For the first step of the guided capstone, I filled out a problem statement worksheet based on the information provided by Big Mountain case study overview. The hypothesis formation is how should Big Mountain Resort selects a better price strategy in order to offset the additional operating costs of $1,540,000 for this season? The context is: Big Mountain Resort is a ski resort offering spectacular views of Glacier National Park and Flathead National Forest, with access to 105 trails. Because Big Mountain Resort has recently installed an additional chair, their operating costs increased by $1,540,000 this season. In response to this extra cost, a better pricing strategy must be made in order to increase revenue for the resort for this season. The Criteria for success is Success for this project = aligning on a detailed rollout map to make a better pricing strategy to achieve an increase revenue of $1,540,000 for Black Mountain Resort this season. The Scope of solution space is: Data from other ski resorts with similar operation conditions. The Constraints within solution space are: Cannot find ski resorts with similar operation conditions; Missing key data, such as other resorts’ revenues; and difficult in acquire key data. e.g. other resorts refuse to disclose their revenue. The Stakeholders to provide key insight are Jimmy Blackburn, the Director of Operations; and Alesha Eisen, the Database Manager. The Key data sources are Ski resorts with similar operation conditions as the Black Mountain Resort; Revenues from these ski resort; The amount of people visited at these ski resorts on weekdays and weekends; Cost (operational, maintenance, etc.) from these ski resorts.

The second step in the data science method is data wrangling, which involves taking raw data and preparing it for processing and analysis. This project aims to guide Big Mountain Resort's pricing and future facility investment plans. In order to do this, some fundamental questions need to be answered via this data-wrangling process. I first got an overview of the ski resort data and the data of our interest. Initially, the ski resort data had 330 rows and 27 columns. I explored the data by looking at the number of missing values, duplicated resort names, geographic information, weekend and weekday ticket price, etc. When I investigated the histograms of numeric features, I noticed that one resort has an incredibly large skiable terrain area. By browsing its official website, an error was found and corrected. The same issue was found for the skiable area, and its data was updated. I also spotted and removed two outlier data of “yearsOpen”. Furthermore, a few columns and rows were dropped because of the severity of missing data (“fastEight”, and “AdultWeekend/AdultWeekday”, respectively). With deeper going through the information of every state’s population and area from the internet (states\_url = '<https://simple.wikipedia.org/w/index.php?title=List_of_U.S._states&oldid=7168473>'), I noticed the state names that contain 'Massachusetts', 'Pennsylvania', or 'Virginia' has square brackets, resulting 4 missing states after comparison. Therefore, I removed the brackets, including their contents, so I could confidently add the population and state area columns to the ski resort data. I generated a scatter plot to visualize the relationship between weekday and weekend prices and further extracted such data for Montana since Big Mountain Resort is located in Montana (Figure 1).

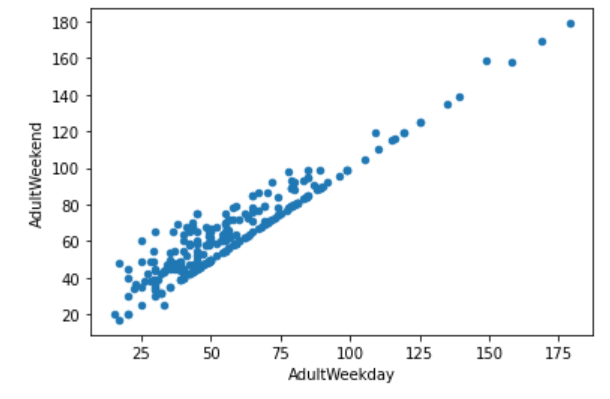


Figure 1. Weekday and weekend price correlations.

During this process, I found that the weekend prices have the least missing values of the two. Hence, I dropped the weekday prices and kept just the rows that had a weekend price. In the end, I double-checked the amount and percentage of missing values by row. Although there were still a few rows containing missing data, at this point, I was hesitant to remove them further because I didn’t know how useful the missing features were in predicting ticket prices, so I decided to keep these rows. Finally, I ran the info on the ski report data again, and there were 277 rows and 25 columns.

The step three of the Data Science Method is the exploratory data analysis (EDA), an approach for summarizing and visualizing the important characteristics and statistical properties of a dataset. Visualizing the data helped me to identify emerging themes. Identifying these trends further helped me to form hypotheses about the data. During this step, I dig deeper of the numerical features in the data, which involved the ski resort’s characterizations, such as the total chairs, skiable area, weekday and weekend prices, etc. After merging with the state info, the numerical data added how many resorts in each state, the total skiable area in each state, how many days are open per year in each state, how many total night skiing areas per state, how many terrain parks in each state, how many resorts per 100k people and 100k square mile in each state (Figure 2).

