Big Mountain Resort recently installed a new chair lift that will increase operating costs by $1,540,000 for the upcoming season. In the past, their pricing strategy was to charge a premium above ticket prices in their market segment which was not capitalizing on the resort's facilities as much as it could. The goal of this project was to use data to find a pricing strategy that would more accurately price the resort amongst its competitors (based on its facilities) and increase revenue by enough to offset the increase in operating costs in the upcoming season.

To solve this problem, the data science team used a single CSV file with 27 features and 330 rows that was provided by Alesha Eisen. Each row corresponded to a different ski resort in the US that was considered part of Big Mountain Resort's market share. About 15-16% of the resorts were missing ticket prices (AdultWeekday, AdultWeekend, or both), and rows missing values in both columns were excluded from the analysis. Since most resorts (and all Montana resorts) had equal weekday and weekend prices and seven resorts were missing weekday prices versus only four for weekend prices, the team focused its analysis on weekend prices and dropped the AdultWeekday variable (rows missing values in the AdultWeekend column were dropped, as well).

Using the main dataset, a summary dataset was created that aggregated some of the features (number of resorts, total skiable area (in acres), total days open per year, total number of terrain parks, and total nightskiing area (in acres)) by state. The team augmented this dataset with each state's population and area in square miles by scraping this information from a Wikipedia page and used this to calculate density statistics (resorts per 100k capita and resorts per 100k km^2).

Initial EDA was conducted on the main dataset by plotting histograms of the data distributions in each column (Figure 1). A few things were noticed. Quite a few columns had a positive skew, and two columns had extreme outlier values (SkiableTerrain_ac and yearsOpen), which were dealt with accordingly. The SkiableTerrain_ac outlier was replaced with a more reasonable value from a reputable source and the yearsOpen outlier was dropped (since it could not be determined what was meant by the obviously wrong value). Another feature (fastEight) was dropped from the dataset since half of the values were missing and all but one of the remaining values were 0.

Additional EDA was conducted by plotting a heatmap (Figure 2) of the features and scatterplots (Figure 3) of the features against the target feature. The heatmap revealed that there was some multicollinearity in the data, and quite a few features had positive correlations with the target feature: fastQuads. Runs, LongestRun_mi, Snow Making_ac, Total_chairs, and Vertical_drop, The scatterplots helped illustrate some of these correlations, as well. Additional features were engineered around resort transportability (total_chairs_runs_ratio, total_chairs_skiable_ratio, fastQuads_runs_ratio, and fastQuads_skiable_ratio), and these showed a negative correlation with our target feature indicating an exclusive vs. mass market effect (Figure 4).

The state summary data was then analyzed to see if all states should be treated the same in the model. A boxplot of average weekend prices showed that most prices ranged from around 25 to over 100 dollars (Figure 5). Looking at the top performers across different features, there wasn't a clear pattern. Some states hosted many resorts, but other states hosted a larger total skiing area. The states with the most total days skiing per season were not necessarily those with the most resorts. Principle components analysis (PCA) was used to find linear combinations of the original features that were uncorrelated with one another and ordered them by the amount of variance they explained. It also allowed the team to visualize the higher dimensional data in two dimensions. The analysis revealed that the first two components accounted for over 75% of the variance, and the first four for over 95% (Figure 6). A scatterplot of the states plotting the first two components and price did not show a clear relationship between the PCA components and price (Figure 7). Based on the findings, there did not seem to be a clear reason to treat any state differently than the rest in our analysis; however, it revealed that there were some useful features in the dataset. Thus, it was joined to the main dataset and new features (around state resort competition) were engineered.

The final dataframe that the model was trained on consisted of 25 features and 277 ski resorts. To prevent overfitting, the data was partitioned into training (70%) and testing (30%) splits. As a baseline, the team evaluated

how accurately it could predict ticket prices by guessing the average price of $63.81. The “model” resulted in a mean absolute error of 19.14 meaning on average one would expect ticket prices to be off by around $19. A Linear Regression pipeline was then created that imputed missing values, scaled the data, selected the best features to use (and optimal number), trained the model, and used five folds cross-validation to estimate model performance while preventing overfitting. The cross-validation mean absolute error was 10.50 (standard deviation: 1.62), which was much better than guessing the mean. Next, a Random Forest pipeline was created that selected the optimal number of trees, decided whether to scale the features, imputed missing values with either the mean or the median, and used five folds cross-validation to estimate the model and prevent overfitting. The cross-validation mean absolute error was 9.64 (standard deviation: 1.35), and since the model’s error was almost $1 lower than the regression (and the model also had less variability), we decided to use the Random Forest. This decision was verified by looking at the mean absolute errors of the test sets (Linear Regression: 11.79; Random Forest: 9.54). 69 trees were the optimal number, imputing the missing values with the median worked better than the mean, and scaling the features did not help. fastQuads, Runs, Snow Making_ac, and vertical_drop were the dominant four features (Figure 8), and after graphing the cross-validation scores against training set size, the team determined that no additional data was needed since performance leveled off around a sample size of 50 (Figure 9).

