Big Mountain Resort Summary

At the starting the data had 330 rows and 27 columns in total. In that data our our own resort Big Mountain Resort was at 151th row. We had to remove the fastEight columns since almost more than half of the values were missing. In the same way the rows that did not have neither Adult Weekday ticket price nor Adult Weekend ticket price had to be removed. Also we had to remove the Adult Weekday prices column since the AdultWeekend column had more information than AdultWeekday column. In the boxplot aside from some relatively expensive ticket prices in California, Colorado, and Utah, most prices appear to lie in a broad band from around 25 to over 100 dollars. Some States show more variability than others. Montana and South Dakota, for example, both show fairly small variability as well as matching weekend and weekday ticket prices. Nevada and Utah, on the other hand, show the most range in prices. Some States, notably North Carolina and Virginia, have weekend prices far higher than weekday prices. At this point we could be inspired from this exploration to consider a few potential groupings of resorts, those with low spread, those with lower averages, and those that charge a premium for weekend tickets. However, we're told that we are taking all resorts to be part of the same market share, we could argue against further segment the resorts. Another effect can also be noted above: some States show a marked difference between weekday and weekend ticket prices. It may make sense to allow a model to take into account not just State but also weekend vs weekday. Finaly we end up with 277 rows and 25 columns.

The purpose of this section is to build machine learning models. At the start we were looking for the state-wide picture for the market. All the data were numerical data. It seems that the absolute state size or population of a state are not relevant to the ticket price, instead the ratio of resorts serving a given population or a given area of a state are more relevants. So we introduce two columns with resorts_per_100kcapita and resorts_per_100ksq_mile in place of state_population and state_area_sq_miles. So we have constructed some potentially useful and business relevant features, derived from summary statistics, for each of the states we're concerned with. We've explored many of these features in turn and found various trends. Some states are higher in some but not in others. Some features will also be more correlated with one another than others. One way to disentangle this intercommencted web of relationship is via PCA. From the heatmap it is very clear that AdultWeekend ticket price has quite a few reasonable correlations. fastQuads stands out, along with Runs and Snow Making_ac. The last one is interesting. Visitors would seem to value more guaranteed snow, which would cost in terms of snow making equipment, which would drive prices and costs up. Of the new features, resort_night_skiing_state_ratio seems the most correlated with ticket price. If this is true, then perhaps seizing a greater share of night skiing capacity is positive for the price a resort can charge.

In machine learning if we keep training our model on all of the data or keep making more and more complex models that fit the data better and better we might have possibility to end up with (1) no data to test the performance of the model or (2) overfitting the model for that sample of data. To avoid the problem (1) one approach we might proceed with is to partition the data in 2 sets of data for training and testing the model. So we train the model for one set of data and test the performance of the model with the other set of data. Here we partition the data 70/30. It is always a good practice to start by checking whether mean is a best guess or not. To check how good the guess is, we use the metrics R-squared, mean absolute error or mean squared error. One important warning I learned here is that while calculating R-squared we need to put the arguments in right order. If we have missing values in the data we need to impute the data and to do that median is a good choice if the distribution is skewed. Also mean is a good choice for that as well. To perform this series of tasks we used the pipeline of the sklearn. By default the pipeline uses 10 features from the data which might be good or bad depending on the data. For every choice of k we need to cross validate. So it is necessary to find the appropriate number of features (k) for the best performance. We came to learn about the Random Forest Model and linear regression model. Since the random forest model had the lower cross-validation mean absolute error we have select it as our best model for this data.

The objective of this section is to gain some insights into what price Big Mountain's facilities might actually support as well as explore the sensitivity of changes to various resort parameters. It is worth to note here that to accomplish the aforementioned task there was an implicit assumption which is "all other resorts are largely setting prices based on how much people value certain facilities". The purpose of this capstone project is to predict the appropriate ticket price for the Big Mountain Resort. Therefore to remove the biasness we need to exclude the Big Mountain information from the data. Our chosen model predicted that the ticket price should be $95.87 instead of $81.00. Next we checked that where Big Mountain stand in the market context and we got a positive feedback. The business has shortlisted some options and we checked whether those options has any impact on the ticket price or not. Some of the options suggested to increase the ticket price by $1.99 and some other options had no impact at all.

We picked the Adult Weekend ticket prices and had to get rid of the Adult weekday ticket prices due to the lack of data for that particular column. We could have found more accurate prediction if we had both of the ticket prices information. Our model predicts that the Big Mountain Resort is charging less based on the assumption that all the other resorts are pricing the ticket correct. It could be true or it could be for the reason that we don't have enough information, for example, the operationg costs.