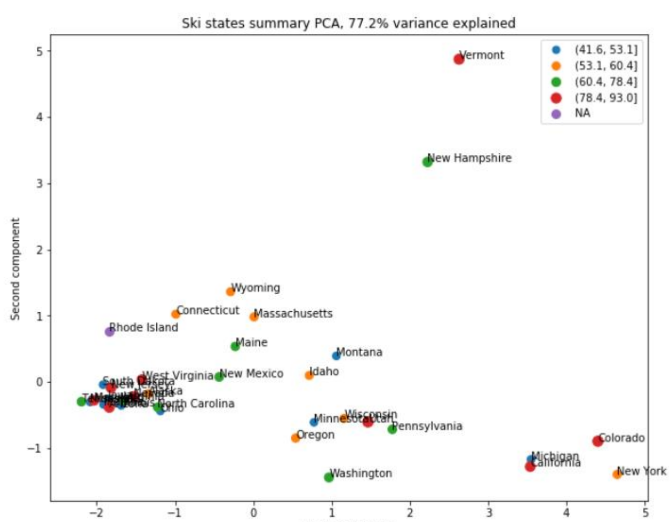
Summary

The current pricing strategy of Big Mountain resort is based on taking an average ticket price across its market segment. Based on the random forest model that we developed using US resort data, Big Mountain resort can increase the adult weekend ticket price and increase revenue in the process. Although Big Mountain Resort ranked high on some of the facilities that our random forest model positively correlated with the ticket price, our ticket price is significantly lower than the modelled ticket price even considering our expected absolute error. The model suggests a room for increase. The modelled ticket priced was $94.22 whereas the current ticket price is $81. Considering an average expected number of visitors over the season of 350,000 our modelled revenue increase using the modelled ticket price and current ticket price is **$4,627,000**.

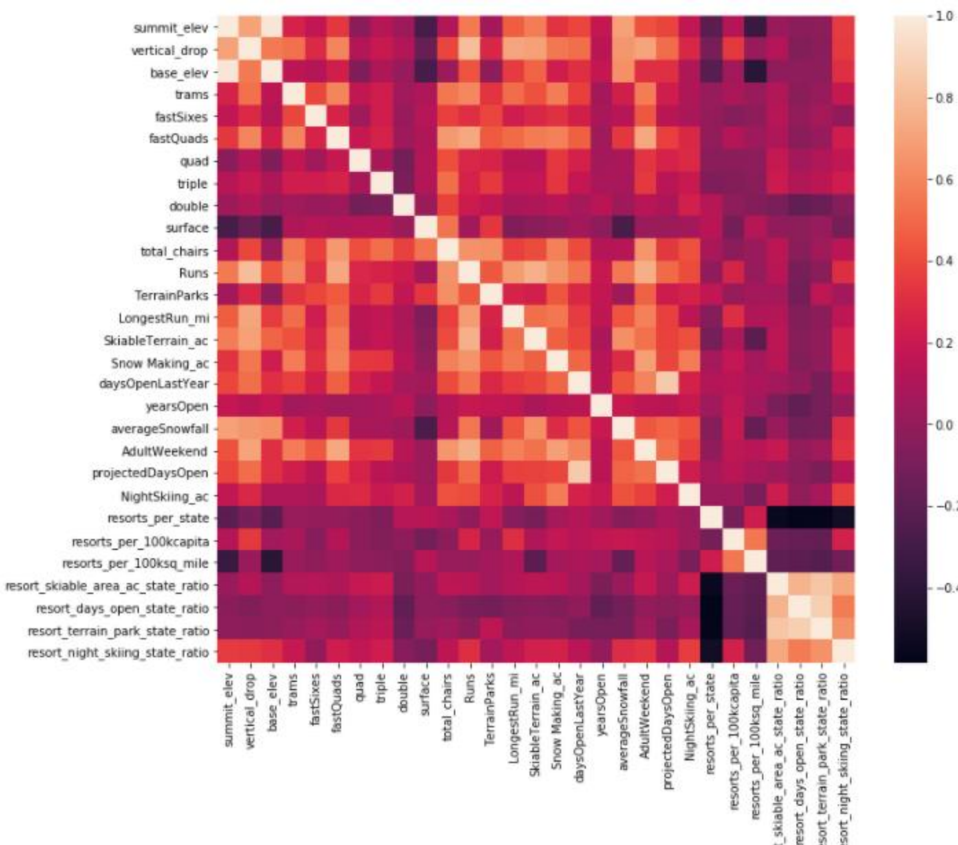
We started our analysis by making sure that our collected data is organized and well defined. The data was then examined for possible duplicated resort data and possible errors and outliers in the data (e.x. silver mounttain's skiable area is significnatly less than what was on the data). The population data was then merged with the resort data. There were two possible target feature, the weekday price and weekend price. In Montana the distribution of weekday and weekend prices is equal. Further examining the data it was then found that the weekend prices have the least missing values of the two. The weekend price was then selected as the target price and resort data (rows) with missing adult weekend price were dropped along with the adult weekday price column.

After preparing the data for analysis, exploratory data analysis was then performed on the data. It was found that Montana came in 3rd when it comes to total state area but is not even on the top 5 when it comes to total state population. Montana is also not in the top five when it comes to the number of resorts per state and skiable state. New York State has the most resorts and has the most night skiable area. Normalizing the resort data per 100k population, Montana came in 4th in resort density on the other hand Vermont dominated in both resorts per capita.

In order to further explore the interconnected relationship between high dimensional data and find patterns in the data, principal component analysis was performed. The data was scaled since it is a variance maximizing exercise. From the pca we found that the first two components account for over 75% of the variance and the first four components account for over 95% of the variance.



From the graph we can see that there is not an obvious patten. The red points representing the upper quartile of price can e seen to the left, the right and up top. There is also a spread of the other quartiles as well. In this representation of the ski summaries for each state which accounts for some 77% of the variance, shows that there is no pattern with price. A feature correlation heatmap was then created to check for possible relationships between variables. We can see from the heat map that the ticket price is highly correlated with the number of fast quads, snow making machine, runs and total chairs. From the scatterplots of ticket price vs feature we can clearly see some of the high correlations we were clearly picking up on. There is a strong positive correlation between vertical_drop, fastQuads, Runs and total_chairs. Normalizing the runs, skiable area, and fastQuads by the number of chair and plotting these features against the ticket price we can see that the relationships seemed counterintuitive. The more chairs a resort has to move people around, relative to the number of runs, ticket price rapidly plummets and stays low. This effect is probably due to the mass market vs exclusive market situation.



Two types of machine learning algorithms were used to model the relationship between variables. The best cases for each algorithm were then compared. The median variables for each feature from the train set was then used to fill the respective missing data for the features. The best case Random forest regression model had a lower cross-validation mean absolute error by almost $1. It also exhibits a lower variability with a lower standard-deviation among other random forest regression models. Further exploring the model we found out that closing one run makes no difference in the revenue therefore if the executive team is interested in further exploring the possibilities of maintaining revenue while decreasing cost the model can be further utilized and explored.