Once the final model was determined, the team fit it on the entire dataset (except for Big Mountain Resort) using cross-validation (mean absolute error: 10.39, standard deviation: 1.47). The model was then used to predict the price of Big Mountain Resort’s tickets. The modeled price was $95.87, which was much higher than the actual price of $81.00. That meant that even with the expected mean absolute error of $10.39, there was room for an increase. In line with this, Big Mountain Resort performed well on nearly all the key variables that the model highlighted (Figure 10).

With 350,000 visitors expected for the upcoming season, each buying an average of 5 tickets, ticket prices would have to be raised by $0.88 / ticket to cover the increase in operating costs from the new lift. While the model suggests raising prices by $15.87, Big Mountain Resort is already the highest priced ticket in Montana by over $10 (Figure 11). Thus, a more conservative approach was recommended. The expected mean absolute error was $10.39, which means that, on average, predicted prices were $10.39 off (some high, some low). The team assumed that the projection was high (given the other prices in Montana), and safely recommended increasing prices to $85.50 (approximately equal to the recommended price - the expected mean absolute error). With the expected value of each customer increasing from $405.00 to $427.50, this increase allows Big Mountain Resort to increase revenue by more than the operating expenses of the new lift even if 4% fewer visitors come to the resort over the course of the season. If there is no reduction in visitors, revenue will increase by $7,875,000 over the current pricing model.

With an eye toward future seasons, the team then evaluated the impact of different scenarios for either cutting costs or increasing revenue. For cutting costs, the team determined that it could immediately close the least used lift without impacting revenue (Figure 12). If it wanted to close more lifts, closing 5 or 8 of the least used lifts would be the best option. For increasing revenue, the team determined that if it increased the vertical drop by adding a run to a point 150 feet lower down (which would require installing an additional chair lift) without adding additional snow making coverage, it could increase revenue by about $3,474,638 over the course of a season (justifiable given the increase in operating costs from the new lift were $1,540,000).

The team was able to conclude that Big Mountain Resort has plenty of room to increase ticket prices over the coming season and has options to both increase revenue and reduce costs in coming years. This year, the resort should increase ticket prices to $85.50 and close its least used lift. Next year, it should increase its vertical drop by 150 feet and install an additional chair lift (which would increase profit by $1,942,500). To improve the accuracy of the model, future work should focus on collecting more data related to the number of visitors that attended the resorts, collect cost information, and create a historical database of each resort by year.

