BIKE SHARING ASSIGNMENT

SUBMISSION

Name: Avishek Sen Gupta

Boom Bikes

Introduction

A bike-sharing system is a service in which bikes are made available for shared use to individuals on a short term basis for a price or free. Many bike share systems allow people to borrow a bike from a "dock" which is usually computer-controlled wherein the user enters the payment information, and the system unlocks it. This bike can then be returned to another dock belonging to the same system.

BoomBikes aspires to understand the demand for shared bikes among the people after this ongoing quarantine situation ends across the nation due to Covid-19. They have planned this to prepare themselves to cater to the people's needs once the situation gets better all around and stand out from other service providers and make huge profits.

Business Understanding and Domain : CRISP-DM #1

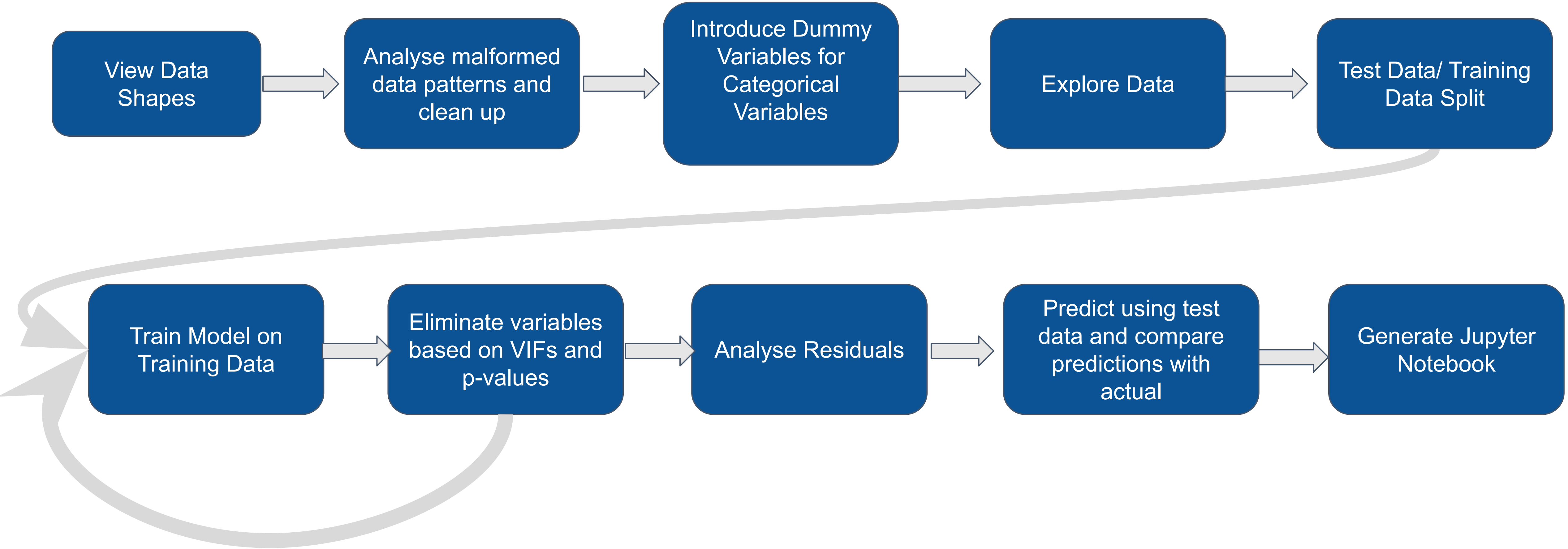
Boom Bikes wants to know:

- Which variables are significant in predicting the demand for shared bikes.
- How well those variables describe the bike demands

Objective

The objective is to model the demand for shared bikes with the available independent variables. It will be used by the management to understand how exactly the demands vary with different features.

