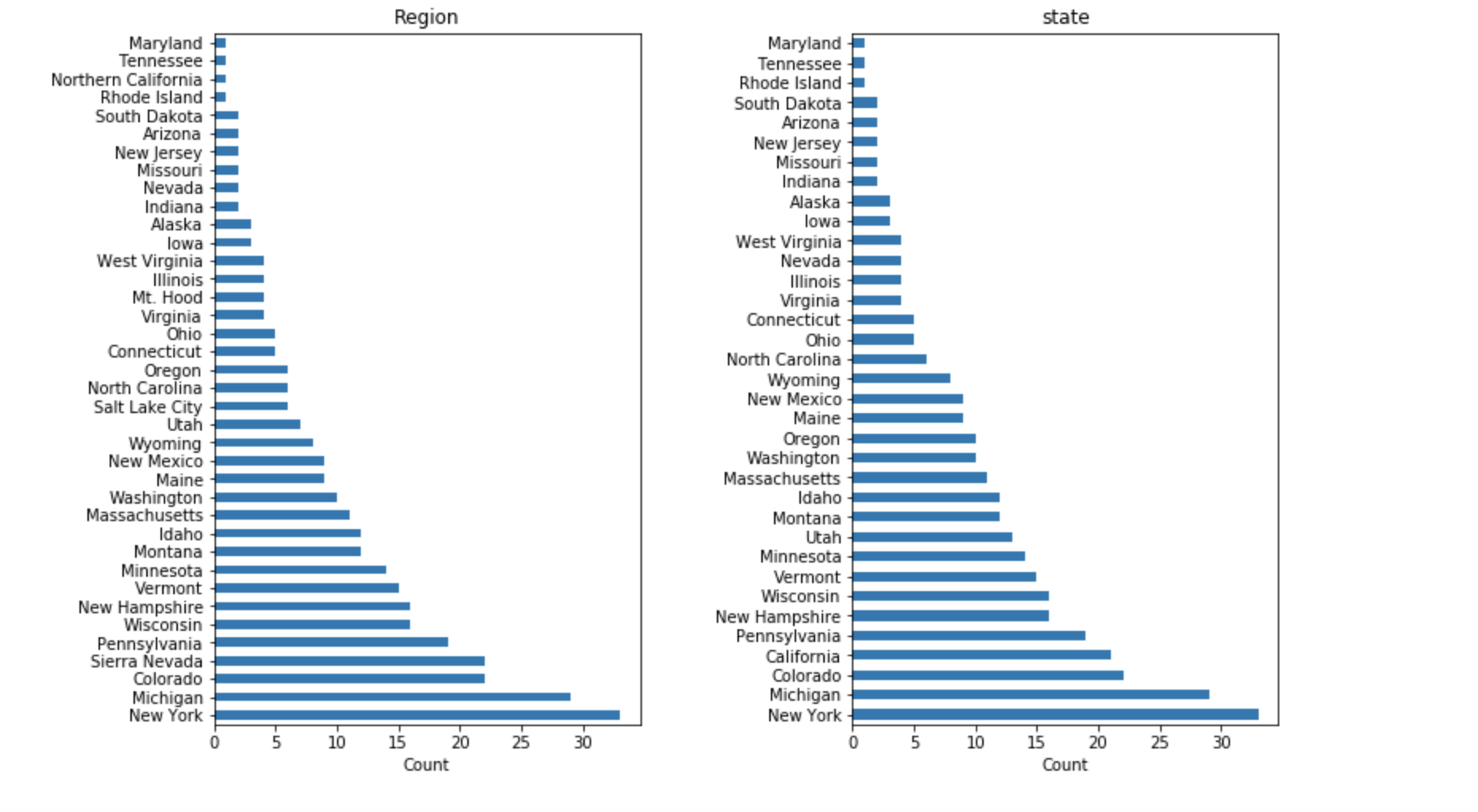
**Big Mountain Resort Report**

by Abel Mekuria

Big Mountain Resort located in Northern Montana. It has recently installed an additional chair lift to help increase the distribution of visitors across the mountain. This increases their operational cost by $1,540,000. However, the resort would like to keep the profit margin to 9.2 %.The resort can do this by increasing the cost of tickets to offset the operational cost. This will help the resort to keep their annual profit to 9.2 % despite an increase of operational cost. In order to keep the resort’s profit margin to 9.2 %, the resort needs to identify how much they can benefit from the additional chairs. The new installed chair will be the company more competitive than its competitors. Using the data from 330 resorts in the US and the Big Mount resort data, it will be important to investigate how increasing the price of the tickets will make them make more profits.

