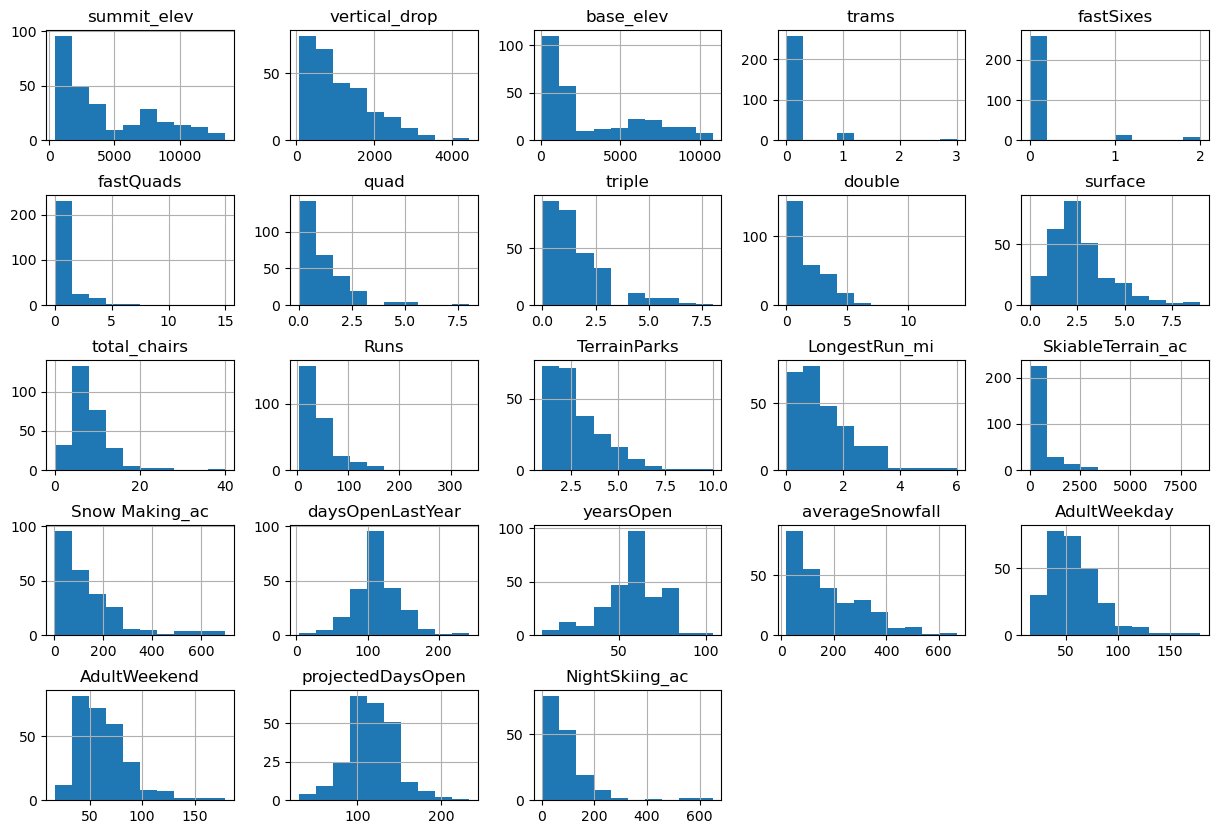
The purpose of the data science project is to determine a good pricing model for ski resort tickets in Big Mountain Ski Resort’s market segment. By analyzing the facilities that Big Mountain Ski Resort have, we seek to build a predictive model for ticket price based on the number and quality of facilities. This model will be used to provide guidance Big Mountain future investment plans.

