Problem Statement:

Big Mountain Resort, located in Montana, is facing a challenge in determining the optimal ticket price for the upcoming ski season. The resort management team is uncertain about how their resort stacks up against competitors in terms of ticket prices, and how to best price their tickets to maximize revenue. In order to make informed business decisions, they require insights into the competitive landscape and a data-driven approach to pricing recommendation.

Data Wrangling:

Our data wrangling process began by identifying missing values and duplicates in the dataset. We found that there were several missing values in some features, but we concluded that there were no duplicates of the resorts. We then plotted the distribution of resorts by state and found that Montana was 13th in the number of resorts. We also plotted the distribution of ticket prices by state and found that California, Colorado, and Utah had the most expensive tickets.

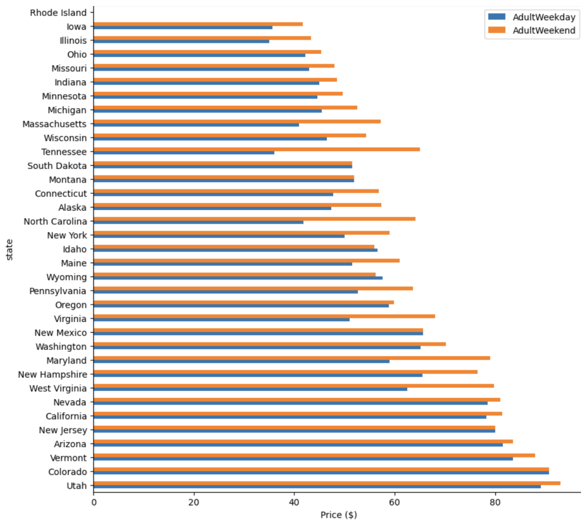
```
Crystal Mountain    2
Alyeska Resort      1
Brandywine          1
Boston Mills        1
Alpine Valley       1
Name: Name, dtype: int64

You have a duplicated resort name: Crystal Mountain.
```

```
Alyeska Resort, Alaska     1
Snow Trails, Ohio          1
Brandywine, Ohio           1
Boston Mills, Ohio         1
Alpine Valley, Ohio        1
dtype: int64
```

Taking a closer look at the duplicates

Next, we analyzed the distributions of each feature and removed some outliers, such as a resort that had been open for 2019 years. We also aggregated the sum of several features to obtain state-wide summary statistics for the market segment. We dropped rows that lacked both weekday and weekend ticket prices, which accounted for 14% of the data. We also merged information about the population of each state with the ski data.




Average ticket price by state for each kind of ticket: Adult Weekend and Adult Weekday

Exploratory Data Analysis:

In our exploratory data analysis, we looked into several features, such as the number of resorts per state, the amount of skiable area per state, the number of days open, and the population of the state. We also created additional features such as ratios of the resorts per state over the state population and resorts per state over the state area in square miles. We found that there was a pattern suggested of a relationship between state and ticket price. Some states, such as Vermont and Colorado, tend to have higher ticket prices on average compared to other states in the dataset. Correlations in features such as fastQuads, Runs and Snow Making_ac stood out.

We used scatterplots to identify features that were strongly associated with ticket prices, such as vertical_drop, fastQuads, Runs, total_chairs, and resorts_per_100kcapita. However, we noted that we should be cautious when interpreting the relationships between some features and ticket price, such as the relationship between the number of chairs relative to the number of runs and ticket price.

Model Preprocessing with Feature Engineering:

Our target feature for prediction was the ticket price, and we were left with 277 rows of data after removing missing values and dropping unnecessary columns. When performing feature selection for modeling, we remained wary of highly correlated features, such as the ratio features that were introduced in the data preprocessing step. We also considered the potential confounding effect of state labels and the possible variations in skiing culture and preferences across different states. Finally, we carefully evaluated the relevance and importance of each feature in predicting the target variable, and avoided overfitting by selecting only the most informative features.

