Guided Capstone Report

Introduction:

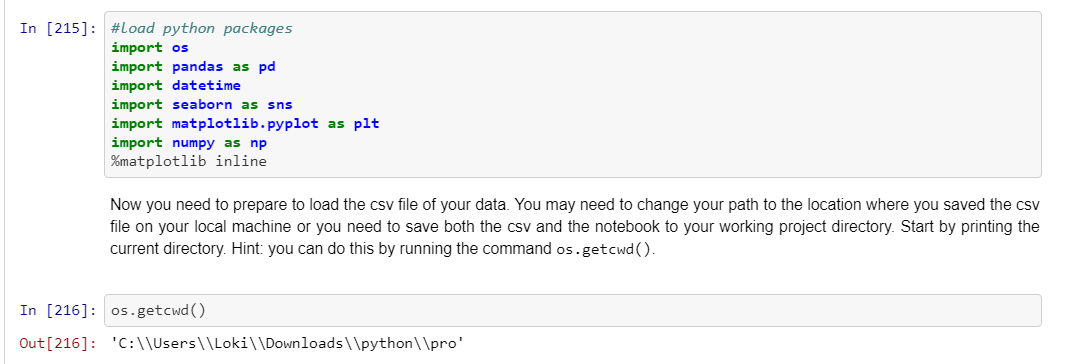
In this Guided Capstone project, we had worked on data collected from a database manager in a single CSV format. We had data that contains information of about 330 resorts in the United States of America, and we have the same columns for all the resorts in the dataset. Below we can see the description for each column.



1.Data Wrangling:

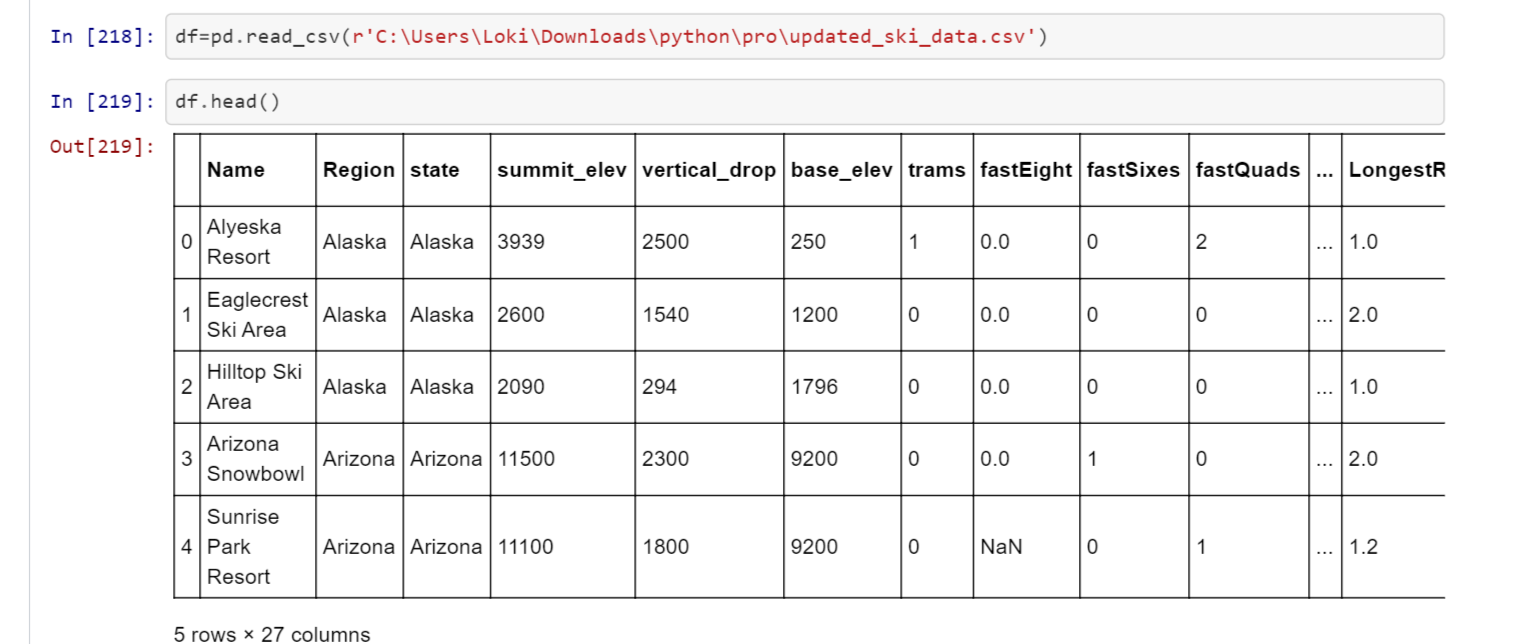
1.1 Data Collection:

This is the very first step in our project, with the help of packages we will be able to load our data which has to be utilized.



