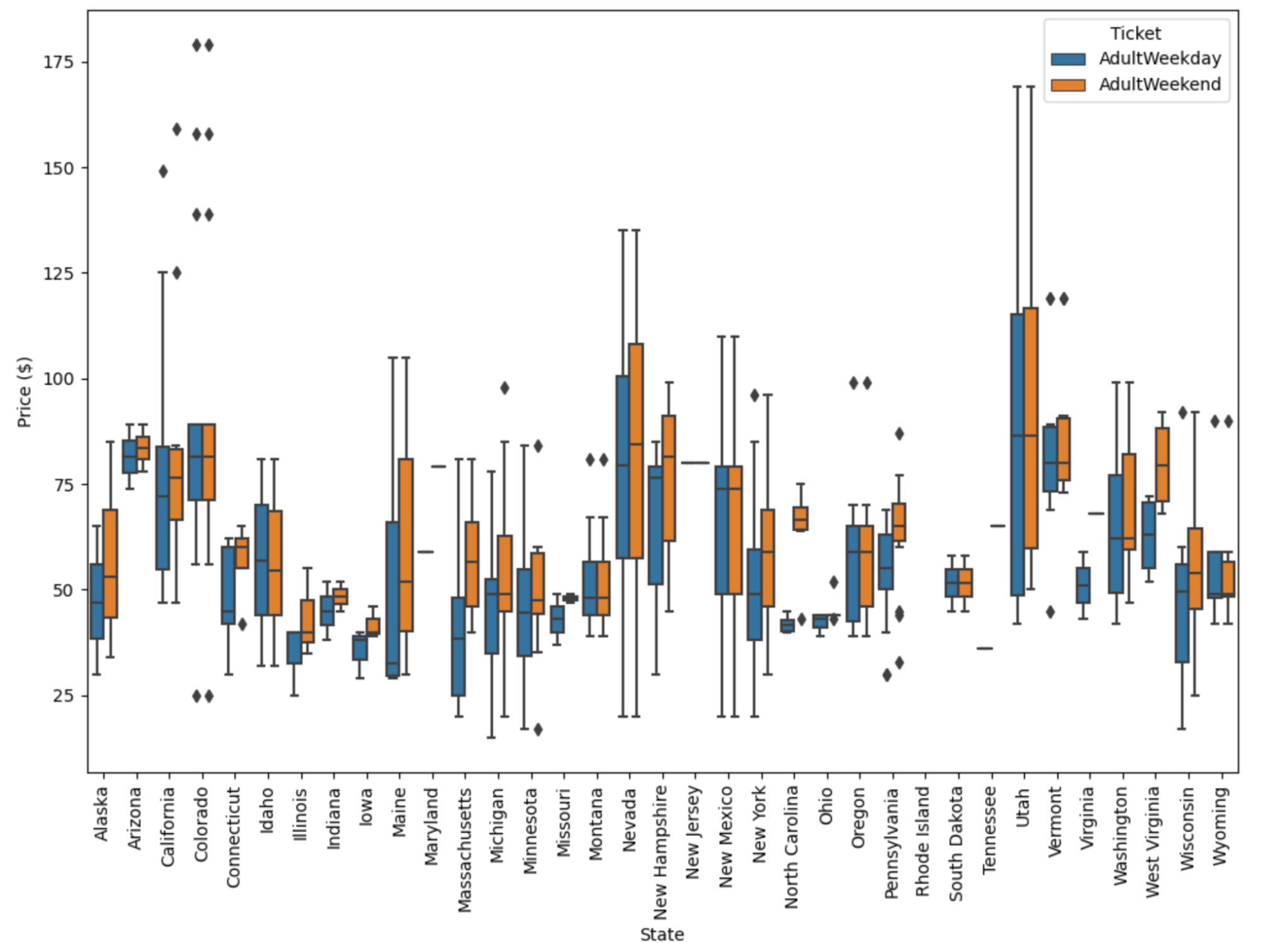
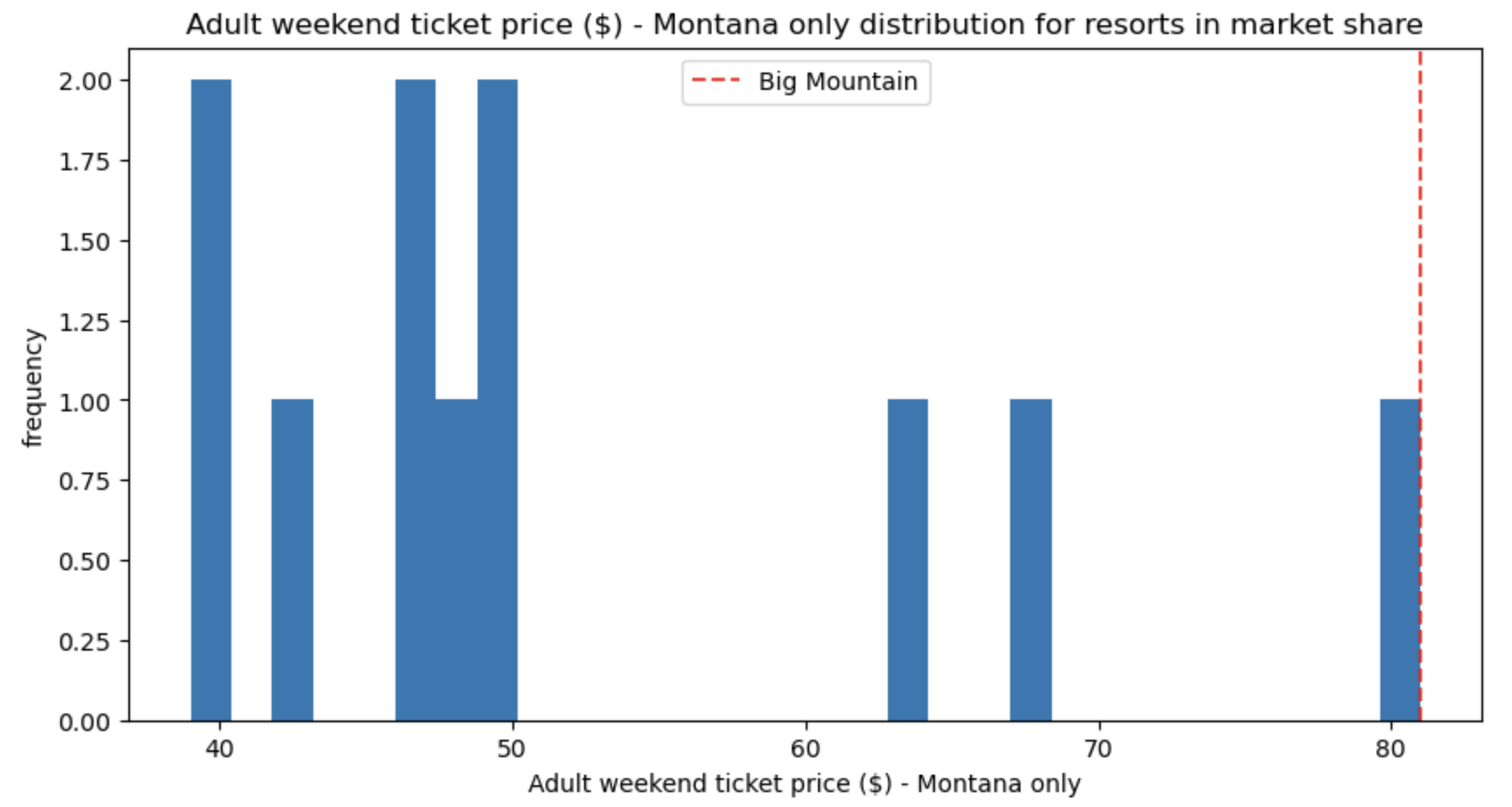
It’s critical in any data analysis project to identify the right problem. Building a model that doesn’t solve the correct problem is useless. In our analysis, we are looking for what opportunities exist for Big Mountain to increase revenue by $1,540,000 this season through cost cutting or increasing ticket prices to offset the additional expenses incurred by adding an additional chair lift. Our focus will be on the value of each facility and how this impacts ticket prices.