Figure 1

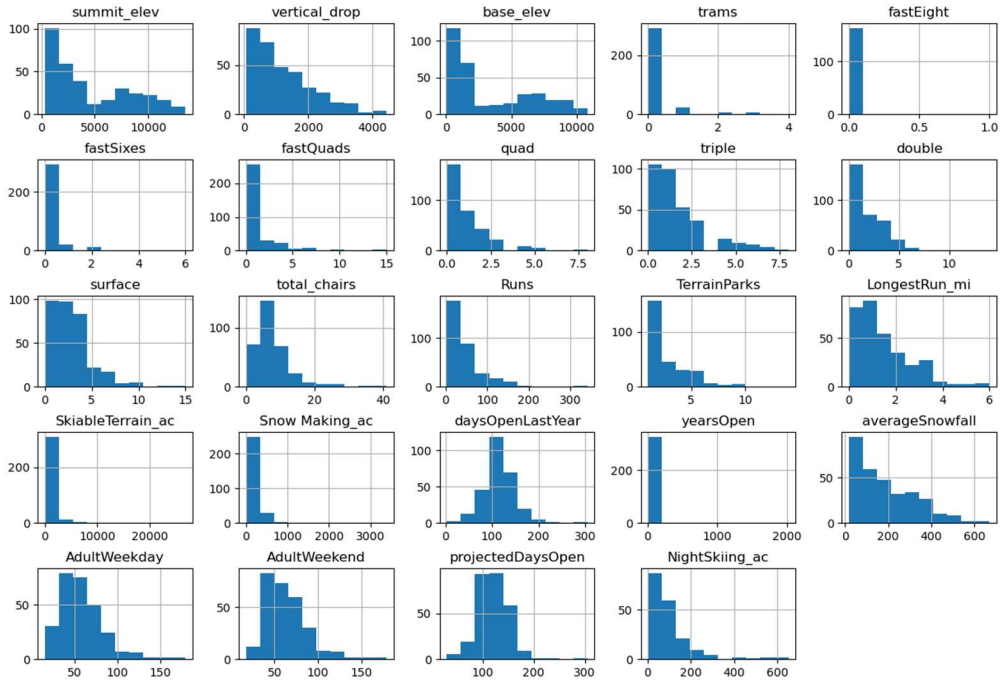


Figure 2

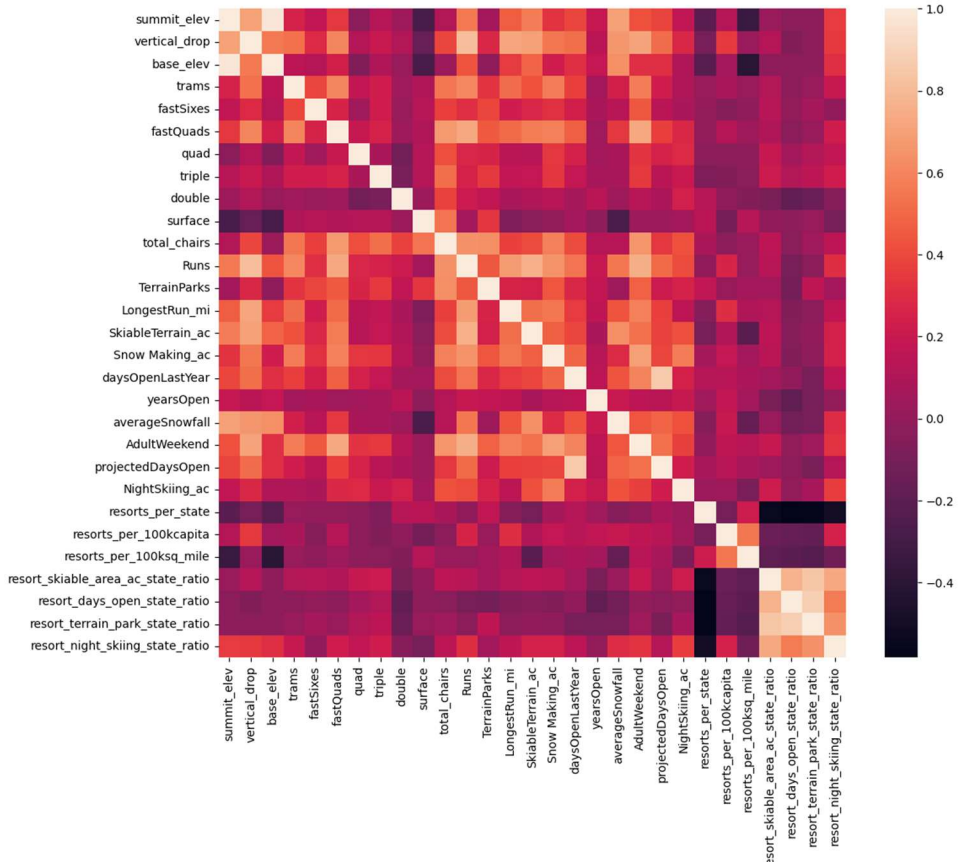