**Data Wrangling**

* The Original data has 330 observations (ski resorts) each with 27 columns (resort data features)
* The Clean data has 277 observations (ski resorts) each with 25 column (resort data features)
* In particular, two columns from the dataset have been removed. 'fastEight' is removed because it has lots of missing values and null values. Lack of information. 'AdultWeekday' has also been removed because it has the same value as 'AdultWeekend' for the state of interest - Montana. 'AdultWeekday' also has more missing data, hence removed in favor of keeping 'AdultWeekend' ticket prices.
* Observations with null values for ticket prices have been removed during the data cleaning process. Certain erroneous data points in 'YearsOpen' and 'Snow\_Making\_Ac' have also been adjusted.
* The cleaned data presents an opportunity to explore and perform analysis to determine features affect ticket prices. Some columns of interest would be 'SkiableTerrain\_ac', 'Snow Making\_ac', 'Runs', 'LongestRun\_mi' vs 'AdultWeekend' prices. A good analysis would be to look at the correlation betweeen the two data series.
* The following subplot shows the distributions of feature values after data wrangling, with a few noticeable skewed distributions.



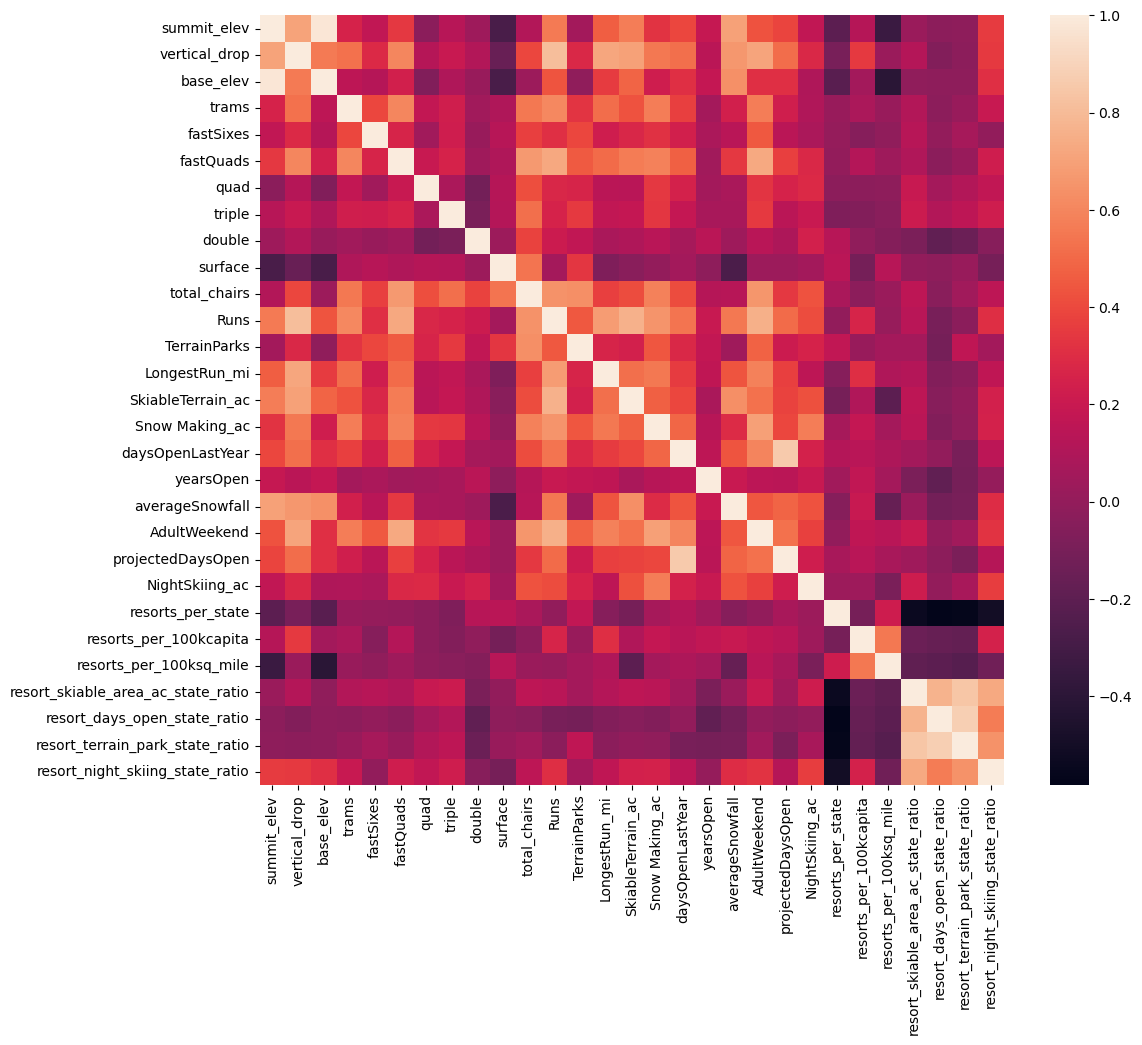
**Exploratory data Analysis**

The data contains 22 numerical data fields(11 integer type, 11 float type) and 3 categorical data fields.