***Data Wrangling***

In this data wrangling section, we cleaned the data and checked if all the important values are available. In addition, we checked for missing values. For instance, the value in the fastEight column was missing. Then we identified Big Mountain resort and evaluated to see if all any values were missing. After that, we examined the categorical features as well. In addition, some outliers affected the output of the histogram plot such as skiable terrain and snowmaking. Based on the weekday and weekend price analysis, the weekday prices are missing a lot of values. Therefore, the weekend price selected for the target feature.

 Figure 1. Distribution of resorts by state and region

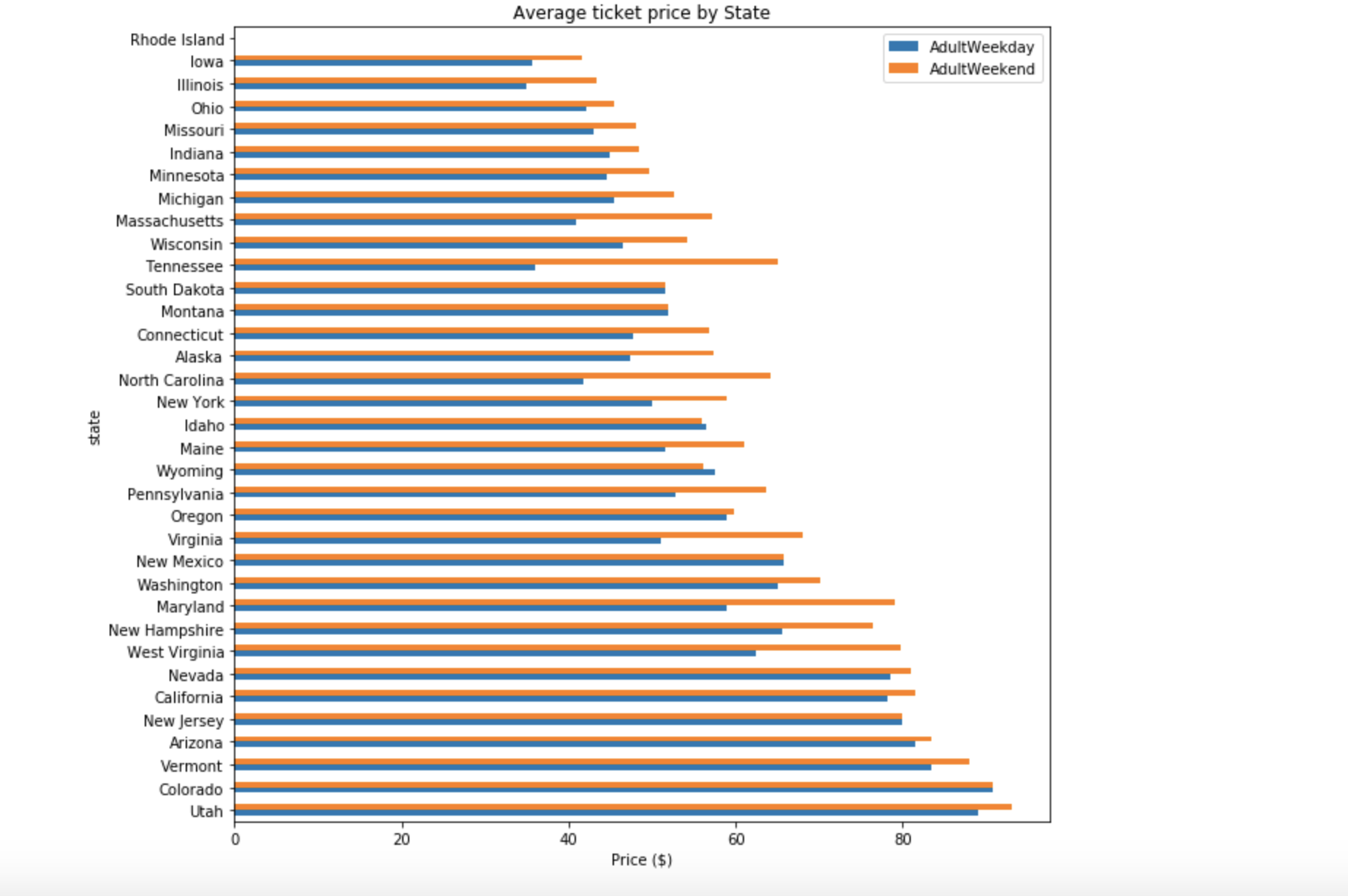


Figure 2. Average ticket price by state