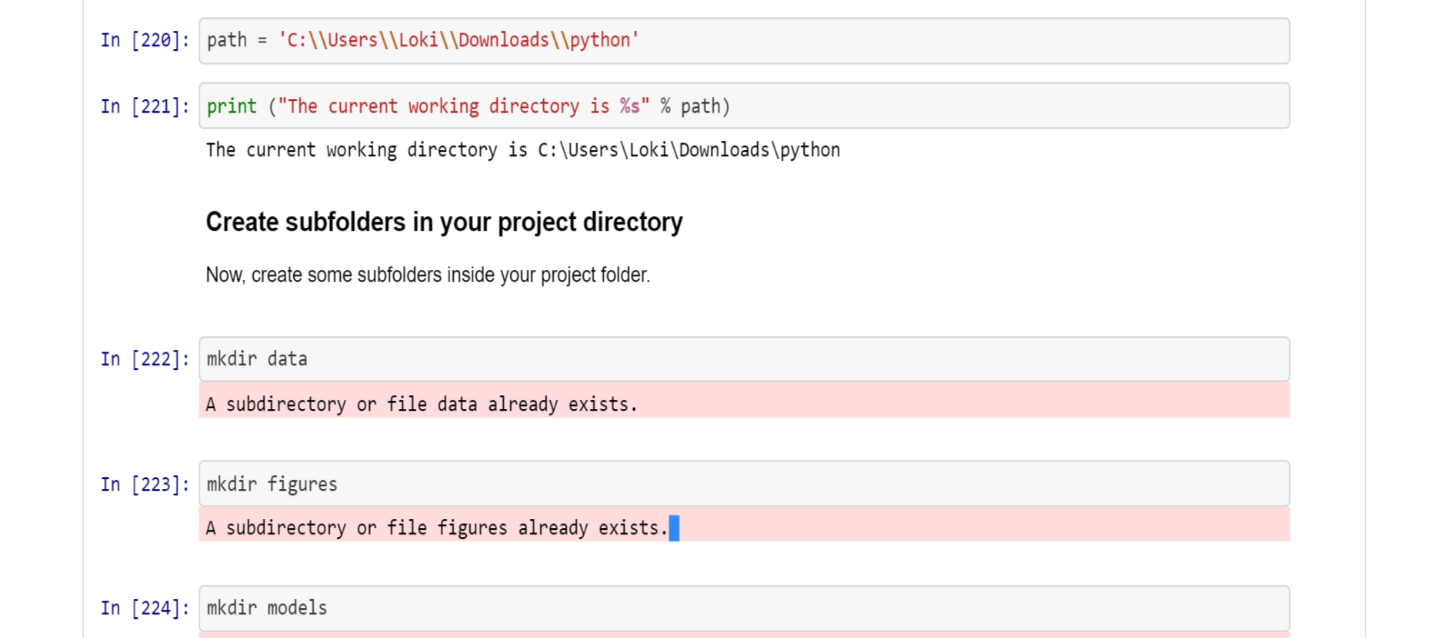
1.2 Data loading:

We will be loading our data with the help of pandas and visualizing it the first five rows of data.



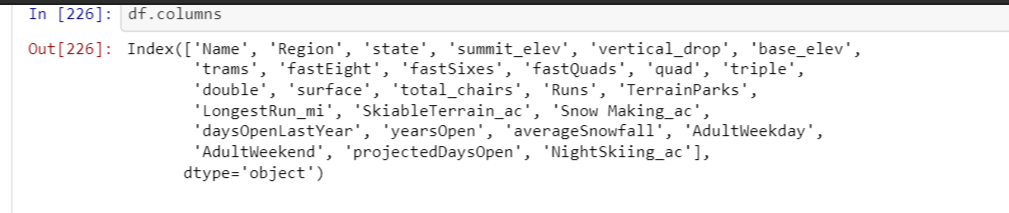
1.3 Data Organization:

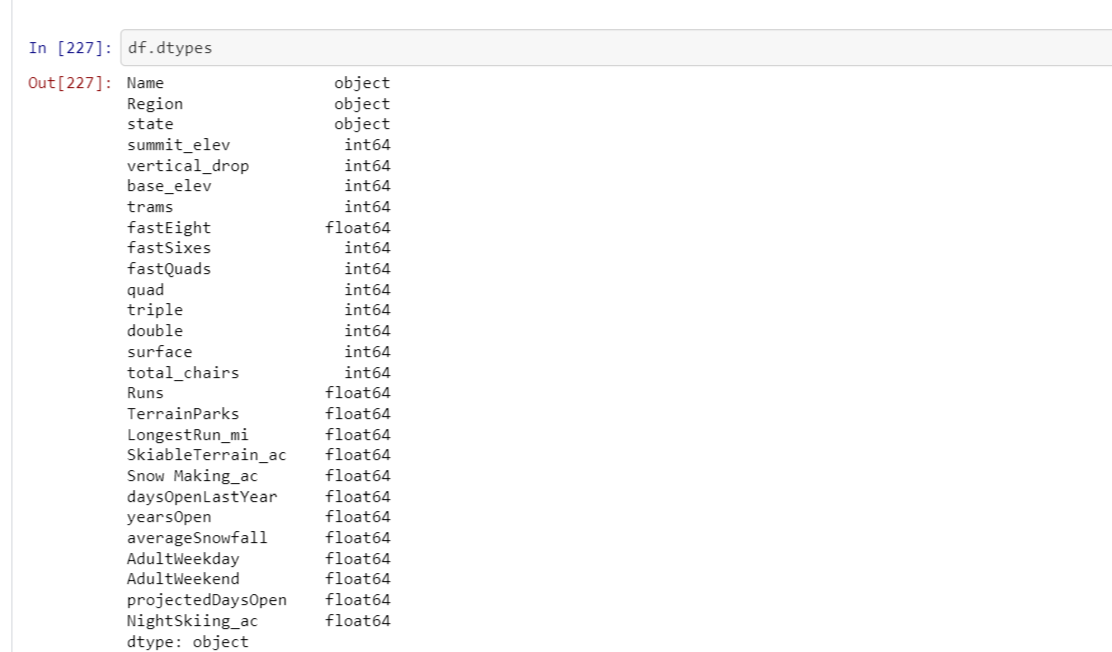
In data organization we will creating a folder with some subfolders to segregate our work.

