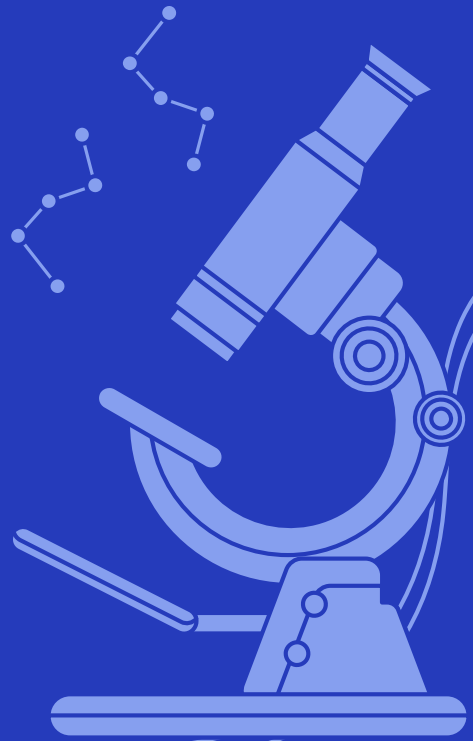
BoomBikes: Linear Regression Analysis

By Prem Kumar Subudhi



Problem Statement

Boom Bikes going loss because of Covid and finding difficult to sustain, need to come up with a business plan to recover from the current situation and book profit in future.

Task

Find the factors on which the demand for these shared bikes depends. Specifically, they want to understand the factors affecting the demand for these shared bikes in the American market. The company wants to know the variables those are significant in predicting the demand for shared bikes and how well those variables describe the bike demands.

The Goal of problem statement

1. Data Understanding and loading, drop un-necessary features that does not contribute for the model.

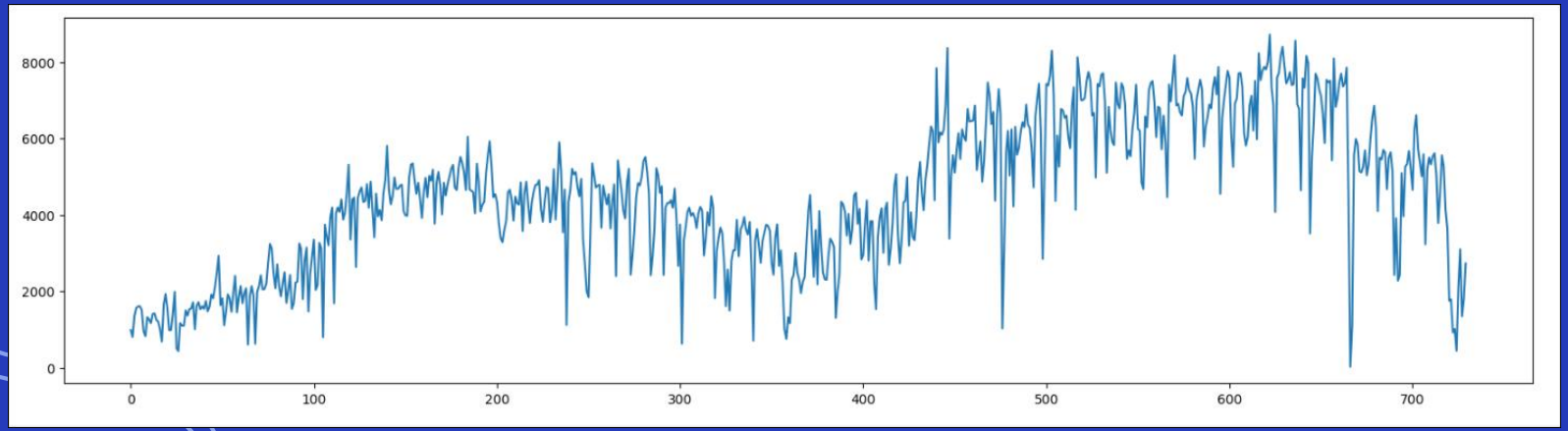
2. Create a linear model with all the features that may contribute or effect the price (i.e CNT feature)

3. The model should be interpretable and easily understandable.

Linear regression model

Simple / multiple linear regression

The feature Cn has
total rental bikes including both casual and registered, which is continuous in nature as shown in below diagram.



