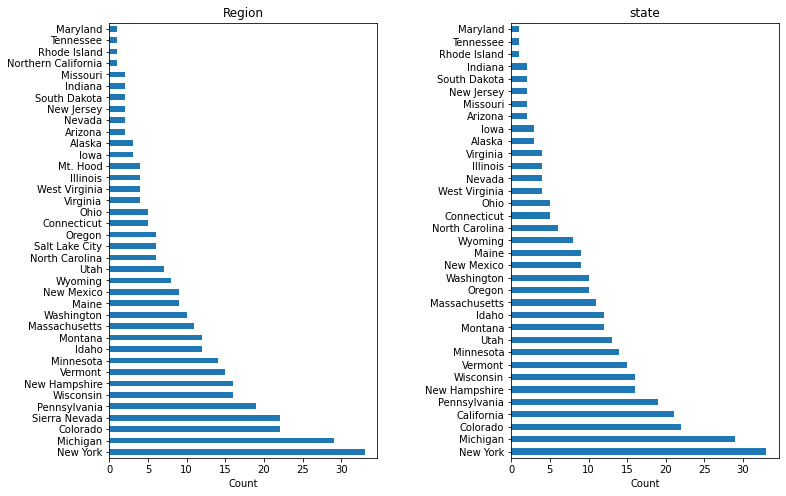
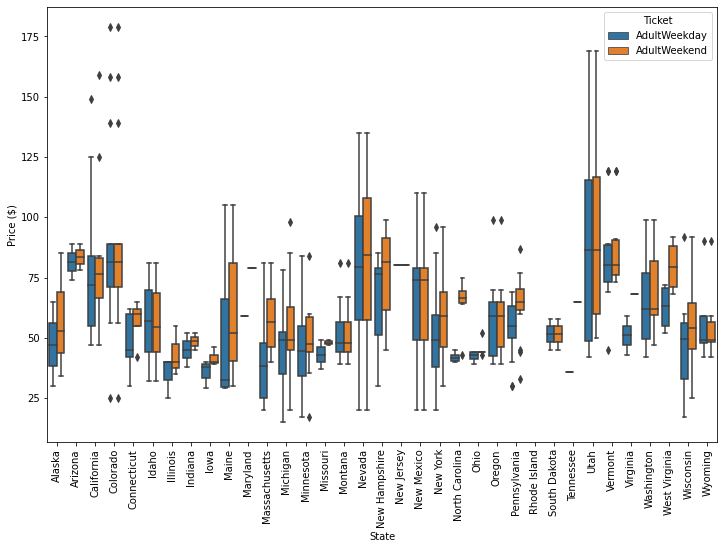
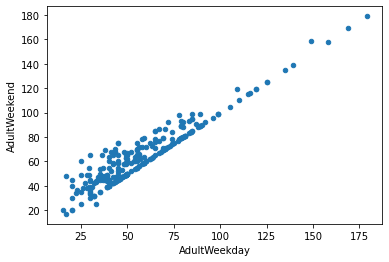
**Introduction:**

Big Mountain Resort offers spectacular views of Glacier National Park and Flathead National Forest, with access to 105 trails. Every year about 350,000 people ski or snowboard at Big Mountain. Currently, Big Mountain Resort bases their ticket pricing just on the market average and this does not reflect the impact of its facilities to the pricing. I am tasked to provide a ticketing strategy to Big Mountain Resort considering it’s facilities and their impact on revenue which will provide insight for further investment strategies to increase the revenue.

When I loaded the dataset I observed that it contains 330 entries with 27 columns. I checked the entry for the Big Mountain Resort and saw that it has no missing values. The distribution of resorts by state revealed that NY has the most of the resorts and Montana is 13th. What if resorts in NY are charging premium prices? This could be a problem since if I use the whole data this could skew my model’s accuracy towards resorts in New York. If I filter resorts in Montana, it would slash the volume of the data that is available to me. One way to check this is to look at ticket price distribution among the states. Except the very expensive ticket prices in CA,CO,UT most prices lie in a broad band between $25-$100.





I filtered the entries that are missing the ticket prices, i.e. weekday and weekend. It turns out that %14 of the entries are missing both ticket prices so I dropped those entries. When I plotted the weekend and weekday prices I realized that they are very close to each other. Thus, I decided to drop the weekday column and removed the entries that have missing values in the weekend column. In addition I dropped fast\_Eight column since >50% of the entries have missing values in this column. With the additional changes in the data set, I am left with a cleaned organized data with 227 entries and 25 columns.

I created a state summary dataset that contains information as resorts per state, resorts per 100k capita, etc. I applied standard scaler to this dataset and applied dimensionality reduction method to observe the amount of variance explained with number of components.



Based on this result we can see that more than 75% of the variance is explained by two components. I plotted these two components as a scatterplot and labeled each point with state names and also included ticket price information as quartiles. There are no clear cluster among states thus I will treat all states equally for price prediction.

When developing a model I did the following steps: imputation for the missing values, scaling, training the model. I used mean or median for imputation, standard scaler for scaler (mean:0, standard deviation:1) and linear regression as model. I realized that mean or median imputation does not make a big difference. To facilitate this process I created a pipeline. Since my dataset contains a number of features, I am worried that it can cause overfitting. Thus I used SelectKBest function to select the most important features. Using GridSearchCV, I identified 8 features explains the variance best.

At the next step, I tested Random Forest Regressor to predict the ticket prices. I created a pipeline that contains imputer, scaler and random forest model. I assessed the model performance with cross validation and performed GridSearchCV for hyper parameter tuning. Hyperparameter tuning revealed that random forest does not need scaling, and median imputation is better than mean. Best value for number of estimators is identified as 69. When I compared the two models, random forest regressor has better performance than the linear regression model.

Most important features identified by both models are:

* fastQuads
* Runs
* Snow Making\_ac
* vertical\_drop