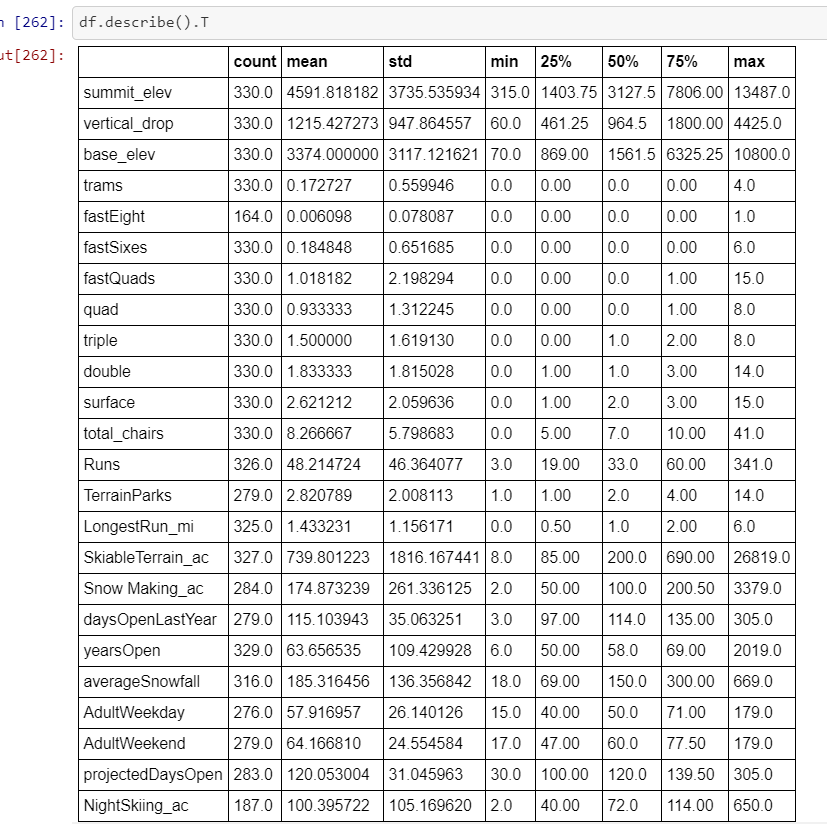


1.4 Data Definition:

In this step we will be visualizing the column names, data types, description of columns, percentage of missing values and range of values.

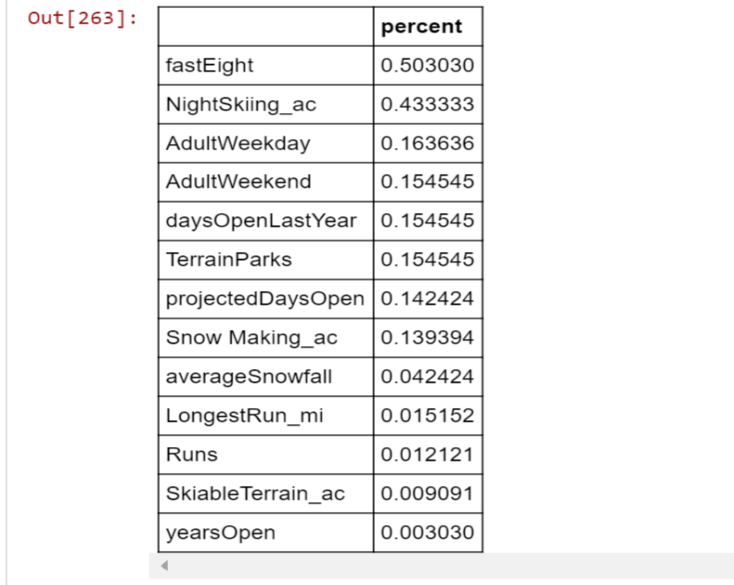




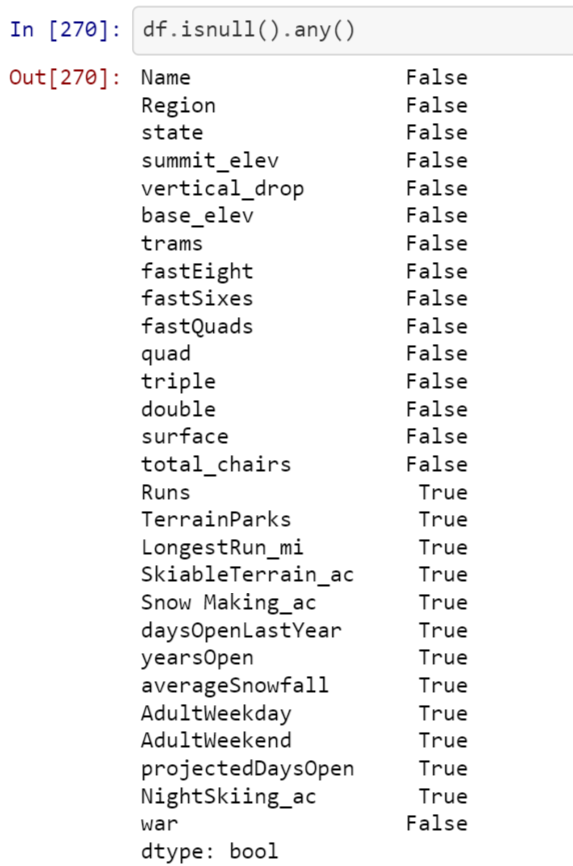


1.5 Data Cleaning:

In this step we will try to find the missing values and fill with them if we find any patterns or fill them with mean, median, forward and backward fill.



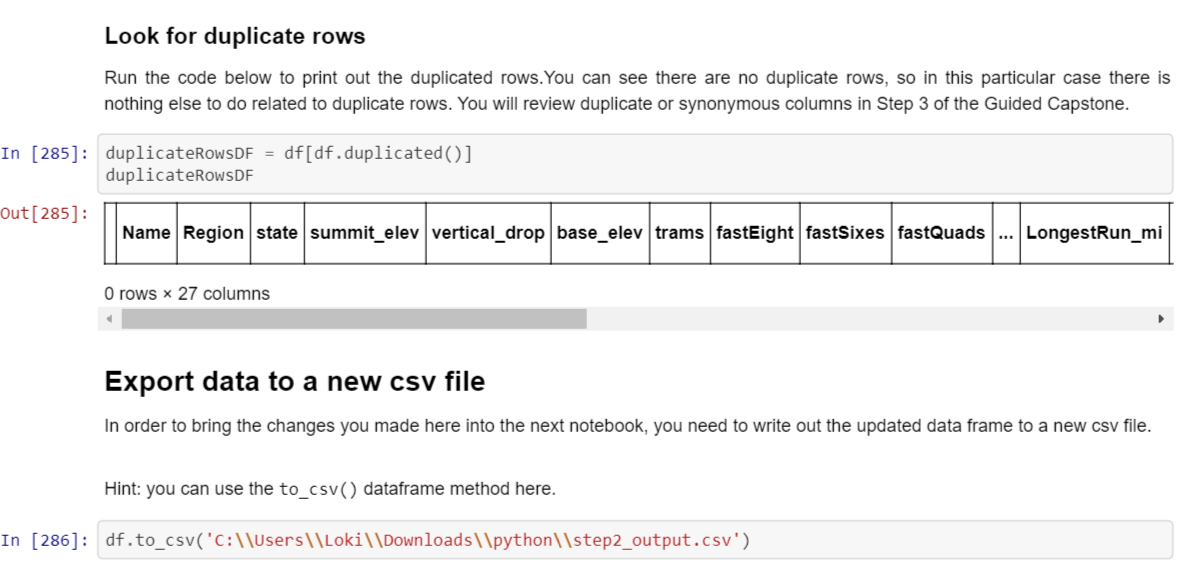
After filling the fastEight columns with the help of sum of other chairs columns and calculation the difference between them and total\_chairs. We had filled the values in fastEight.