***Exploratory Data Analysis***

In this section, which is exploratory data analysis, there were different numerical and categorical features. The state-wide data summarized and categorized by orders. This allows us to see what the state wide market looks like. In the subsection of EDA, visualization of high dimensional data by using functions such as scale and PCA transformation. This enables us to obtain the derived features. In order to identify a relationship between state and ticket price at a higher level, a correlations heatmap has been used. However, the heatmap can't provide a specific relationship between two variables. Therefore, scatterplots of numeric against ticket price.

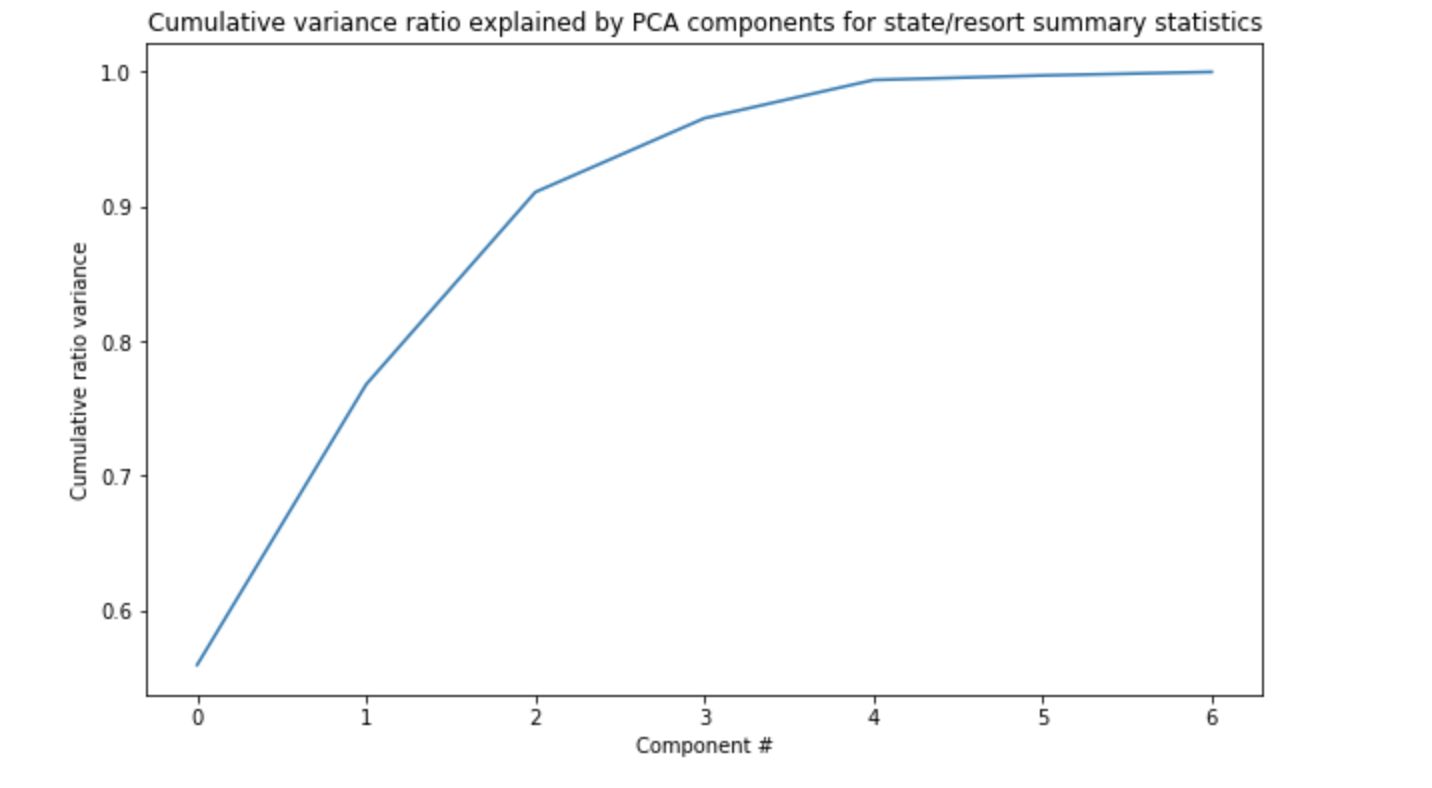


Figure 3. cumulative variance ratio estimated by principal components analysis (PCA)