Figure 3

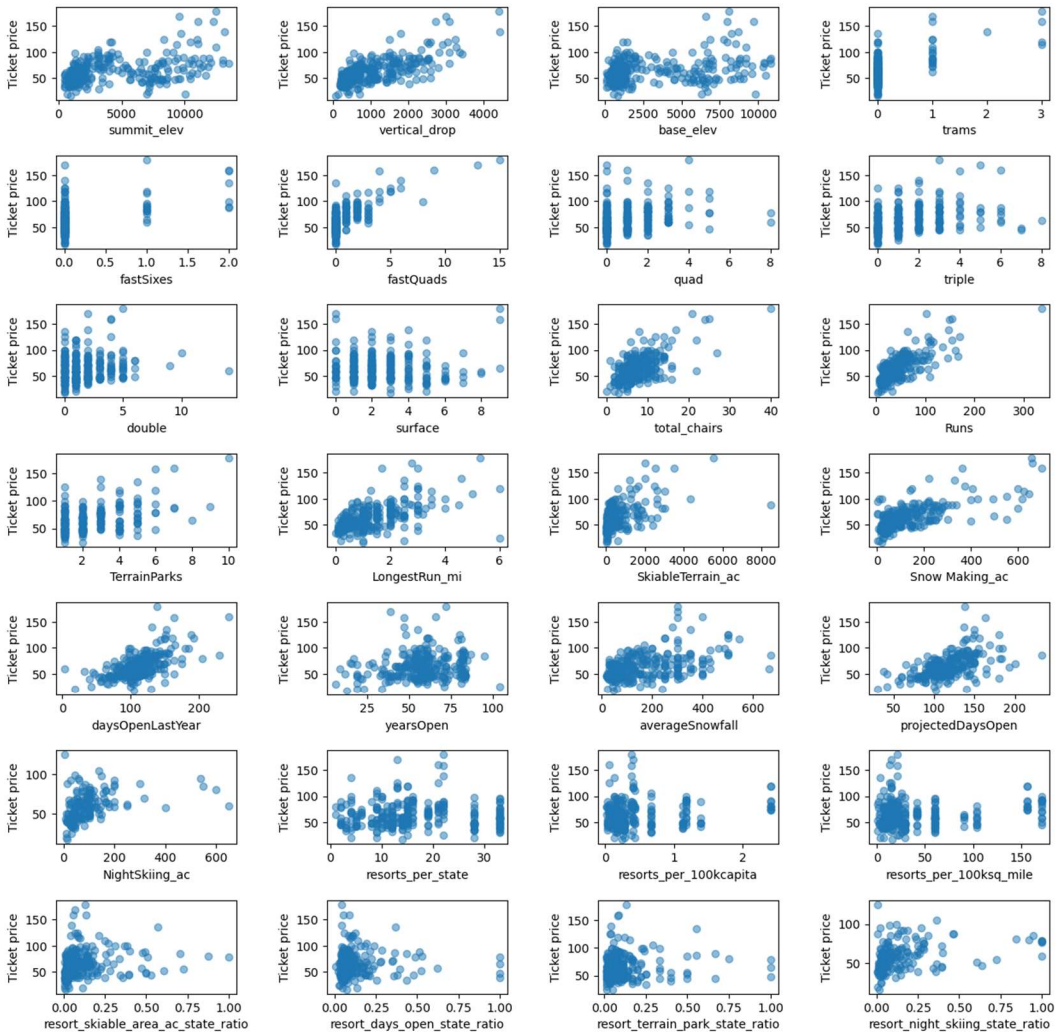


Figure 4

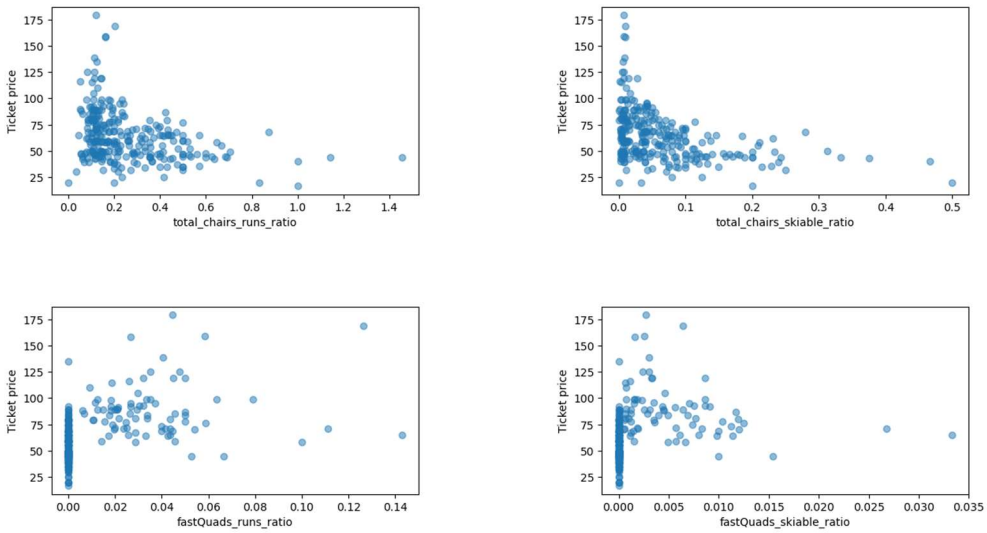