By looking at the different numerical features of the state, we cannot come to a definitive conclusion about which specific feature of the state data affects ticket prices. However, it makes intuitive sense that the competitive landscape of ski resorts within states will affect ticket prices. This prompts the creation of new features which standardizes across states for subsequent modeling. The 2 new data columns are resorts\_per\_100kcapita and resorts\_per\_100ksq\_mile.

I should be wary of the correlation between features when performing feature selection for modeling. A great tool to use is PCA. This technique will find linear combinations of the original features that are uncorrelated with one another and order them by the amount of variance they explain. Using a heatmap to look at the target feature (AdultWeekend prices) vs all other numerical features is a great way to spot the correlations. Due to the ability to breakdown how specific or combination of state features affect ticket prices, we could actually treat all states equally and build a pricing model that considers all states together.

Feature Correlation Map below:

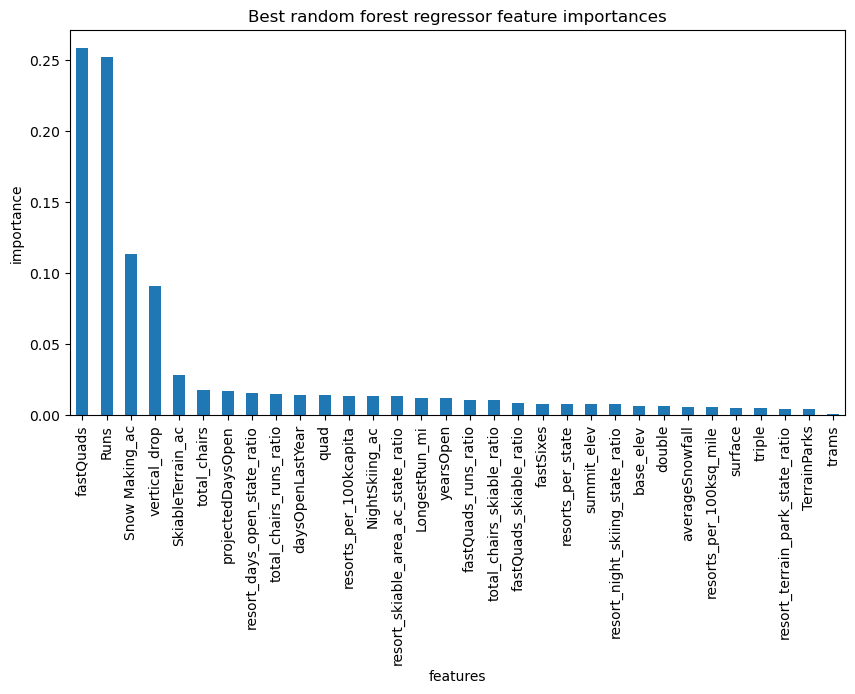


**Model Preprocessing with feature engineering**

The workbook starts by introducing some metrics used to measure modelling performance such as Mean Absolute Error, Mean Squared error, and R-squared. These values are then calculated using the 'DummyRegressor' non-model to provide a baseline. I learned to split data into training and testing sets. The performance is then improved using a linear model from a Mean Absolute Error of 19 to 9. The SKlearn Pipeline is then introduced and the linear model is improved by determining the best number for 'selectKbest' feature. To avoid the use of arbitrary numbers, the cross-validation technique is introduced and used upon our data set to review k=8 is a good value that offers the lowest variance. The improved linear model offers some insight to the features that affect the ticket prices the most - with the summary stats below:

vertical\_drop 10.767857 Snow Making\_ac 6.290074 total\_chairs 5.794156 fastQuads 5.745626 Runs 5.370555 LongestRun\_mi 0.181814 trams -4.142024 SkiableTerrain\_ac -5.249780

The random forest regressor was later tried, using cross validation to fit the test data. The model returns the most dominant 4 features to be fastQuads, Runs, Snow Making\_ac, vertical\_drop which is consistent with the linear model, in a different order. The random forest regressor has a lower cross-validation mean absolute error and exhibits less variability. Since both models have consistent results, the random forest regressor should be selected going forward due to the lower MAE.