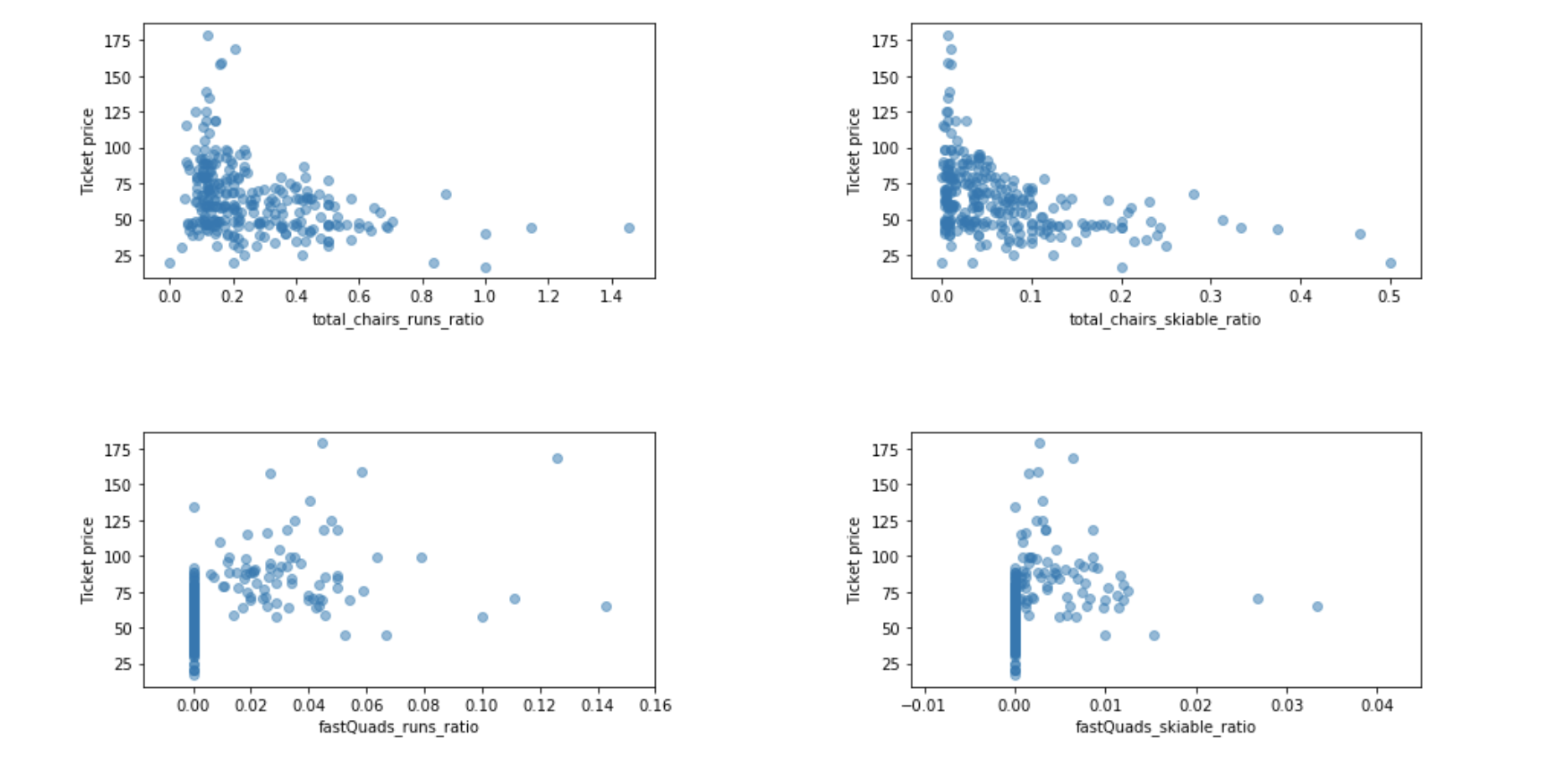
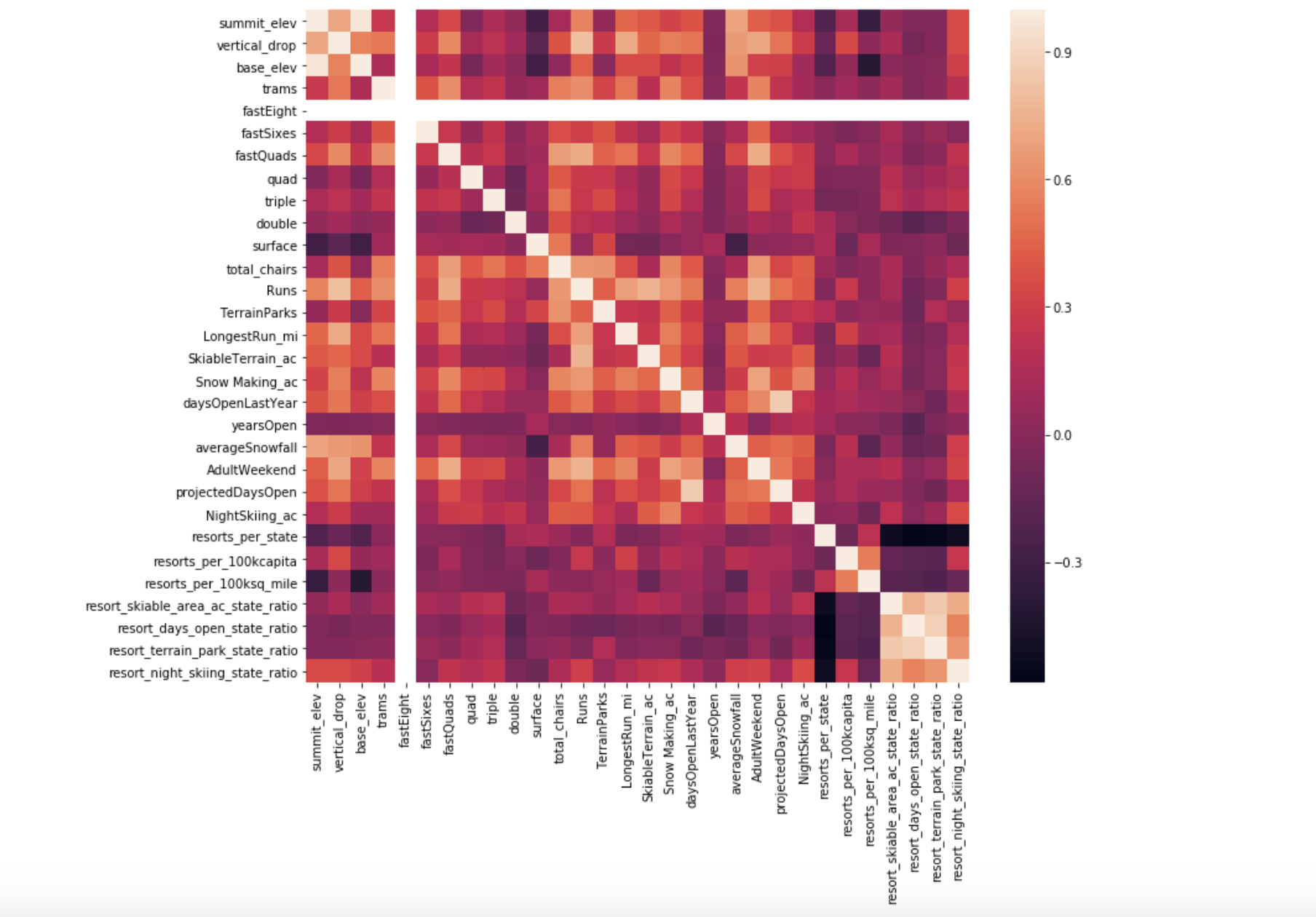
Figure 4. Scatterplots of numeric features against ticket price

