## Data Analysis

Python's pandas library was used to load the dataset and perform analysis on the data. The original data consisted of 330 rows and 27 columns. The data was analyzed for data quality, missing values, types of features, relationship between features.

Some of the key steps performed in the data analysis and assessment are listed below:

- Identify the percentage of missing values for each column.
- Identify the distribution of Resorts by Region and State.
- Identify the distribution of Ticket price by State.
- Generate summary statistics for all the numeric features; plot distribution of features to check if these look plausible or skewed..
- Derive state-wide summary statistics for the data. Information on state-wide supply and demand can be useful in determining the pricing strategy.

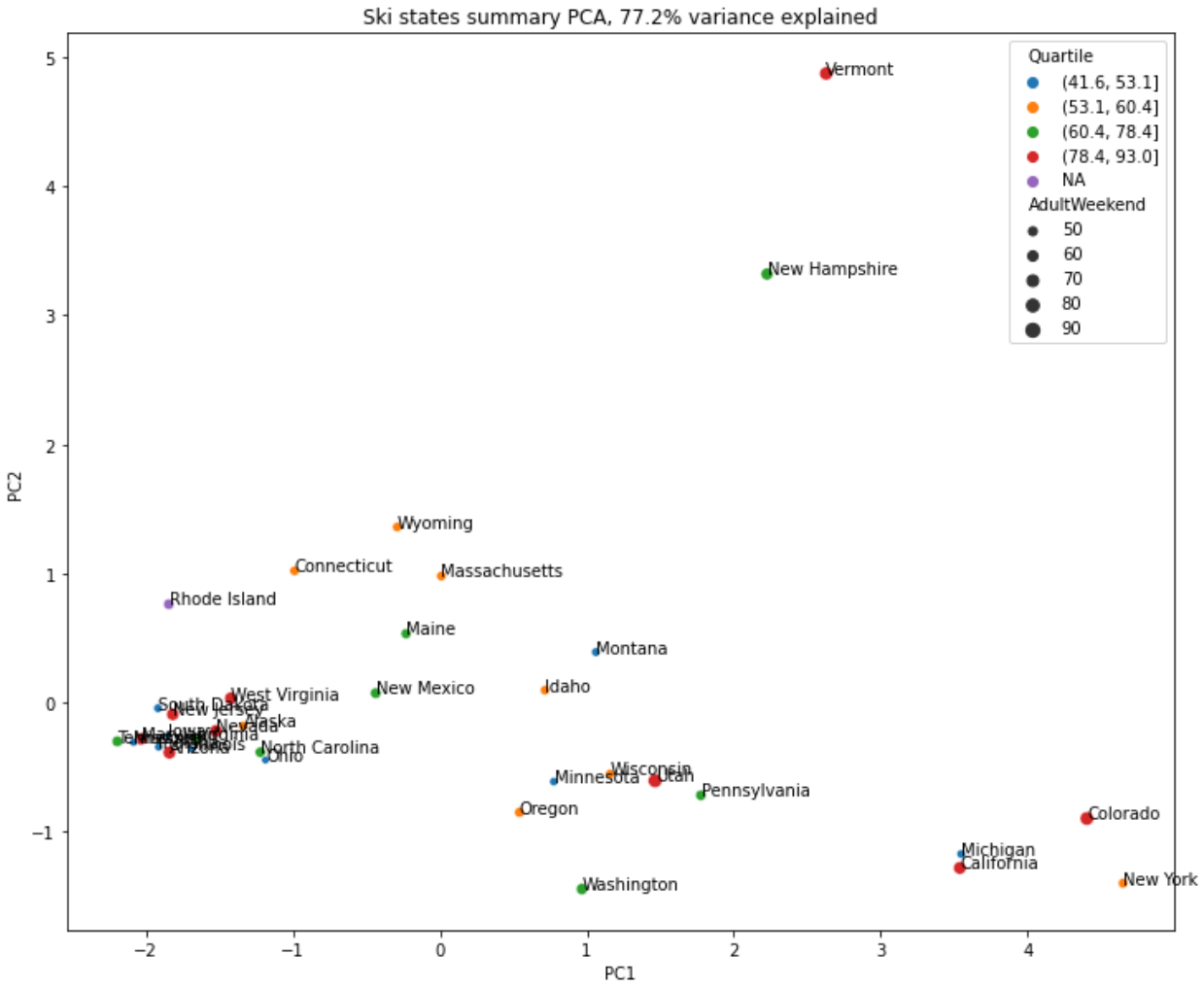
In addition, the population and area data for the US states was also analyzed to derive some additional insights such as:

- Number of resorts per state
- Total skiable area
- Total night skiing area
- Total days open

The exploration showed that there are big states which are not necessarily populous; there are states that host many resorts but other states host a large skiing area. And the states with the most total days skiing per season are not necessarily those with the most resorts.

