**Bike Rental Predictions**

Bheema Sai Teja

September 2019

**Contents**

1. **Introduction:**
   1. Problem Statement……………………………………………….3
   2. Data………………………………………………………………3
2. **Methodology:**
   1. Pre-Processing……………………………………………………5
      1. Exploratory Data Analysis…………………………………6
      2. Missing Values analysis……………………………………9
      3. Outliers Analysis…………………………………………...9
      4. Feature Selection…………………………………………..11

2.1.5 Feature Scaling…………………………………………….13

* 1. Modelling………………………………………………………..15
     1. Decision Tree……………………………………………….15
     2. Random Forest……………………………………………...16
     3. Linear Regression…………………………………………..16

1. **Conclusion:**
   1. Model Evaluation………………………………………………….17
      1. Mean Absolute Error (MAE)………………………………18
      2. Mean Squared Error (MSE)……………………………….19
   2. Model Selection…………………………………………………….20

**Appendix A -** Python Code…………………………………………..20

**References……………………………………………………………….** 23

# Chapter-1

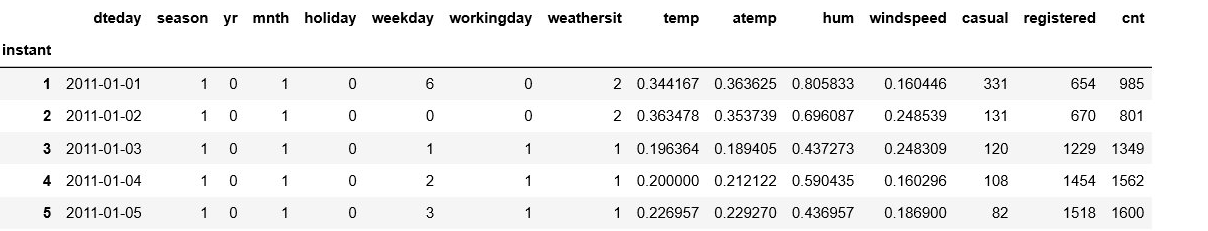
## **INTRODUCTION**

**1.1 Problem Statement**

A Bike rental company has collected bike rental data on daily basis. The aim of this case is to predict the bike rental count every day based on environmental and seasonal settings.

**1.2 Data**

The task is to develop a regression model that will predict the count of bike rentals every day based on the environment and weather. Here’s the data set.



The details of the data attributes are:

**Instant** - Record index

**Dteday**  - Date

**Season**  - Season (1:springer, 2:summer, 3:fall, 4:winter)

**Yr**  - Year (0: 2011, 1:2012)

**Mnth** - Month (1 to 12)

**Hr**  - Hour (0 to 23)

**Holiday** - weather day is holiday or not (extracted fromHoliday Schedule)

**weekday** - Day of the week

**workingday** - If day is neither weekend nor holiday is 1, otherwise is 0.

**Weathersit**  - (extracted from Freemeteo)

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

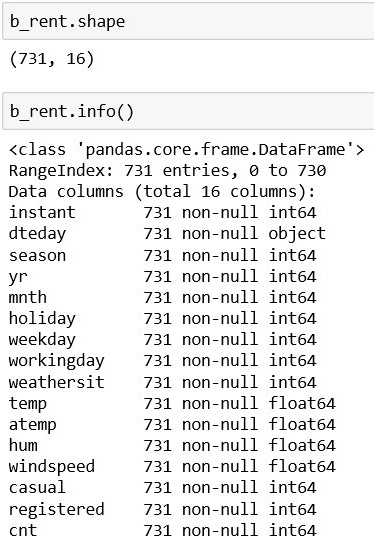
## Chapter-2

# METHODOLOGY

**2.1** **Pre-Processing**

Any predictive modelling requires that we look at the data before we start modelling. However, in data mining terms *looking at data* refers to so much more than just looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as **Exploratory Data Analysis**. To start this process we will first try and look at all the distributions of the variables.

