# Guided Capstone Project Report

## 1) Problem identification

The company Big Mountain Resort provides facilities such as lifts, T-bars, and magic carpet for skiers and riders of all levels and abilities. This company has recently installed an additional chair lift which increases their operating costs by $1,540,000 this season. So, the company must decide which pricing strategy to use. In this case, the price is considered above the average price of resorts in its market segment. However, there are some limitations. The business wants some guidance on how to select a better value for their ticket price. They are also considering several changes that they hope will either cut costs without undermining the ticket price or will support an even higher ticket price.

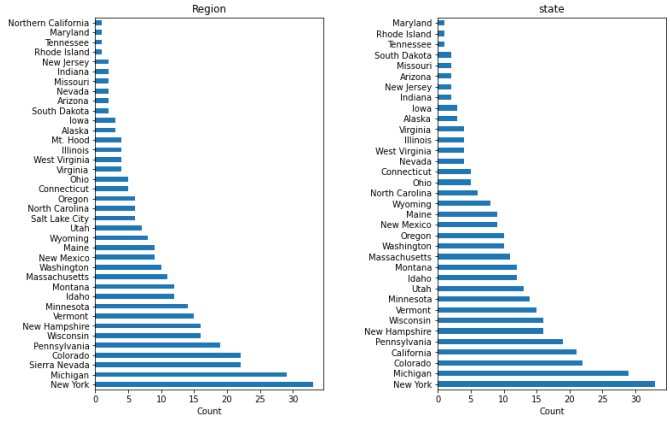
## 2) Data wrangling

The number of samples and features are identified as 330 and 27, respectively. The strategies used for the imputing missing values are dropping the rows with missing values, replacing suspicious values with correct values, dropping row with suspicious values, and dropping columns with missing values. Between two target features, AdultWeekend and AdultWeekday, AdultWeekend has fewer missing values, so it is considered as target variable for the modeling and other one is dropped. After the cleaning process, the dataset has 277 rows and 25 features.

## 3) Exploratory Data Analysis

In this step, the place of Montana among other states is examined by order of each of the summary statistics. Montana is less densely populated compared to some states, but the resort's state of Montana was in the top five for size. In addition, the result shows the resort's state of Alaska is the largest among all states. In another examination, it was found that New York state with 33 resorts is in first place in order to the number of resorts, however, Montana with 12 resorts was in 12th place(Figure 1).

Figure 1 The distribution of number of resorts in each region(left) and state(right) separately.

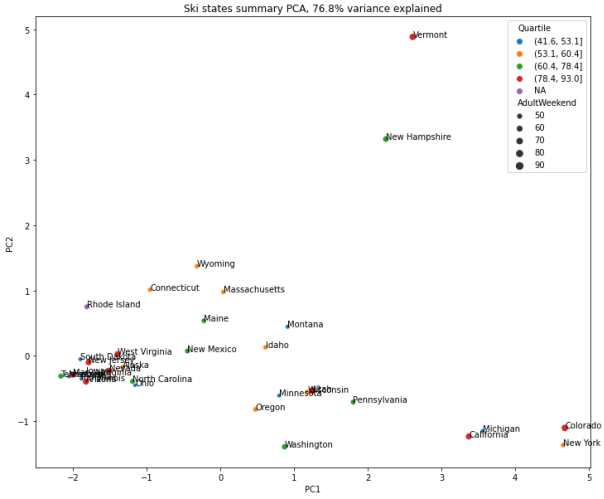


As a resort might not be suitable for skiing so we need to identify the skiing area of each state. The state New York despite having most resorts compare to other states, is not into the top five of skiable area but in term of the area of skiing available at night, is in the first place. However, Montana is in the fourth place for total skiable area but is not among the top five states with the most of night skiing area. Colorado seems to have a name for skiing; it's in the top five for resorts and in top place for total skiable area and total days open.

The result above shows big states are not necessarily the most populous, states with many resorts do not necessarily host larger total skiing area, states with the most total days skiing per season are not necessarily those with the most resorts. In this step, we are interested in measures that explain the skiing competitive landscape of each state. So, measures of resort density like the ratio of resorts serving a given population or a given area are calculated. The states Vermont and Montana are at the first and fourth place in terms of resorts per capita. The state New Hampshire is in the first place in terms of resorts per resorts per area.

After plotting cumulative variance ratio explained by PCA components for state/resort summary statistics, 75% of the variance can be explained by the first two components and the first four components explain over 95% of the variance. Then, the first two derived features are plotted adding average ticket price for each state.

Figure 2 Ski states summary PCA.




The plot shows there is a spread of states across the first component, but Vermont and New Hampshire might be off on their own a little in the second dimension, although they're really no more extreme than New York and Colorado are in the first dimension.

**4) Pre-Processing and Training Data**

In this step, after dividing the dataset into two parts train and test, data imputation is done using two techniques mean and median. Then, two models linear regression and random forest are considered as

candidates and their performance are compared to a simple baseline model. In this step, a pipeline of all steps imputing, scaling, and linear model/randomforest is built to move faster but with confidence. In order to have best performance, cross-validation and GridSearchCV with different parameters are used for estimating models performance and  hyperparameters tunning. The most important features resulted of these two techniques are similar but in different order. The selected model is random forest model having lower mean absolute error using cross-validation than the linear regression model's.

**5) Modeling**

The model selected in the last step is refitted on the train + test dataset excluding the Big Mountain Resort. The predicted price is $96.32, while the actual price is $81.00. Even with the expected mean absolute error of $10.41, this suggests there is room for an increase. However, the result is looked carefully by visualizing the distribution of ticket price of all resorts and Montana only and important features in modeling including vertical drop area covered by snow makers, total number of chairs, number of fast quads, longest run length, number of trams, and skiable terrain area. After examination, the business has shortlisted some scenarios which shows if the resort is adding a chair lift, support for ticket price will be increased by $0.29 and over the season, the resort could be expected to amount to $507246.37.