Guided Capstone Project Report

Introduction/Problem Identification

Big Mountain Resort acquired a new chairlift to attract more skiers and snowboarders. What features must they include with the added chairlift to maintain a 09.2% or higher net profit? Linear regression model will extract important features to consider as well as predict the price for Big Mountain Resort. Data on over 300 Ski Resorts across the nation is given to analyze. With the information provided Weekend prices are predicted for Big Mountain Resort as well as considerations to improve the lift and surrounding attributes for Big Mountain Resort.

Data Preprocessing/Data Cleaning

Missing values were accessed in multiple categories. Values that are logical to put as zero were filled with zeros. The categories that were filled as zeros were , FastEight Chairlifts (FastEight), night skiing area (NightSkiing_ac), and Terrain parks (TerrainParks).

Missing values that were filled with the mean in this case were Skiable terrain area (Skiableterrain_ac), Runs, yearsopen, AverageSnowfall, daysopenlastyear, longestrun_mi, projectedaysopen, and snow making_ac. Mean was used to not change the distribution of the data collected which would affect our training model.

Duplicate rows were considered since this dataset did not have any, no deletion was necessary.

Categories with less than 1% of missing values the rows were omitted from analysis.

Data Exploration

Exploration of variables is necessary to see the distribution as well as determining variables that are necessary for the linear regression model.