Checking the shape of data and attributes data types:



And then renaming the necessary columns for clear understanding and converting the attributes data types if necessary.



**2.1.1 Exploratory Data Analysis:**

Organizing, Plotting and Summarizing the data.

By visualising the data we can know the data is distributed and the we can detect the pattern in the data. According to the data, Fig. 2.1 is bike rental count in each season most bike rentals are in fall whereas the least in the spring. Fig. 2.2 is the bike rental counts in every month of a year.

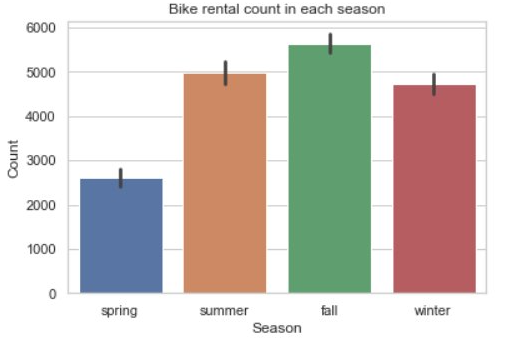


Fig. 2.1

As we can see most of the bike rentals are in the months of June and September and the least are in the beginning of the year i.e., January & February.

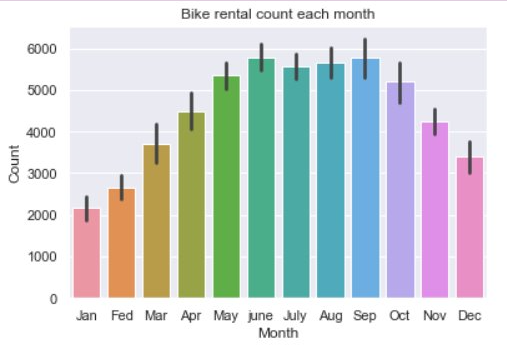


Fig. 2.2

As the given data is for years 2011 and 2012, in the Fig. 2.3 we can see the bike rental counts in the both the years.

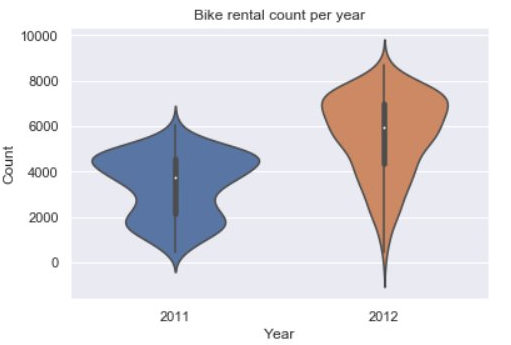


Fig. 2.3

And then the plot of bike rentals on the holidays and the working days in the Fig. 2.4

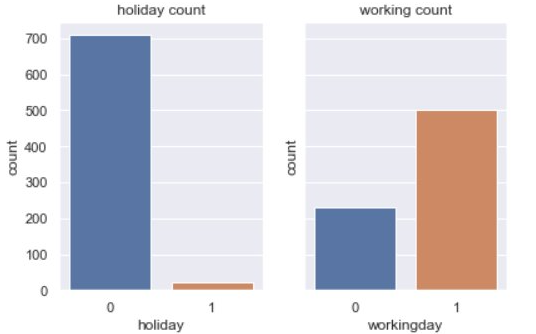


Fig. 2.4 Bike rental count weather the day is holiday or working day.

As we have the rental counts based on the day of the week in Fig. 2.5 we can see bike rentals on each day of the week.

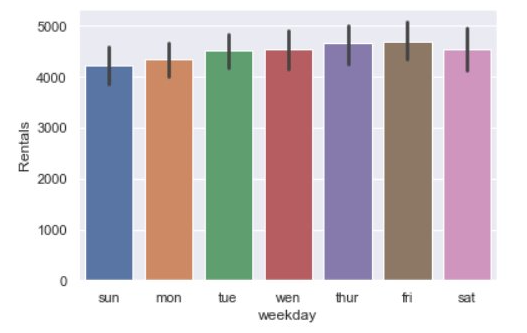


Fig. 2.5 Bike rentals count on each day of a week.

