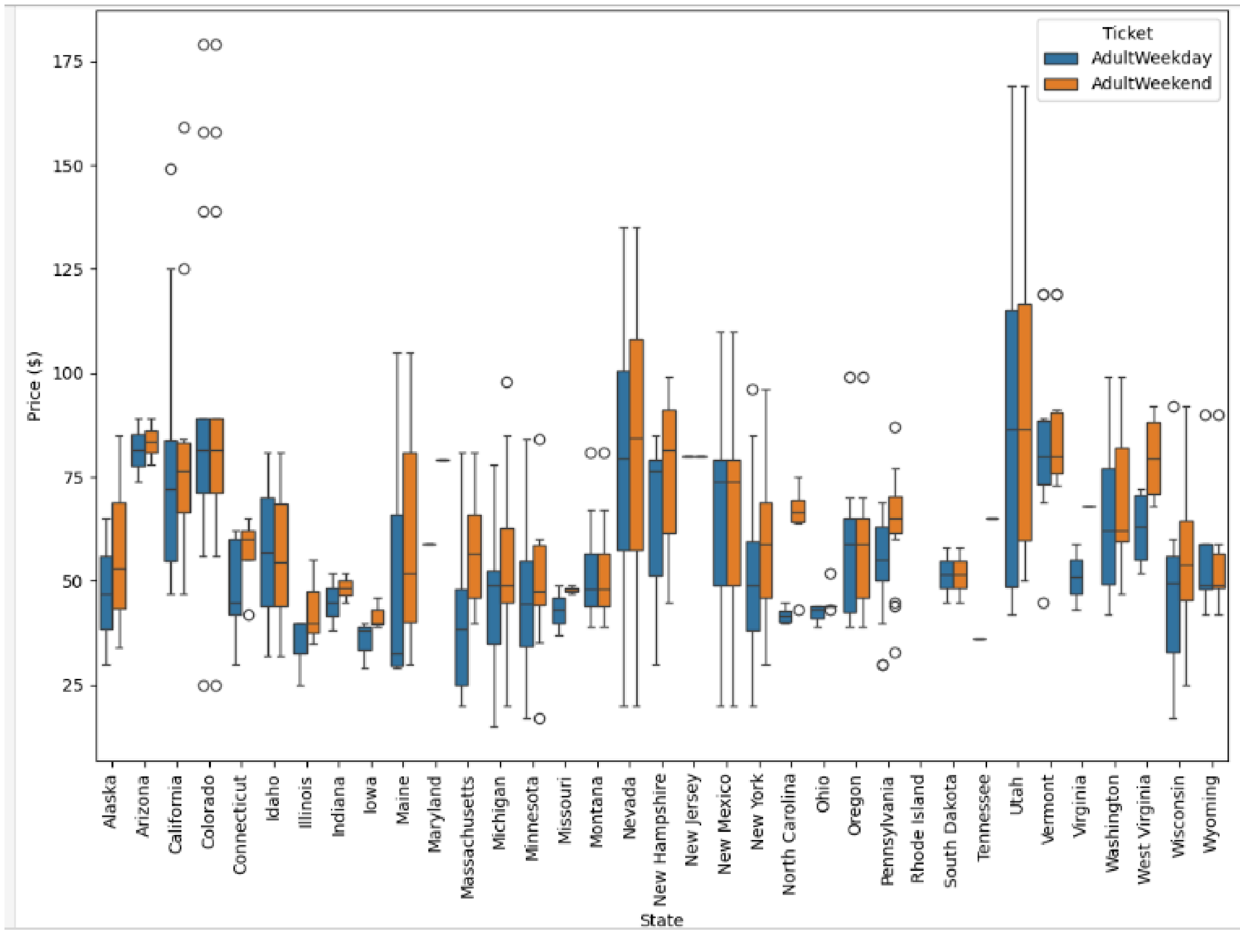
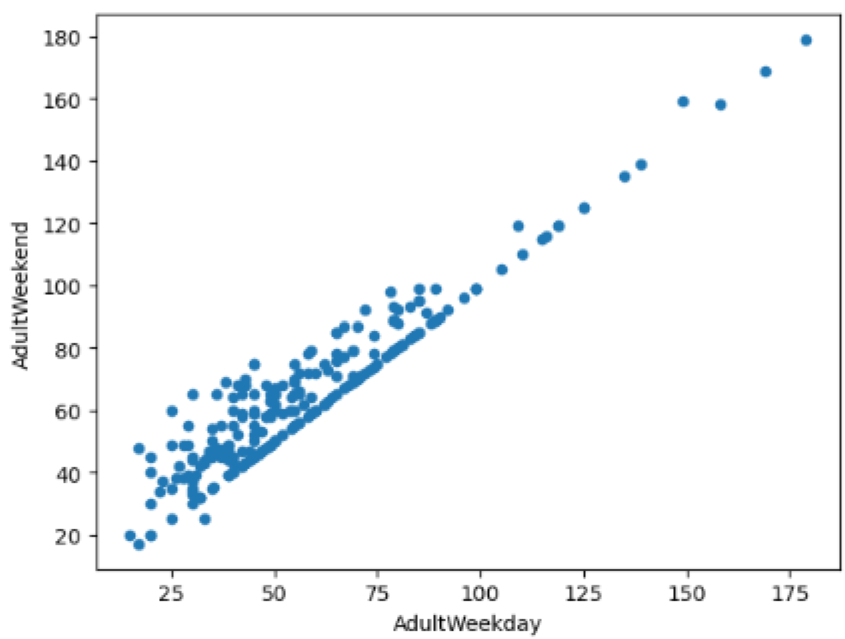
Guided Capstone Project Report