The rest of the missing values are filled with the help mean which is one of the best way to fill the missing values.



After filling those values we will try to look for duplicate rows and later save in a CSV file.

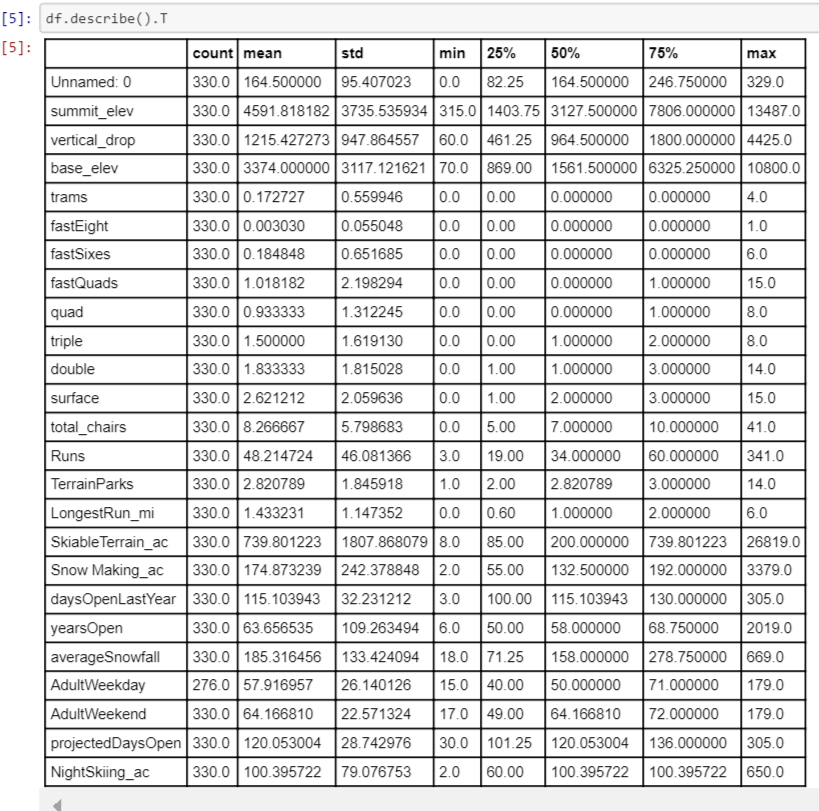


2. **Exploratory Data Analysis:**

**In the exploratory data analysis we will try to build a data profile tables and plots, eliminate the highly correlated columns and eliminate outliers and anamolies.**

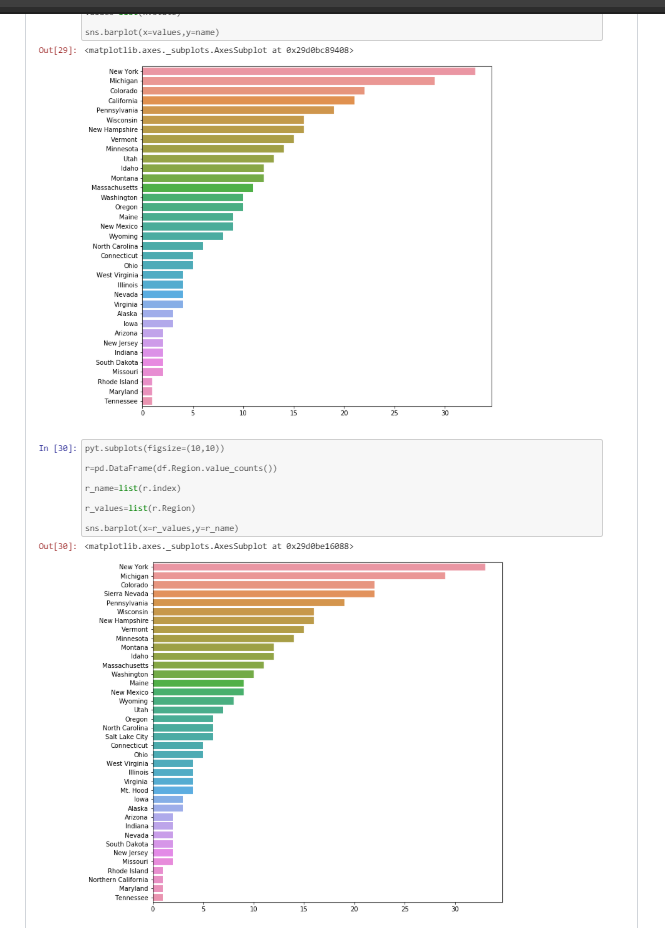
**2.1 Build data profile tables and plots:**