**2.1.2 Missing Values Analysis:**

In predictive modelling we have to look into data weather the data is clean, does it have any missing values. If it does have any missing then those values have to be replaced by mathematical methods if the number is insignificant they can be removed.



We don’t have any missing values so, we are good to go.

**2.1.3 Outlier Analysis:**

Outliers in the data may occur due to poor measurement quality or some external reasons. In a simple way we can detect outliers by plotting box plots of the different variables in the data set. In Fig. 2.6 we can see the distributions of the four main predictors.

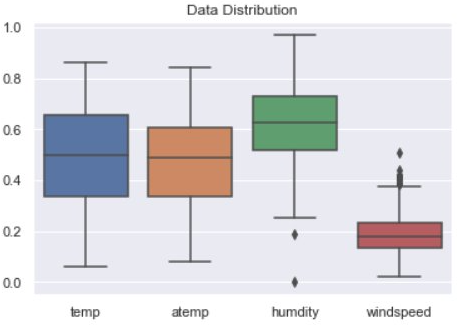


Fig. 2.6

By observing the plots we can say that both the humidity and windspeed have the outliers may due to the adverse weather conditions whatever may be the reason having outliers in the data set while predictive modelling may result in wrong predictions so, these outliers should be removed. In Fig. 2.7 we can see that the casual column has outliers but here it does not matter because we have to predict these columns they are the target variables.

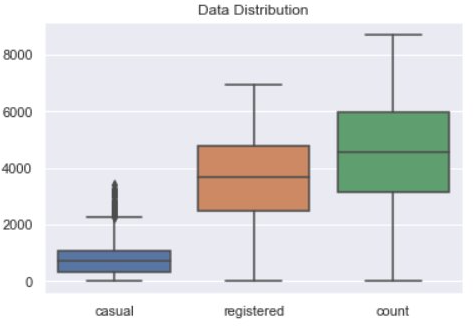
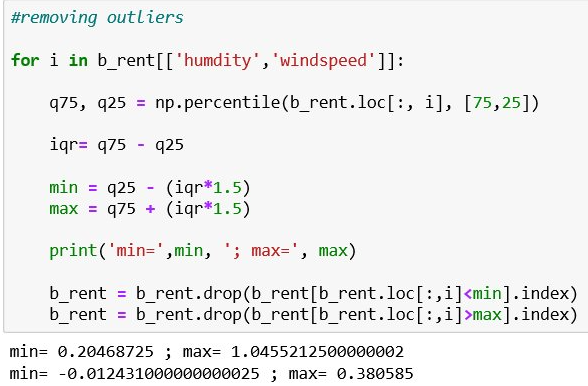


Fig. 2.7

Here is the method to remove the outliers from the data with the 25th and 75th percentile calculating the MIN and MAX then removing all the data points less than MIN and greater than MAX.

Code in Python:



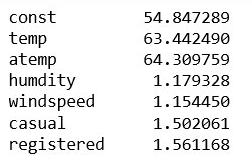
**2.1.4 Feature Selection:**

In the predictive modelling feature selection is about selecting the independent variables which will be helpful in predicting the target variable. It is also know as Dimensionality Reduction. For numerical data we can use correlation plot for categorical data we use chi-squared test.



Fig. 2.8

And the multi-collinearity is checked by Variance inflation factor:



From the correlation plot (Fig. 2.8) and VIF temp & atemp are highly correlated to each other so, having both of them is useless so, we have to remove atemp.

For the categorical data we use chi-squared test in which we can compare variables with each other or with the target variable this is a hypothesis test we get p-value. If the p-value is greater than critical value then we can reject the null hypothesis.