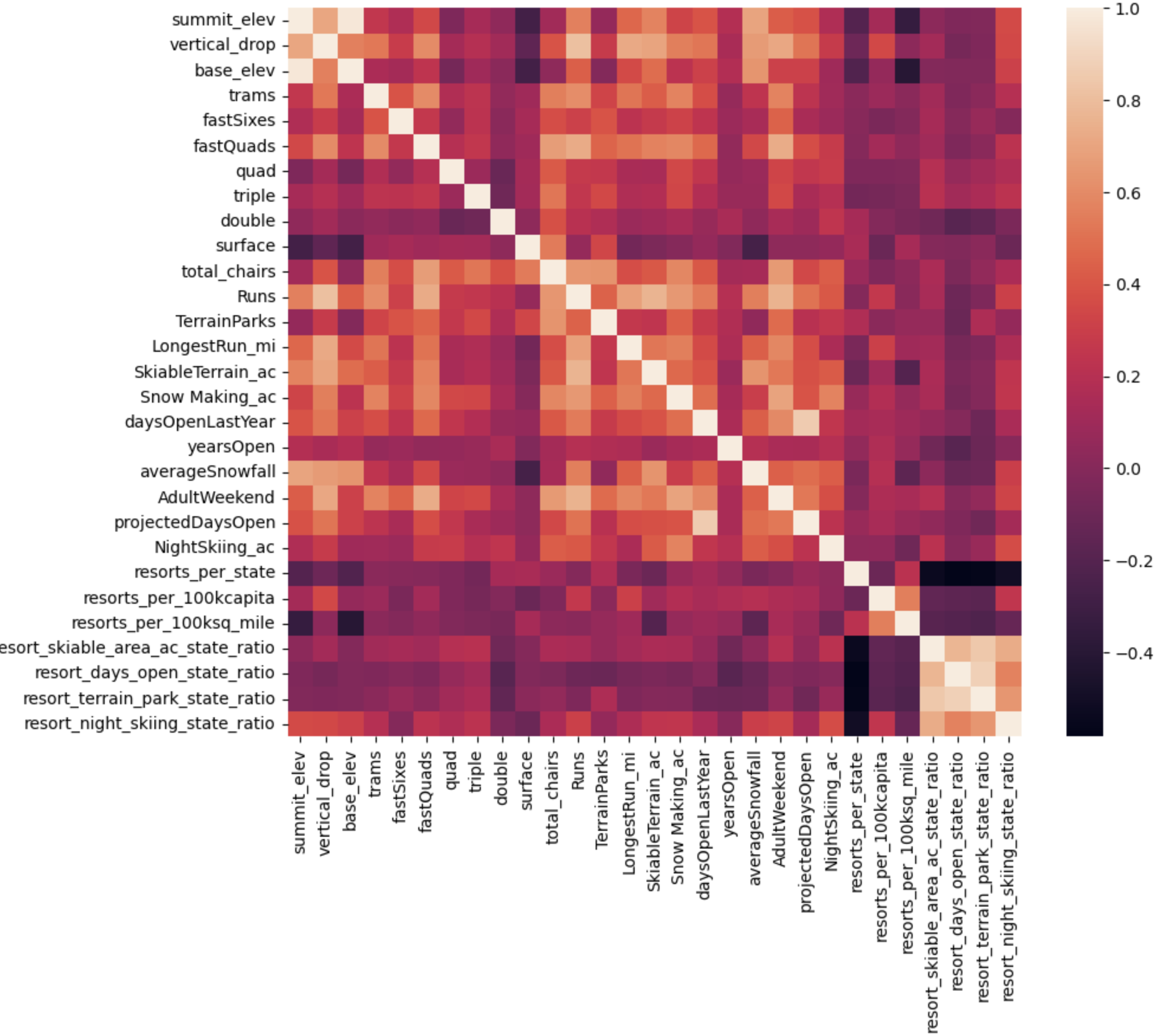
Now that our problem is defined, I did a basic review of the provided data for accuracy, tidiness, and completeness. This involved identifying missing values and determining how to impute the missing values or drop them. I found that the dataset contained two types of ticket price: an adult weekday price and an adult weekend price and on checking missing values, I dropped all records with no pricing data. To better understand the distribution of ticket prices, I calculated the mean ticket price of ‘AdultWeekday’ and ‘AdultWeekend’ by state and visualized it with a barchart and boxplot to help identify outliers.



I chose to aggregate some statewide data (resorts per state, terrain parks, skiable terrain area, days open last year, and night skiing area) to help understand how these features impact ticket pricing strategies. I also pulled some population data from wikipedia to provide additional context.