Problem Solving Workflow



Data Understanding: CRISP-DM #2

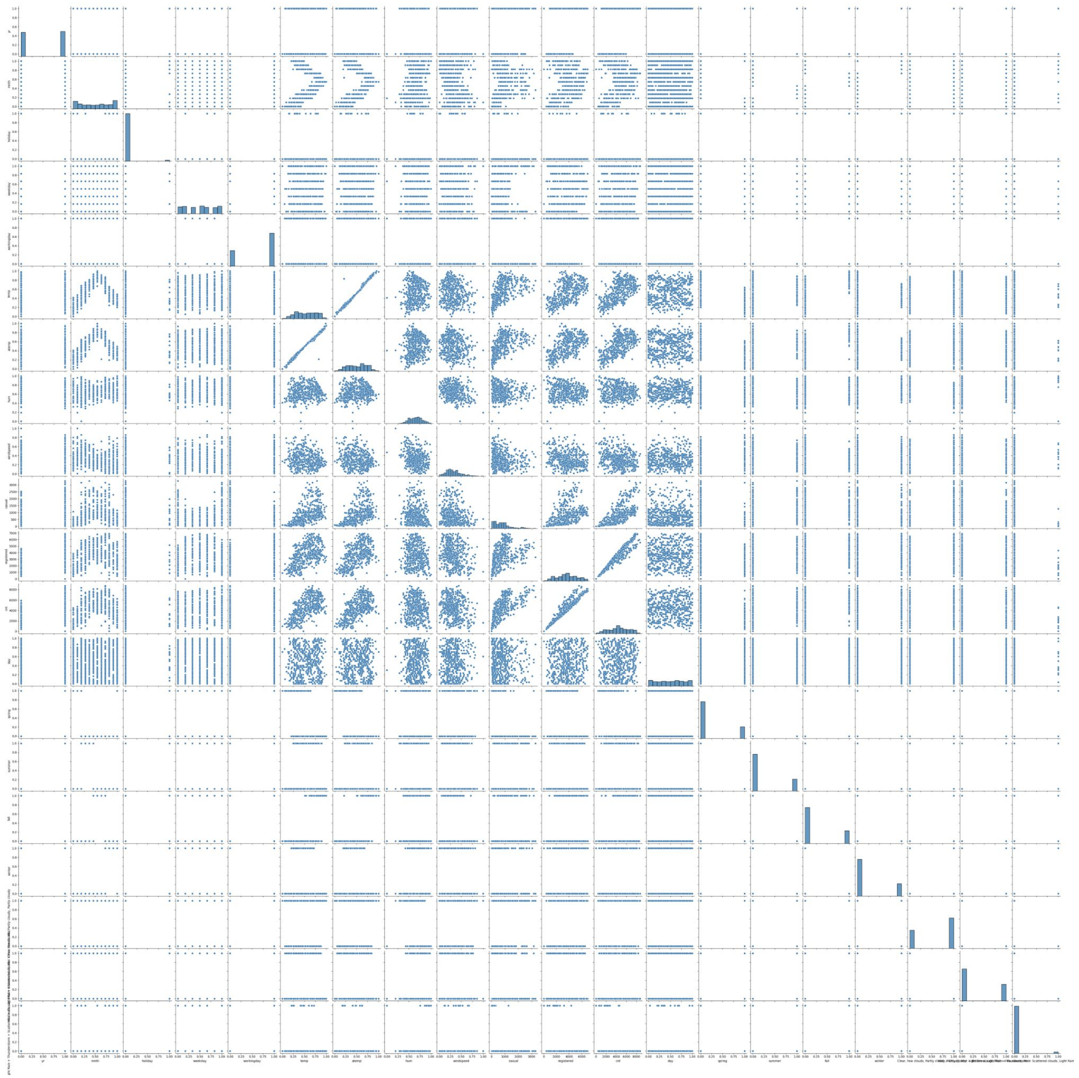
To understand the shape of the data and its cleanliness, we perform a couple of preliminary checks on the *day* dataset.