Linear regression model evaluation

The measure R-squared

Mercury is the closest planet to the Sun and **the smallest one in the entire Solar System**

Adjusted R-squared

Venus has a beautiful name and is **the second planet from the Sun**. It's hot and its atmosphere is poisonous

Residual analysis

Earth is the third planet from the Sun and **the only one that harbors life** in the Solar System

Assumptions of linear regression

Linearity & normality

The change in independent variable increase proportionately with independent variables. If the residuals are not normally distributed, it might affect the accuracy of confidence intervals and hypothesis tests

Independence

All the features used in the model are assumed to be independent with each other, otherwise the model will introduce errors.

Homoscedasticity

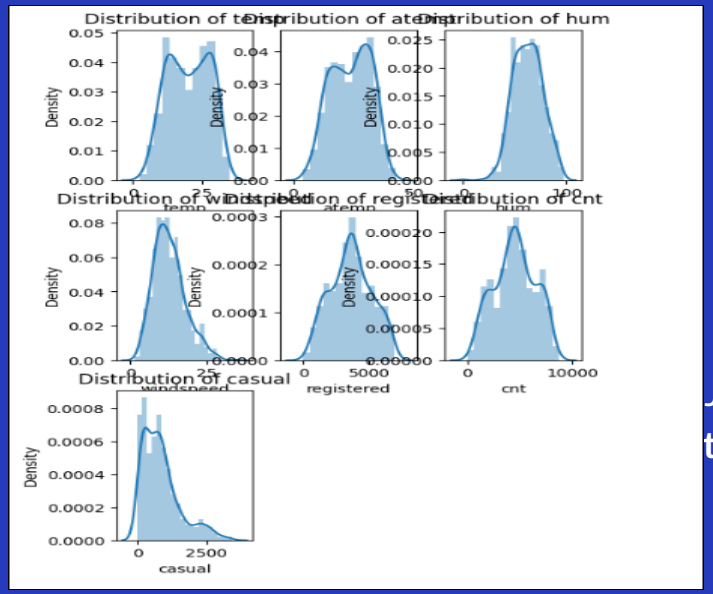
When this assumption is met, it implies that the model's predictions are equally accurate across all levels of the independent variable(s).

No multicollinearity

Above independence nature leaded to non multicollinearity, the model will become weak when it has multi-collinearity introduced.

Distribution Plot

Visualizations



Analysis

Despite being red, Mars is **actually a cold place.** It is full of iron oxide dust

