Guided Capstone Project Report

**Data Wrangling**

The data contains 330 ski resorts across 35 states which also include the target of our study, Big Mountain resort in Montana. There are some missing values in about a third of the available data elements including our target variables ie weekend and weekday ticket prices(averaged). About 14% of our records have missing price data either weekend or weekday.

There are statewide features that may be helpful as predictors of prices so we selected some and aggregated. They include, TerrainParks, SkiableTerrain\_ac, daysOpenLastYear and NightSkiing\_ac.

After getting these statewide aggregates we added state population and area data as this could definitely affect demand and hence price within the state. We then dropped records without price data.

**Exploratory Data Analysis**

The state data with the summary features are further augmented to add these measures relative to population size and area. This is more meaningful to make comparisons across states. Used Principal Component Analysis to reduce the number of features under consideration by correcting for existing features with high correlation. Using the state as index was a convenient way to deal with this categorical feature. Using scaling we were able to conform our features to the same approximate range instead of the wide variations in our heteregenous data.

We joined state level summary data and the resort level data and continued a review of the available features looking for correlation with our target variable ie AdultWeekend price. Explored the use of scatterplots and heatmaps to visualize these relationships.

**Preprocessing and Training**

Name, Region and State are not useful as features to predict Adult Weekend prices and removed from consideration in our dataset. We established a baseline model performance by taking the mean of predicted prices for each row and calculated metrics such as R2, mean squared error and mean absolute error.

**Modeling**

Various modelling scenarios suggests the management of Big Mountain Resort could make decisions about shutting down a certain of runs to recover the operating expenses incurred for those runs. A knowledge of how much each potential closing of a run would recover would help decide the optimal number of runs if any to close to maximize profits.

An application user interface such as a web application that provided the ability to input values for specific features and execute the model for those features based on the values input by the user either in absolute terms or by specifying a delta would allow recalculation and prediction of the price and the overall profitability. This is particularly true if we are able to get additional data on how changes or deltas in these features would affect cost.