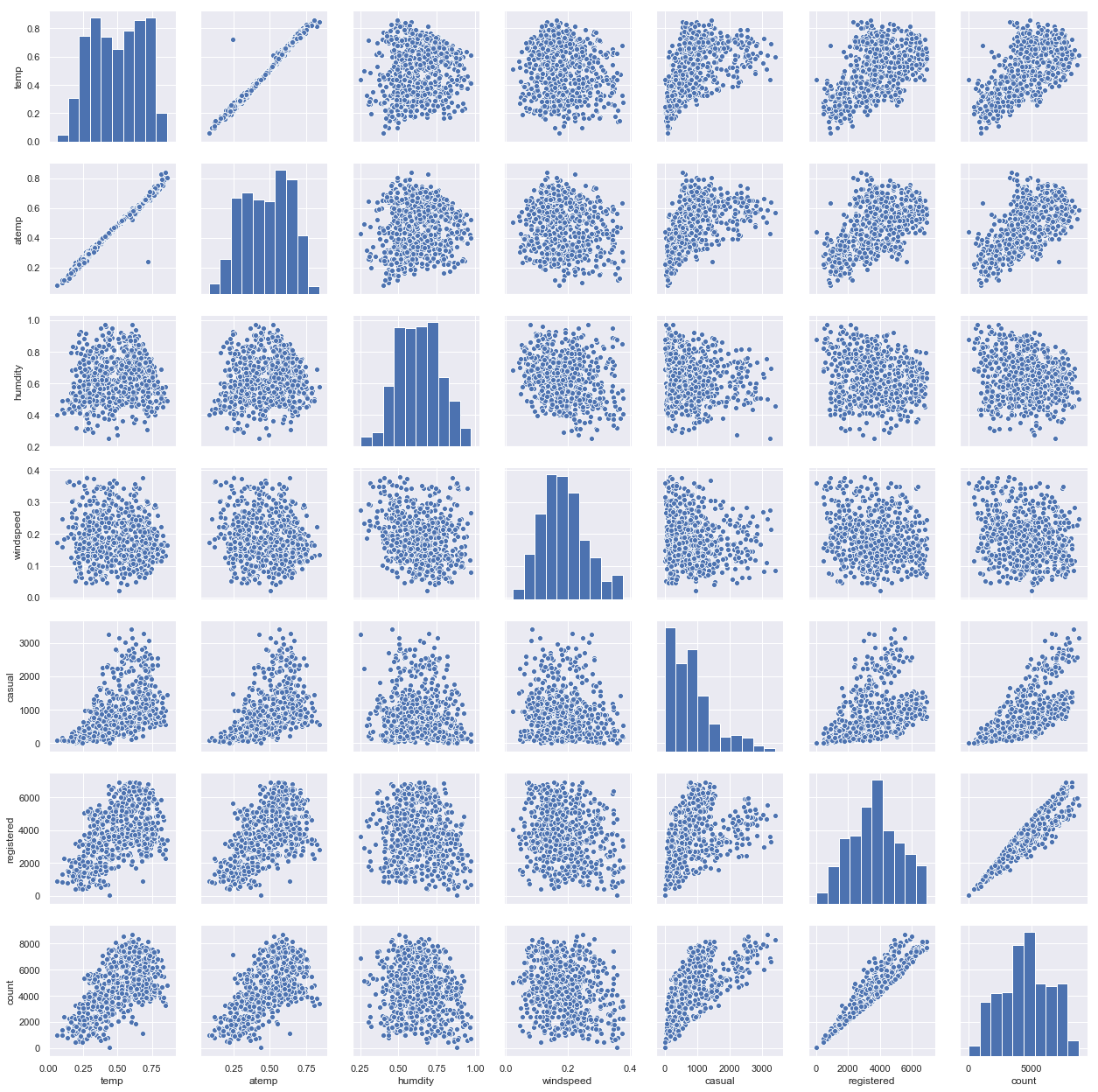
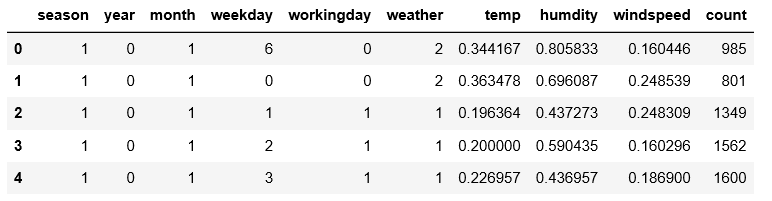


Fig. 2.9 Data distribution and correlation between variables

Removing casual and registered because they are what we have to predict. Instant and date

are removed because they are not useful in predicting.