Patterns

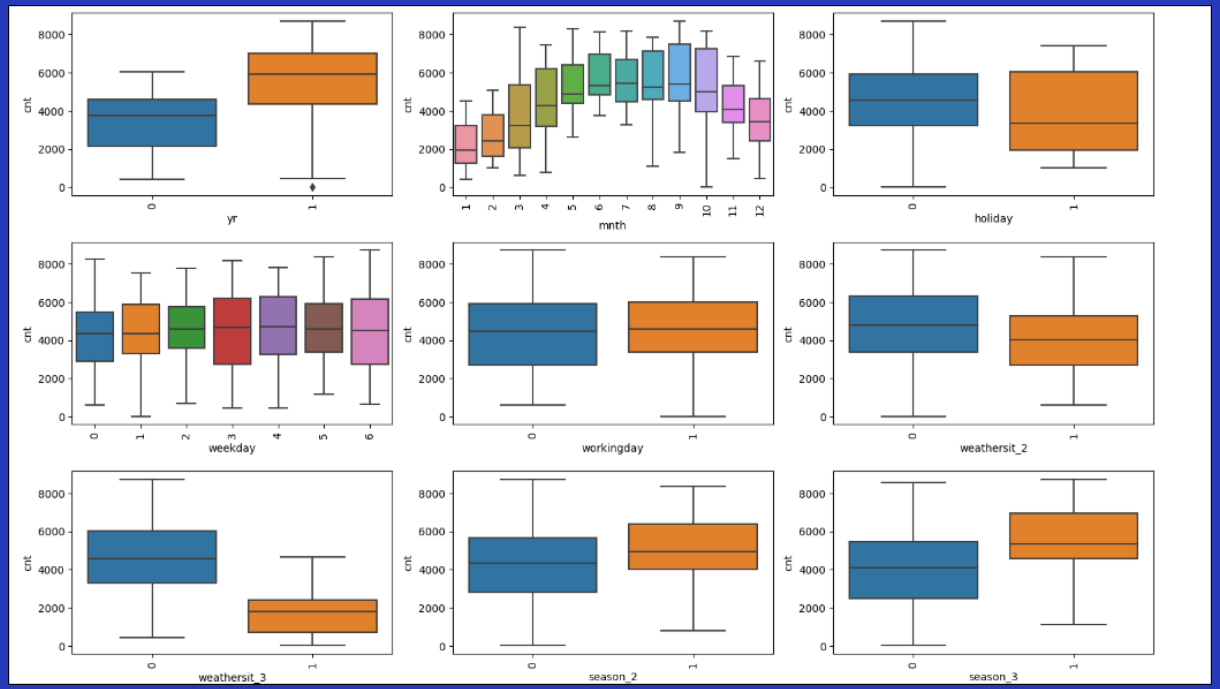
Earth is the third planet from the Sun and **the only one that harbors life**

