Big Mountain Resort is planning on adding an additional chair lift to their ski resorts which will add an additional cost of $1,540,000 to their normal operating costs for this current season. The problem is that they are currently basing their price on the average cost of other resorts and simply just adding a premium on top of the average and they want to have a more data driven approach to deciding on their pricing. They also want to make decisions on other changes that will lower their operating costs while adding value so that they can charge more for a ticket or at least cut costs and have the ticket prices remain the same.

The data was gathered from a csv file and we had to drop some of the data to null values or/or make changes to incorrect data. We also checked to see if any duplicate names of resorts appeared and checked to see if Region and State are always the same as it appeared they were initially. We found that there were a few unique regions that were not named the same as the state which helped us to realize that a duplicate resort name was actually a different resort from a different state. Each state was shown to have different levels of variance in ticket price and some of the variables had suspicious outliers that were investigated and removed if proven to be incorrect such as skiable terrain in one resort. This was all done to ensure the data is accurate and correct conclusions can be drawn from them. We added population and area data onto our ski data to know more about how population volume and ski resort density could affect ticket prices.

In Data exploration, we explored which of the states were the top in a few categories such as state area, population, resorts per state, resort density and more. We decided that calculating resort density and population density would be of value and created new columns with this info. We then scaled the data so that every column was on the same scale and therefore could be understood better by a machine learning algorithm. This sets the mean of each column to 0 with a standard deviation of 1. After this, we used PCA to determine how much of the data's variance could be captured with a certain amount of components. It was shown that with just the first 2 components, we accounted for over 75% of the variance. We calculated the average ticket price per state. We had to remove some null values for price since earlier on, we left some of the resorts that only had 1 out of the 2 listed prices in the data. Rhode Island had a null value and its price was replaced by the mean ticket price. A heat map was created to show how each component correlated with one another showing high correlations and now correlations on several pairs.

In the data preprocessing and training, we split our data into a test group and a train group. 30 % of the data was used for the test and 70% for the train data. Each group was given an x and a y variable where x represents the features we are using (independent variables) and y is the dependent variable which in our case was the price of the tickets. Our target resort was not included in the training data. We removed data that was not numeric as this data cannot be read by the machine learning algorithm. The first test was to see how good just taking the mean of the prices would be at predicting prices just to use it as a comparison baseline. We created formulas and then later just used the built in formulas from sklearn for determining variance of components by using R2, mean absolute error and mean squared error. These are metrics used to determine how well our model is performing. Since we used the mean of prices as our model,

these metrics revealed that we had room for improvement on the model as was expected. After this, we filled in NA values with the median price and later scaled this data to be more easily read and understood by our model. A linear regression model was trained using this updated and imputed data and we retested the model using the same metrics mentioned above. We were able to over 80% of the variance on the train set but got lower scores on the test group, showing we were on the right track but perhaps overfitting on the data a bit. This process was repeated but using the mean to impute data this time. The results were not much different from using the median. We finally decided to use a method that finds the most valuable features that impact the price otherwise known as selectKBest but this actually showed a worse prediction. We had to use cross validation which folds the training data into different ways essentially making small training data subsets to test different combinations. The last strategy was the random forest model which showed to have lower variability and the metrics suggested it was a better model. We used this as the chosen model.

After creating the model, we refitted it using all of the data but excluding Big Mountain since it is our target resort we are trying to predict the price for. Using this fitted model, we predict that big mountain can be charging 95.87 for a ticket which is higher than the current $81 they are charging. Even with the margin for error, they can still safely charge more for a ticket. After reviewing a handful of the features that were shown to affect ticket price, Big Mountain was in the upper range of all of them, showing that a premium price was likely justified. We created a function to see how adding or subtracting from each of these features would make changes to ticket price. I would recommend that Big Mountain raises their ticket price to $90 before making any additional changes. This is $5 less than the predicted price but gives a bit of wiggle room considering the margin for error of about $10. I recommend they also remove the 5 least popular runs since this won't reduce ticket price much but will reduce operation costs. By adding the new lift and by increasing the vertical drop of the tallest slope, ticket price is predicted to go up 2$ which alone would account for a $3,474,638 increase in revenue. This would more than cover the cost of the new lift and likey cover the costs of the other changes. When we also factor in the $9 increase to the base price before making changes and the reduction in operation costs from removing runs, this will put them even more in the positive. We do not have data for the reduction in cost from removing runs, but with the expected 350,000 visitors and average 5 tickets per visitor that would be an expected increase of nearly 16 million dollars on top of the 3.4 million dollar increase. That's nearly 20 million dollars in total! It's hard to predict what changes to average tickets bought and average visitors will happen with a change in ticket price, but we predict that adding the additional chair lift and higher peak will justify the price increase and will not reduce tickets bought and visitors.