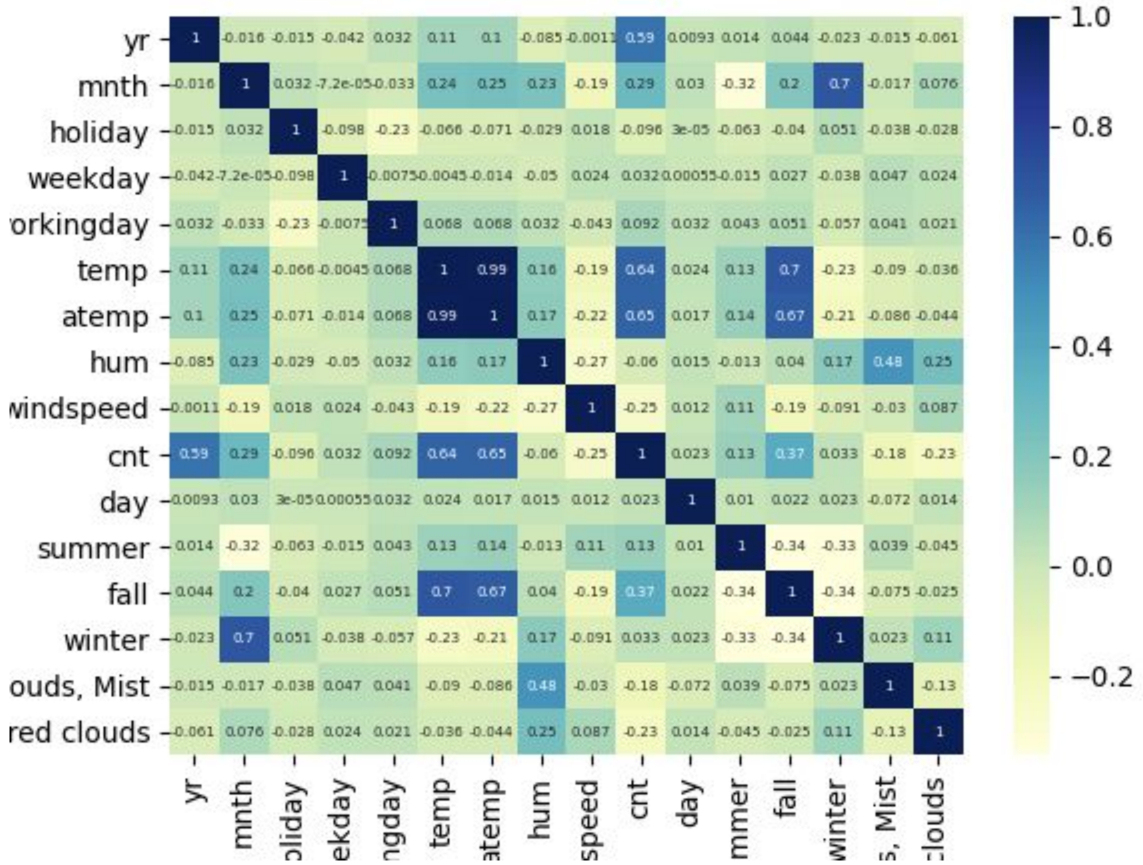
- There were **no null or empty entries** found. Thus, no imputation is necessary.
- One interesting point to note is that the data point **"Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog"** was not present in any of the data vectors.
- The included **data dictionary** provided domain understanding of the various columns.

These are the patterns observed in the malformed data:

- **Registered** and **Casual** counts should not be used to determine total demand, because they will probably be heavily correlated with the total demand.
- Separate analysis needs to be done to determine dependency of these values on the factors, and since total demand is the target variable in question, we should not use these as predictor variables.

Data Understanding: CRISP-DM #2





Data Understanding: CRISP-DM #2

These are some of the observations from doing some Exploratory Data Analysis on the set:

- *cnt* has a definite positive correlation with *temp*, *atemp*
  cnt has a definite positive correlation with *casual*
- *cnt* has a strong positive correlation with *registered*
- *cnt* has a correlation with `mnth`, but it is not linear
- If it's a holiday, *cnt* seems to be lower, considering higher percentiles

Of course, these are all from visual inspection, and the actual metrics gathered from building the **Linear Model** will provide further insight.