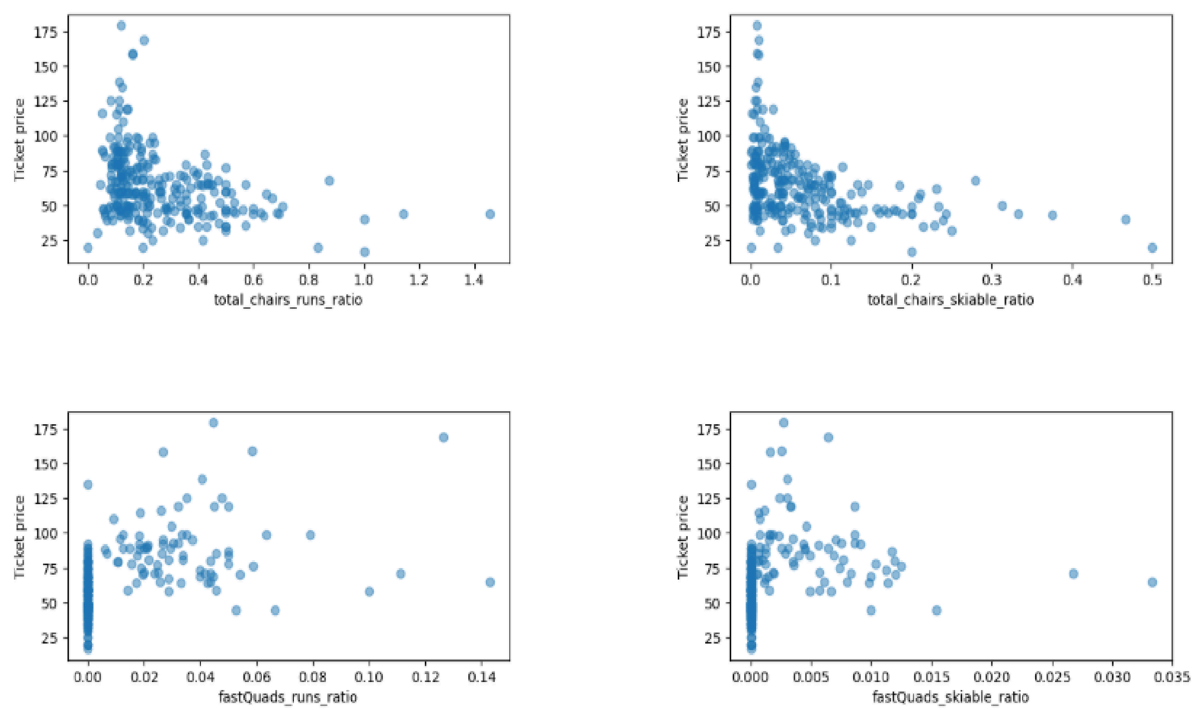
- Problem Statement:
  - What are the most frequently used aspects of the resort that we can look at to select a ticket price value that supports the increase in Big Mountain Resort's operating cost of 1.5M this season?
- Data Wrangling:
  - Our goal in wrangling the data was to find a target feature to focus on in order to determine a ticket price value to support Big Mountain Resort's increase in operations cost. Loading the Ski Resort Data from the CSV file, we had 330 rows and 27 columns to look at and after auditing the data we had a bunch of missing values and duplicates in our data. Exploring the relationship between region and state which were our categorical features, we've came to the conclusion that there are 38 different region and 35 different states in our data and one of the resorts Crystal Mountain was in 2 different states. Since our focus is finding a ticket price value for Big Mountain Resort we want to focus on Montana which is where the resort is located and has the 12th highest ticket price value. After cleaning the data and dropping rows with missing values we were left with 277 rows and 25 columns. After looking at the boxplot distribution of ticket prices between the weekday and the weekend values in Montana we see that the prices were pretty much equal but looking at the scatter plot distribution for the states in the ski data we have come to the conclusion that our target feature would be to look at the weekend prices since they were higher overall.