After removing all the unnecessary columns this is how the clean data set look like:



**2.1.5 Feature Selection:**

In predictive modelling we will first try and look at all the probability distributions of the variables. Most analysis like regression, require the data to be normally distributed. We can visualize that in a glance by looking at the probability distributions or probability density functions of the variable.

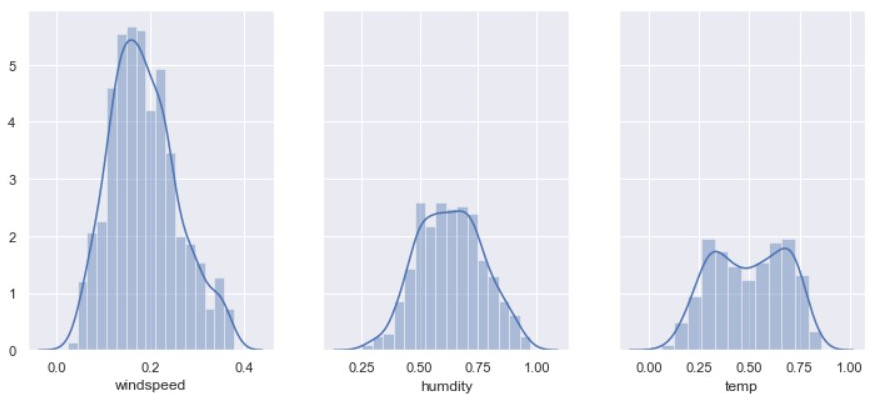


Fig. 2.10 Probability distribution

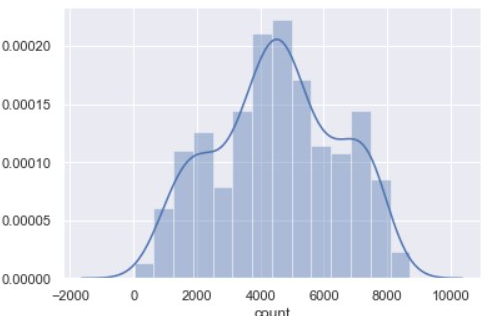
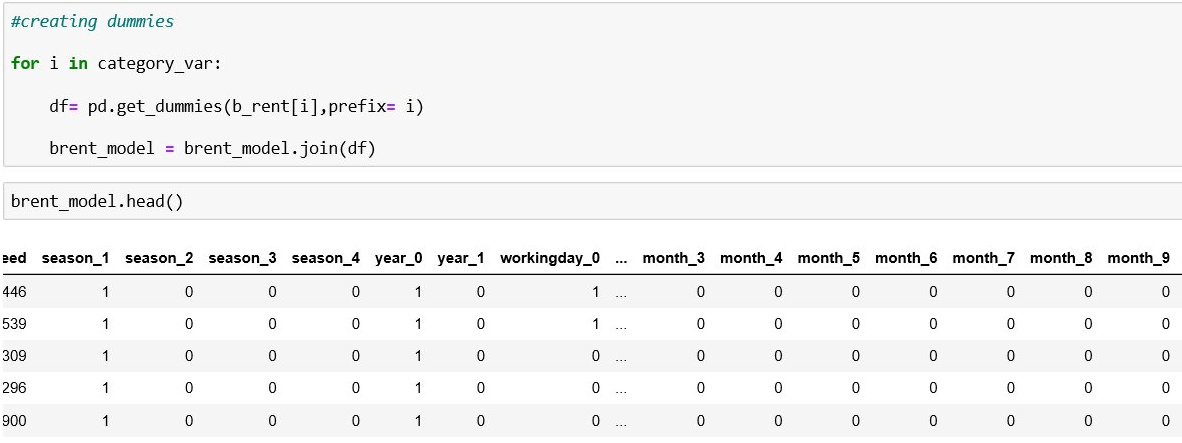
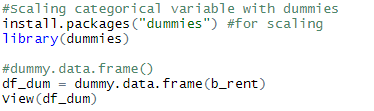


Fig. 2.11

And the rest all categorical variables should be stored as dummies so that they can be good in prediction model.



