BIKE RENTING in Python

CHANDRA MOHAN SRIVASTAVA

21/11/2019

**INDEX**

**1 Introduction   
 1.1 Problem Statement  
 1.2 Data**

**2. Methodology**

**2.1 Pre-Processing.**

**2.2 Missing Value Analysis**

**2.3 Data Visualization and Exploratory Data Analysis**

**2.4 Feature Selection**

**3. Model Selection and Evaluation**

**3.1 Linear Regression**

**3.2 Decision Tree**

**3.3 Random Forest**

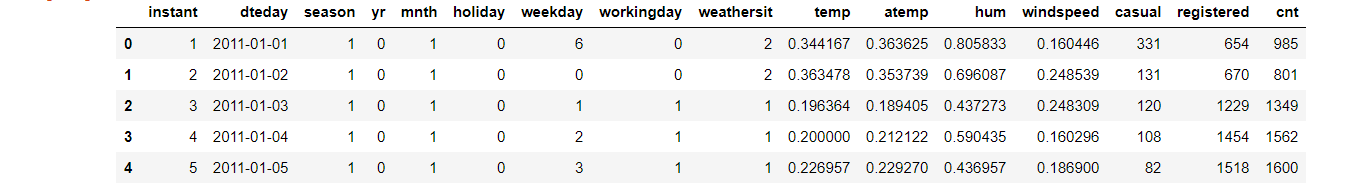
1. **Introduction** 
   1. **Problem Statement**

Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able rent a bike from a one location and return it to a different place on an as-needed basis. Currently, there are over 500 bike-sharing programs around the world.

The data generated by these systems makes them attractive for researchers because the duration of travel, departure location, arrival location, and time elapsed is explicitly recorded. Bike sharing systems therefore function as a sensor network, which can be used for studying mobility in a city. In this competition, participants are asked to combine historical usage patterns with weather data in order to forecast bike rental demand in the Capital Bikeshare program in Washington, D.C.

1.2 **Data**

The task is to build a regression model which will predict the number of bike rented on the basis of predictor variables. Following is a snippet of what the data  
looks like:



Description of the attributes :

instant – index which will be removed

dteday – date in yy/mm/dd format

yr- year : 0 = 2011 , 1=2012

mnth – month which ranges from 1 to 12 (January to December)

holiday – 0 = no holiday(including weekends) , 1 = holiday  
weekday – ranges from 0 to 6 where 0 represents Sunday and 6 Saturday and the numbers in between represents Monday to Friday in chronicle manner.  
workingday – 0 = not a working day and 1 = working day

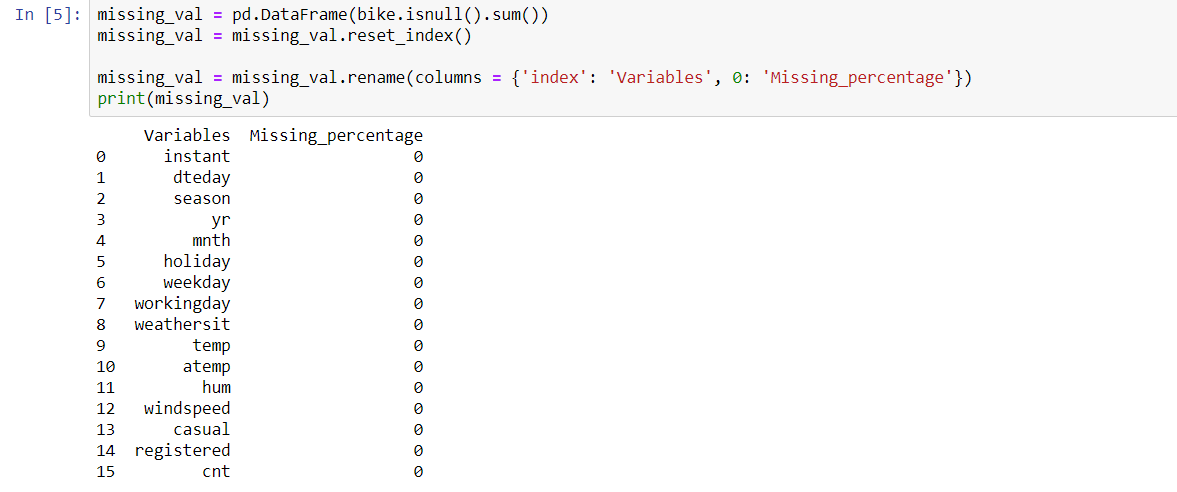
weathersit – ranges from 1 to 4 .

1: Clear, Few clouds, Partly cloudy, Partly cloudy  
2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist  
3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds  
4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

temp: Normalized temperature in Celsius. The values are derived via (t-t\_min)/(t\_max-t\_min), t\_min=-8, t\_max=+39 (only in hourly scale) atemp: Normalized feeling temperature in Celsius. The values are derived via (t-t\_min)/(t\_maxt\_min), t\_min=-16, t\_max=+50 (only in hourly scale) hum: Normalized humidity. The values are divided to 100 (max) windspeed: Normalized wind speed. The values are divided to 67 (max) casual: count of casual users

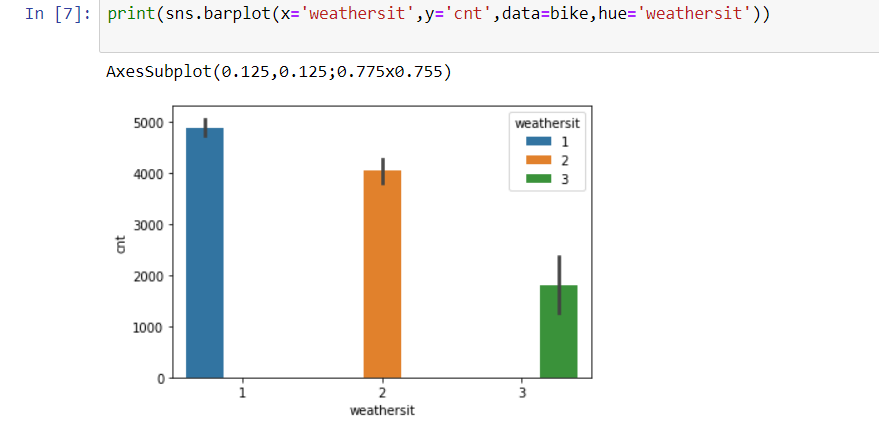
registered: count of registered users   
cnt: count of total rental bikes including both casual and registered

