Problem Statement

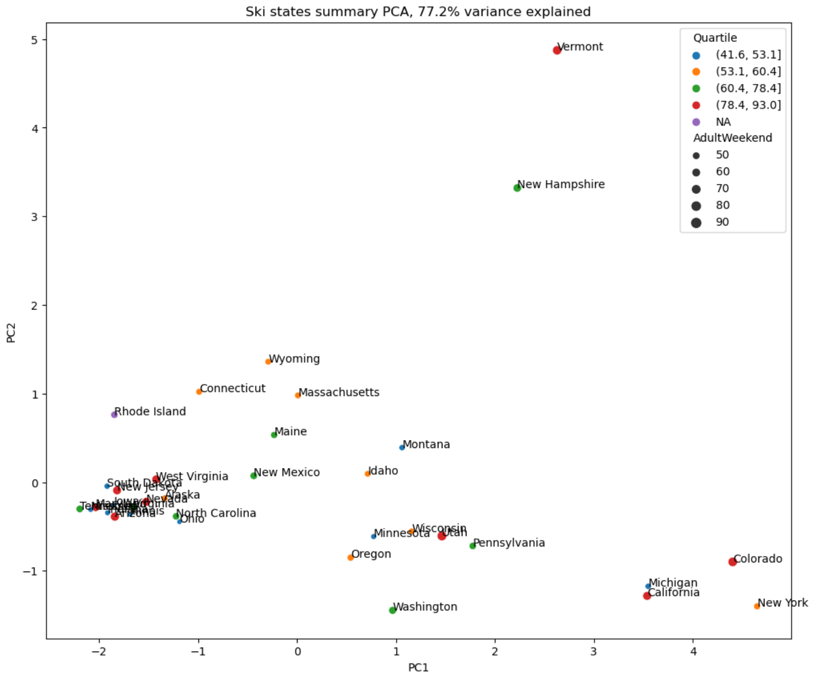
What opportunities exist in operations or sales for the Big Mountain resort to maintain their annual profit margin of atleast 10% in the upcoming tourist season given the increase in operating costs by $1.54 million dollars due to installation of the new chair lift?

Data Wrangling

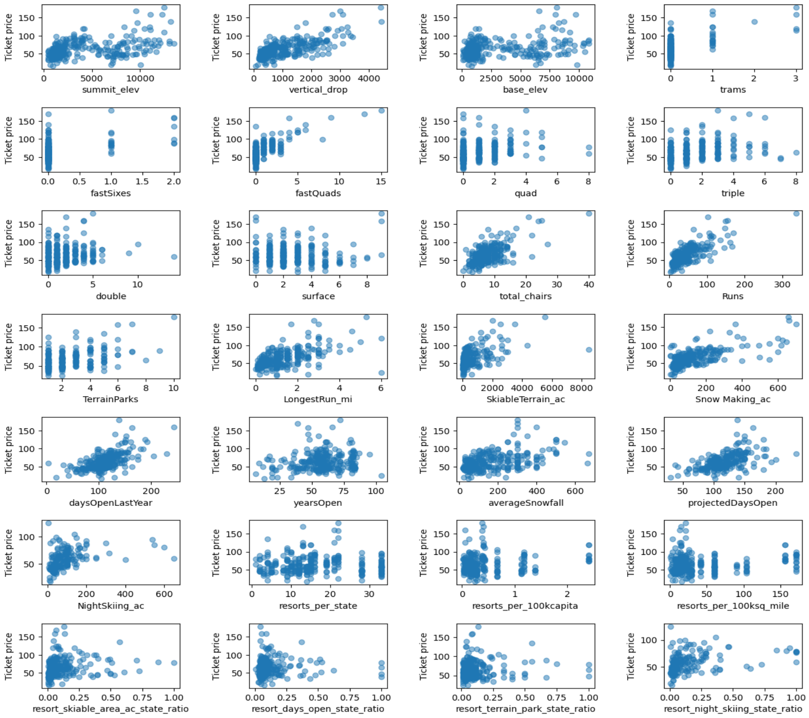
I was provided with ski resort data in csv format that contained information about 330 ski resorts across the United States in the form of 26 total features including the resorts' ticket prices. In this data, I first determined the number of missing values by column and found that the 'fastEight' column had the most at 166 and even the 'AdultWeekday' and 'AdultWeekend' columns had missing values. Next, I created a series of plots to visualize key patterns and trends in the data. Specifically, I plotted the distribution of resorts by state and region, distribution of ticket price by state, and the distribution of feature values. Through the last plot, I found multiple records with erroneous values that could be modified without compromising the integrity of the data, hence I proceeded to fix them. Next, I fixed the issue of our target features, 'AdultWeekday' and 'AdultWeekend' ticket prices having no price data by first removing records where both features were null, then I removed the 'AdultWeekday' column because it contained more null values than 'AdultWeekend' and lastly I removed the records where 'AdultWeekend' was null. I also created an additional data frame 'state_summary' by combining certain features from the existing 'ski_data' dataframe and a dataset I found on wikipedia. I cleaned up this dataframe's state column to include only the official state names.At the end of this step, there were 277 records and 25 total features in the 'ski_data' dataframe.

Exploratory Data Analysis

For exploratory data analysis, I explored major columns of the 'state_summary' with the states as index to get an idea of the top states in these statistics. In order to visualize the high dimensional data, I first scaled then performed PCA transformation on the 'state_summary' dataframe. I then plotted the cumulative variance ratio and observed that the first two components accounted for 75% of the variance.