**It is the very first step in this process, we will be able to visualize the data with the help of describe method.**



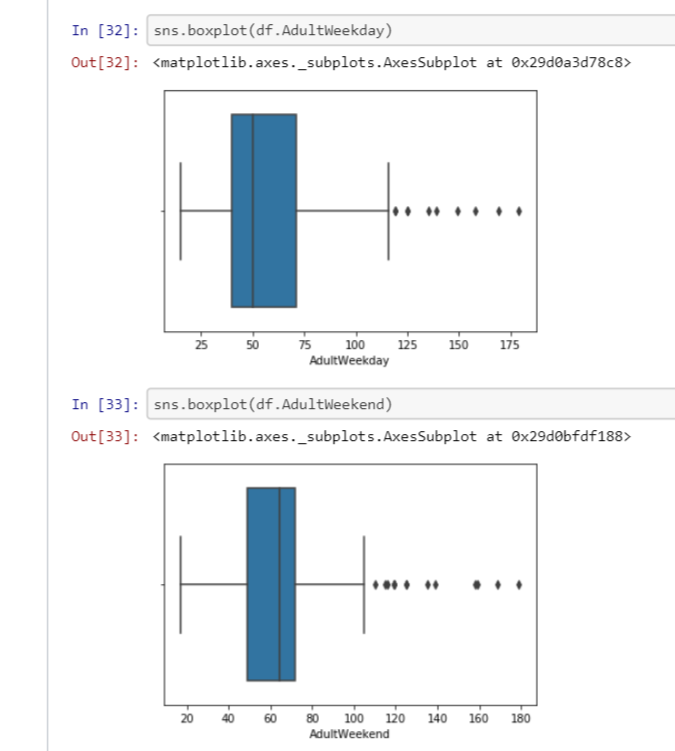
2.2 Eliminating Highly correlated features:

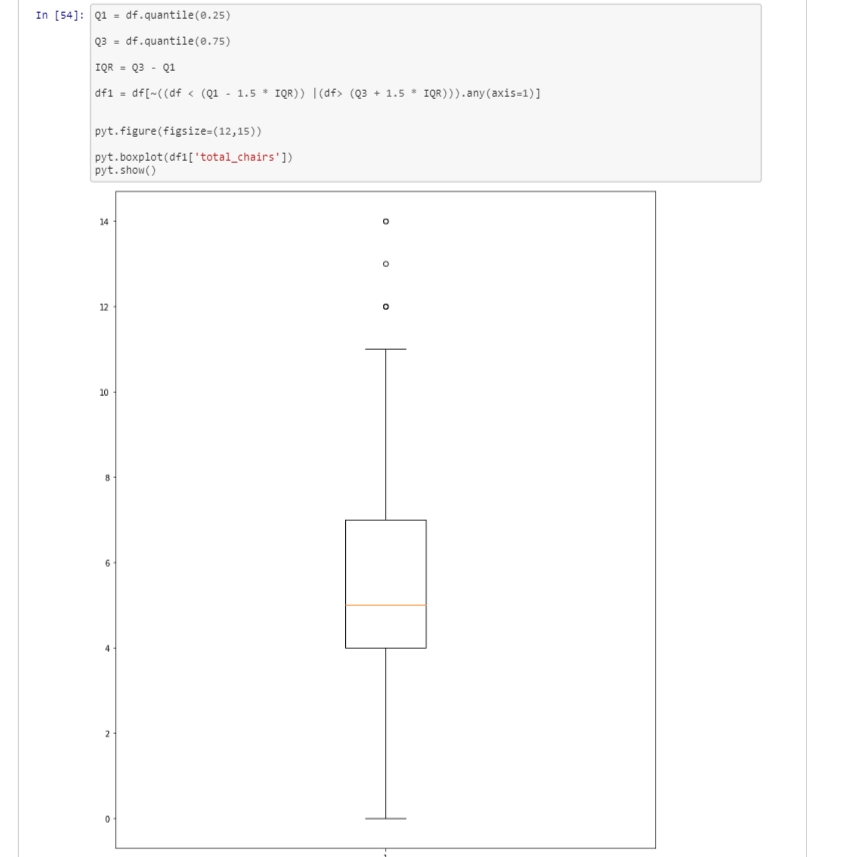
This is very important step in this project because without eliminating highly correlated features will show an impact in model accuracy.



2.3 Anamolies & Outliers:

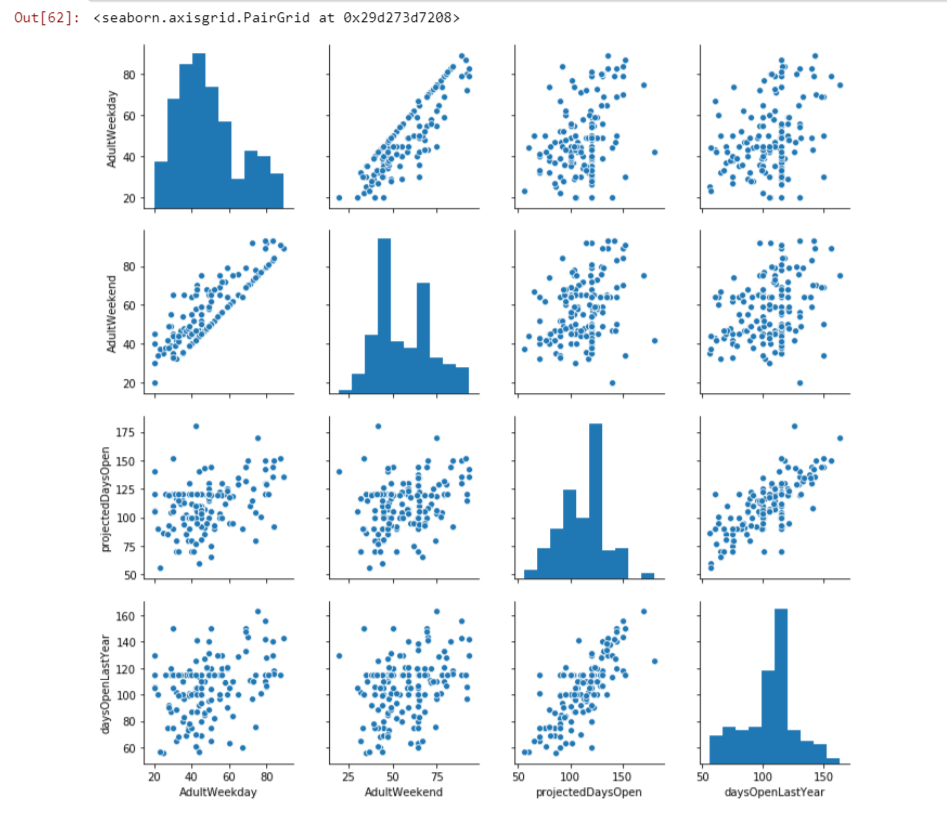
In this step we will identify the outliers in the data by using boxplots and try to eliminate those values.