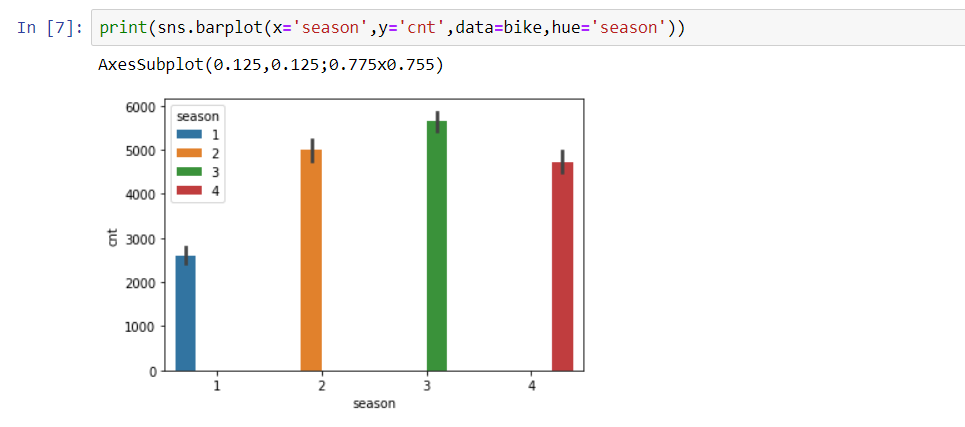
2 **Methodology**

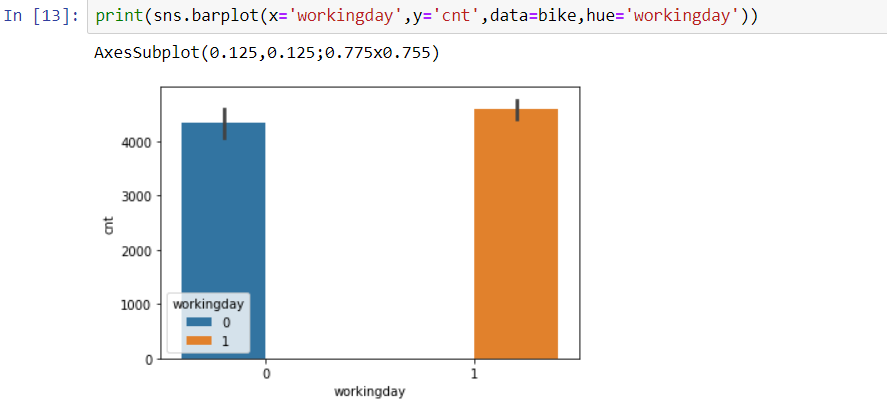
**2.1 Missing Value Analysis**:   
 It is important to get the hang of the data by checking the number of missing   
 values. In the data set presented, there is none as depicted below.  


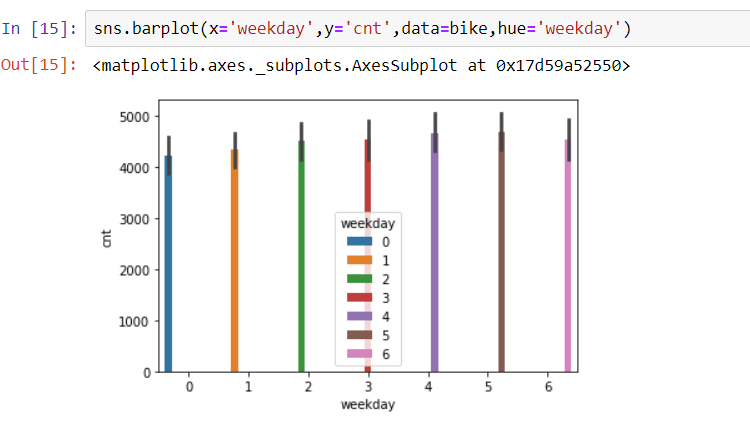
**2.2 Pre-Processing**

1. Before diving into model selection , it is important that we get the hang of the data.  
 In the data set , we have a number of continuous variable which only contains discretized   
 values only. Thus , those attributes gets converted to factor form.