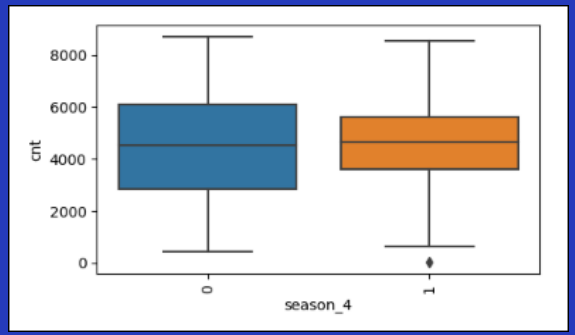
Assessment

Jupiter is a gas giant and **the biggest planet** in the entire Solar System

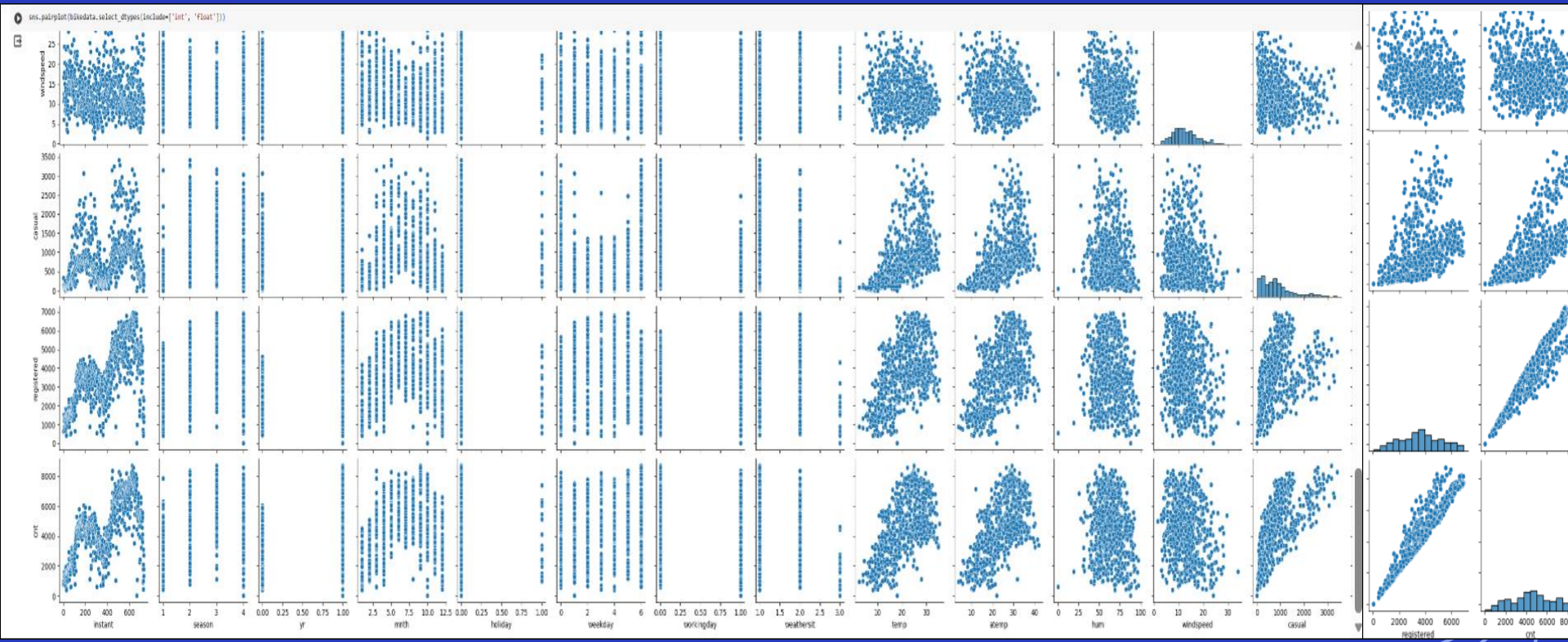
Data splits