Figure 5

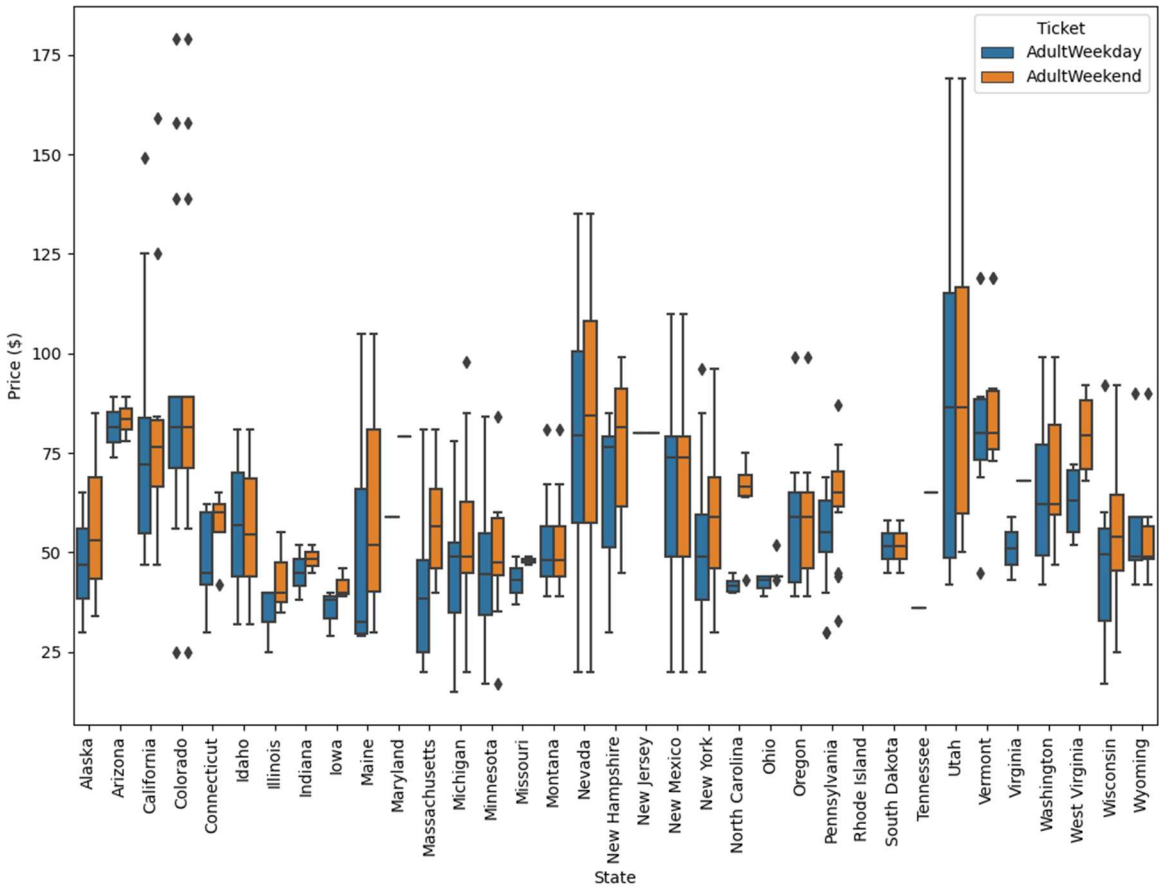


Figure 6

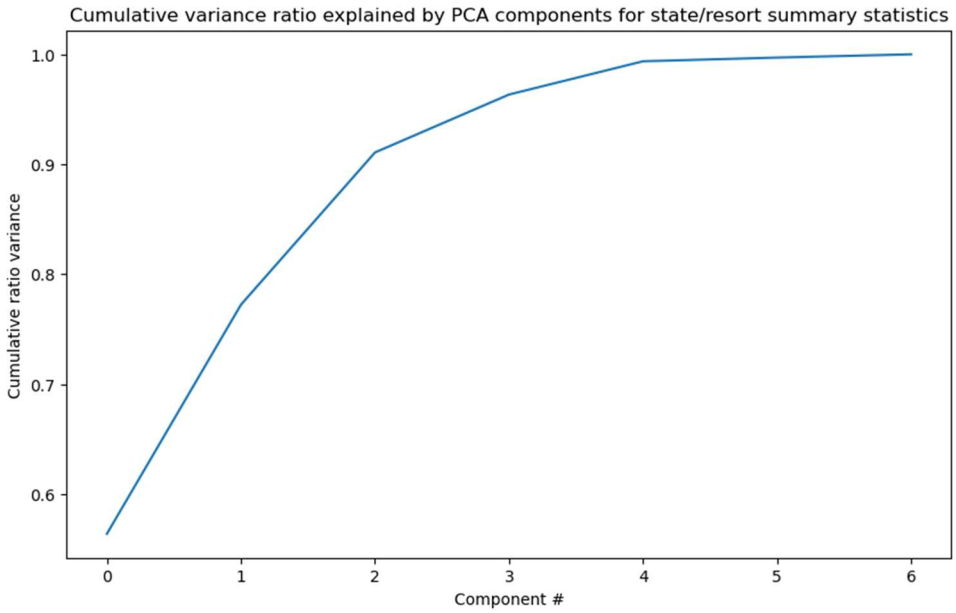