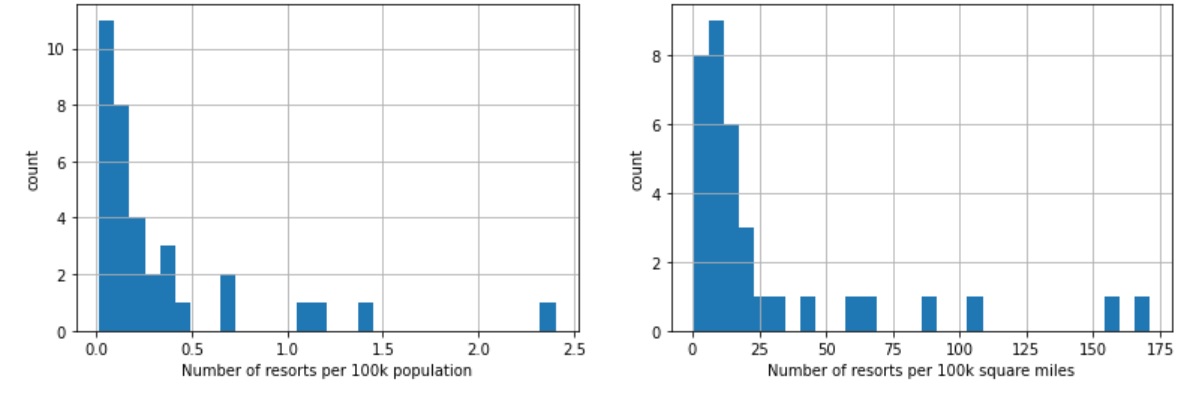


Figure 2. Number of resorts per 100k population (left) and per 100 square miles (right).

The categorical features in the data are the resort names, the region, and the state where the resorts are located. There is no direct relationship between the state and its ticket price. However, some state-related categories were analyzed in this work to demonstrate the relationship alternatively. For example, resort\_night\_skiing\_state\_ratio was the ratio of the night skiing area of the resort and the total state night skiing area, which seems the most correlated with the ticket price. In addition, resorts\_per\_100kcapita is the info regarding the number of resorts in the state per 100k population. When this value is low, there is quite a variability in the ticket price, although it's capable of going relatively high. Ticket prices may drop a little before climbing upwards as the number of resorts per capita increases. Ticket prices could rise with the number of resorts serving a population because it indicates a popular area for skiing with plenty of demand. The lower ticket price when fewer resorts serve a population may similarly be because it's a less popular state for skiing. The high price for some resorts when resorts are rare (relative to the population size) may indicate areas where a small number of resorts can benefit from a monopoly effect. Since our ultimate goal in this project is to make a pricing strategy, any analysis done in this “exploratory data analysis” with the result revealing some correlations with the ticket price, is worth digging deep into in subsequent modeling. Our target resort- Big Mountain Resort, is located in Montana, the third largest state area, but not in the top five of state population nor the number of resorts per state. However, it has the fourth biggest total skiable area and the fourth resort density (per capita, not per area). It looks like Montana is still in a decent place nationwide, and it is easy to make people think Montana has fewer but larger resorts, so I readily compare prices with those larger resorts. Remember that the target features and their relationships must be comparable with the Big Mountain Resort to perform any feature selection for modeling.

The next step is data pre-processing, which is an important step because it involves the process of removing things like out-of-value ranges and impossible combinations from your dataset. Analyzing data that hasn’t been properly screened can lead to misleading or incorrect results. This step also includes the process of splitting the dataset into testing and training subsets. Since predicting the adult weekend ticket price is my primary goal, and I have records with missing values. I need to predict these missing values and try different models to pick the best one with a machine-learning approach. After extracting the Big Mountain data, I partitioned the data into 70/30 training and testing splits. I wondered how good the mean is as a predictor, so I built a linear regression model and used R square as the metrics that summarize the difference between predicted and actual values. In order to impute the missing values, I first used the median and scaled the data; then, I made the predictions using the model I trained on the train split. This simple linear model explained over 80% of the variance on the train set and over 70% on the test set; the lower value on the test set made me suspect the model was overfitting. Moreover, by calculating the mean absolute error on the training and testing splits, I had a better estimate of a ticket price than using the mean as the predictor. Then instead of the median, I imputed missing values with the mean, and I got a similar result as using the median. Still, I need to pick the model with the best test set performance; the cross-validation can provide the best estimate of a model’s performance and whether its performance on the test split was consistent with this estimate. The cross-validation requires a range of k to investigate. A plot of k value vs. R square helped to pick up the good value for k. The model coefficient revealed the positive and negative features associated with the price. Repeatedly checking the performance wasn’t a good practice. Therefore, a random forest regressor was tried, in which a pipeline was defined to assess performance using cross-validation. The random forest model marginally improved upon the default CV results and the dominant top four features are in common with the linear model (Figure 3).