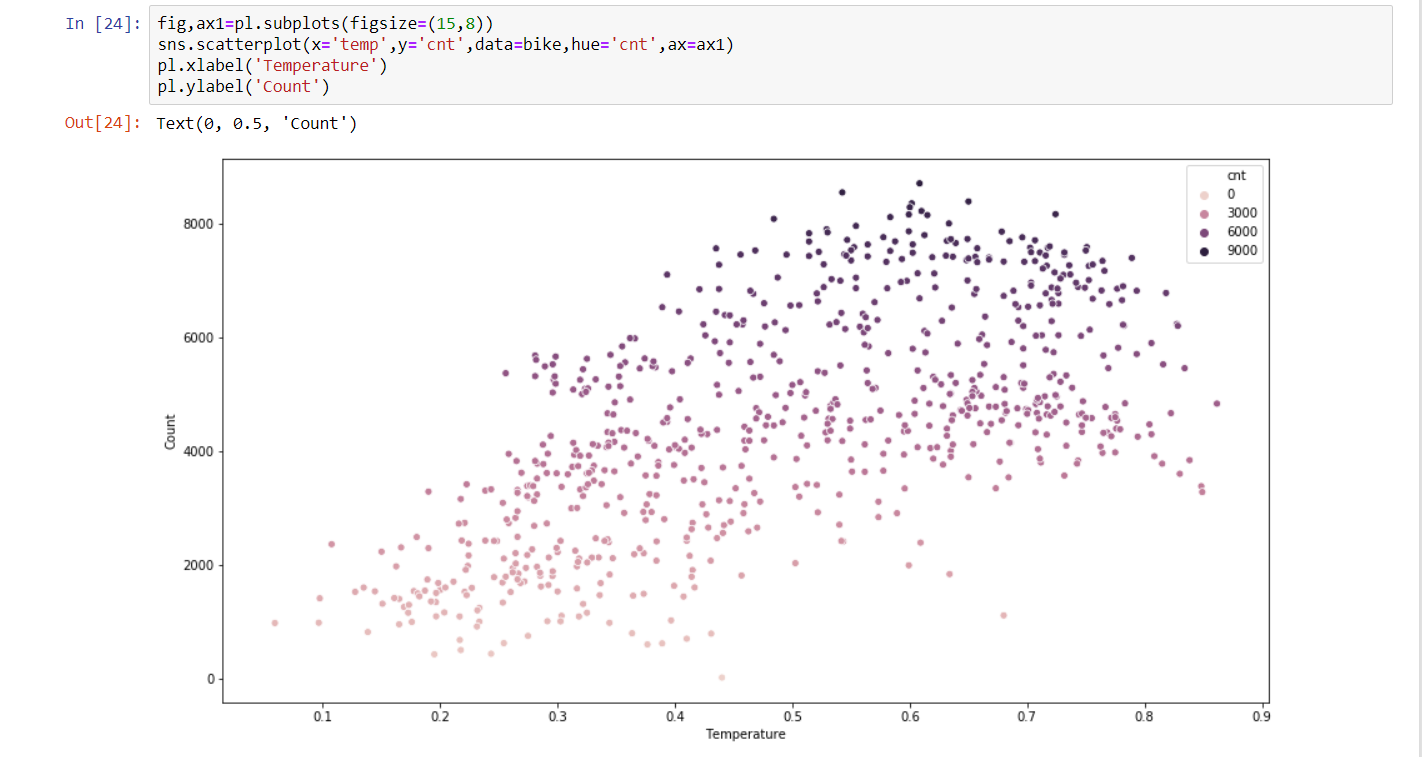


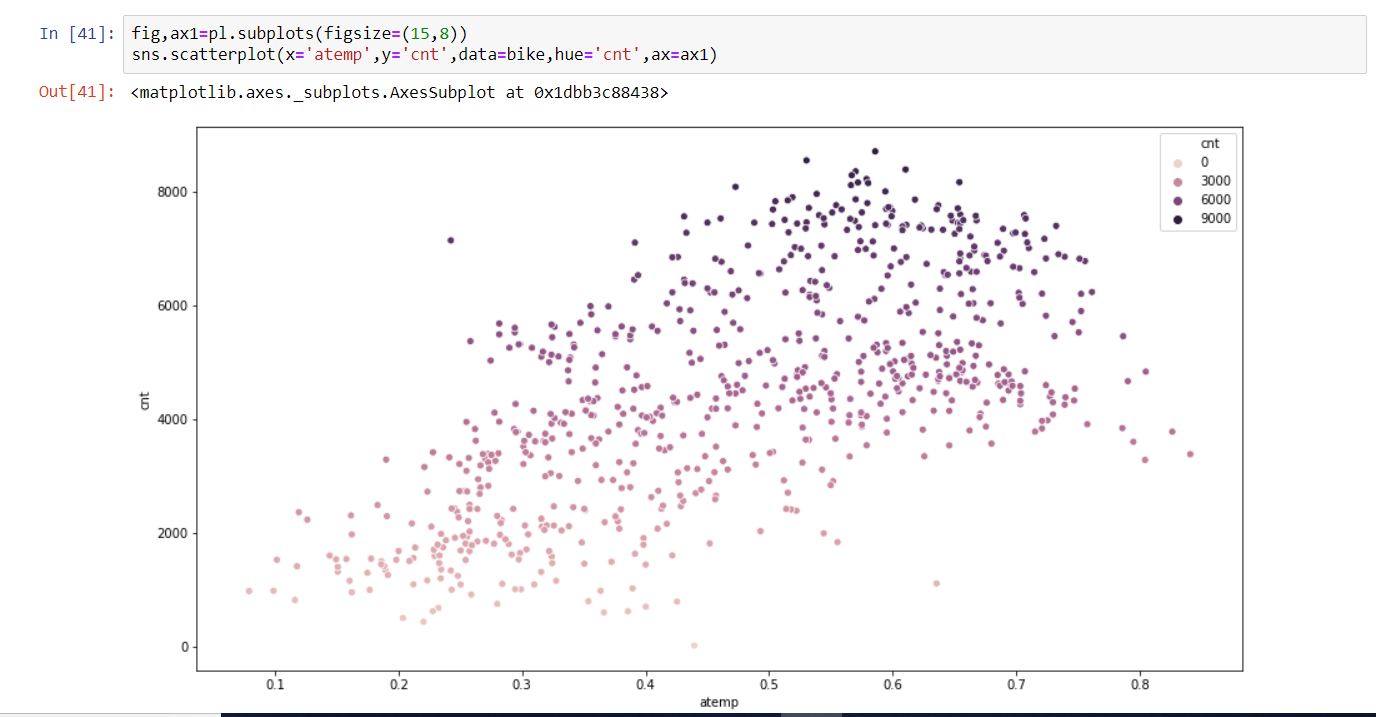
Thus variables such as weekday , season, holiday , workingday , weathersit gets discretized.

1. Also the variable instant will get deleted as it contributes no information.
2. We’ll also see later that cnt vs temp and cnt vs atemp bears the same relation – which   
   implies that one of the variable will get removed for dimensionality reduction.

**2.3 Data Visualization and Exploratory Data Analysis** Another process which occurs before Model selection is Data Visualization of the   
 target variable with the predictor variables.

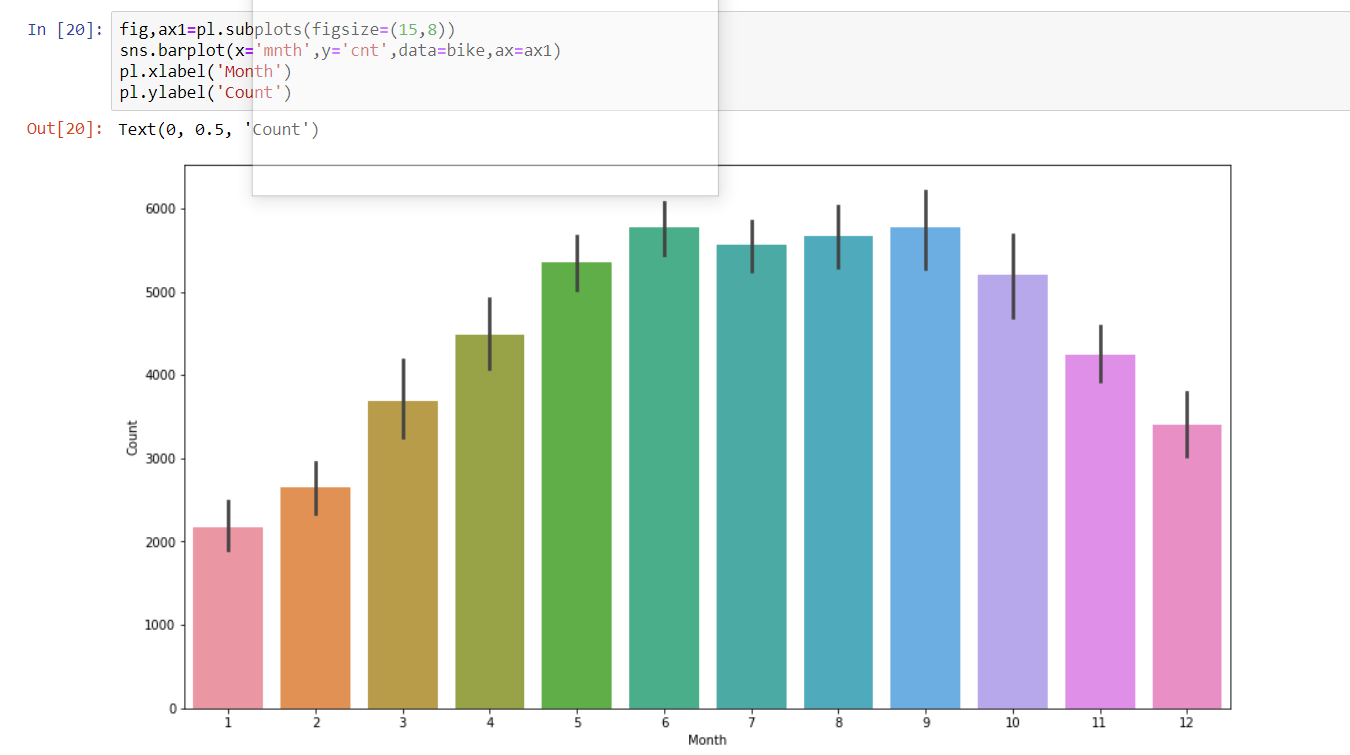
First, we will plot the graph of tmp vs cnt and atemp vs cnt in order to get some   
 insights.