Figure 7

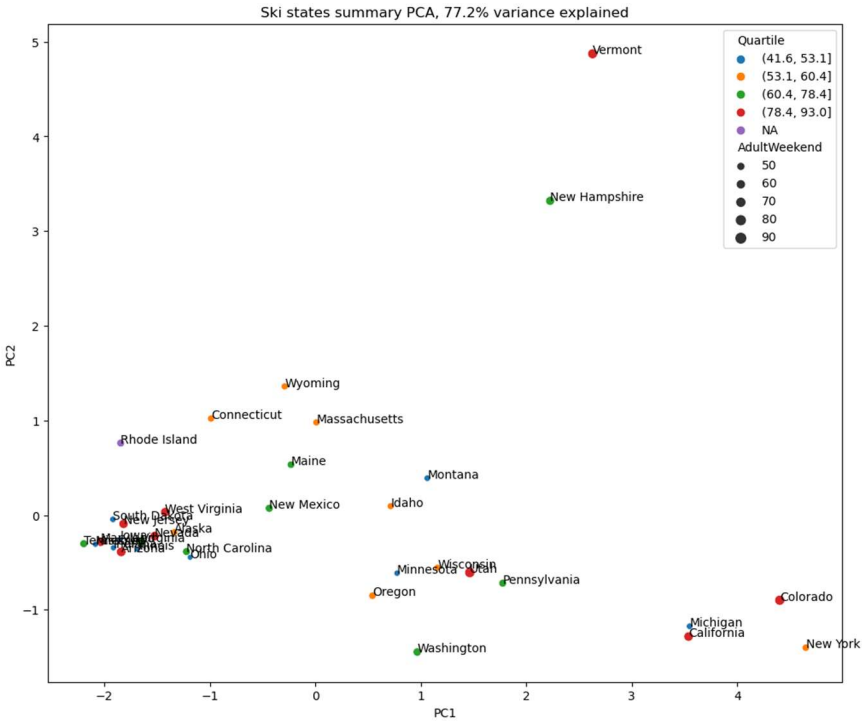


Figure 8

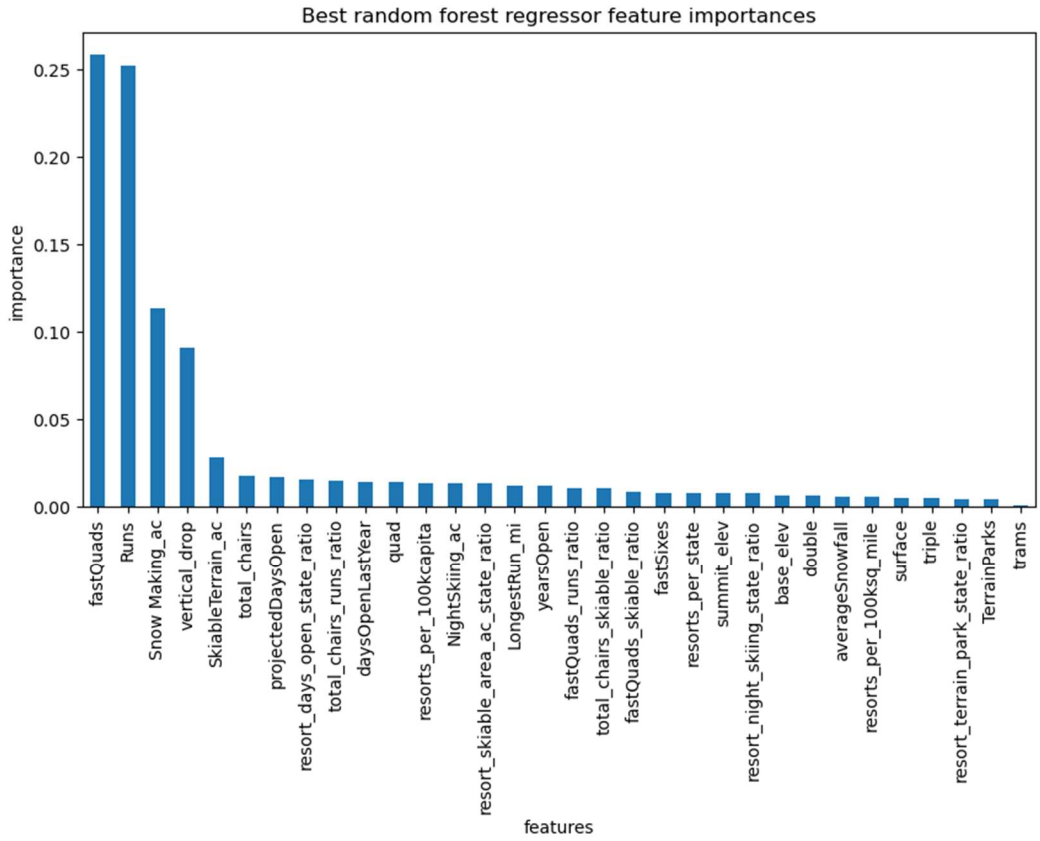