Created dummy variables for Categorical variables

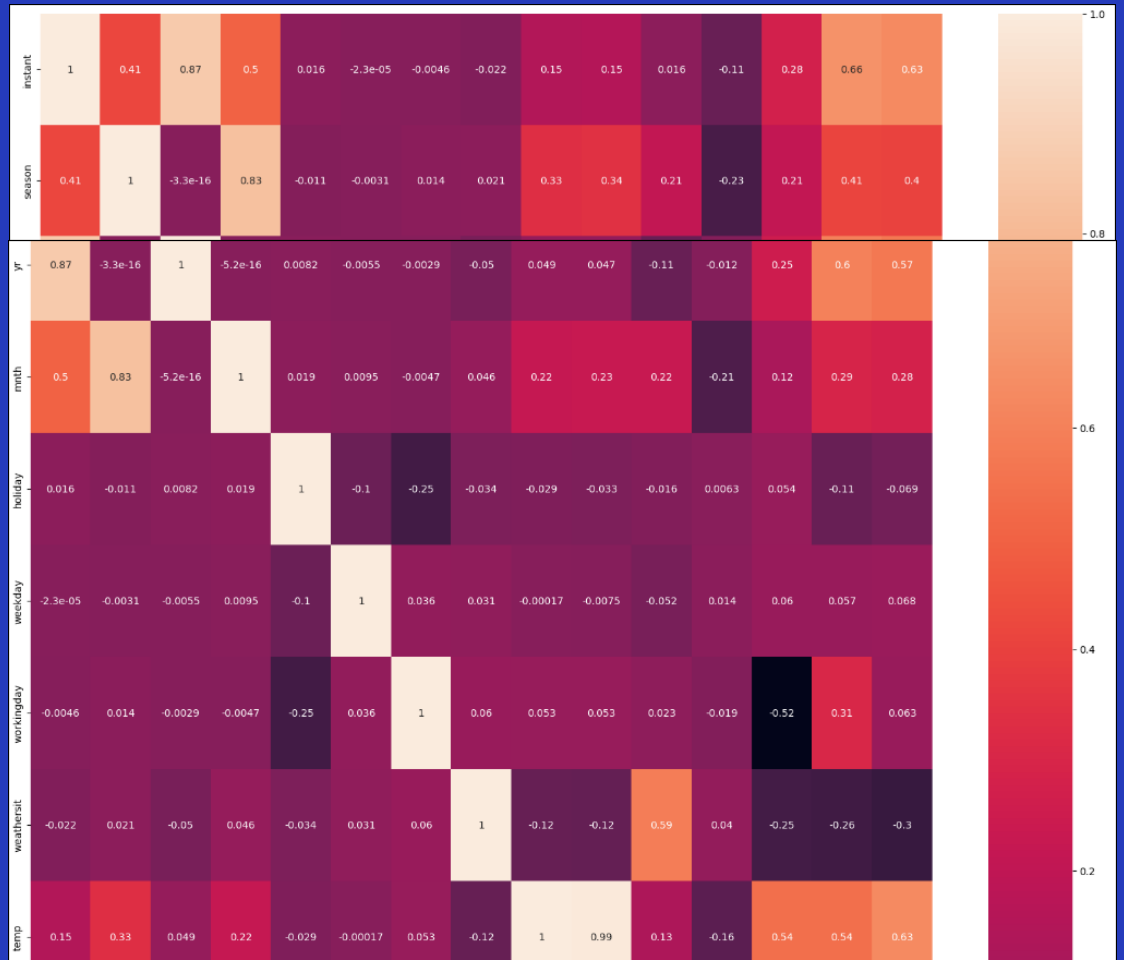

cnt
yr
mnth
holiday
weekday
workingday
weathersit_2
weathersit_3
season_2
season_3