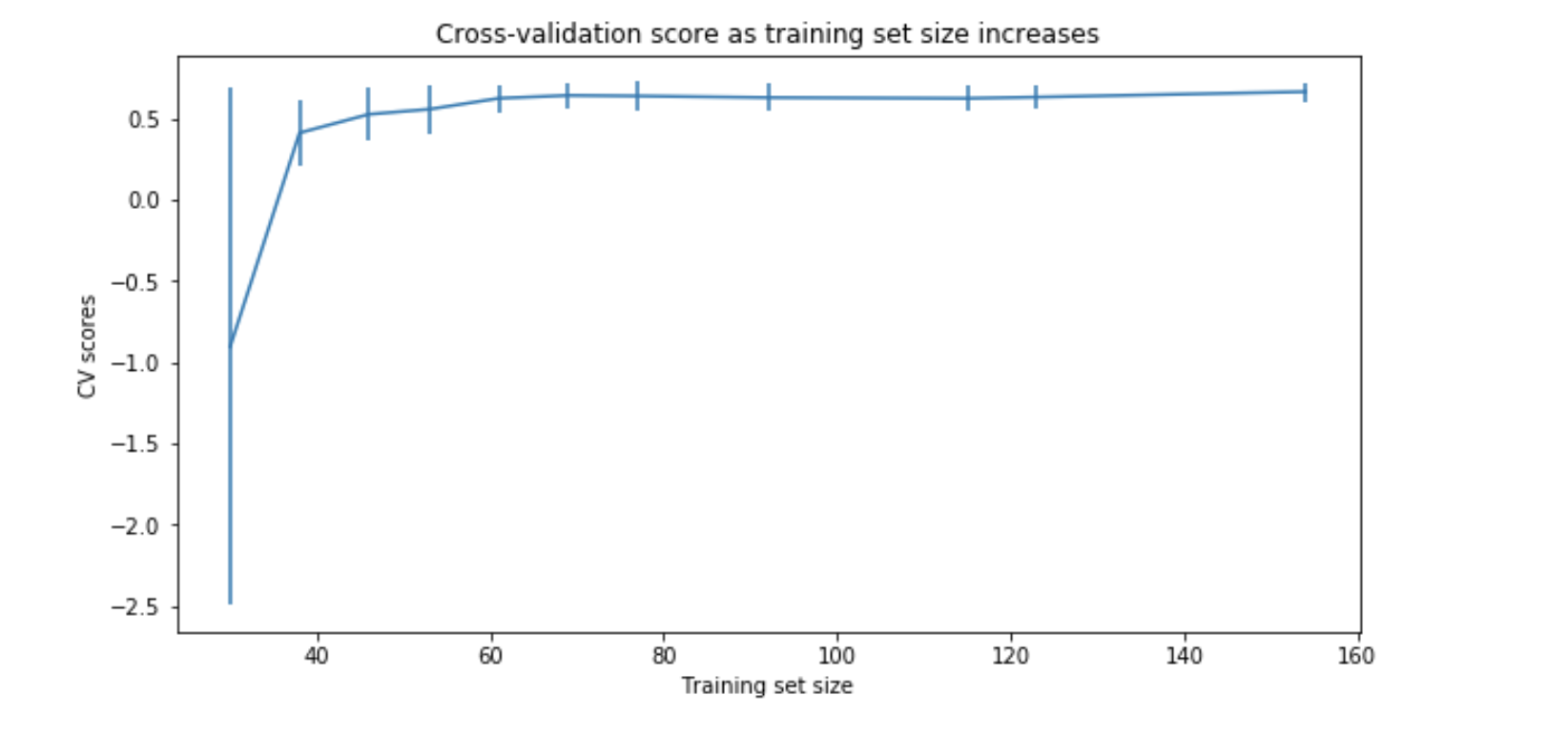
Figure 5. Feature correlation heatmap

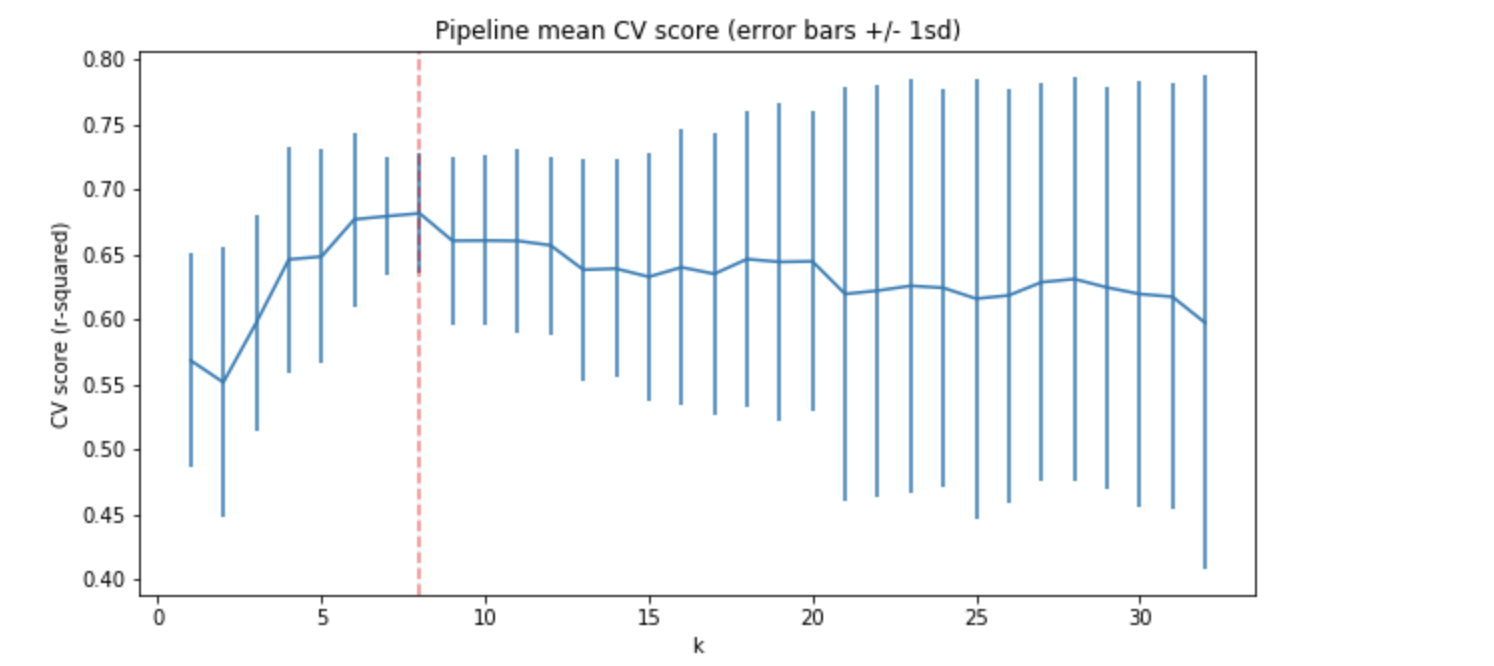
***Overview of Pre-processing and Training Data Development***

Data pre-processing implies the process of removing data that are out-of-value ranges and impossible combinations from your dataset. The pre-processing and training step involves the process of dividing the dataset into testing and training subsets.

Training and testing splits of data will help without letting a model learn anything about the test split that you have a somewhat independent assessment of how your model might perform. A test split is very useful as a final check on expected future performance. These are many ways of assessing how good one set of values agrees with another, which brings us to the subject of metrics. We used the following steps to assign missing values, scale the features, train a model, and calculate model performance. Besides the metrics, we have used pipelines, which are important and useful components of sklearn. Since the pipeline was overfitting, we refined the linear model by finding the k-value. A baseline of the average price used for simple estimation. Then the cross-validation was particularly important as it showed us how to choose and test the models based on the same dataset. The various test and cross-validation datasets also used to remove bias data and improve the model performance.

Another model we have used is the random forest regression model. A random forest regression model is an ensemble of randomized predicted values at point by the -th tree, where are independent random variables. Lastly, we compared the random forest regression model with the linear regression model. The random forest regression was chosen because a lower cross-validation means absolute error by almost $1. It also exhibits less variability.