Upon further analysis with the variance, I found that the pricing model should be based on all states without treating any one state specially. Lastly, I made a series of scatterplots of numeric features against ticket price and I found that features such as vertical_drop, fastQuads, Runs, total_chairs, and resorts_per100kcapita are potentially useful.

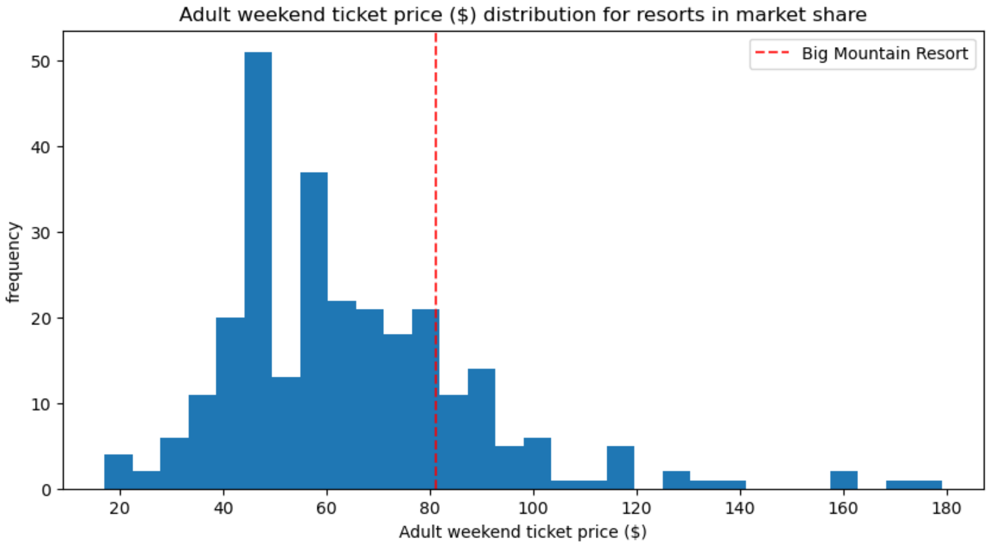


Model Preprocessing

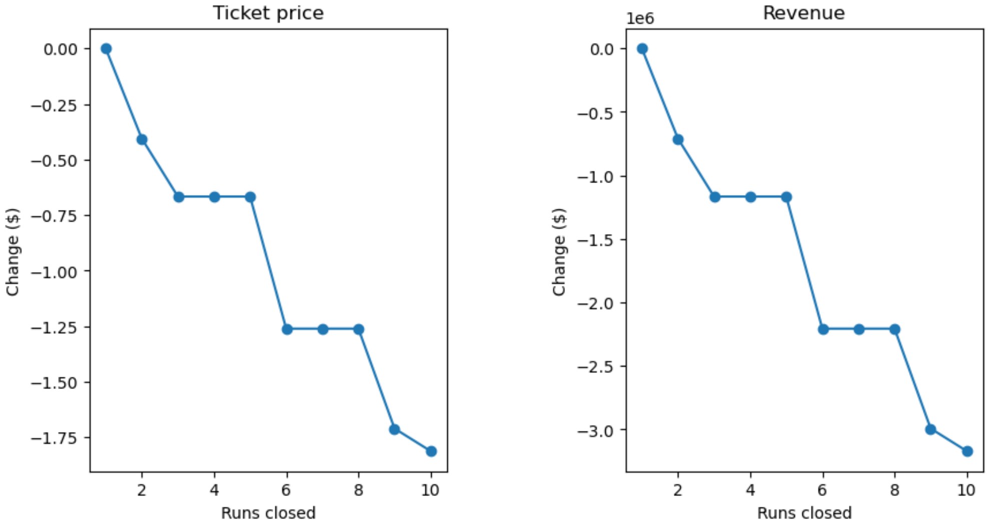
In order to determine the best model to determine the ticket price, the data needed to first be preprocessed. First, I separated the Big Mountain resort record from the other records. Then I performed a train/test split on the data so that several models can be trained and tested on it. The first model trained was the dummy regressor which simply used the sample mean as a predictor. By most major metrics (R-squared, mean absolute error, mean squared error), it was not a good model. The mean absolute error indiciated the predicted ticket price based on this model may be off by $19. Next, I imputed missing feature values, first with median and then with mean, and then assessed the model performance by the metrics, however there was no significant difference between either method. Lastly, I fit the linear regression and the random forest model then measured their performance. According to both models, the top four features in relevance to the ticket price were fastQuads, Runs, Snow making_ac, and vertical_drop. By all metrics, the random forest model performed better as it had a lower cross-validation mean absolute error by almost $1.

Modeling

In the modelling step, I refit the random forest model on all data except Big Mountain then used it to predict the price for Big Mountain resort. The model suggested a price of $95.87 while the actual price is $81. Even with the absolute mean error of $10.39, there is scope for increasing the price. I then plotted the current price for the resort in the market context.



I also modeled four additional scenarios for the ticket price in response to changing specific features in the dataset. The plot for the first scenario of closing down up to 10 of the least used run is presented below.



Pricing Recommendation

Based on the random forest model, the recommended price is $95.87. However, accounting for the mean absolute error of $10.39, I would recommend the ticket price to be around $85.48 given the current features of the resort.

Future scope of work

One major deficiency in the data is the lack of total operating costs for each resort which hinders a complete analysis of the data in several ways. To assess total profit margin, modelling the ticket price in relation to the features and their operating costs would paint a better picture.