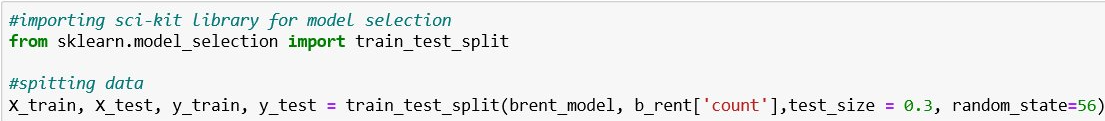
In R:



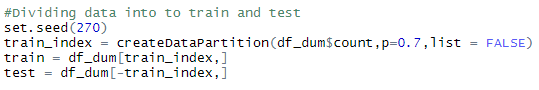
**2.2 Modelling:**

In modelling we first have to split the clean dataset to train-set and test-set and then develop different models and evaluate them by metrics.

In Python:

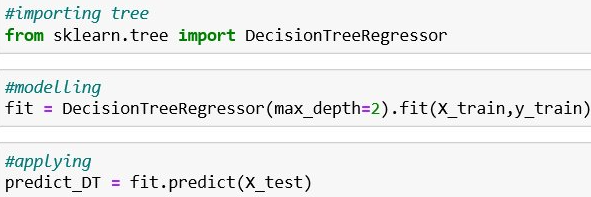


In R:

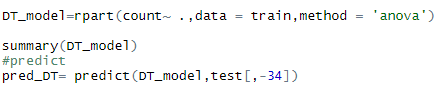


**2.2.1 Decision Tree:**

In Python:

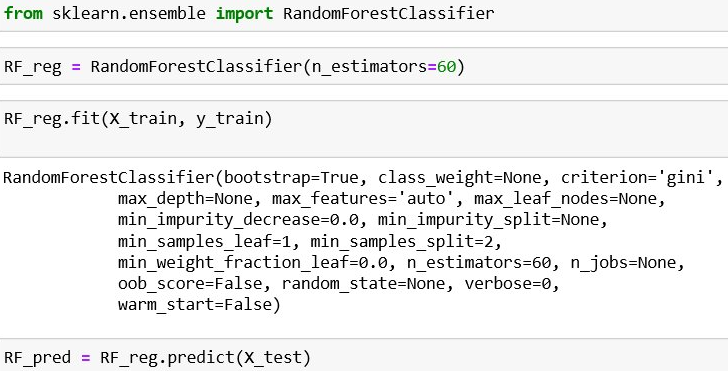


In R:

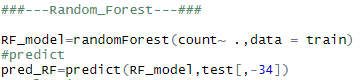


**2.2.2 Random Forest:**

In Python:

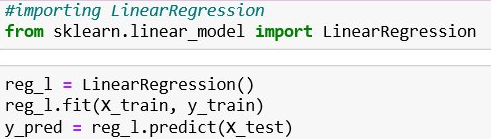


In R:

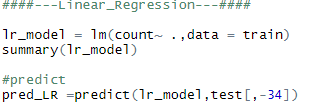


**2.2.3 Linear Regression:**

In Python:



In R:



# Chapter-3

## CONCLUSION

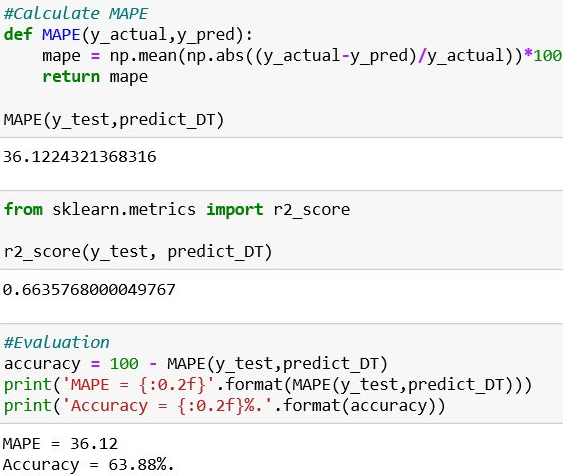
**3.1 Model Evaluation:**

For Evaluating a regression model we have regression metrics they are :-

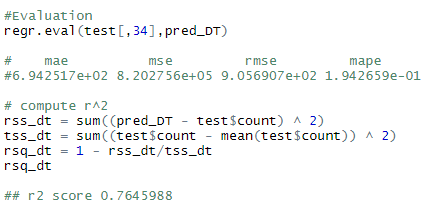
**3.1.1 Mean Absolute Percentage Error (MAPE)**

**3.1.2 Root Mean Squared Error (RMSE)**

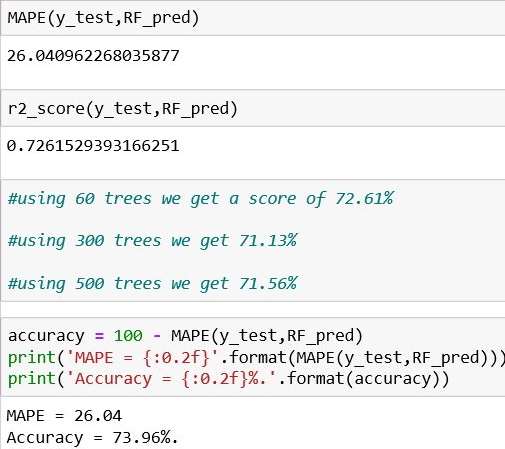
**Decision Tree:** (In Python)



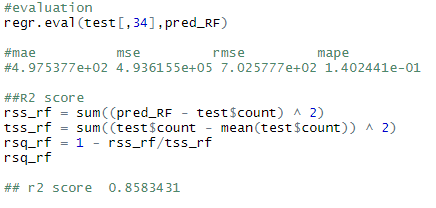
In R:



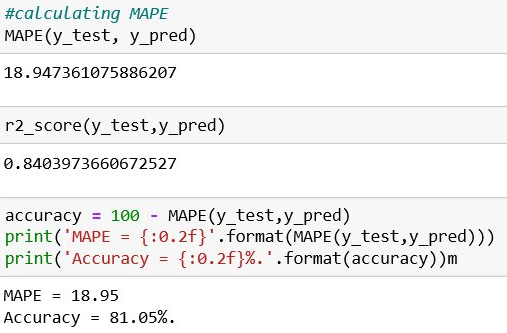
**Random Forest:** (In Python)



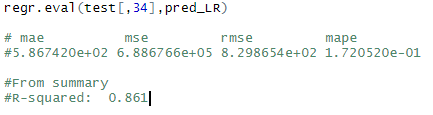
In R:



**Linear Regression: (In Python)**



In R:



**3.2 Model Selection:**

From above all,

Linear Regression better for predicting rental counts as it has better accuracy of 81.05% with r2\_score of 0.84.