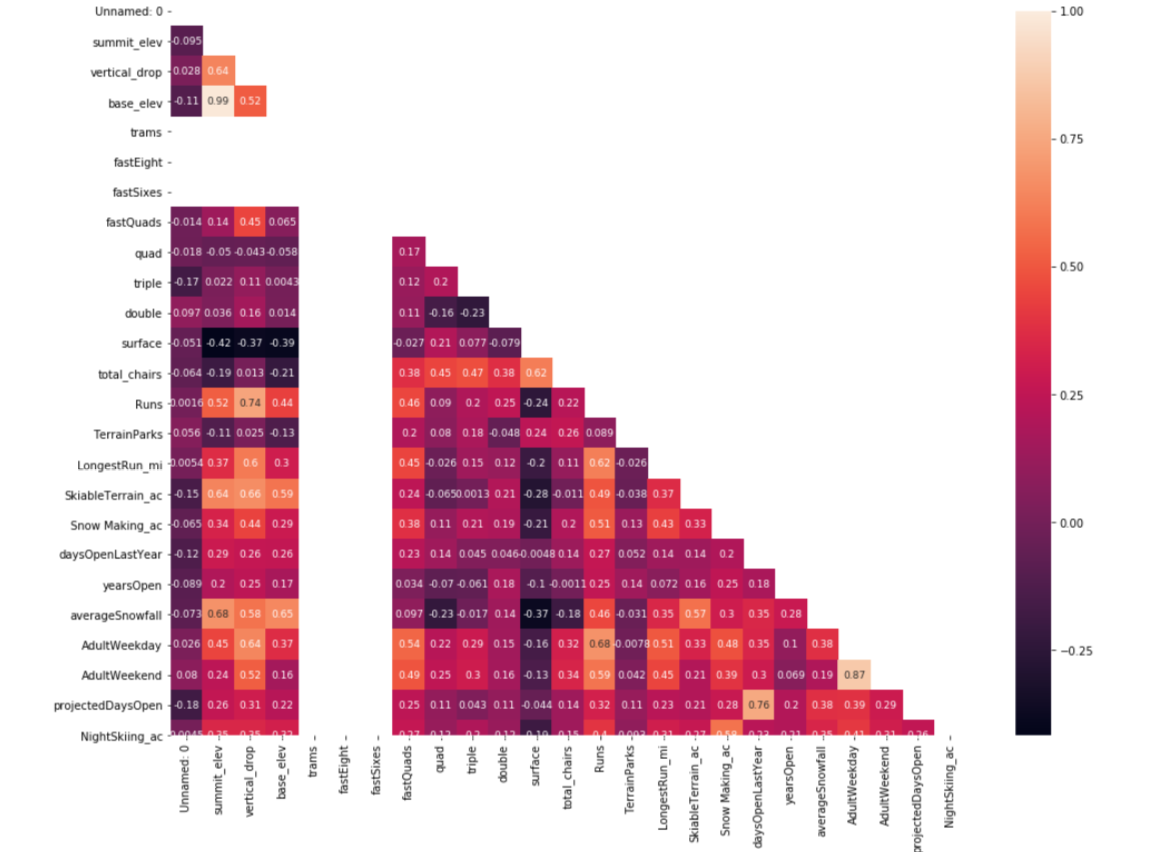




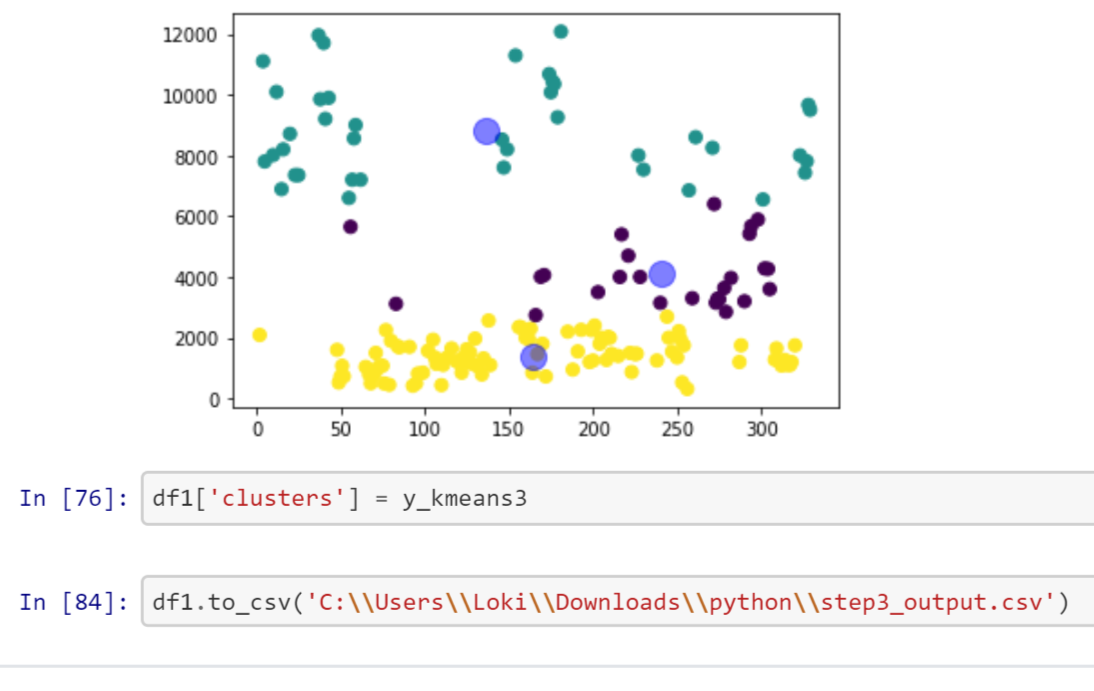
2.4 Explore data relationships:

In this step we will try to explore the relations ships among various attributes with the help of pair plot and heatmap. Heatmap helps to identify columns that suffer with multi-collinearity.