To further evaluate if any of the States had any special meaning, the technique of principal component analysis (PCA) was used to find linear combinations of original features that are uncorrelated with one another. These derived features were used to view the data in a lower dimension.

The data was then plotted using a scatter plot, with PC1 along the x-axis, PC2 along the y-axis, size as the average ticket price, and hue as the quartile. The plot did not show any patterns, the quartiles were spread across the graph and there was no obvious pattern observed related to the price.



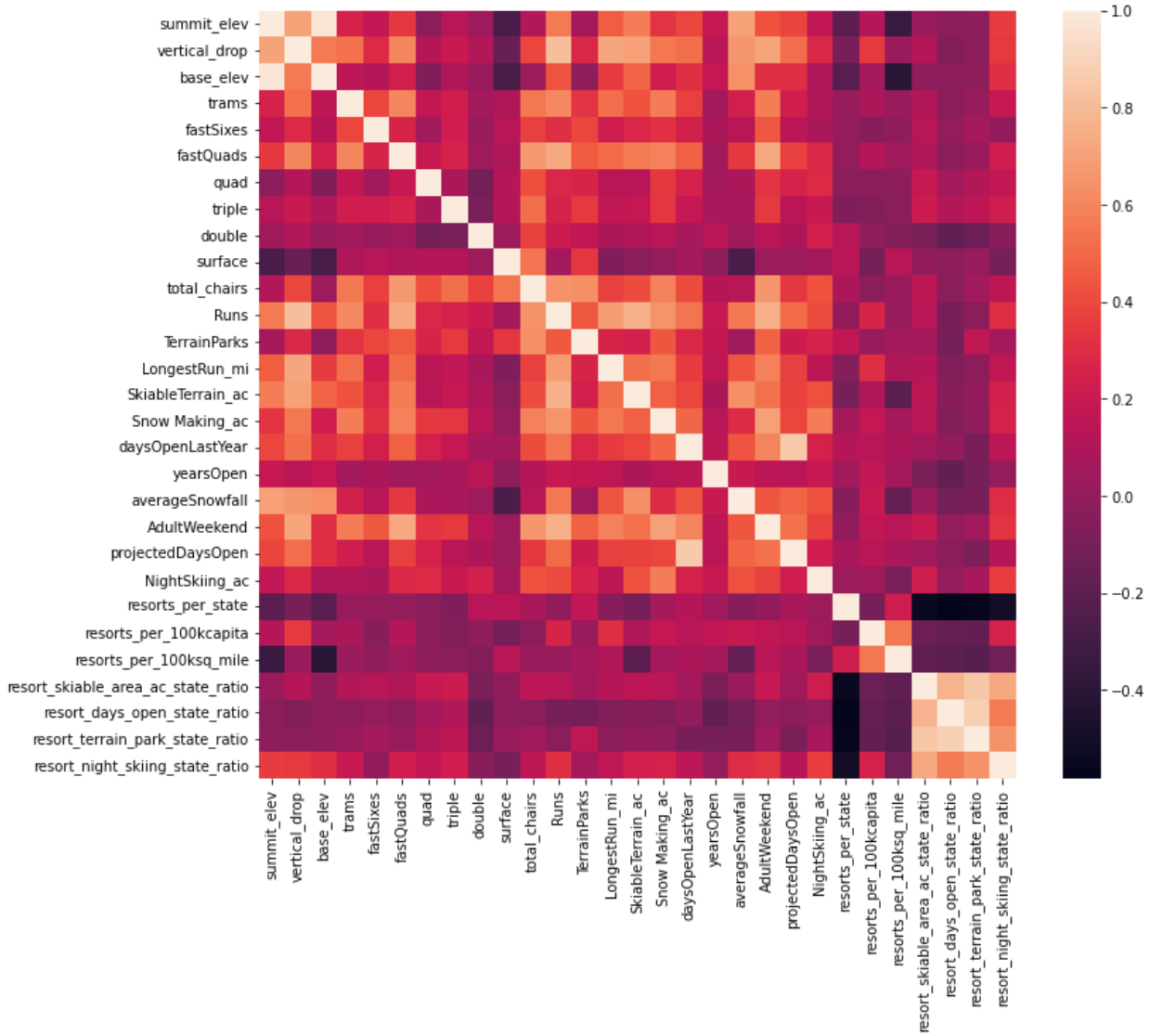

As the state-wide summary is not showing any patterns with the price, we can consider the states together without treating any one state specially.