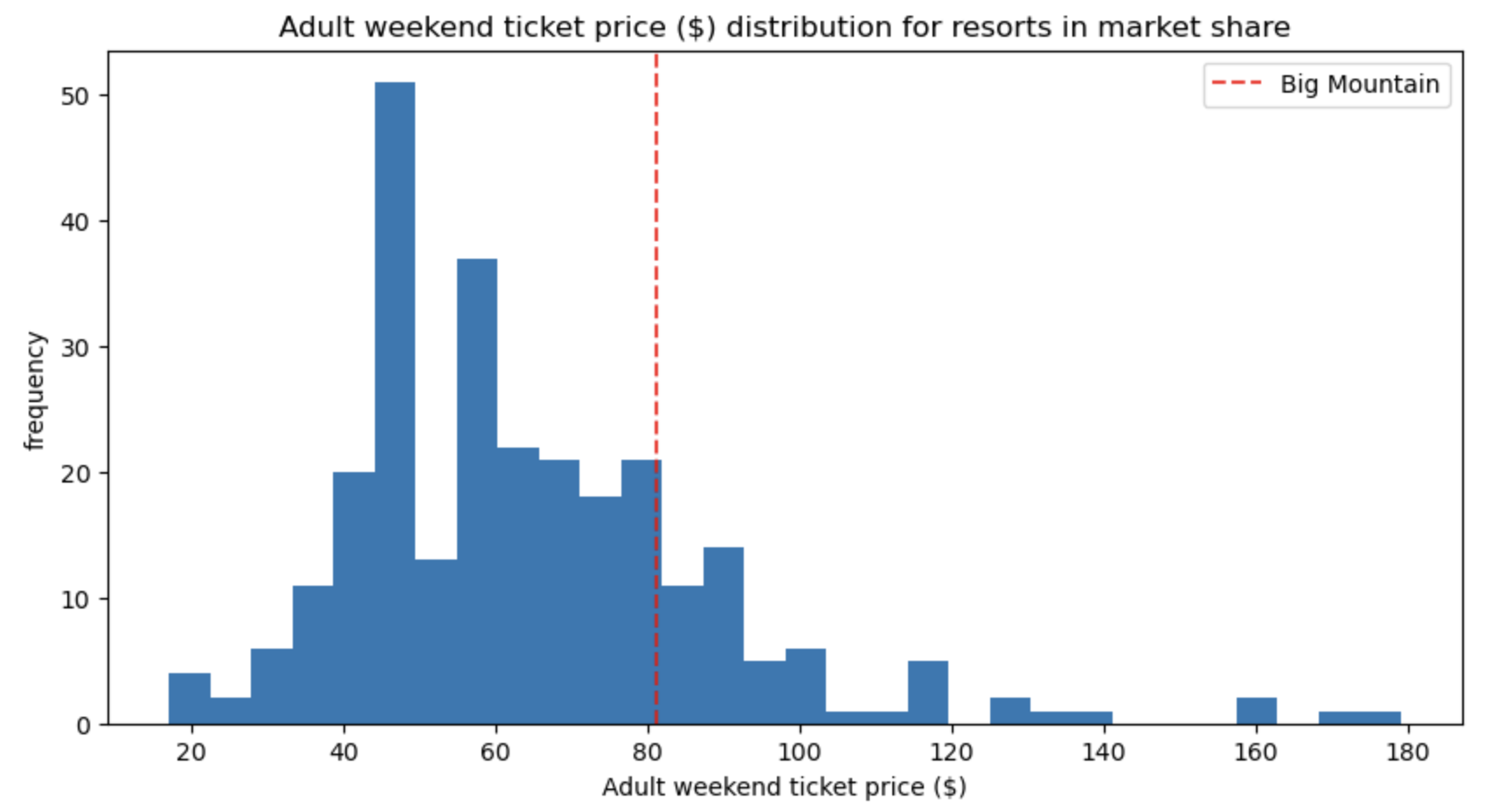
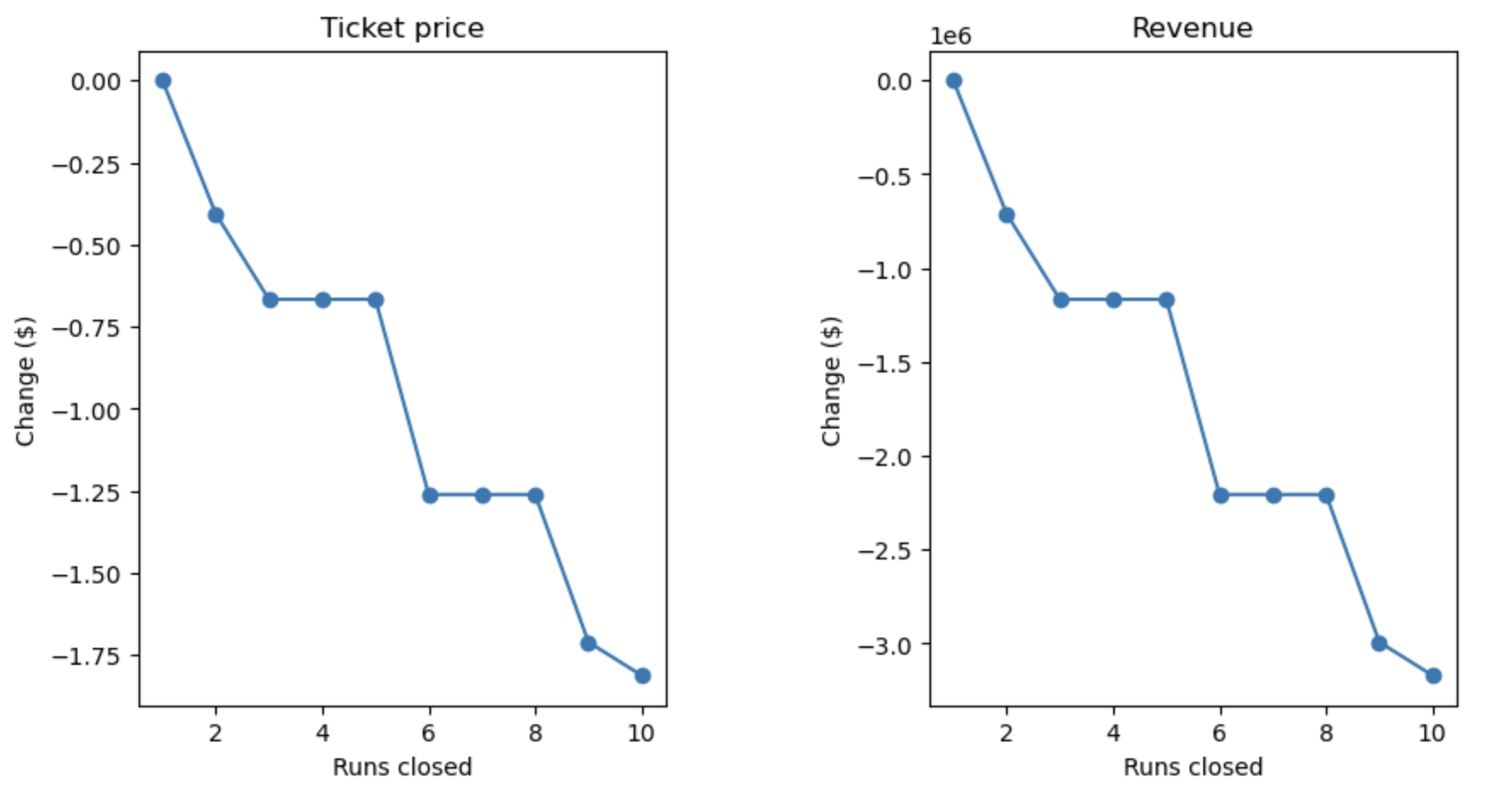
Most ticket prices fall between $25 to over $100 dollars. The current weekday and weekend prices for our resort, Big Mountain, were seen to be equal, at $81. This ticket price is at the upper end of the ticket prices in Montana, and so it makes sense to treat weekend and weekday prices the same. Thus, since there were a few more weekend prices than weekday, I dropped the weekday prices and chose the weekend price to be the target ticket price.

Continuing to explore the data I was unable to identify a clear pattern suggestive of a relationship between state and ticket price. Due to this, state labels were treated equally towards building a pricing model that considers all states together, without treating any one particularly special. Real insights were gained after calling a seaborn correlation heatmap on the original dataset that identified which features were more positively and negatively associated with ticket price. From here it is immediately clear that ticket price is heavily impacted by these primary features: FastQuads, Runs, Vertical Drop, Total Chairs; and then moderately impacted by the secondary features: Longest Run, Skiable Acres, Snowmaking Acres, and Night Skiing. It seems that the four primary features are in fact best for modeling ticket price.



One relationship to be wary of is the relationship between ‘AdultWeekend’ ticket price and ‘resorts\_per\_100kcapita.’ There is a lot of variability in ticket price and can be attributed to several reasons. Also, moving forward I think it's prudent to stay wary of comparing prices to what ratio a ski resort has of any feature statewide. This is particularly noticeable in New England, but across the dataset having the best resort for a category in a state isn't necessarily indicative that you're the highest priced resort in the actual "market".