Comparison of States data and Regions data was done in a bar plot of the states and the number of resorts within each state. In the figures below since the distribution is the same, we can eliminate Regions variable.
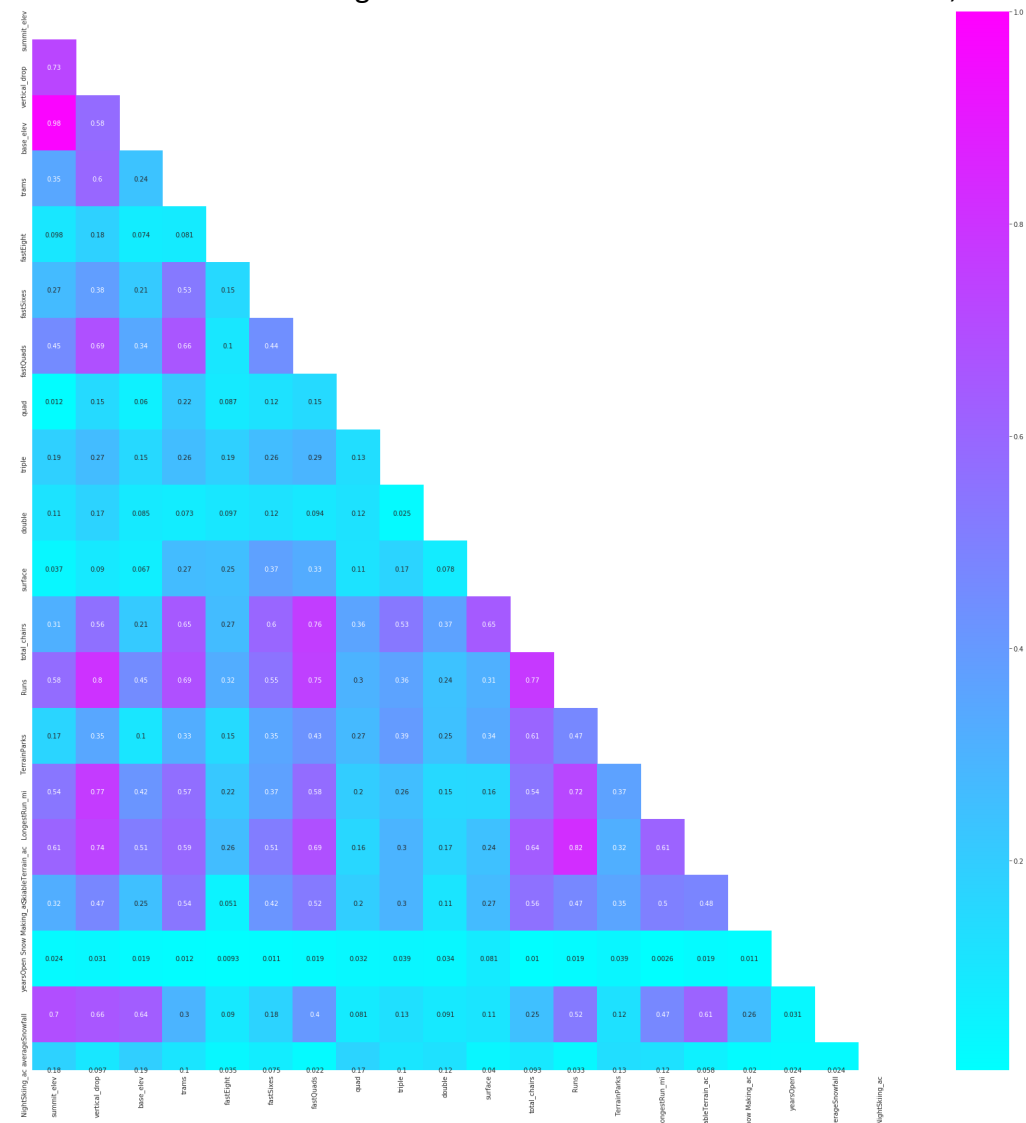
Using a correlation matrix to extract any variables that have collinearity which means having a correlation higher than 95% will be taken out of our model. In this case 'base_elev' has a high correlation so this will be dropped from our analysis.
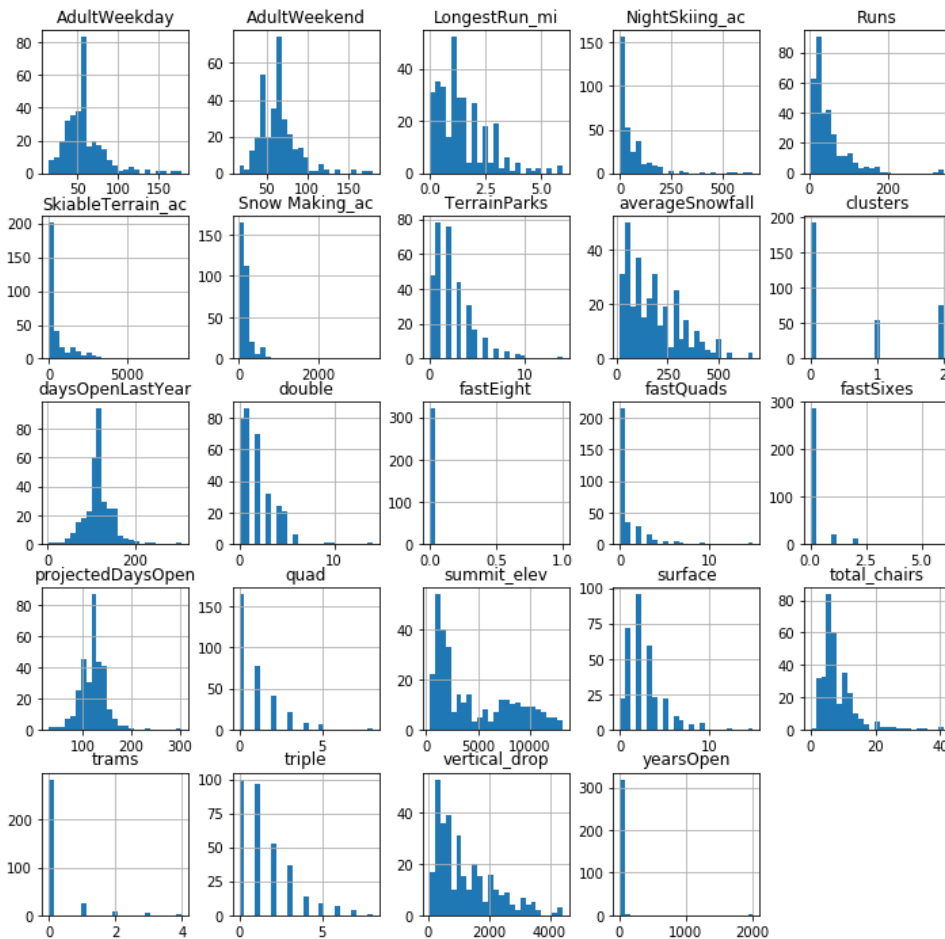
figure 3
Figure 3 refers to all the distributions with in each variable. Since a normal distribution is required for regression models standardize procedures are implemented so the data can be compared equally

Model Description and selection.
Before model training and a testing variables were considered again for our models. Since Ski resorts cannot change their location sates were removed from the data set.   The resulting data size was 322 different ski resorts with 25 different category types for processing.   Linear regression model is used to predict AdultWeekend prices.
The linear regression model produces an 81% explained variance with a 5.84 Mean absolute error. Which means that the predicted value and the true value has an error of about $5.84.

Next Steps
The model predicted AdultWeekend prices should be 63.94 and our actual price is 81.00.   The features that contributed to the model most were AdultWeekday, yearsopen, SkiableTerrain_ac, Runs, surface, quad, fastquads, triple, and daysOpenLastYear. These variable are also other variables to consider when Big Mountain adds a chair lift. To attract more customers they can lower prices and create more runs which will in turn create a larger skiable terrain area.   Further analysis with the features listed will be needed to see if it is a viable option for Big Mountain to attract more customers. Since Pricing is a quick way to attract more customers, they can create deals to attract customers to come on the weekends.