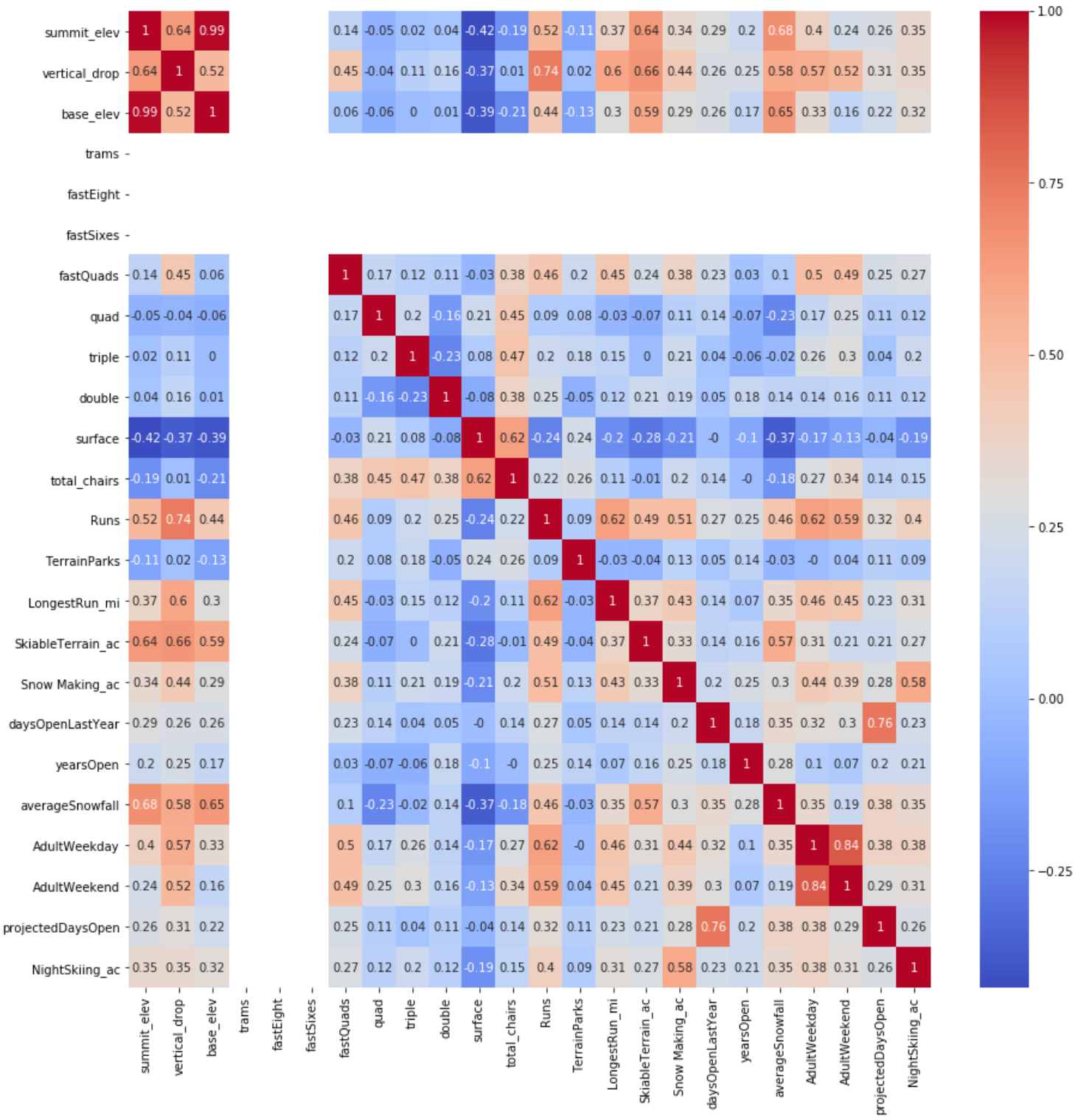


AdultWeekday and **AdultWeekend** confirm the presence of high values but these are meaningful values. **AdultWeekend** has been taken as the Dependent Variable (y) for the purpose of the analysis. It is a bit skewed but is close to normally distributed.