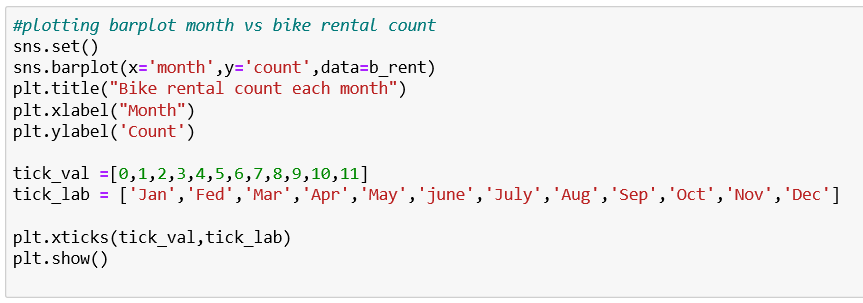
**Appendix A:**

Python Code for figures

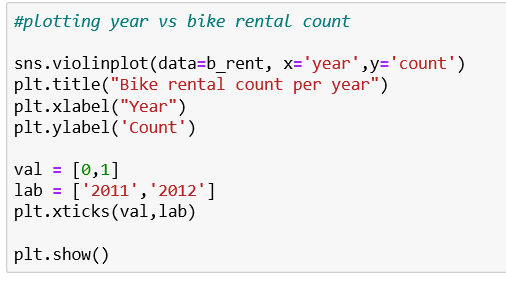
**Fig. 2.1**



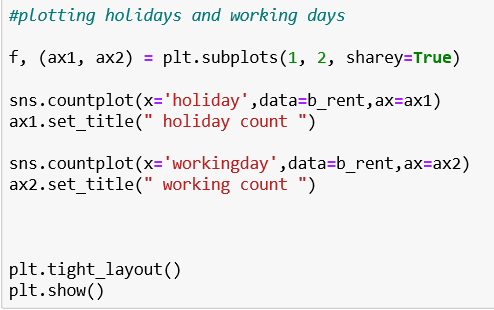
**Fig. 2.2**



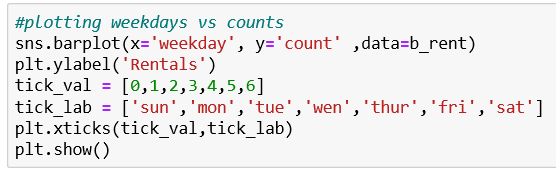
**Fig 2.3**



**Fig. 2.4**



**Fig 2.5**



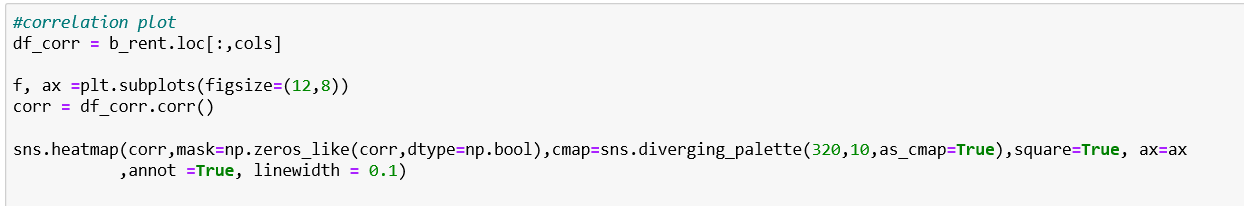
**Fig 2.6**



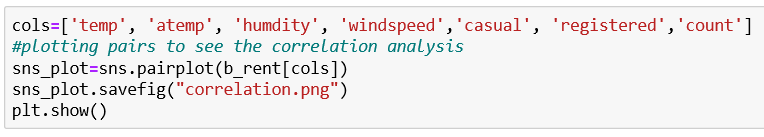
**Fig. 2.7**



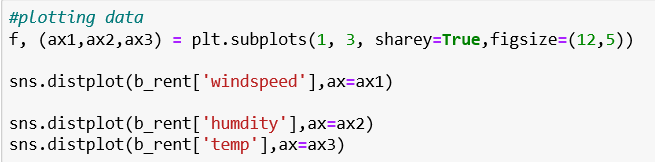
**Fig. 2.8**



**Fig. 2.9**



**Fig. 2.10**



**References:**

For calculating r2 score in R:

<https://stackoverflow.com/questions/40901445/function-to-calculate-r2-r-squared-in-r>

Various types of plots in seaborn:

<https://seaborn.pydata.org/tutorial/categorical.html>

For performing Ridge Regression:

<https://www.datacamp.com/community/tutorials/tutorial-ridge-lasso-elastic-net>

For dummies in R:

<https://www.r-bloggers.com/r-the-dummies-package/>