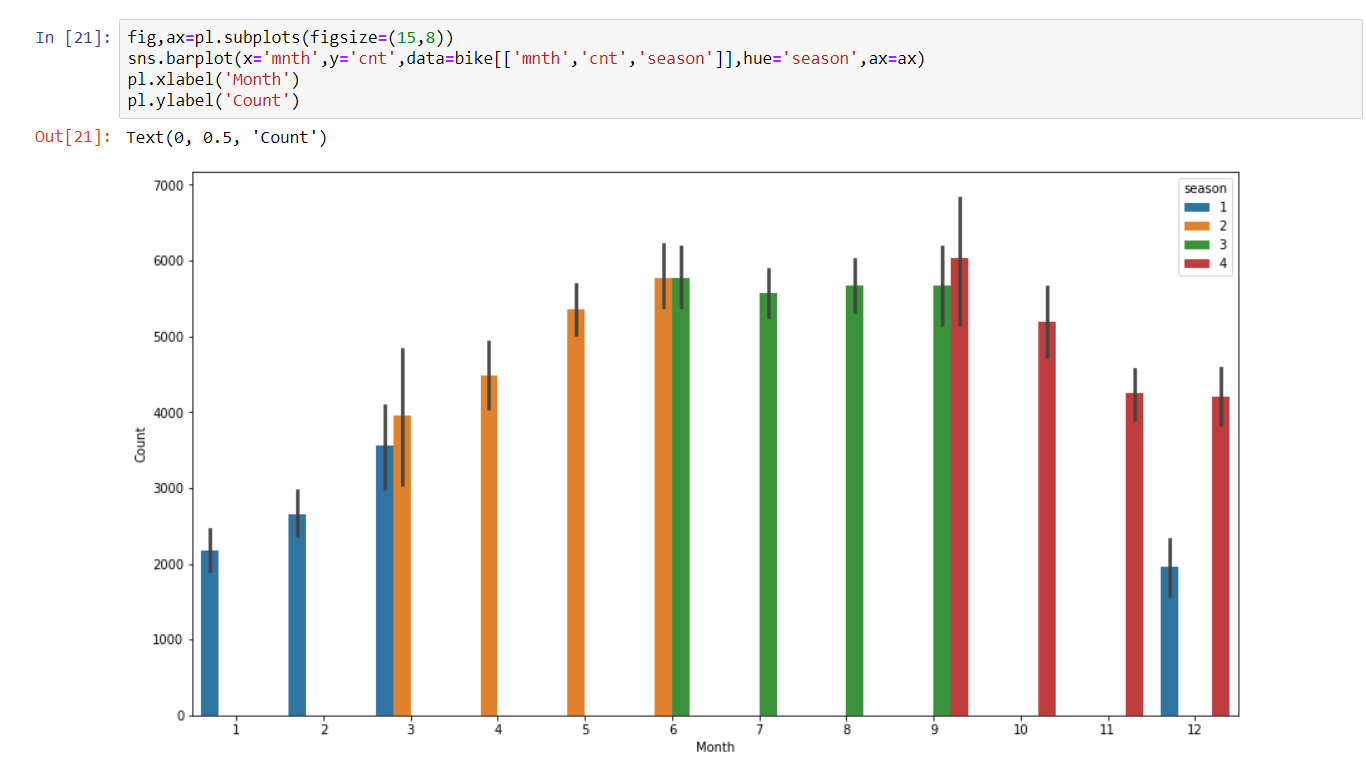


We can see that atemp vs cnt and temp vs cnt bears the same relationship. Thus we will drop atemp(dropping temp will also produce the same result).

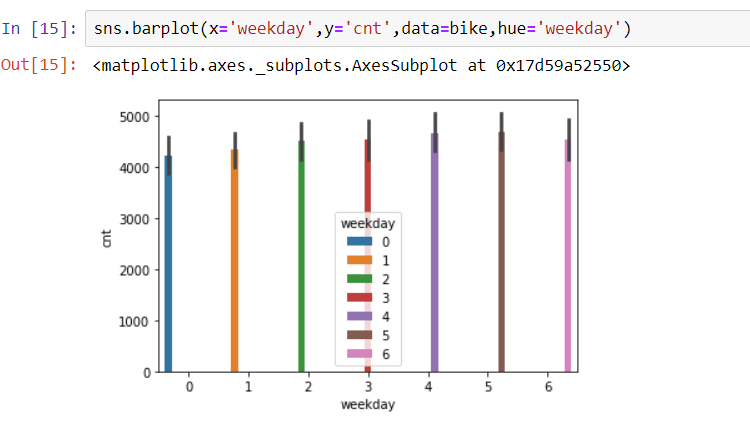
One can below see the bike count is related with other variables such as mnth :



From the above graph , we can see that bike count increases with progression of the year linearly , falls down in july , up til September and then decreases downwards linearly.

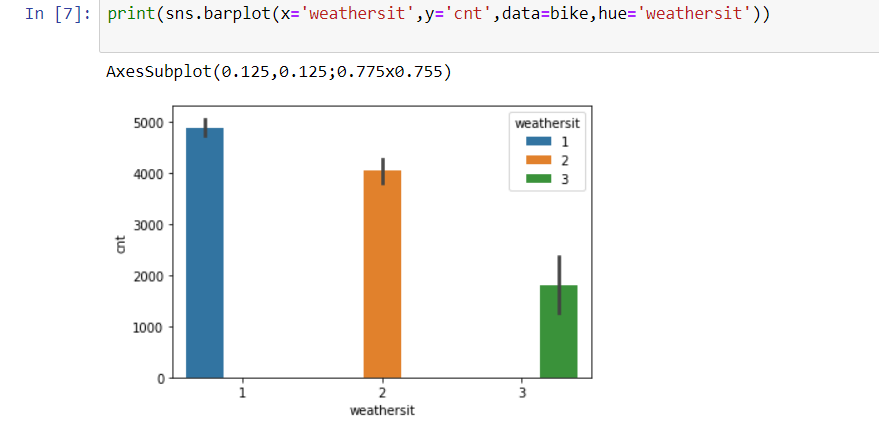


Following above is a graph which shows that the bike count is steady during the season fall   
 and has the maximum average of the bike count which means bike count is maximum used during  
 the fall. Then comes , winter(4) , summer(2), and then spring(1).



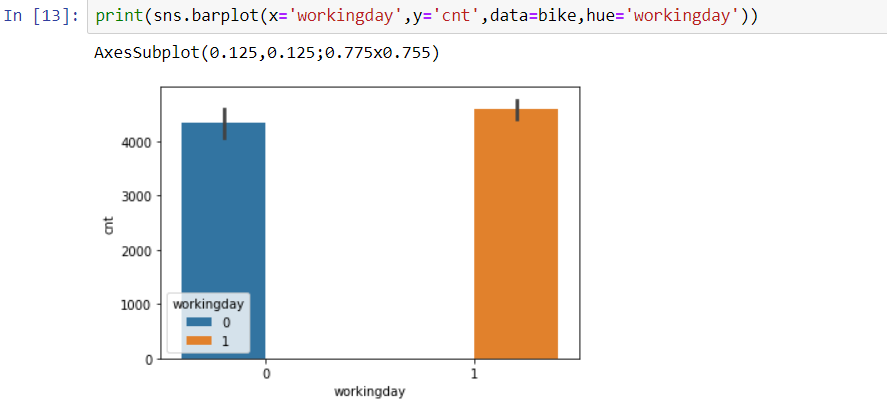
We can see that the maximum bike count is maximum every for Friday and Saturday , which makes sense as it’s the weekend. But , the variance around the mean isn’t that large , which means that the bike distribution is steady throughout the weekend.