Correlation Matrix

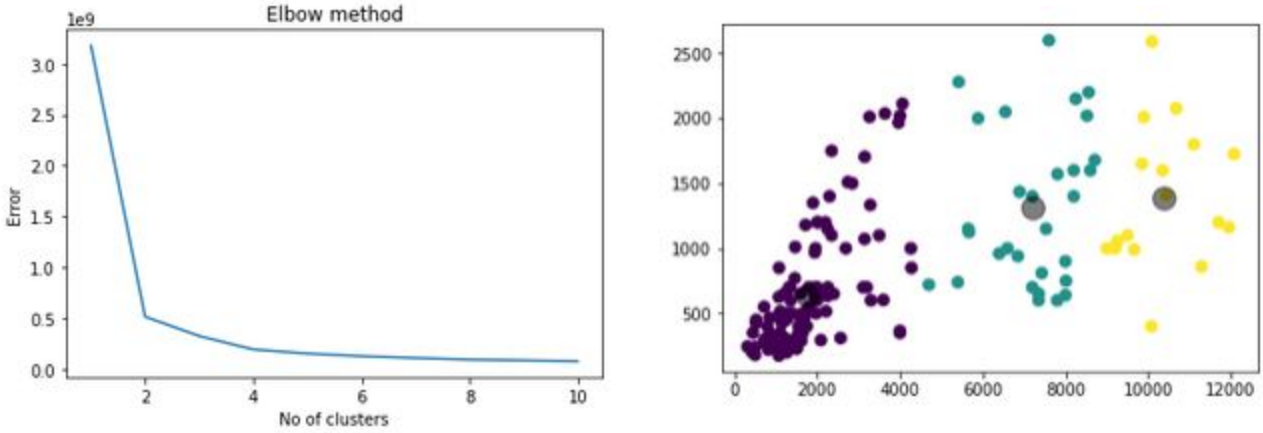
Summit_elev is highly correlated with **vertical_drop** and **base_elev. AdultWeekday** and **AdultWeekend** are highly correlated. **DaysOpenLastYear** is highly correlated with **projectedDaysOpen.** Correlations more than 0.4 are considerably significant. The red ones represent high correlations.



Feature Creation using k-means clustering

K-means clustering was used to segment the data into clusters based on similarities. Three clusters were identified based on the elbow plot. Also, the clusters are represented by different colours in the

right side plot. A cluster column was included in the main data frame to identify which resort belongs to which cluster. Big Mountain Resort falls in cluster 2.



III. Data Preparation and Training Data Development

Initially dummies were created for the 'state' variables. But it was seen that these variables were not of much help for the price prediction. Also, 'state' is something that is out of the resort's locus of control. So it was dropped.

Subsequently the features and the price (AdultWeekend) were separated into X and Y data frames.

The data frame X (containing the features) was then scaled using StandardScaler from the sklearn package. This is necessary to normalise the magnitude of various features so that they are on a similar level. Similarly, the AdultWeekend (y) was flattened and stored into an array.

Further, the data (both X and y) was split into train and test data sets. Train data had 75% of the values and test data had 25% of the values. The purpose was to feed the training data to the models and obtain predictions on the test data.

IV. Modelling

The train data was fed to various Linear Regression models and the models were evaluated based on the predictions and actual values and several iterations of modelling was done.

The main metrics considered for model comparison were explained variance and mean absolute error. The best model had the maximum value for explained variance and minimum value for mean absolute error.

V. Observations and Road Ahead

The model which was finalised was explaining approx 90% of the variance in the price with the mean absolute error of 5.87 which was the least in various iterations.