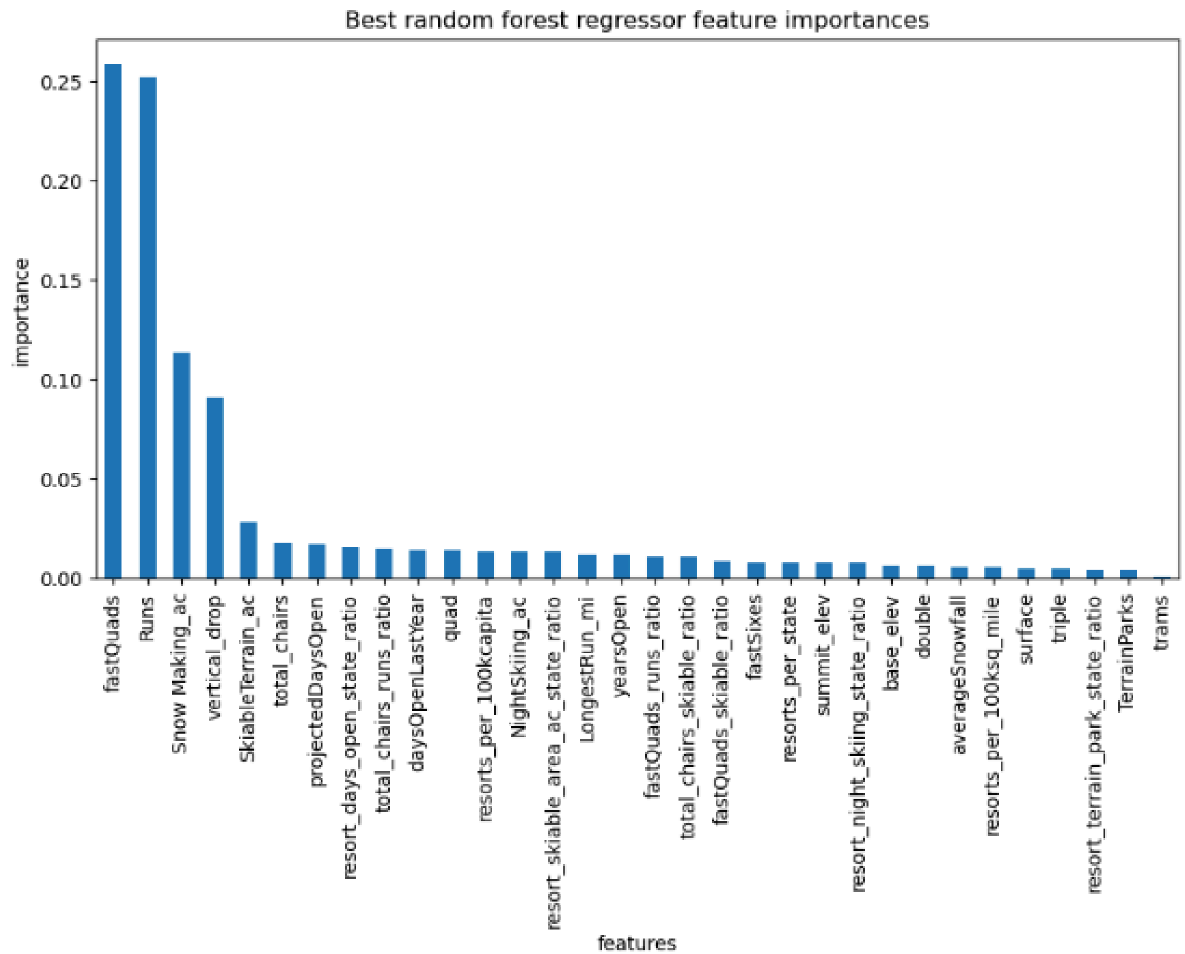
Ticket
AdultWeekday
AdultWeekend
Price ($)
175
150
125
100
75
50
25
Alaska
Arizona
California
Colorado
Connecticut
Idaho
Illinois
Indiana
Iowa
Maine
Maryland
Massachusetts
Michigan
Minnesota
Missouri
Montana
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
Ohio
Oregon
Pennsylvania
Rhode Island
South Dakota
Tennessee
Utah
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming
State