cnt
season_4

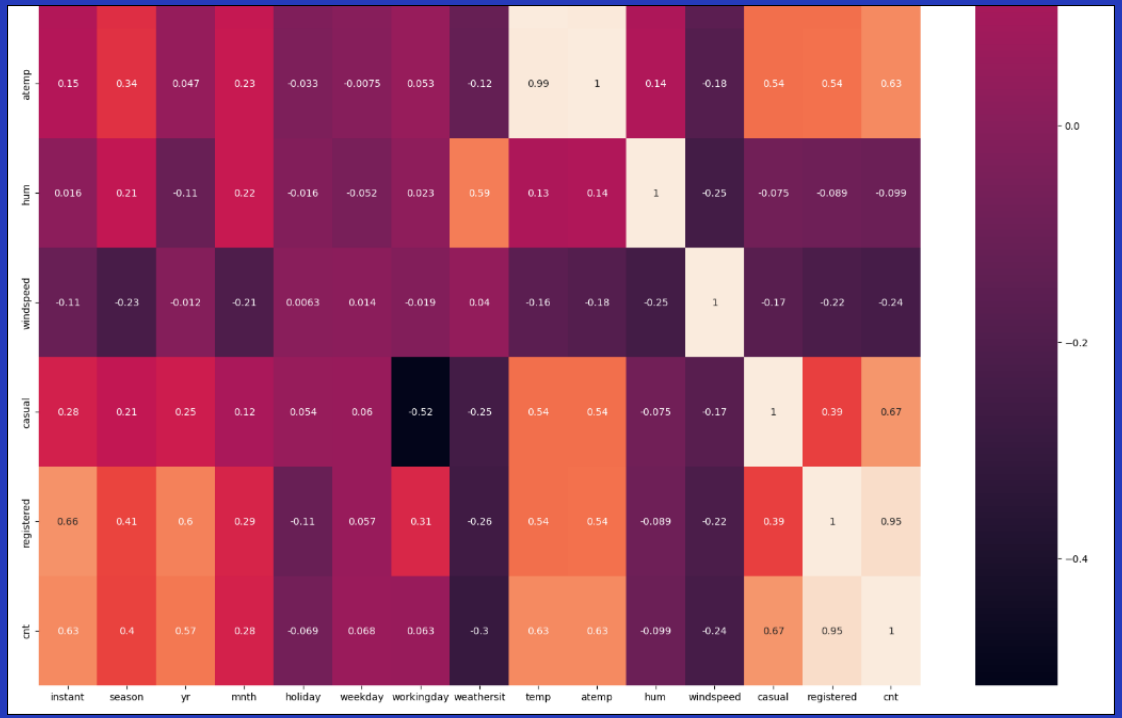
Seaborne Pairplot



Heat Map


instant
season
yr
mnth
holiday
weekday
workingday
weathersit
temp
1 0.41 0.87 0.5 0.016 -2.3e-05 -0.0046 -0.022 0.15 0.15 0.016 -0.11 0.28 0.66 0.63
0.41 1 -3.3e-16 0.83 -0.011 -0.0031 0.014 0.021 0.33 0.34 0.21 -0.23 0.21 0.41 0.4
0.87 -3.3e-16 1 -5.2e-16 0.0082 -0.0055 -0.0029 -0.05 0.049 0.047 -0.11 -0.012 0.25 0.6 0.57
0.5 0.83 -5.2e-16 1 0.019 0.0095 -0.0047 0.046 0.22 0.23 0.22 -0.21 0.12 0.29 0.28
0.016 -0.011 0.0082 0.019 1 -0.1 -0.25 -0.034 -0.029 -0.033 -0.016 0.0063 0.054 -0.11 -0.069
-2.3e-05 -0.0031 -0.0055 0.0095 -0.1 1 0.036 0.031 -0.00017 -0.0075 -0.052 0.014 0.06 0.057 0.068
-0.0046 0.014 -0.0029 -0.0047 -0.25 0.036 1 0.06 0.053 0.053 0.023 -0.019 -0.52 0.31 0.063
-0.022 0.021 -0.05 0.046 -0.034 0.031 0.06 1 -0.12 -0.12 0.59 0.04 -0.25 -0.26 -0.3
0.15 0.33 0.049 0.22 -0.029 -0.00017 0.053 -0.12 1 0.99 0.13 -0.16 0.54 0.54 0.63
1.0
0.8
0.6
0.4
0.2

Heat Map


atemp
hum
windspeed
casual
registered
cnt
0.15 0.34 0.047 0.23 -0.033 -0.0075 0.053 -0.12 0.99 1 0.14 -0.18 0.54 0.54 0.63
0.016 0.21 -0.11 0.22 -0.016 -0.052 0.023 0.59 0.13 0.14 1 -0.25 -0.075 -0.089 -0.099
-0.11 -0.23 -0.012 -0.21 0.0063 0.014 -0.019 0.04 -0.16 -0.18 -0.25 1 -0.17 -0.22 -0.24
0.28 0.21 0.25 0.12 0.054 0.06 -0.52 -0.25 0.54 0.54 -0.075 -0.17 1 0.39 0.67
0.66 0.41 0.6 0.29 -0.11 0.057 0.31 -0.26 0.54 0.54 -0.089 -0.22 0.39 1 0.95
0.63 0.4 0.57 0.28 -0.069 0.068 0.063 -0.3 0.63 0.63 -0.099 -0.24 0.67 0.95 1
instant
season
yr
mnth
holiday
weekday
workingday
weathersit
temp
atemp
hum
windspeed
casual
registered
cnt
0.0
−0.2
−0.4

Linear regression analysis: key steps

Data preparation

Convert the Categorical variables to dummy variables and remove the first column

Step 01

Step 02

Model building

Regresssion model Y=constant+sumof(Beta0*In dependent variables)