Data Preparation: CRISP-DM #3

Initial Data Preparation

- Categorical variables were replaced with dummy variables. Specifically, the columns *weathersit*, *season* columns were converted. The original columns were not used in the analysis.
- The dummy variables so obtained were renamed to be more understandable.
- The *registered* and *casual* columns were dropped because of reasons explained in the previous slide.
- The day of the month was extracted from the *daydte* column.

Test Data / Training Data Split

The testing data and training data were obtained by splitting the master data set, with the training set size being 70% of the total set.

Scaling

The Training data set was scaled to bring variables to the same scale. Specifically, *temp*, *atemp*, *hum*, *windspeed*, *day*, *dayofweek*, and *mnth* were scaled.

Modelling: CRISP-DM #4

```
------------------------------------------------------------------------------------------------------
Model Summary
------------------------------------------------------------------------------------------------------
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.832
Model:                            OLS   Adj. R-squared:                  0.828
Method:                 Least Squares   F-statistic:                     224.5
Date:                Wed, 07 Jul 2021   Prob (F-statistic):          5.38e-185
Time:                        19:12:31   Log-Likelihood:                -4132.2
No. Observations:                 510   AIC:                             8288.
Df Residuals:                     498   BIC:                             8339.
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================================================================================================
                                                                                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------------------------------------------------------------
const                                                                               1632.3824    250.899      6.506      0.000    1139.431    2125.333
yr                                                                                  2019.4785     72.856     27.719      0.000    1876.336    2162.622
holiday                                                                             -667.7096    229.622     -2.908      0.004   -1118.856    -216.563
weekday                                                                              415.2930    108.047      3.844      0.000     203.009     627.577
temp                                                                                4298.7699    296.976     14.475      0.000    3715.289    4882.251
hum                                                                                -1082.7695    337.483     -3.208      0.001   -1745.835    -419.704
windspeed                                                                          -1583.3941    230.598     -6.866      0.000   -2036.459   -1130.329
summer                                                                              1025.7656    132.582      7.737      0.000     765.276    1286.255
fall                                                                                 647.1607    177.921      3.637      0.000     297.593     996.729
winter                                                                              1425.0695    113.168     12.593      0.000    1202.725    1647.414
Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist                        -501.4583     94.627     -5.299      0.000    -687.376    -315.541
Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds -2150.3118    236.521     -9.091      0.000   -2615.013   -1685.610
==============================================================================
Omnibus:                       77.673   Durbin-Watson:                   2.034
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              187.954
Skew:                          -0.787   Prob(JB):                     1.54e-41
Kurtosis:                       5.523   Cond. No.                         20.0
==============================================================================
```

Modelling: CRISP-DM #4

```
-----------------------------------------------------------------------------------------------------
Variance Inflation Factors
-----------------------------------------------------------------------------------------------------
                                                                                     Features    VIF
0                                                                                       const  49.11
8                                                                                        fall   4.78
4                                                                                        temp   3.50
7                                                                                      summer   2.54
5                                                                                         hum   1.89
9                                                                                      winter   1.87
10                              Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist   1.57
11  Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds   1.25
6                                                                                   windspeed   1.19
1                                                                                          yr   1.03
```