**Pricing recommendation**

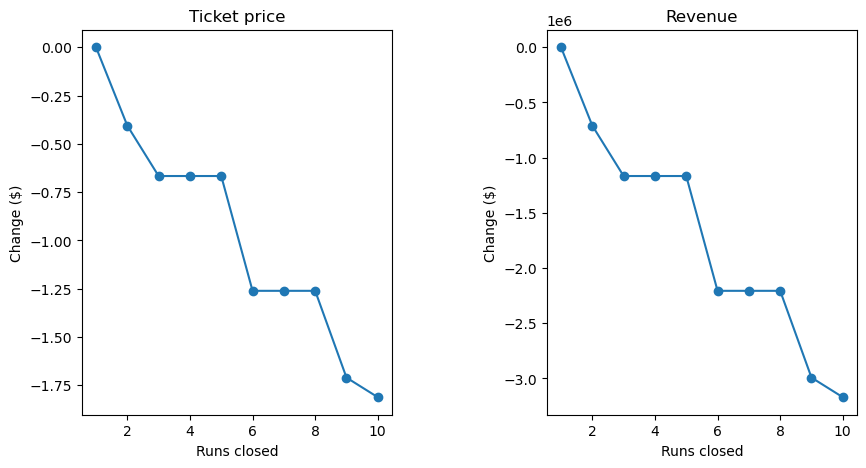
Big Mountain currently charges $81. Assuming all other resorts are fairly priced, the model predicts Big Mountain should be charging $95.87 with an error of $10.39, showing room for a ticket price increase. I would use the model to suggest to business leadership that Big Mountain is undercharging for it's facilities relative to the market since Big Mountain ranks very high in many of the key features such as vertical\_drop, Snow Making\_ac, total\_chairs, fastQuads, Runs, LongestRun\_mi, SkiableTerrain\_ac.

**Recommendation Scenarios**

In the scenario of adding a run, increasing drop by 150ft and installing a new chair lift, Big Mountain should be able to increase the ticket prices by $1.99. This will create additional revenue of $3474638 per season. If the cost of a new chair lift is lower than the revenue, then this may be a good upfront investment for the business.

Other scenarios such as increasing the snow making\_ac and increasing the longest run by 0.2 mile to boast 3.5 miles length, requiring an additional snow making coverage of 4 acres actually makes no difference to the predicted ticket price. These scenarios should not be recommended based on the model.

Regarding run closures, closing 1 run makes no difference. Closing 2 and 3 runs affect the prices negatively by almost $0.75. However, there is no difference between closing 3, 4 or 5 runs. Closing 6, 7 or 8 runs causes a large drop in ticket price of $1.25. Closing 9 and 10 runs drops the ticket price by almost $1.75.



**Further work**

The price data in our data set assumes all the other resorts are priced fairly. However, in reality, ticket prices can be very noisy and subjective - some will be overpriced and some underpriced. Not knowing exactly that, our model only shows that Big Mountain Resort is currently underpricng relative to the other ski resorts. Perhaps, other resorts invest more in marketing to justify an overpriced ticket, which is not shown since we lack any costs data. It would be useful to look at data of total profits rather than just revenue.

It would be great to build a front-end UI for the predict\_increase function. Business leaders or future business analysts would be able to simply use the front-end UI and enter their inputs and the model will output the ticket price and revenue changes.

**Conclusion**

Our model suggests that Big Mountain should be increasing its ticket price given its current facilities. Big Mountain Ski Resort ranks highly in many key features that correlate with ticket prices which gives it strong pricing power. To improve revenue further, management could increase the vertical drop by adding a run to a point 150 feet lower down, install an additional chair lift to bring skiers back up, without additional snow making coverage. Our model predicts that this increases ticket prices by $1.99. If the costs associated with this operation is less than the revenue it brings, then this should theoretically increase Big Mountain’s profits.