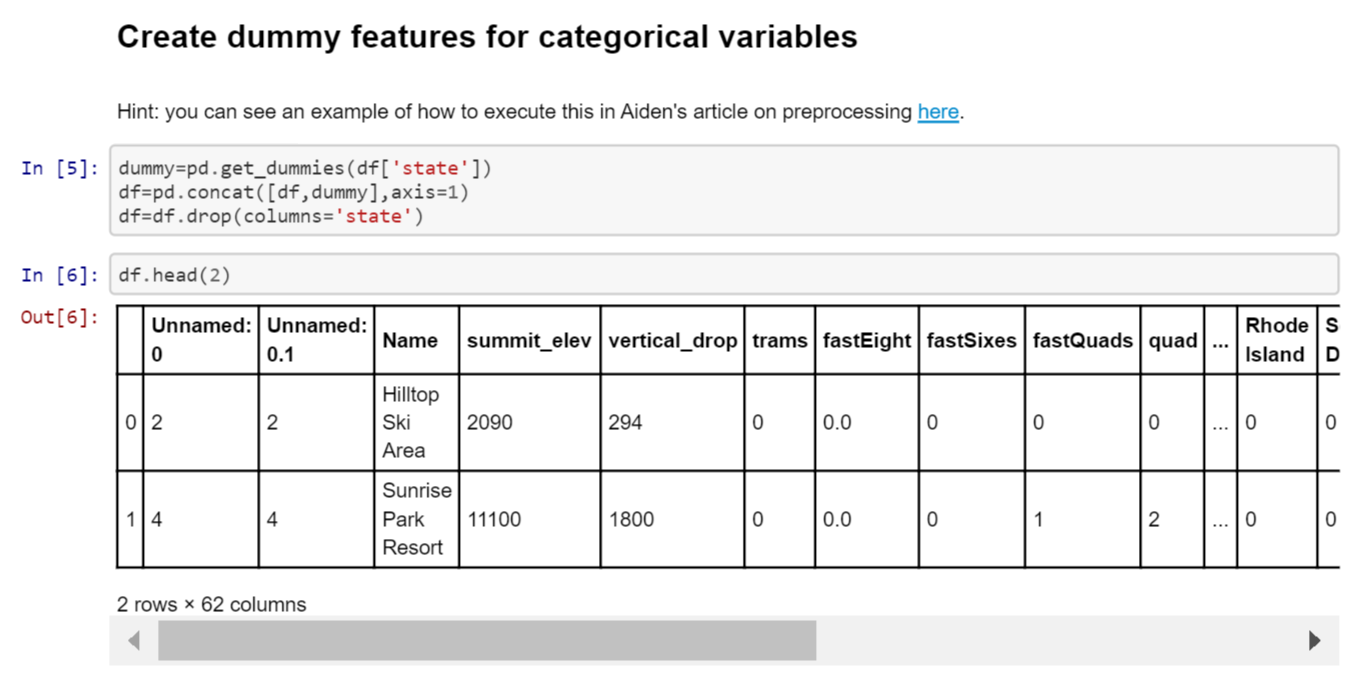


The vertical drop at the location with respect to the summit elevation of it is plotted below. As can be clearly seen the resorts can be broadly divided into 3 clusters.

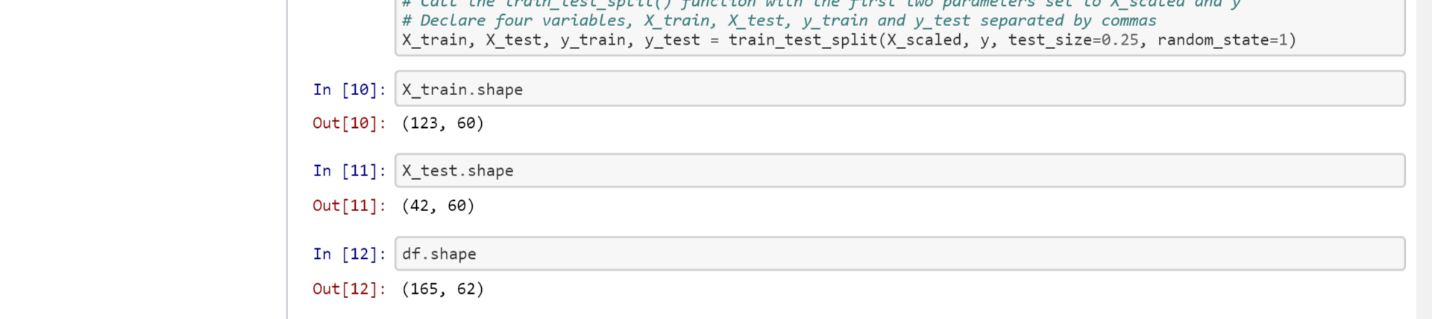


3.Pre-Processing and Training Development:

In this step we will try to create dummy variables for the categorical variables and later we apply standard scaler for numeric variables.



After preprocessing and splitting the testing and training data with a test\_size of 0.30 we got the shapes as the following:



4.Modelling:

In this modelling phase we will try to eliminate some of the features in dataset and apply linear regression. The model which gives more accuracy is chosen as the final model.