Figure 9

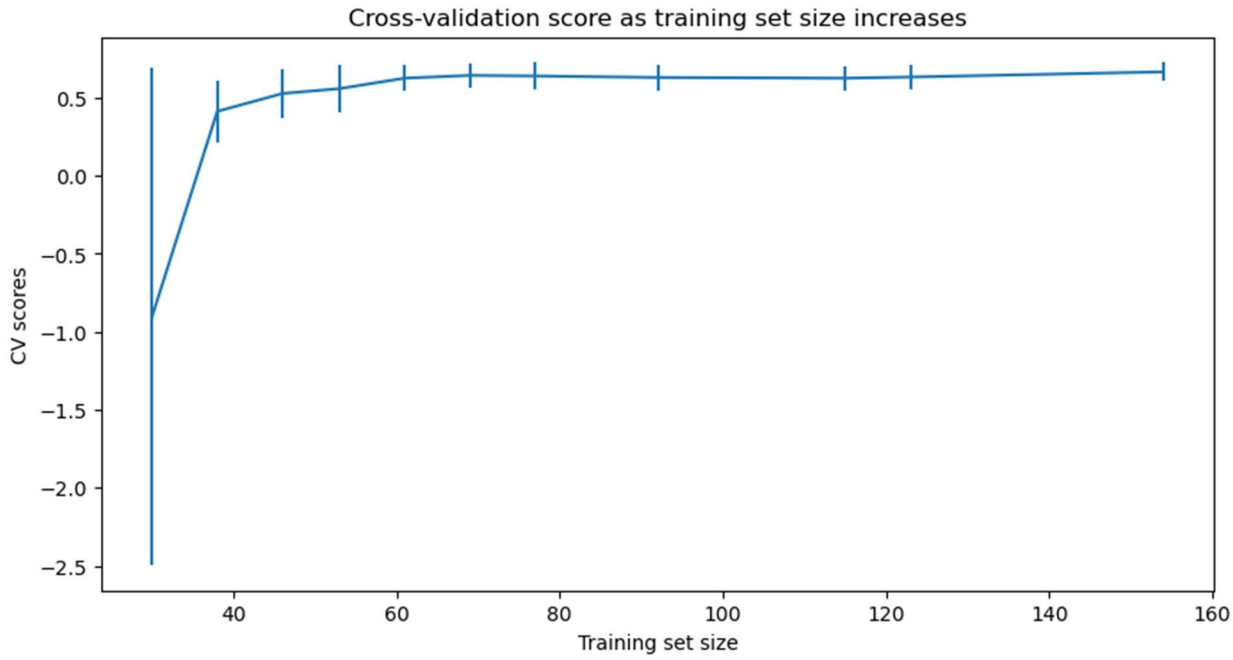
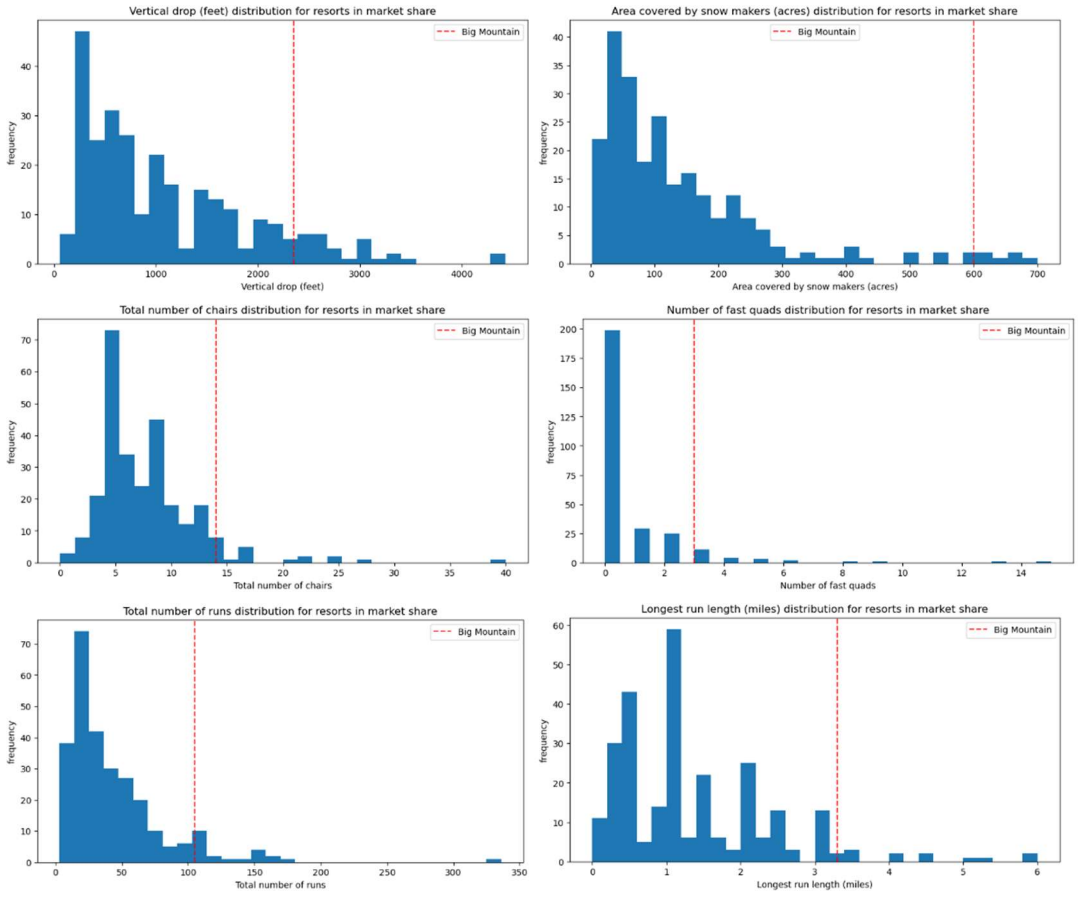