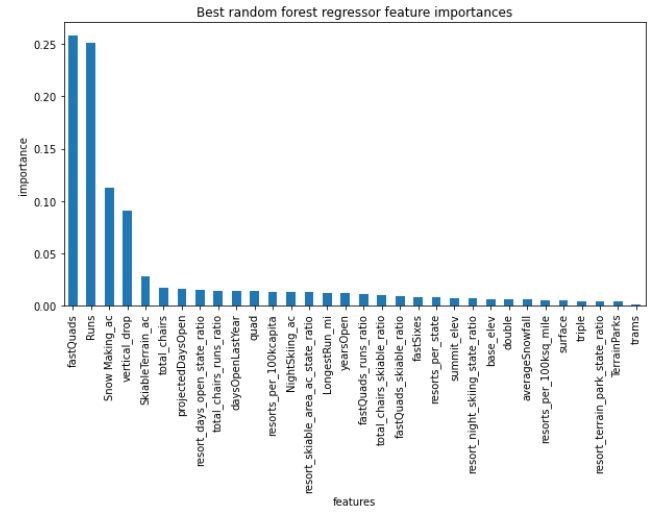
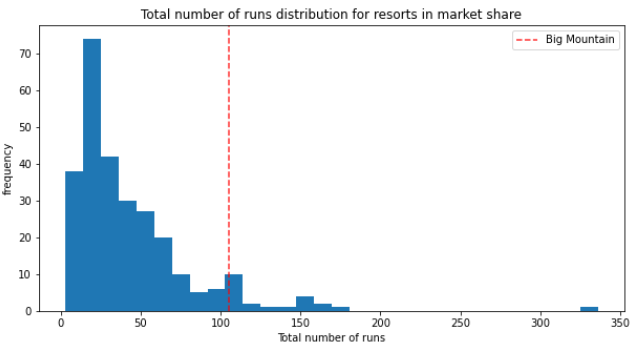
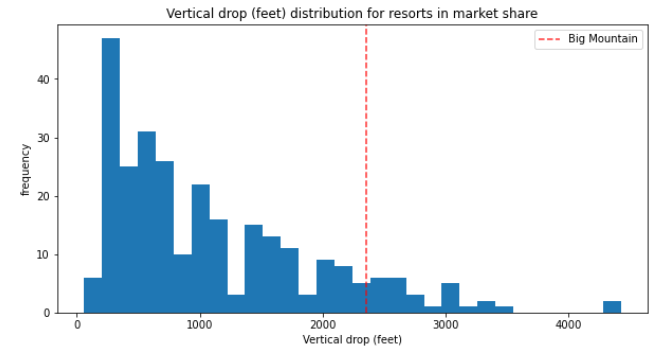
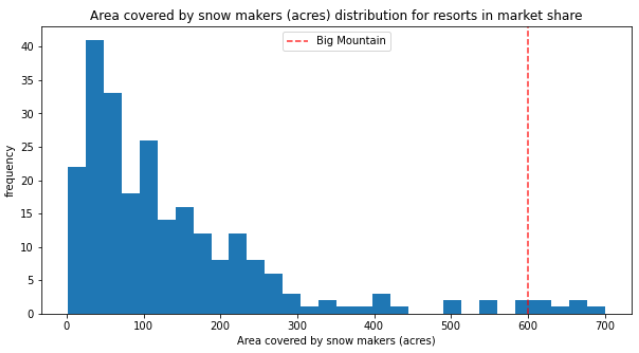
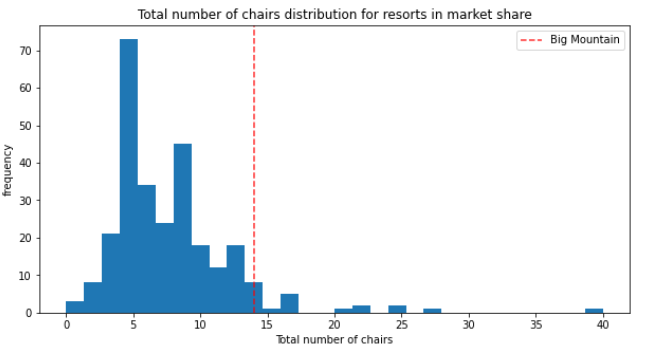