AdultWeekend
180
160
140
120
100
80
60
40
20
25
50
75
100
125
150
175
AdultWeekday

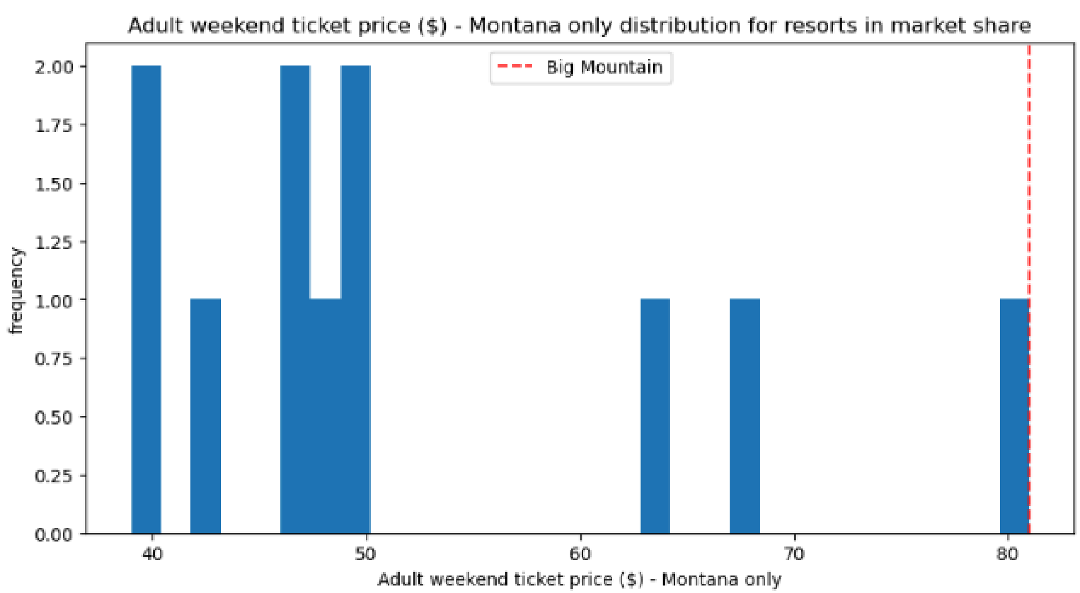
- Exploratory Data Analysis:
    - Our numerical features that we want to focus on in the data included the ticket prices for each resort and the categorical features we want to focus on were the region and state of the resorts. There wasn't a clear pattern with the relationship of state and ticket price but we could say that the skiable area affects the ticket price as well as night skiing availability. This leads us to keep an eye on the skiable area in our modelling for ticket price. The number of total chairs relative to the amount of runs the resort has is something to be wary of in our modelling because these features have an effect on the ticket pricing. As far as state labeling in the data it might be better to treat all states equally since the ticket pricing varies for each resort depending on multiple factors like skiable area, resorts per state, the amount of chairlifts and the days open.
    - ```
      state
      Colorado      43682.0
      Utah          30508.0
      California    25948.0
      Montana       21410.0
      Idaho         16396.0
      Name: state_total_skiable_area_ac, dtype: float64
      ```
    - 