With the data in a good format for analysis, I set the baseline for performance by identifying the average (mean) ticket price. This was used to compare the subsequent ML models to measure their predictive effectiveness. I used the R^2 metric to determine the amount of variance beyond that of using just the mean as well as the Mean Absolute Error (MAE) and Mean Squared Error (MSE) to summarize the difference between the predicted and actual values. The MAE resulted in being off on average by $19 if I guessed the ticket price based on an average of known ticket prices for the test data. The MSE resulted in being off by $24. After determining the baseline performance, I then took steps to create a linear model and a Random Forest model. Vertical drop, Snow Making, Total chairs, Fast Quads, Runs, Longest run , Trams, and SkiableTerrain were selected as the best estimators for the linear model. Vertical drop was the biggest positive feature with a coefficient of 10.767857 and Snow Making was the second biggest at 6.290074. The dominant top four features for random forest regressor were (in order): Fast Quads, Runs, Snow Making, and Vertical drop. This is encouraging as these are also top features with the linear model. The results of these models were a final mean absolute error score for the linear regression model of 11.79 and the random forest regression model of 9.53. The Random Forest model also exhibited less variability, so this is the model I propose we use going forward.

Four additional scenarios were presented to evaluate whether to cut costs or increase revenue via a ticket price increase : 1) Permanently closing down up to 10 of the least used runs. This doesn't impact any other resort statistics. 2) Increase the vertical drop by adding a run to a point 150 feet lower down but requiring the installation of an additional chair lift to bring skiers back up, without additional snow making coverage. 3) Same as number 2, but adding 2 acres of snow making cover. 4) Increase the longest run by 0.2 mile to boast 3.5 miles length, requiring an additional snow making coverage of 4 acres. If we wanted to look into cutting costs, the scenario of closing down runs can be considered. The model says closing one run makes no difference. Closing 2 and 3 reduces support for ticket price and so revenue. If we close down 3 runs, it seems we may as well close down 4 or 5 as there's no further loss in ticket price. Increasing the closures down to 6 or more leads to a large drop in ticket price. Since I don’t know the impact of run closures on operating costs, we would need operating cost data to explore this scenario further. Scenario 2 displays the impact of vertical drop having a positive correlation with ticket prices. This supports a ticket price increase of $8.61 which would generate an additional $15 million in revenue. Assuming another chair lift would increase operating costs by $1.54 million, this scenario suggests that addition of a chair lift to increase the vertical drop would be worth exploring. Scenario 3 is very similar to scenario 2, but with the addition of 2 acres of added snow making. The model suggested support for an increase of $9.90 resulting in an additional $17 million in revenue. Again, I would need to know the operating costs associated with the increase in snow making equipment, but it is probably a safe assumption that it is less than the additional $2 million that scenario 3 predicts compared to scenario 2. Scenario 4 investigated whether or not increasing the longest run to 3.5 miles and adding additional snow making coverage of 4 acres would predict an increase in price. These changes had no impact on an increase in ticket price. It further suggests that the longest run feature is lower down in the feature importance list using the Random Forest model.

As a reminder, Big Mountain is charging $81 for a lift ticket for adults on weekends. In order to cover the additional $1,540,000 in operating expenses, a small increase on average of $0.88 per lift ticket (on the basis of each visitor on average buying 5 day tickets and assuming 350,000 visitors this season) would cover the additional operating costs.The model suggests that a ticket price of $95.87 could be supported. Even taking into account the expected mean absolute error of $10.39, this still suggests there is room for an increase.

Looking into the future and how to improve our predictive capabilities, the addition of operating costs and visitor satisfaction scores would help refine the model to be more accurate. Not knowing the operating costs prevents a true understanding of the profit margin if any. Understanding how visitors feel or value each of the facilities at the resort would also help weigh the feature importance providing a more holistic view of the chosen features. Big Mountain ranks fairly high on many of the league charts of facilities offered, but the modeled price comes in even higher than the current price.

This might suggest that we are undercharging because we aren’t weighting the importance of our facilities properly. It’s reasonable to expect that some resorts will be overpriced and some underpriced. Or if resorts are pretty good at pricing strategies, it could be that our model is simply lacking some key data? The addition of operating costs and visitor satisfaction scores could help refine the pricing model further.