cnt vs weathersit:



Maximum for weather 1 – clear skies, and lowest for slightly severe weather i.e weathersit 3 and no case of the harshest weather

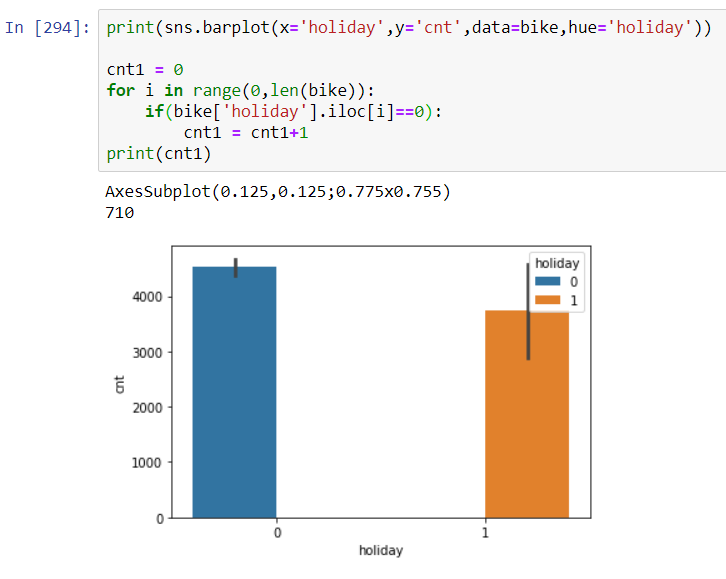
Cnt vs workingday



Bike distribution is almost similar for working day and non working day.

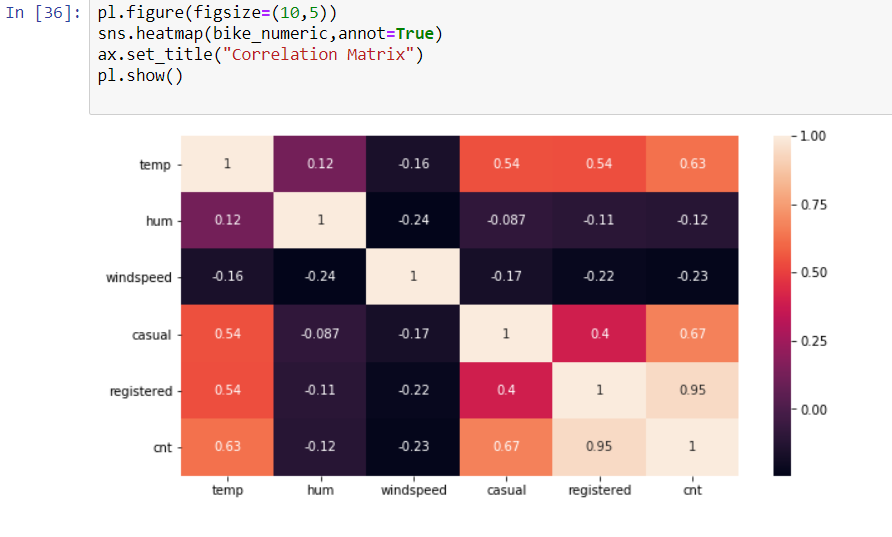
As there are more working days rows(500 rows) than non working days(total no of rows = 231 ,Saturdays and Sundays including) , their maximum being close together indicates that bikecount rises sharply during the non working days i.e workingday = 0)

Also , there are 21 holidays i.e. holiday = 1 and 710 no holiday(710 i.e Saturday and Sundays) , yet their maximum being close indicates that bike rent count skyrockets during holidays.



2.4 **Feature Selection:**

Very important to find multicollinearity present in the dataset whose snippet is given below:



From the above heatmap , we can observe that casual and registered are heavily co-related to cnt.

Therefore we will drop them out of the bike dataframe.

Also , the other reason is that casual and registered is directly related to cnt.



Also , dteday columns also gets removed

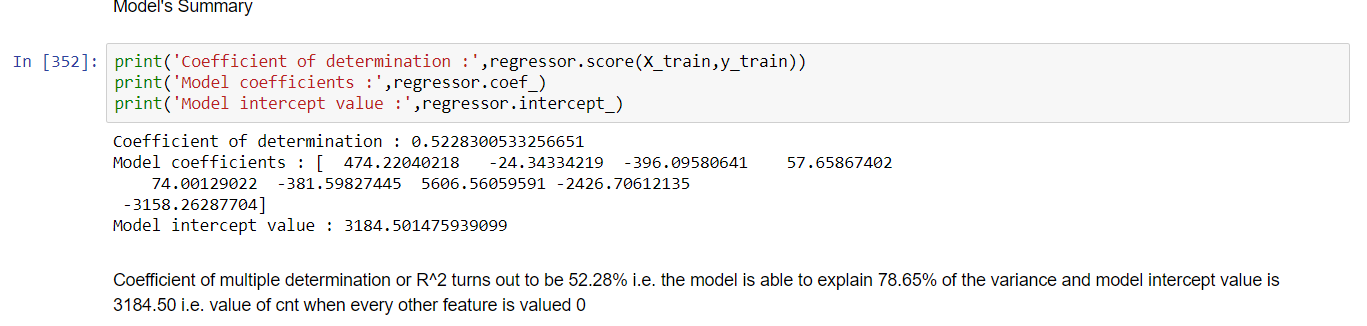


3. **Model Selection and Evaluation:**

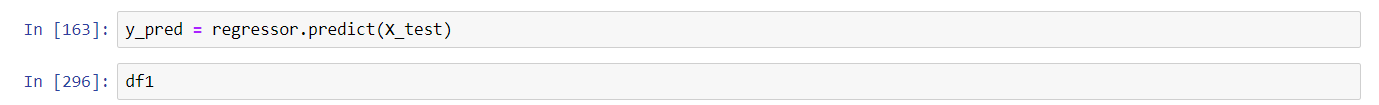
**3.1 Model Selection** In this case , as we have to predict a continuous variable , the prediction will be   
 performed with the following models:

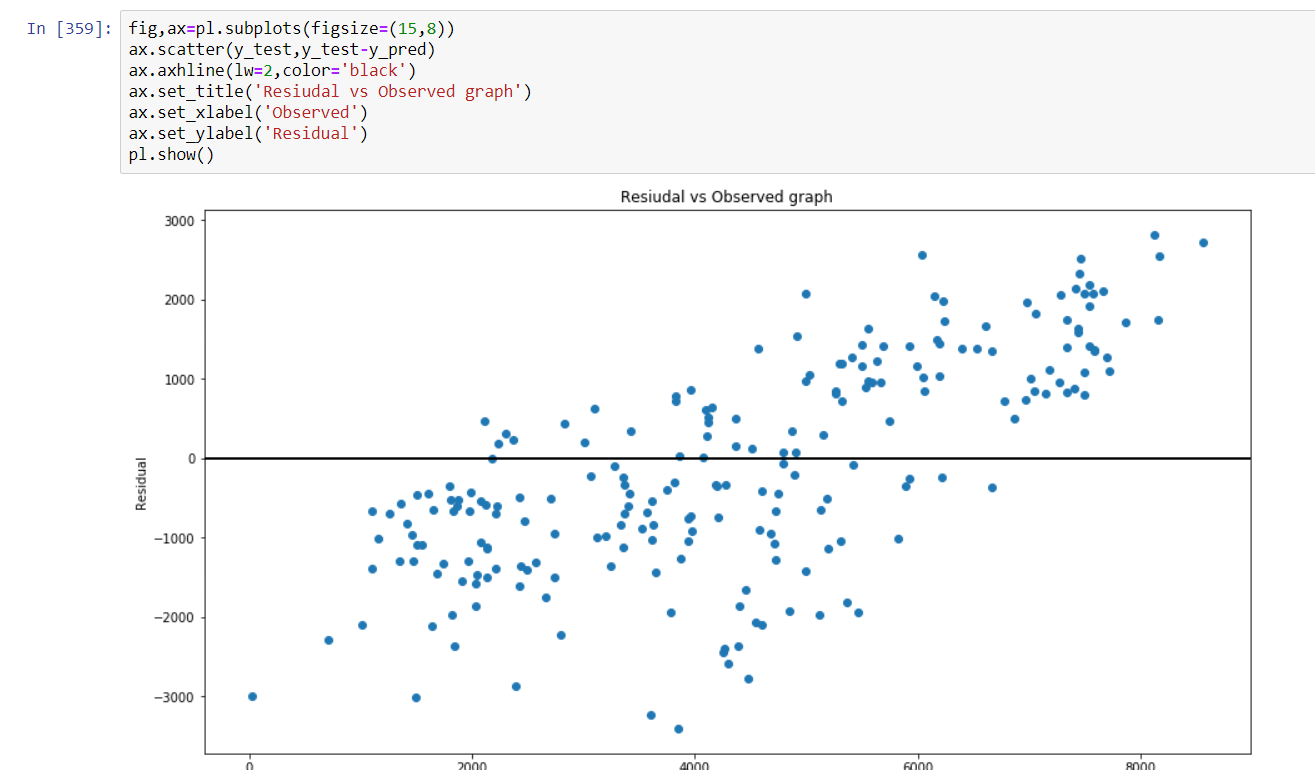
1. Linear Regression.  
 2. Decision Tree Regression.  
 3. Random Forest Regression

3.1.1 Linear Regression:  
 Following line of code is written to perform linear regression in Python:

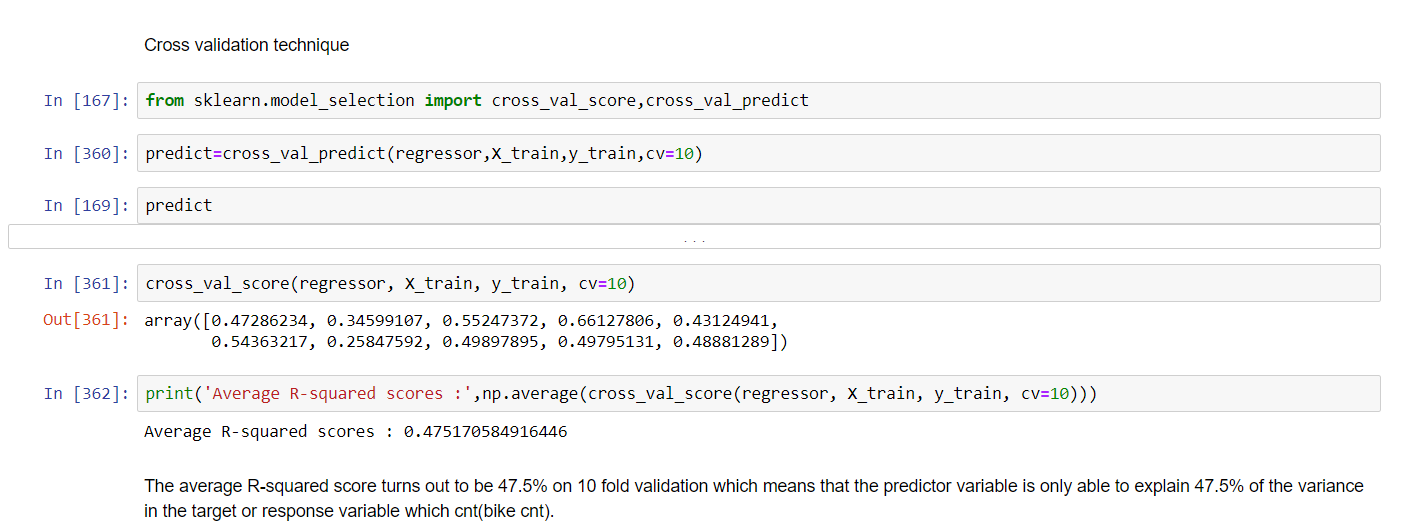


Coefficient of determination reveals that variance of 52.28% is explained by the model.

The model is made to predict via the following line code: 

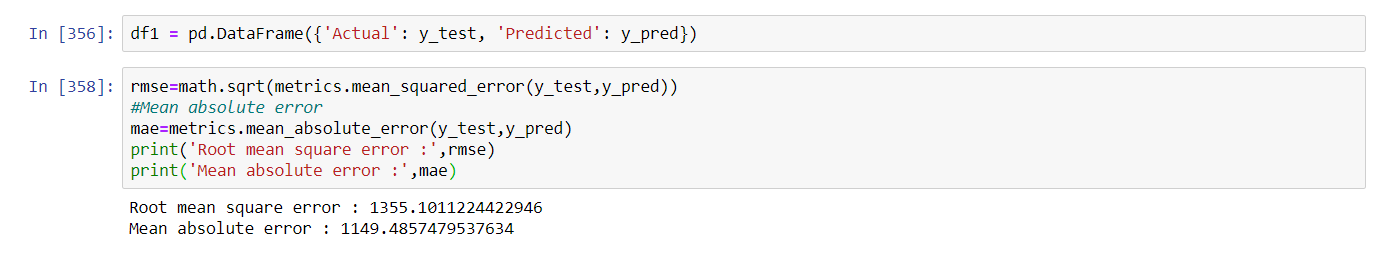
Residual vs Observed for linear regression  


The model is made to then made to go through cross validation :

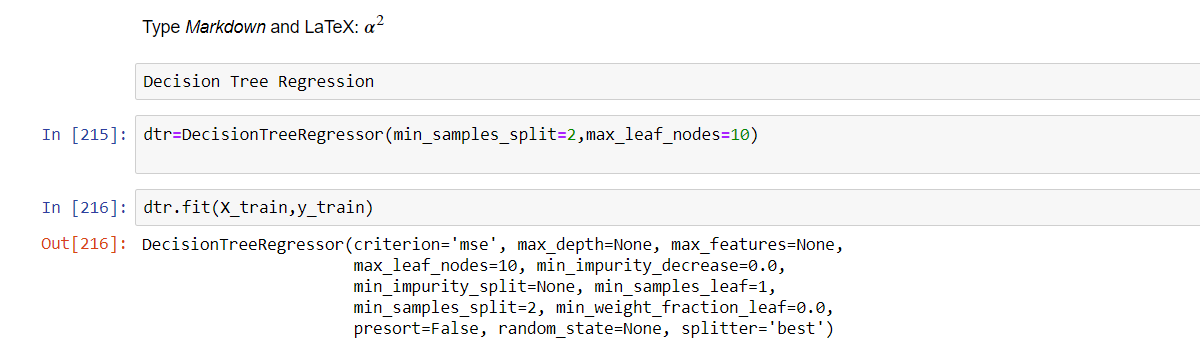


The average variance explained by the model is 47.5% which implies that the model created above is very close to the average or the correct linear regression model.