Figure 6. Cross-validation score test 

****Figure 7. Hyperparameter search using GridsearchCV

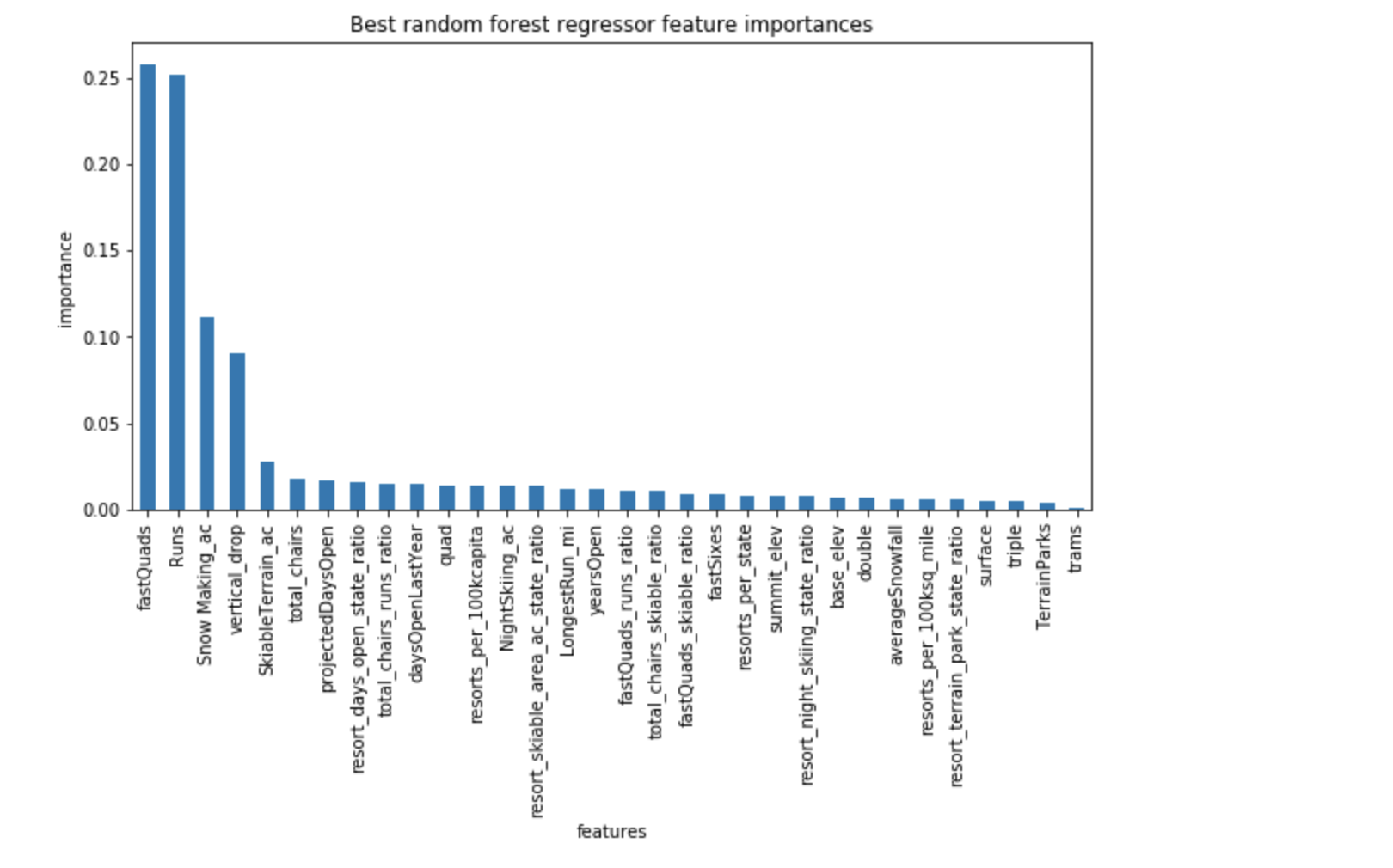


Figure 8.Random forest regression model

***Modeling and Summary***

In this section we used the model we developed to assess Big Mountain’s ideal ticket price and how would that change in various scenarios. The forest regressor suggests a ticket price of $94.22. Four modeling scenarios that have been applied are closing the least used runs, increasing the vertical drop by adding a run to a point 150 feet lower down, increasing the longest run by .2 miles, and increasing 2 acres of snow making. Scenario one shows that closing one run does not make a difference. However, if the number of closing runs increases, the drop will be higher. On the other hand, increasing the vertical drop by 150 feet and adding 2 acres of snow increases support for ticket price by $ 2. This could be expected to amount to $3,474,638 for over the season.

There are future works that need to be done to make the analysis more efficient. For instance, adding more data to improve the model. Besides, we should examine the revenue even more using different scenarios. In contrast, Big Mountain resort should apply the suggested model scenarios and use the outcome to make the model better.