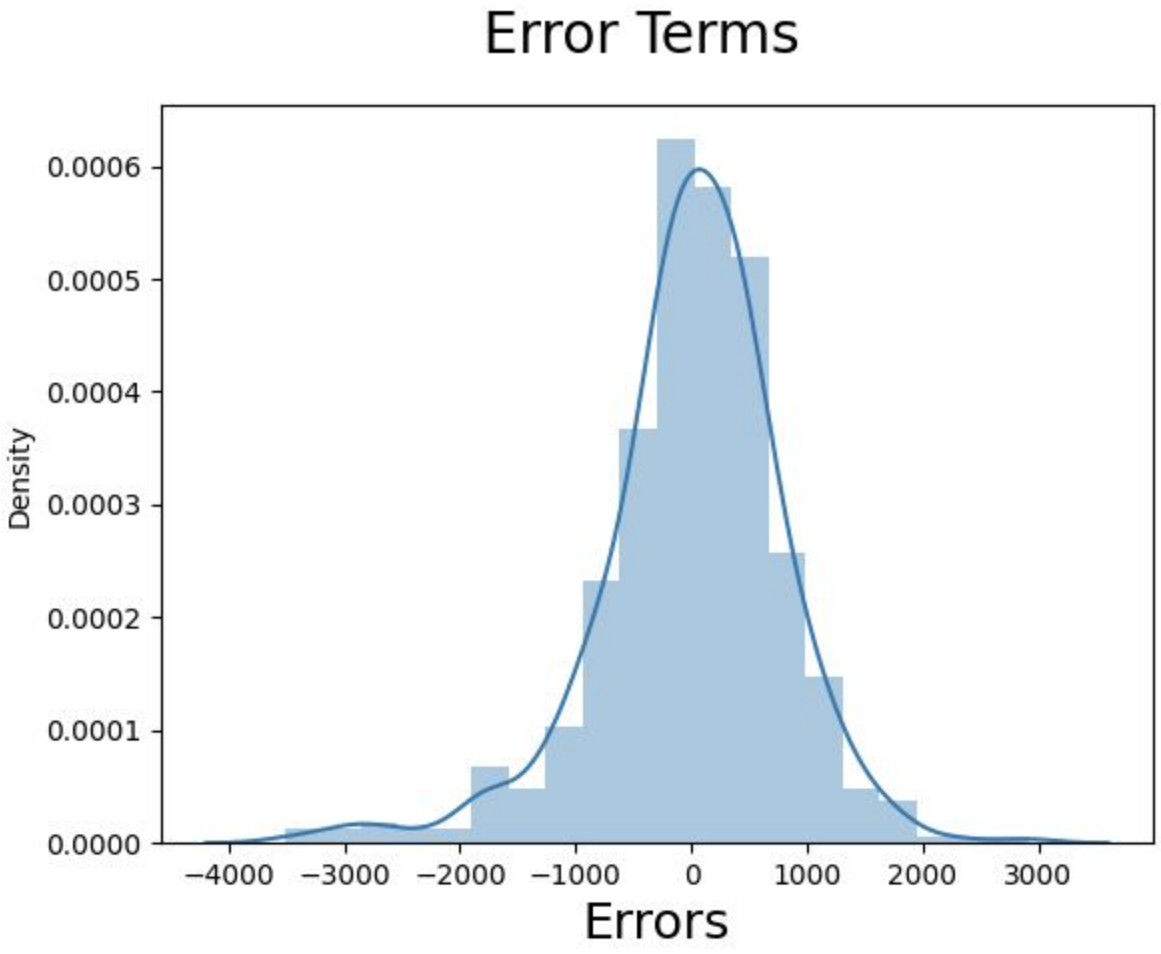
Notes

- The following predictor variables were dropped in the following order:
    - *atemp*
    - *mnth*
    - *day*
    - *workingday*

At each step, the Variance Inflation Factors were calculated, and compared against the p-values of the predictor variables, to make a decision on which variables needed to be dropped.

The final model explains 83% of the variation of the observed data, as can be seen from the **R-Squared** and the **Adjusted R-Squared** metrics.