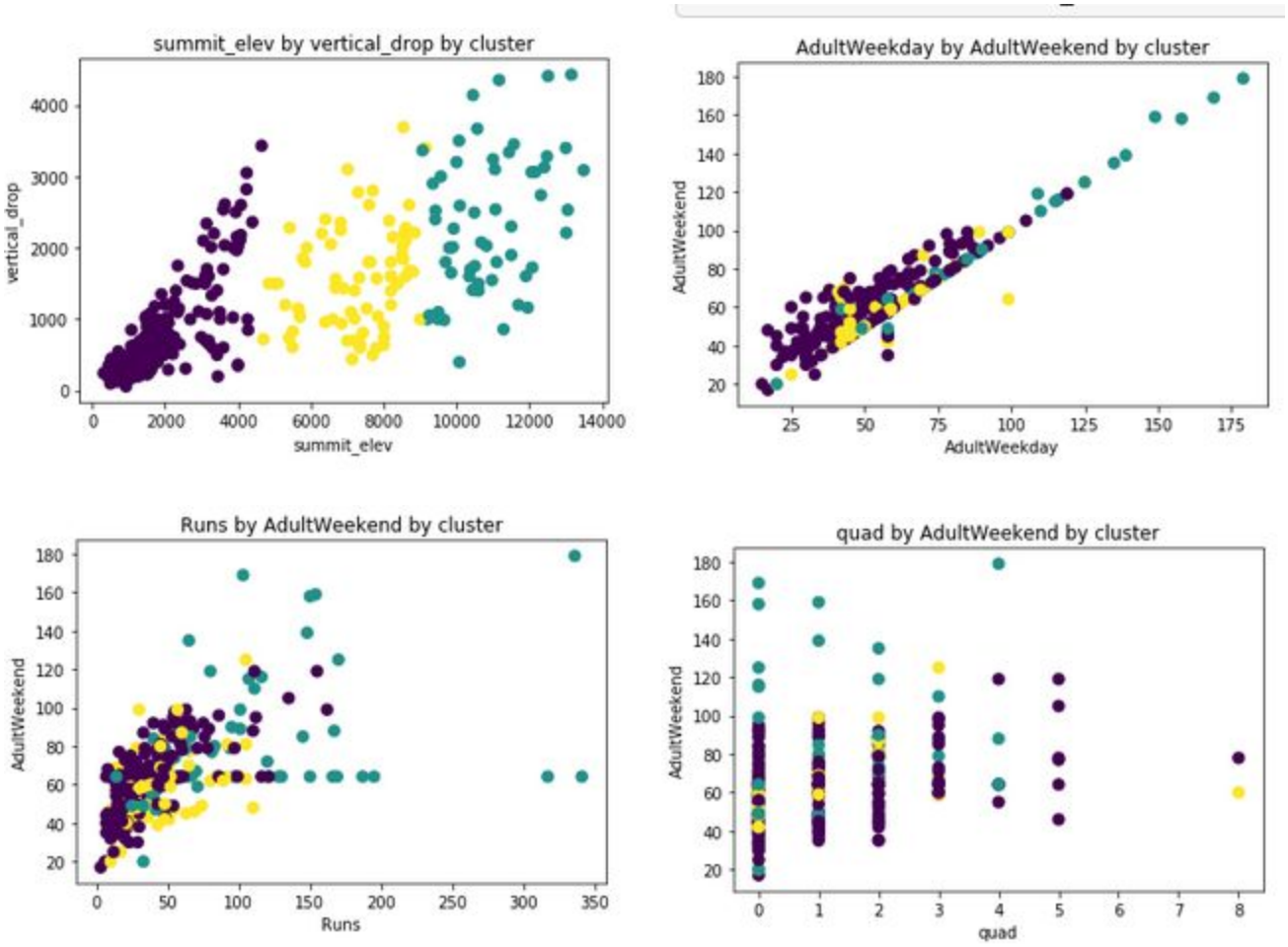
The top features impacting the price (AdultWeekend) were:

	Coefficient
AdultWeekday	20.229646
summit_elev	4.933824
base_elev	3.754470
vertical_drop	3.432703
Runs	1.706320
quad	1.696857
averageSnowfall	1.619698
clusters	1.205591
fastSixes	1.072904
surface	1.001387
triple	0.982694
NightSkiing_ac	0.943038
TerrainParks	0.896765
fastQuads	0.882700
trams	0.810258
total_chairs	0.776649
daysOpenLastYear	0.526324
Snow Making_ac	0.490824
double	0.309183
LongestRun_mi	0.302387
SkiableTerrain_ac	0.298677
yearsOpen	0.265337
projectedDaysOpen	0.203966
fastEight	0.020445

According to the data, **AdultWeekday, summit_elev, base_elev, vertical_drop, Runs** are the top 5 determinants of **AdultWeekend (Price).**

Interestingly, for Big Mountain Resort the **actual price** per ticket (AdultWeekend) was **$81** and the **predicted price** was **$90.81**, which could mean that $90.81 can be charged per ticket to **optimise revenues**. This price is predicted according to the features Big Mountain Resort has and in comparison with its peers (other resorts).

Some interesting features are as under:



Recommendations

The resort has good features. It is positioned very well as compared to its peers in Montana in terms of NightSkiing facilities, chair lifts (3 person), Runs, age of the resort, Average Snowfall, Snow Making Machine area, Total Chairs etc. Because it has these great features, the price can be optimised. The resort can charge **$90.81 instead of $81.** This way the revenue shall increase, even though we have assumed the same number of guests visiting the resort this year as in the last year, i.e., 350,000 guests.

Revenue (price $81) = $28,350,000

Revenue (price $90.81) = $31,783,500

Incremental Revenue = $3,433,500 (only AdultWeekend)

Capex (additional Chairlift) = $1,540,000

Profit after Capex = $3,433,500 - $1,540,000 = $1,893,500

So, the capex is covered and adequate profit margin is available assuming all other costs remind he same.