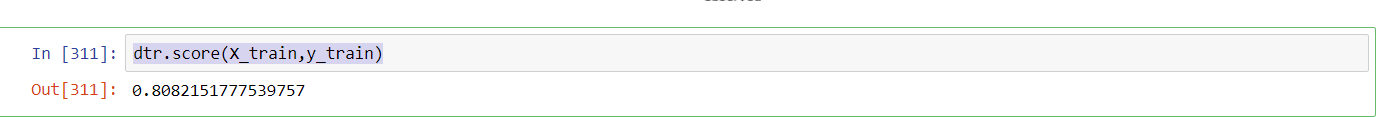
The mae and rmse is given below :



3.2.2 **Decision Tree** Following line of code is used to perform Decision Tree in R :



The accuracy of the model is 80.82%



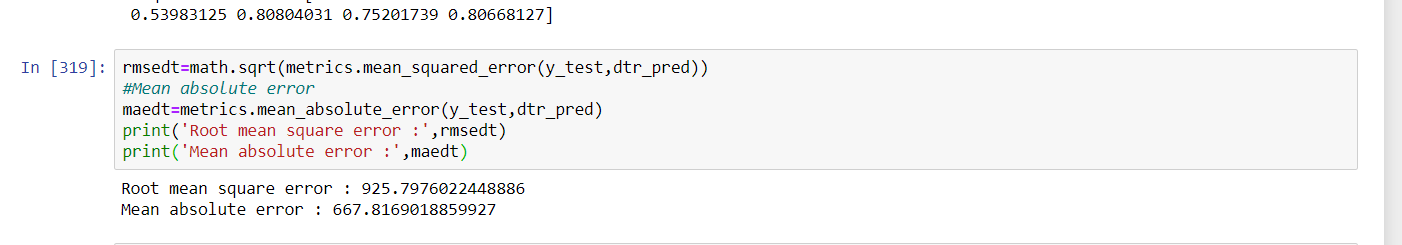
Prediction for the test model :

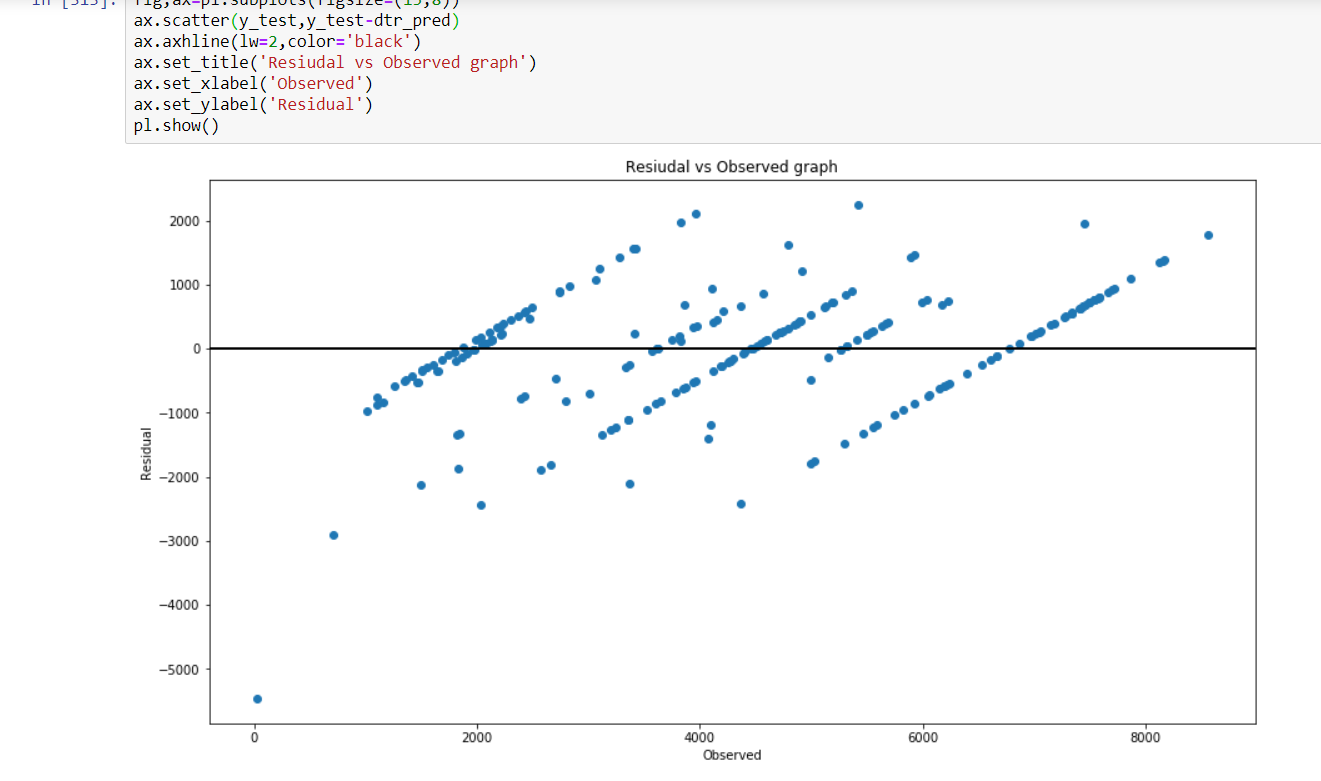


The residual vs observed graph looks like this :

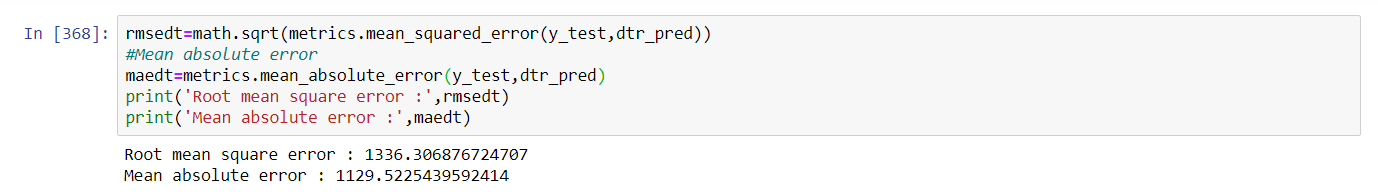
Cross validation technique

Following is the mae and rmse :



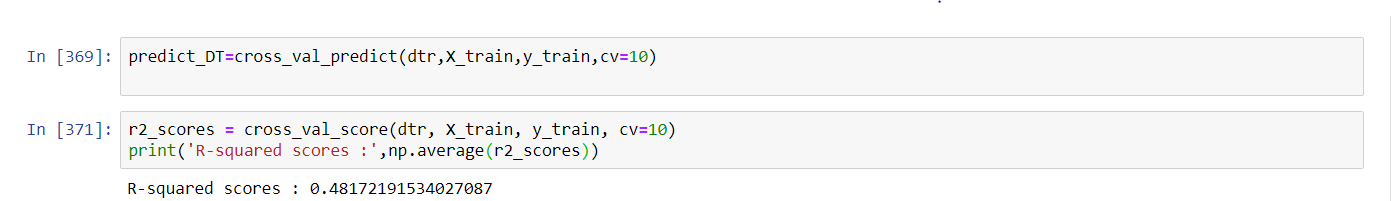


Following is the mae and rmse :



The mae and rmse value are quite similar to the regression model.