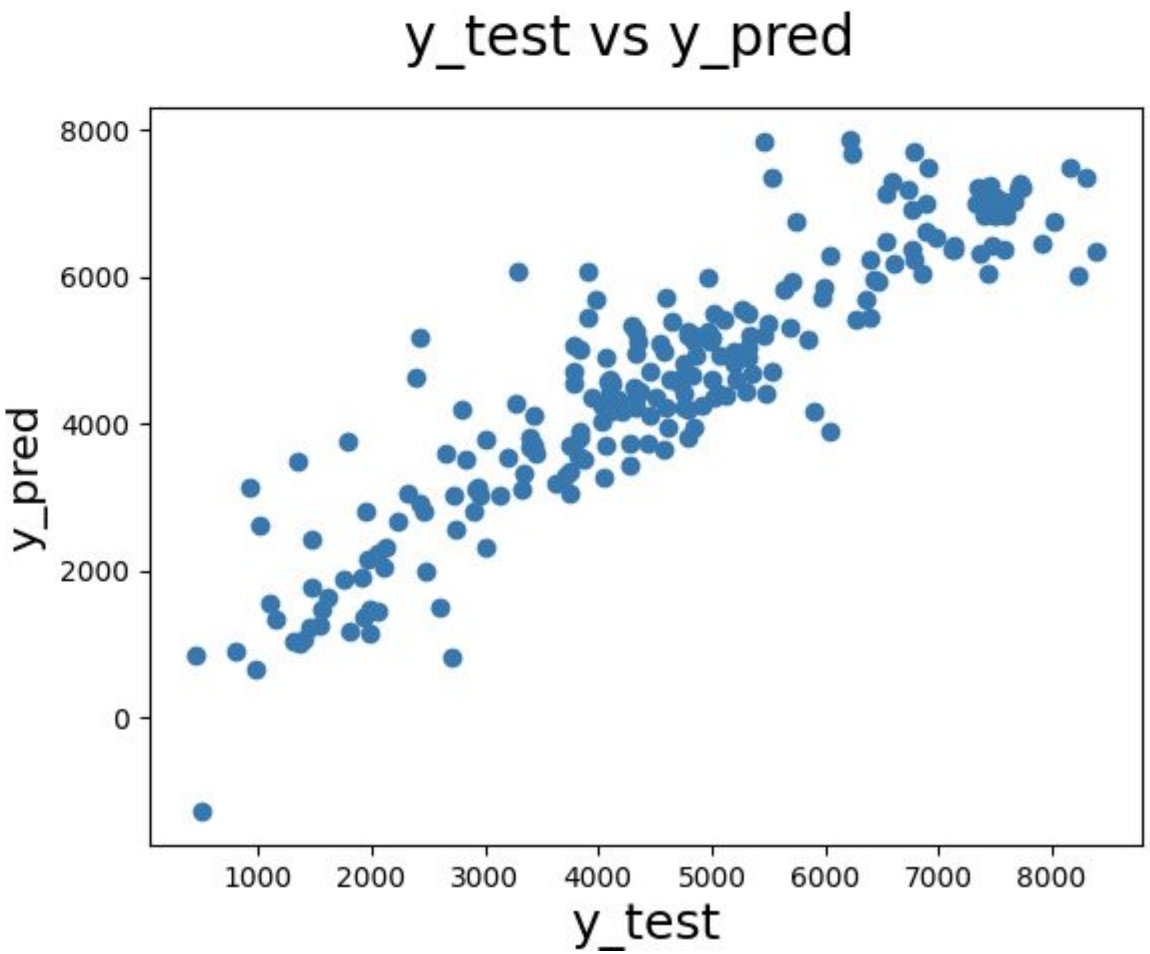
Evaluation: CRISP-DM #5



For model evaluation, the error terms were charted in a histogram for the Training set, to ensure that the residuals are **homoscedastic**.

As is noted from the graph, **the error terms form a good approximation to a normal distribution**, and thus, we can conclude that they are normally distributed.

Evaluation: CRISP-DM #5



For further model validation, **predictions** were made using the Linear Model on the **Test data set**, and the predicted values vs. the actual values were graphed.

As is noted from the graph, the **y_pred** and the **y_test** values follow approximately a 45-degree line, indicating that the **model can generalise well to the Test data set**.

Conclusions

This summarises the **linear relationship** obtained between **total demand** and the relevant **predictor variables**.

rental_bike_demand = 1632.382×**C** + 2019.478×**year** − 667.709×**holiday** + 415.293×**weekday** + 4298.769 ×**temperature** −1082.769×**humidity** − 1583.394×**windspeed** + 1025.765×**summer** + 647.160×**fall** + 1425.0694 ×**winter** − 501.458×**WEATHER_2** − 2150.311×**WEATHER_3**

where:
WEATHER_2 = "Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist)"
WEATHER_3 = "Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds"
and the other variables are appropriately scaled.