Model evaluation

R2_Score of the model expected to be high

Step 03

Step 04

Results

Predicted result from the model

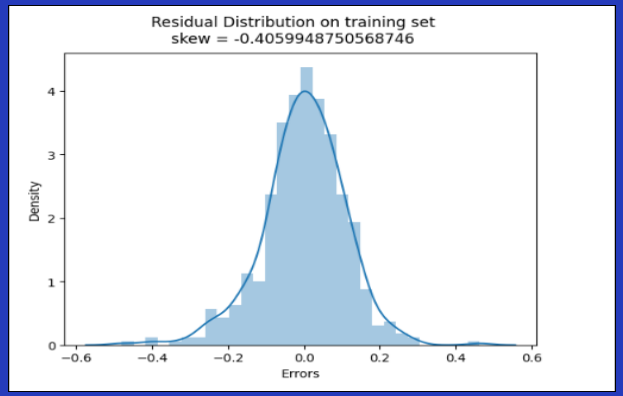
Linear regression key information

Aspect of linear regression	Description/Information
Purpose	Predictive modeling, understanding relationships, hypothesis testing, etc.
Assumptions	Linearity, independence, homoscedasticity, normality, no multicollinearity
Model fit metrics	R-squared, adjusted R-squared, F-statistic, p-values
Data preparation	Collect, clean, and preprocess data. Handle missing values and outliers
Exploratory Data Analysis (EDA)	Visualize data through scatter plots, histograms, correlation matrices

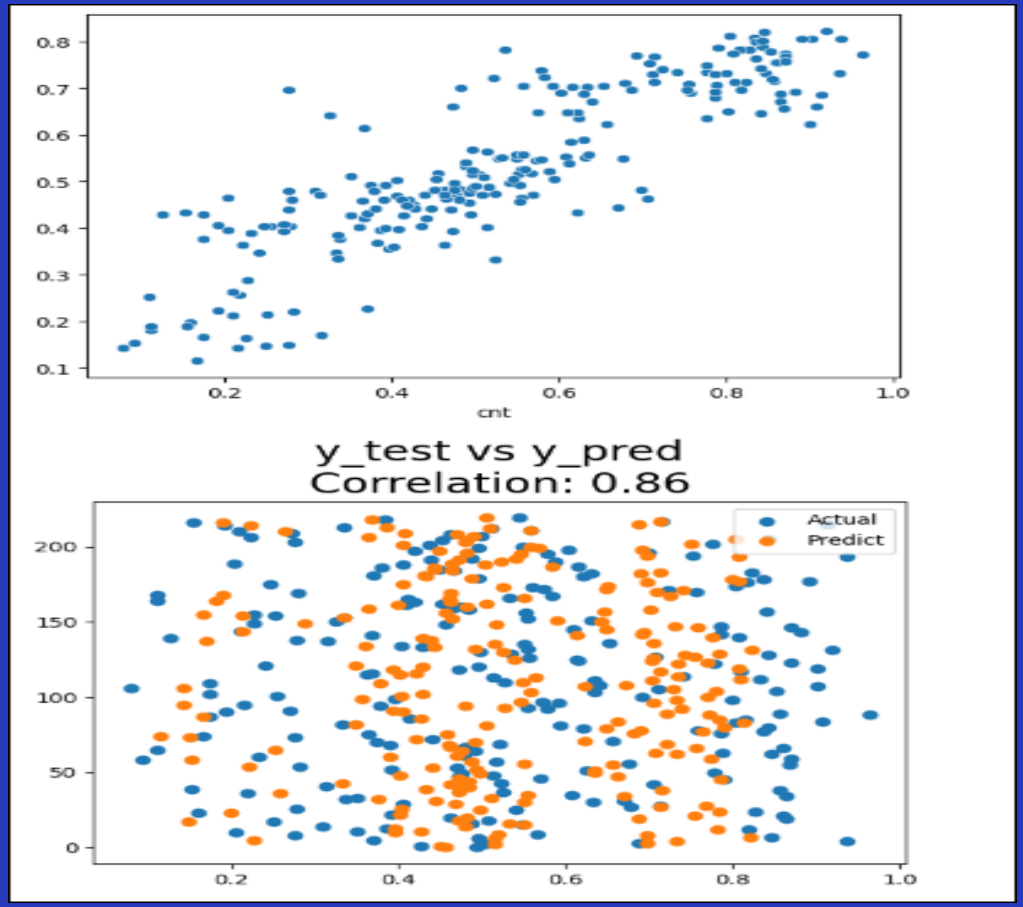
Diagnostic tools for model validation

Residual analysis

Residual plot



scatter diagram for the test set vs predicted.