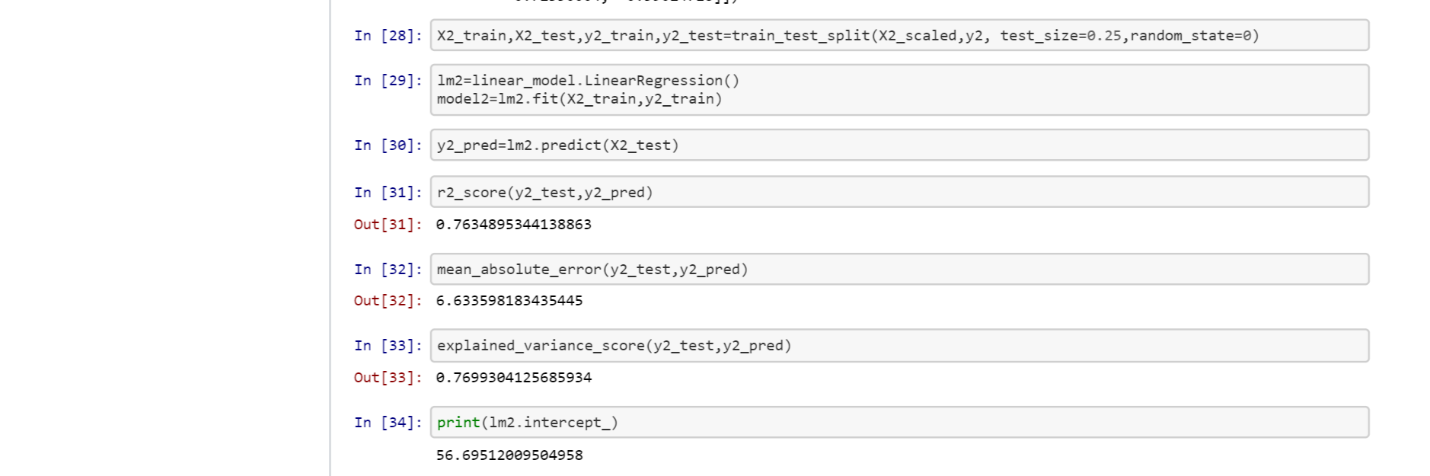
4.1 Model 1:

In this model we will try to eliminate categorial variables and implement the linear regression.



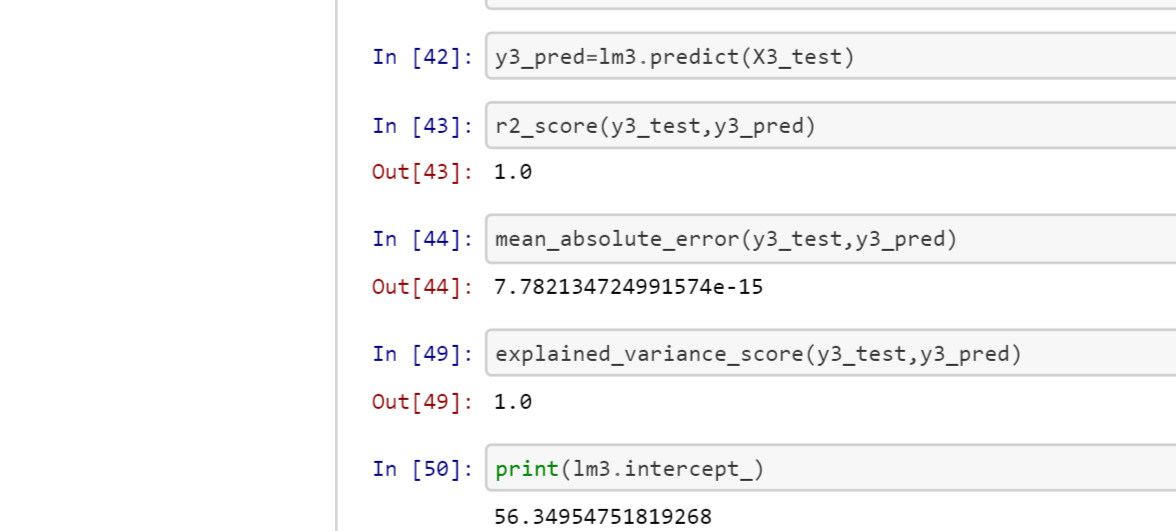
4.2 Model 2:

In this model we will try to eliminate the state features from the dataset to improve the model accuracy for linear regression. The model produces the results as the following:



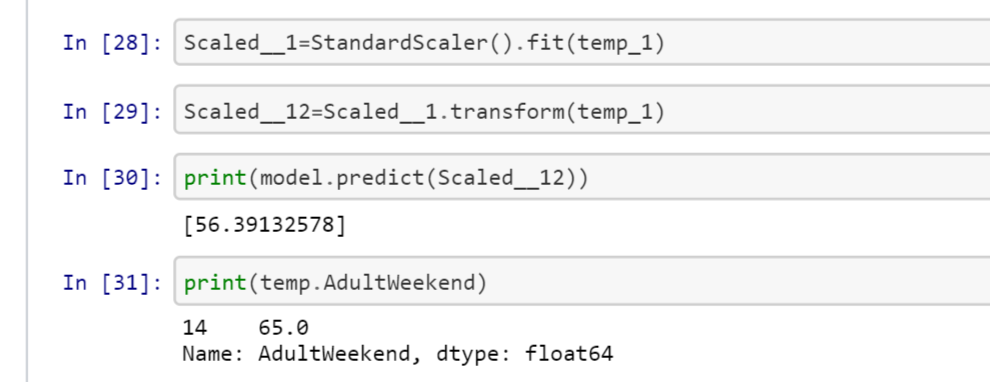
4.3 Model 3:

In this step we will try to eliminate the state, summit\_elevation, base\_elevation columns from dataset and predict the model accuracy using linear regression. It gives the following results:



5. Documentation:

In this process we will choose model3 which gives good accuracy among the three and try to predict the price of adult weekend.



After predicting the prices for Adult Weekend we will try to plot with Adultweekend and projecteddaysopen using lmplot.