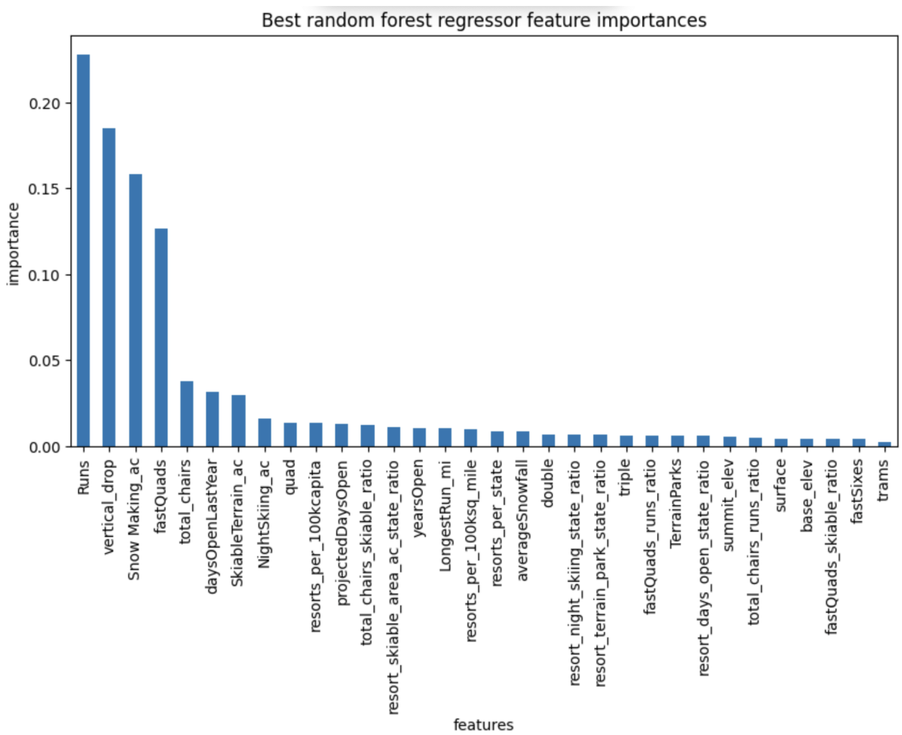
Algorithms used to build the model with evaluation metric:

When performing feature selection for modeling, we should remain wary of highly correlated features, such as the ratio features that were introduced in the data preprocessing step. We should also consider the potential confounding effect of state labels and the possible variations in skiing culture and preferences across different states. Finally, we should carefully evaluate the relevance and importance of each feature in predicting the target variable, and avoid overfitting by selecting only the most informative features.In this notebook, the goal was to predict the adult weekend ticket price using available data. Initially, the average ticket price was used as a baseline to gain an idea of performance.

It was found that this approach did not capture any variance in the data, but it still served as a useful benchmark for more complex models. A linear regression model was then built and features were identified that strongly influenced ticket prices, including vertical drop and the area covered by snow making equipment.

Surprisingly, skiable terrain area was found to be negatively associated with ticket prices, possibly due to larger resorts having to stretch chairlift capacity thinner. Cross-validation was used to estimate the model's performance on unseen data, and the linear regression model exhibited good consistency between the cross-validation results and performance on the test set.

A random forest regressor was also tried with various preprocessing steps, and it was found to perform slightly better than the linear model. However, the chosen model going forward was the linear regression model due to its simplicity and interpretability, and its ability to provide useful insights to inform business decisions.The limitations of this work stem largely from the lack of comprehensive cost information. The only cost data provided was the additional operating cost of the new chair lift, and without more detailed cost information, it is difficult to assess the impact of cost on the optimal ticket price. The model may also be limited by a lack of data on other factors that could influence ticket price, such as marketing strategies, competitor pricing, and seasonality. It is also possible that the model is overestimating the price that can be supported in the market due to the limitations of the data.



Taking a look at the features and their importance to our RF model.

Future Scope of Work:

In conclusion, our analysis provides valuable insights into the competitive landscape of ski resorts in the United States and provides a data-driven approach to pricing recommendations for Big Mountain Resort. However, there are several areas for future research, such as incorporating more data on customer demographics and preferences, conducting market research to understand consumer behavior and willingness to pay, and analyzing the impact of external factors such as weather patterns and economic conditions on ticket prices.

Overall, our analysis demonstrates the importance of data-driven decision-making in business and highlights the potential benefits of using advanced analytics and machine learning techniques to gain a competitive advantage in the market. We hope that our findings will be useful to ski resort managers and other business leaders who are seeking to optimize their pricing strategies and improve their bottom line.