- Model Preprocessing with feature engineering:
    - In the pre-processing phase we found that the mean average price of our training data was $63. After fitting a dummy regressor to our training data we found that it produced the same result as our mean average price and calculating the mean absolute error showed that we would be off around $19 if we were to predict the ticket price based on an average of known values.
      ```

- ```
  #Code task 4#
  #Calculate the mean of `y_train`
  train_mean = y_train.mean()
  train_mean
  ```
- ```
  63.811088082901556
  ```
- ```
  #Code task 5#
  #Fit the dummy regressor on the training data
  #Hint, call its `.fit()` method with `X_train` and `y_train` as arguments
  #Then print the object's `constant_` attribute and verify it's the same as the mean above
  dumb_reg = DummyRegressor(strategy='mean')
  dumb_reg.fit(X_train, y_train)
  dumb_reg.constant_
  ```
- ```
  array([[63.81108808]])
  ```

- Algorithms used to build the model with evaluation metric:
    - When we built a linear model we found that our model explained over 80% of the variance on the train set and over 70% on the test set which could suggest overfitting. On average we would expect to estimate a ticket price around $9 from the real price which is better than just guessing using the average. The linear model's performance result from cross-validation on the test split suggests that without using the same random state for initializing the cross-validation folds, our actual numbers will be different from our training data. Once we did a cross-validation test for multiple k-values we found that 8 was a good choice and found that vertical drop was the most useful feature which is consistent with what we saw in our exploratory analysis with skiable area being a factor in ticket pricing. People seem to want to pay more for skiable areas covered by snow making equipment because of the guaranteed skiing but the overall skiable terrain was negative for this model possibly because resorts can charge a lesser ticket price due to the large amount of visitors they acquire. Visitor numbers were missing from our data so we cannot be sure that this is the case. We built a Random Forest Model and found that fastQuads, Runs, Snow making terrain, and vertical drop were our top features in common with the linear model.
    - ```
      SkiableTerrain_ac    -5.249780
      trams                -4.142024
      LongestRun_mi         0.181814
      Runs                  5.370555
      fastQuads             5.745626
      total_chairs          5.794156
      Snow Making_ac        6.290074
      vertical_drop        10.767857
      dtype: float64
      ```
      ```

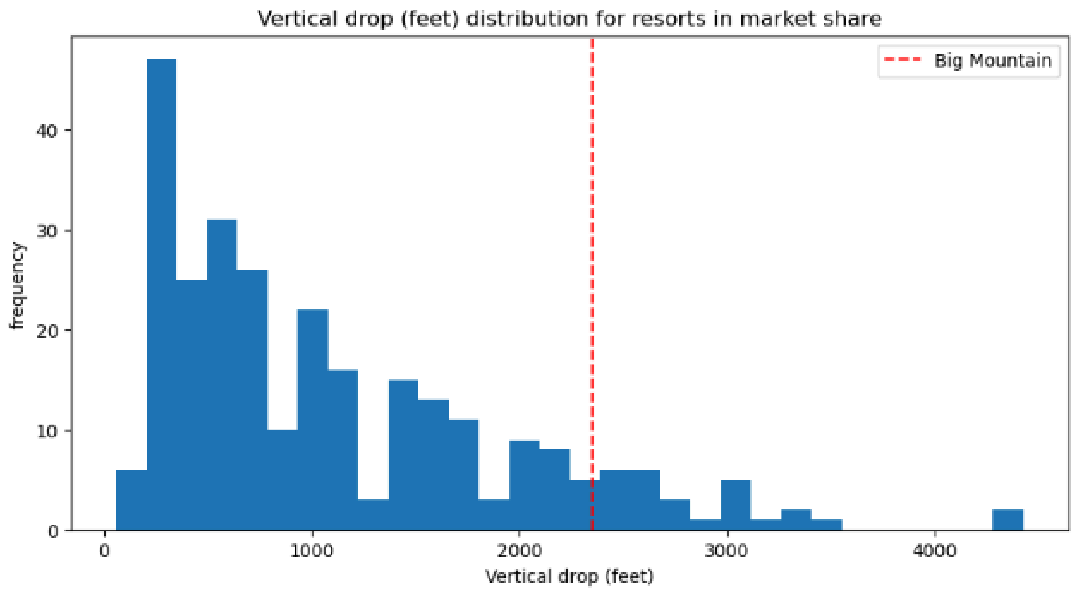


- Winning model and scenario modeling:
    - We decided that the Random Forest Model was the best model for our test data since it provided performance consistent with the cross-validation results. It had a lower cross-validation mean absolute error of almost $1 and it also showed to have less variability. After doing a data quantity assessment it seems to show that we have plenty of data for our model.