The state summary features were merged into the ski resort data. Some additional features were created to derive useful insights:

- ratio of resort skiable area to total state skiable area
- ratio of resort days open to total state days open
- ratio of resort terrain park count to total state terrain park count
- ratio of resort night skiing area to total state night skiing area

A correlation heat map was plotted, the below observations were made from this heat map:

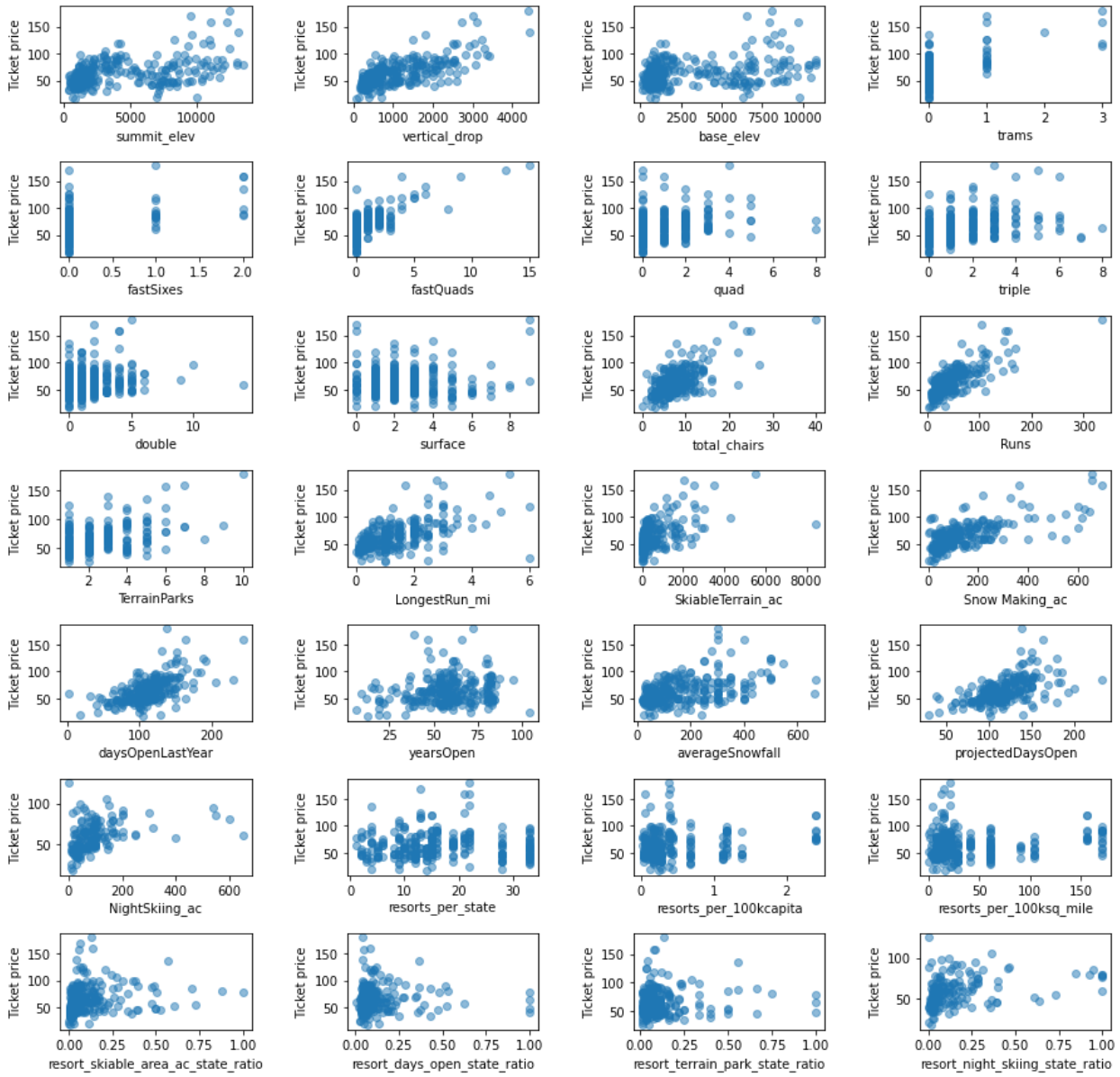
- The target feature, AdultWeekend (ticket price) seemed well correlated with the features:
    - fasQuads, Runs, Snow Making_ac, total_chairs, vertical drop
    - Of the new features, resort_night_skiing_state_ratio, seems the most correlated with ticket price.