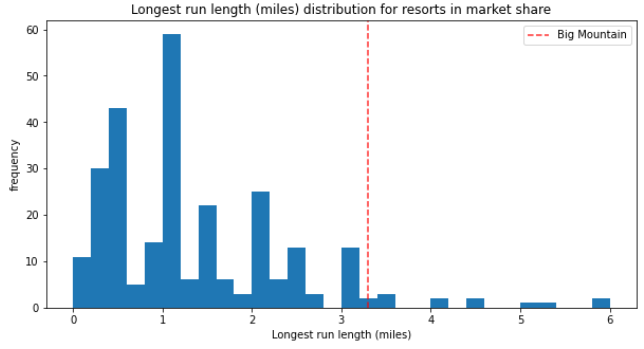
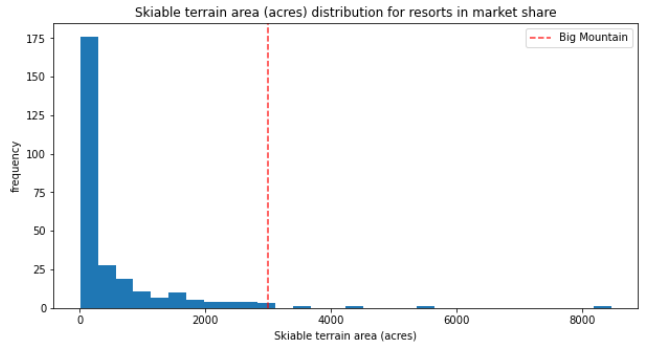
Figure 3. Best random forest regressor feature importances.

When comes to picking up the better model, linear or random forest? I decided to use the mean absolute error using cross-validation to help me choose. The mean absolute error in the linear regression model is 11.793465668669327 with a standard deviation of 1.6220608976799646, while the mean absolute error in the random forest regression model is 9.537730050637332 with a standard deviation of 1.3528565172191818. The random forest model has a lower cross-validation mean absolute error and exhibits less variability. Verifying performance on the test set produces performance consistent with the cross-validation results.

Modeling is the fifth step in the Data Science Method, it involves both model training and selection, as well as model deployment. During this step, I leveraged my cleaned and processed data to make predictive insights. Originally, the Big Mountain Resort's price for AdultWeekend was $95.87, which is $14.87 higher than the predicted price. After looking closely at where the Big Mountains at overall among some important features. I found out that the Big Mountain was doing well for vertical drop and the number of runs, very high up the league table of snowmaking area, having the highest number of total chair and skiable terrain area, having the longest runs, and three fast quads, whereas most resorts don’t have any (Figure 4).





   
Figure 4. Advantages of Big Mountain Resort: data of vertical drop, number of runs, snowmaking area, total chair numbers, skiable terrain area, and runs.

Based on these advantages and the original price was much lower than the predicted one, I could approach the business leadership there was room for a price increment. If Big Mountain was going to add a run, increasing the vertical drop by 150 feet, and installing an additional chairlift, this scenario would be supported from the model for ticket price increased by $ 1.99, which lead to the final price of $97.86. Over the season, this could be expected to amount to $3474638. This scenario is the one I would recommend for further consideration. In terms of closing any least used runs, two plots were created side by side to predict the ticket price change for each number of runs completed and the associated predicted revenue with the assumption that each of the expected visitors buys five tickets. This model indicates closing one run makes no difference. Closing 2 and 3 successively reduces support for ticket prices, and so does the revenue. If Big Mountain closes down three runs, it seems they may as well close down 4 or 5, as there's no further loss in the ticket price. If increasing the closures down to 6 or more leads to a significant drop in the ticket price and the revenue.

However, even I used this model to make the price prediction, there are deficiencies in the data that hampered a better work. The deficiencies include: a lot of missing values; the only price data available is the ticket price; no operating cost data is available; no maintenance cost data is available; no inspection or licensing data is available; no operating cost, maintenance, inspection, and licensing data are available from other resorts; no data on the expected number of visitors from other resorts. If I add a new chair lift, other information, such as additional snowing-making coverage, maintenance, licensing and inspection cost, will be helpful besides the operating cost. This modeled price was optimistic but also doubtful because the validity of the model lay in the assumption that other resorts accurately set their prices according to what the market (the ticket-buying public) supported. The fact that Big Mountain resort seemed to be charging much less than predicted suggested the Big Mountain resort might be undercharging. If the Big Mountain resort was mispricing, this might also happen to other results. So it's reasonable to expect some resorts to be "overpriced" and some "underpriced." If resorts want to be accurate at pricing strategies, key data such as operating and maintenance costs must be provided and also accurate. I would perform a scenario study, which means running some combinations of parameters as scenarios and presenting the result as a map or contour for the business analytics and management team. This map would provide guidance for the general decision. As for detailed analysis, I will prefer to package the model into an application with a basic graphic user interface, which allows the user to change parameters and generate output without needing to rebuild the model.