- 

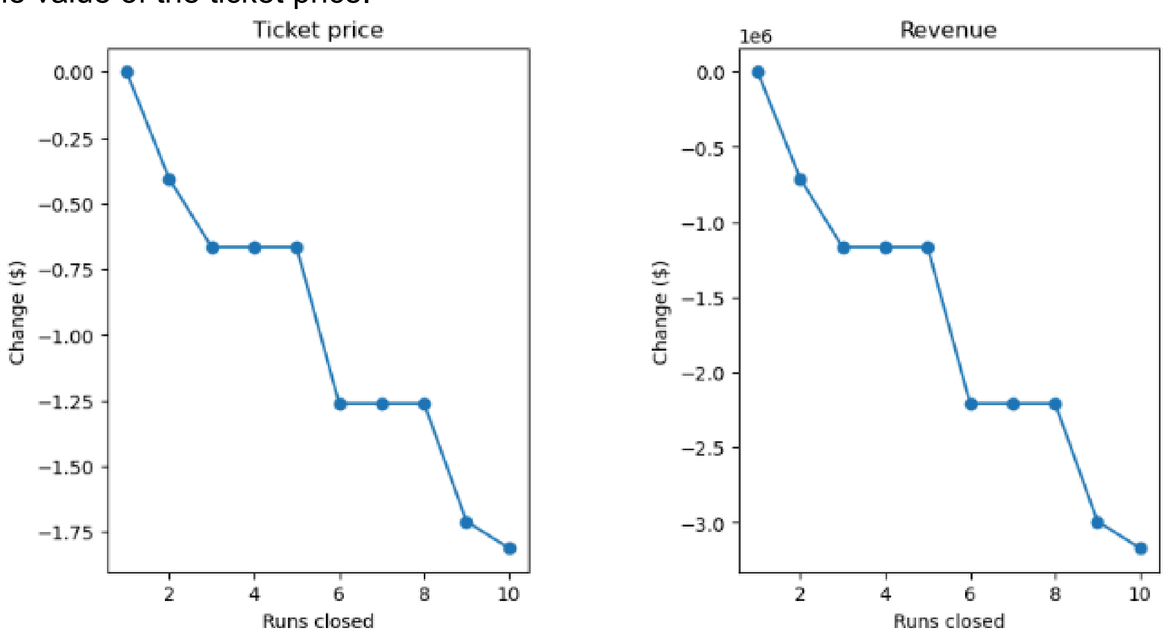
- Big Mountain Resort's weekend price that they charge is $81.00 which is the highest in the state. After refitting the Random Forest Model using all of our test data, we found the modelled price to be $95.87. Even with the expected mean absolute error of $10.39, this suggests that there is room for an increase of price. The validity of our model lies in the assumption that other resorts are setting their prices based on what the market supports, so the fact that our resort seems to be charging much less than what's predicted suggests our resort is undercharging for its ticket prices. Big Mountain Resort has amongst the highest number of total chairs compared to other resorts and also one of the largest skiable areas by snow making equipment. Big Mountain Resort is also doing well in having one of the biggest vertical drops which visitors are willing to pay a higher ticket price for.
  

-



○

- Pricing recommendation:
    - Recommendations would be to increase the vertical drop by adding a run to a point 150 feet lower but it will require the installation of an additional chair lift to bring skiers back up without additional snow making coverage. This would increase the ticket price per person by $2.00 giving us a total of about a $350,000 increase in revenue for the season. Closing one run would not affect the ticket price value but closing 2 or 3 will reduce the ticket price and revenue. Future improvements could be to add more snow making coverage because the model suggested this to be a positive feature to increase the value of the ticket price.



○

- Conclusion:
    - The model results show that we can increase the ticket price by at least $10 given the data provided. The modelled price was much higher than the actual price due to Big Mountain Resort basing its pricing on average market value which shouldn't come as a surprise to the business executives as they have already suspected this to be the case of undercharging. Visitor information being missing from our data limited our results. Looking at the number of runs and total number of chairs, it might be effective to keep the additional amount of these to a minimum as it could have a possible negative effect on the ticket pricing in regard to the amount of visitors that the resort attracts but the data also shows adding more fast quads might be beneficial since the resort covers a wide area.

- Future Scope of Work:
    - If the business executives find this model to be useful we could automate the model's use and create an excel plugin or make the model highly parameter-driven with clear documentation for business analysts to use and explore.