Scatter plots of the numeric features against the ticket price were also plotted. These showed some correlations that were previously observed with the heat map. The scatter plot also showed an interesting observation about the ‘resorts_per_100kcapita’ feature: when the value of this feature is low, there is quite a variability in ticket price, and it can go quite high.

A few additional features representing chairs to runs ratio('total_chairs_runs_ratio'), chairs to skiable area ratio('total_chairs_skiable_ratio'), fastquads to run ratio('fastQuads_runs_ratio'), fastquads to skiable area ratio('fastQuads_skiable_ratio') were also computed.

- The scatterplot of the 'total_chairs_runs_ratio' against the ticket price shows that if there are fewer chairs to operate then the ticket price is high, but a side effect of this could be that with fewer chairs only a limited number of visitors can be served.
- The scatterplot of the 'fastQuads_skiable_ratio' against the ticket price shows that if resort covers a wide area then getting a small number of fast quads may be beneficial to ticket price.




## Modeling

The preprocessed dataset was used to build a pricing model for the Big Mountain Resort, it was first ensured that the Big Mountain Resort data was removed from this data set.

The data set was split into a Training data set and Test set so that the model can be developed using the training data set and trained on the test data.

First, the strategy based on the average price was used as a baseline model for comparison. The results obtained were as follows:

- The mean absolute error (MAE) on the train data set was 17.92 and the MAE on the test data set was computed to be 19.14

Next, the LinearRegression model was used to fit the data and make predictions. The technique of cross-validation was used to partition the training set into 'm' folds, train the model on m-1 of these folds, and calculate performance on the fold not used in training. This procedure was repeated m times, with a different fold held back each time.

Cross-validation allowed us to build m models on m sets of data with m estimates without having to touch the test set.

The GridSearchCV function, along with the pipeline object was used to implement the technique of cross-validation, imputing missing values, scaling the features, and identifying the best features that represent the data. On evaluating the model the value of k for the number of best features was identified as 8, and the 8 best features determined by the model were as follows:

- vertical_drop
- Snow Making_ac
- total_chairs
- fastQuads
- Runs
- LongestRun_mi
- trams
- SkiableTerrain_ac

Another model that was evaluated was the RandomForestRegressor. The model was evaluated to identify a best estimate for the number of trees, and it was tested with both mean and median for imputing missing values, and with and without the StandardScaler for scaling the features. On evaluating the model using the GridSearchCV function it was determined that imputing with the median helps, but scaling the features does not help.

The top four features identified by the RandomForestRegressor model were as follows:

- fastQuads
- Runs
- Snow Making_ac
- vertical_drop

For a final selection, the results of the LinearRegression were compared with the Random Regressor.

- The LinearRegresson had a cross-validation mean absolute error (MAE) of 10.45 and, and the the MAE on the test data set was 11.79
- The RamdomRegressor had a cross-validation mean absolute error(MAE) of 9.64 and the MAE on the test data set was 9.54

***Thus the RamdomRegressor model performed better than the LinearRegresson model.***

## Recommendations based on the optimal model

The RandForestRegressor model was used to gain insights into determining the ideal ticket price for the Big Mountain resorts.