Figure 10



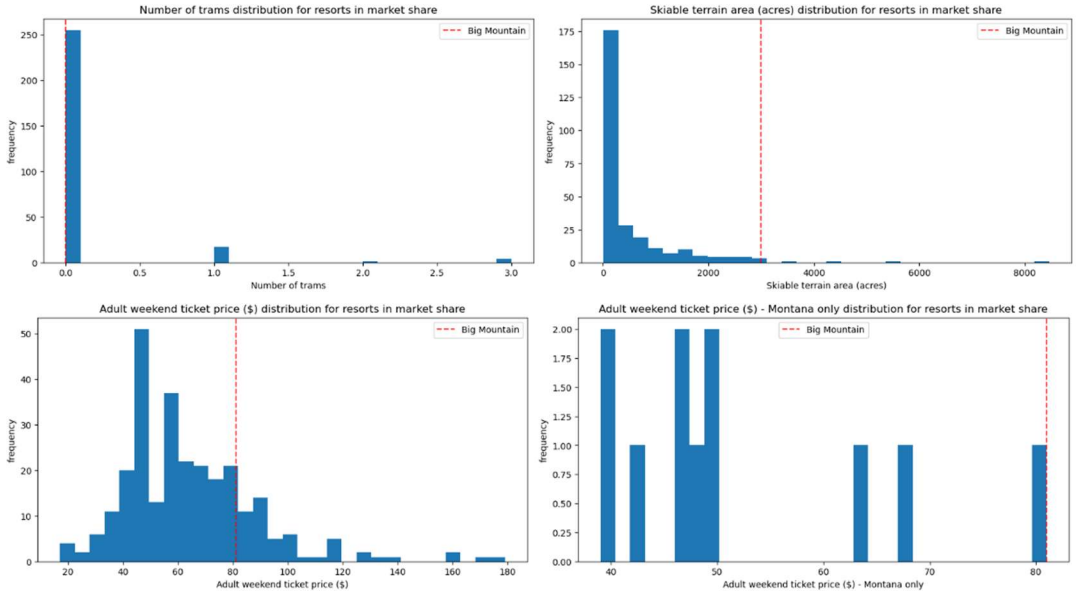


Figure 11

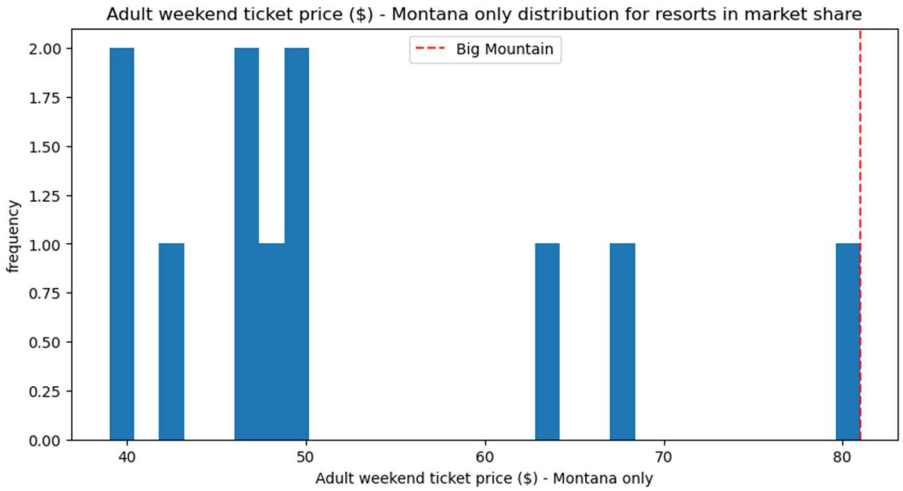


Figure 12