Cross validation technique

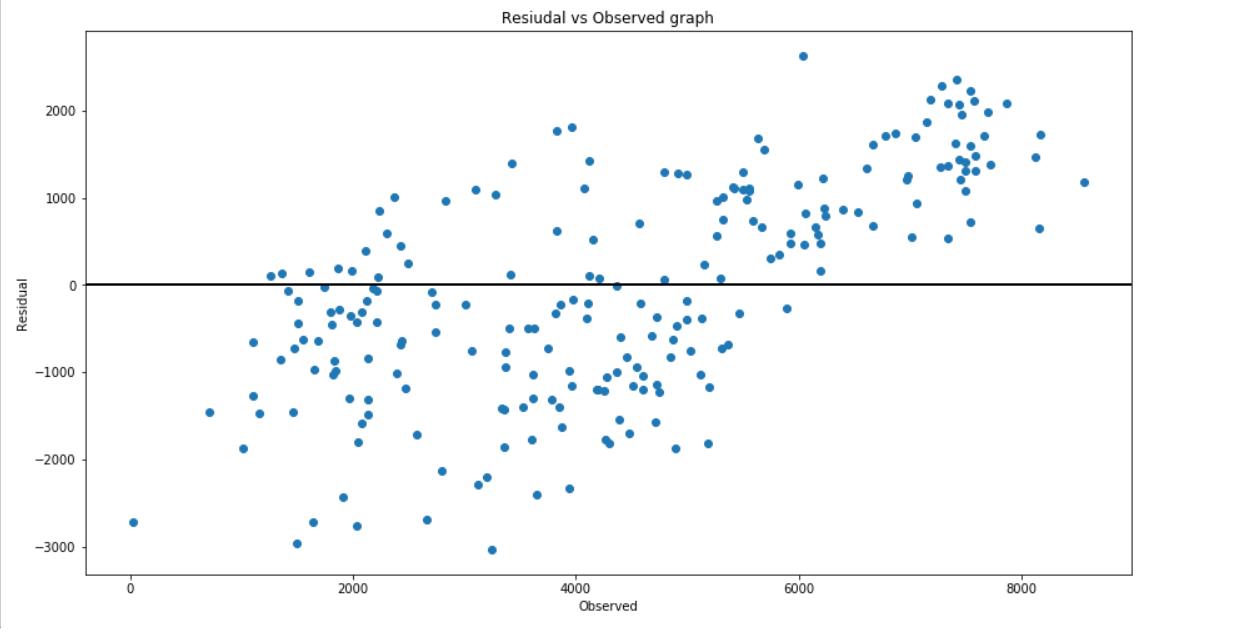


**3.3.3 Random Forest Regression**

Following line of code is used to generate a model with Random Forest

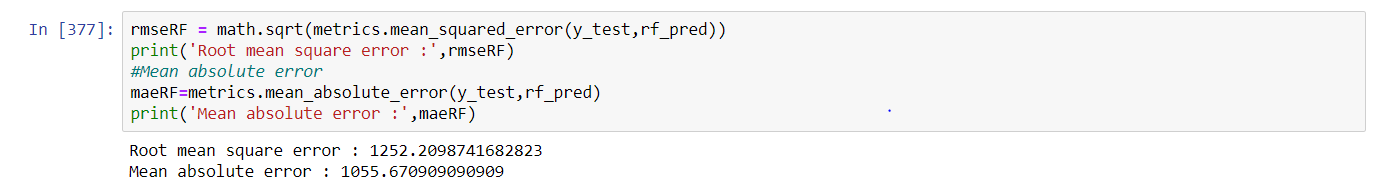


The model’s accuracy is 94.23%.

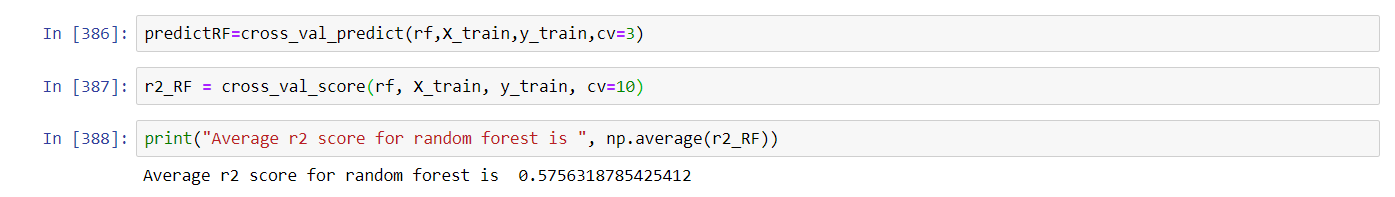


We can observe from the above graph that the model’s accuracy for random forest model is highest compared to the one generated by Decision Tree Model.

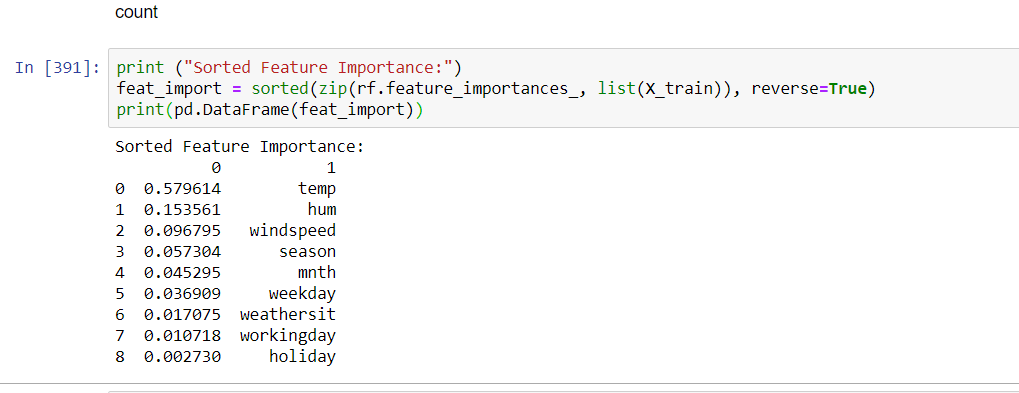
RMSE AND MAE values

​

Cross validation:



As the RF model has the lows RMSE AND MAE , and the highest accuracy , we will select Random Forest model as the most accurate one.



We can see that the most important feature is temperature followed by humidity and then windspeed- which is environment and seasonal conditions.