The model was refit to all the available data, excluding the Big Mountain data. The Big Mountain data was excluded as we want to predict Big Mountain's ticket price based on data from all the other resorts, and we don't want Big Mountain's current ticket price to bias the prediction.

**On fitting the model to all the available data (excluding Big Mountain), the model's prediction for the ticket price was computed to be $95.87. (**Note that Big Mountain currently charges $81.00 per ticket.)

For the features that were identified as significant features for the model , a distribution of each of these features was plotted for all the resorts.

On analyzing the position of the Big Mountain resort in these distributions, it was observed that the resort compares well on most of these features, and in some cases is also high up in the standings compared to the other resorts.

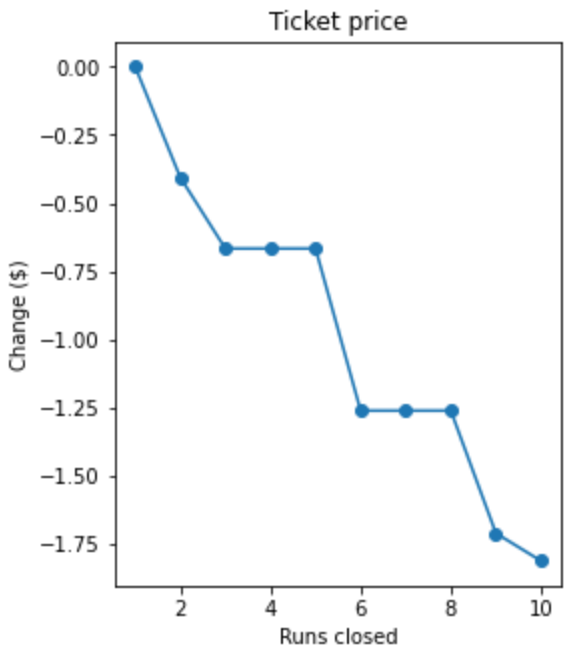
***This analysis suggests that given all the facilities the Big Mountain resort provides, the resort has been most likely undercharging on the ticket price.***

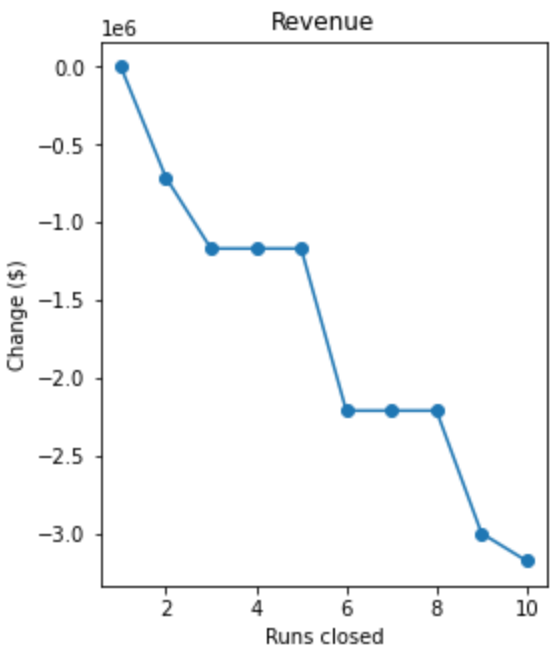
The Big Mountain resort management has been considering the below measures to help with cost cutting or increase of revenue:

1. Permanently closing down up to 10 of the least used runs.
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.

Each of these scenarios were run against the model to identify the impact on ticket price and revenue. Below were the observations:

- Closing down 1 run makes no difference, closing 2 and 3 runs successively will result in a drop in ticket price, closing 4 or 5 runs has the same effect as closing 3 runs. Closing 6 runs or more results in further drop in ticket price and overall revenue.




- If the resort decides on increasing the vertical drop by 150 feet, and installs an additional char lift, this will **increase the ticket price by $1.99** and the **revenue will increase by $3,47,4368.**
- In addition to the above if the resort also adds 2 acres of snow making, this will result in a **very small increase as compared to point # 2 above.**
- If the resort decides to increase its longest run by 0.2 miles and guarantee its snow coverage by adding 4 acres of snow making, **this will make no difference** to the ticket price whatsoever.