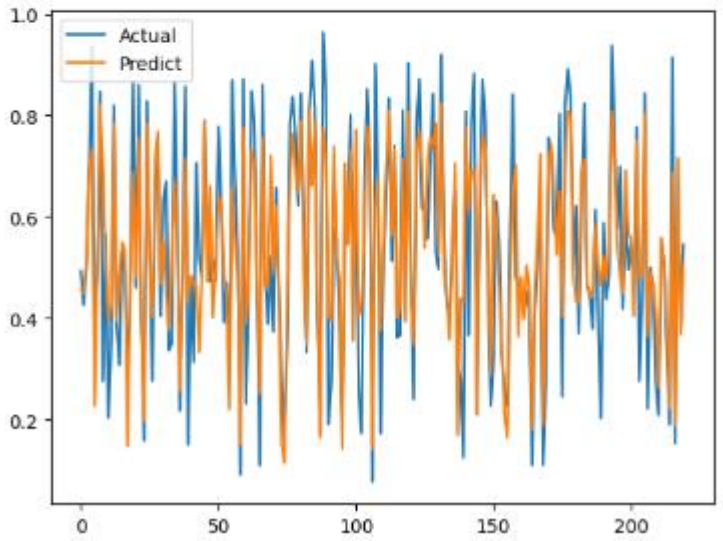


Predicted values from the model

Plot the CNT test-actual vs test-predicted

```
plt.plot(list(y2_test),label='Actual')
plt.plot(list(y2_test_pred), label='Predict')
plt.legend()
```

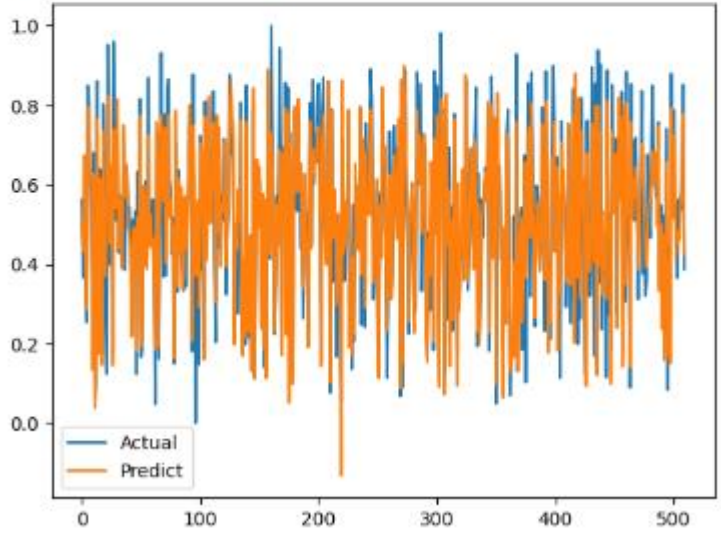
```
<matplotlib.legend.Legend at 0x78b504173610>
```



Plot the CNT trained-actual vs trained-predicted

```
plt.plot(list(y1_train),label='Actual')
plt.plot(y1_train_pred, label='Predict')
plt.legend()
```

```
<matplotlib.legend.Legend at 0x78b510b90a00>
```



The model created :

CNT = 0.26 + .24 * yr + 0.04 * weekday -0.15 * windspeed + 0.27 * season_2 + 0.31 * season_3 + 0.23 * season_4 -0.07 * weathersit_2 + -0.28 * weathersit_3

(As holiday has high P-value it will be excluded.)

OLS vs XGBoost

Train R2 Value ::>> 0.919340607472633
Test R2 Value: 0.8516
Feature Importances:
yr 0.366905
mnth 0.112416
holiday 0.012894
weekday 0.010726
workingday 0.019876
atemp 0.208465
hum 0.039786
windspeed 0.021618
season_2 0.011544
season_3 0.004960
season_4 0.080601
weathersit_2 0.031936
weathersit_3 0.078272
dtype: float32

The XG boost method give the R2 value better than compared to OLS result.

End