BikeSharing Case Study

Member :
Jeeva

Table of Contents

1. Problem Statement
2. Data Description
3. Data Cleaning
4. EDA (Exploratory Data Analysis)
5. Date Preparation
6. Modeling
7. Conclusion

Problem Statement

A bike-sharing system is a service that provides bikes for shared use on a short-term basis, either for a fee or free of charge. The company wants to understand the factors affecting the demand for these shared bikes to optimize usage and increase profits.

Data Description

The US **BombBikes** Sharing dataset contains information about bike rental from **2018-2019**. It includes of **16 columns**, such as
The **temperature**,**date**, **number of rented bikes**, **weather conditions**, and other factors that may influence bike rental demand.
This dataset contains **730 rows** and **16 columns** of the data.

Data Description

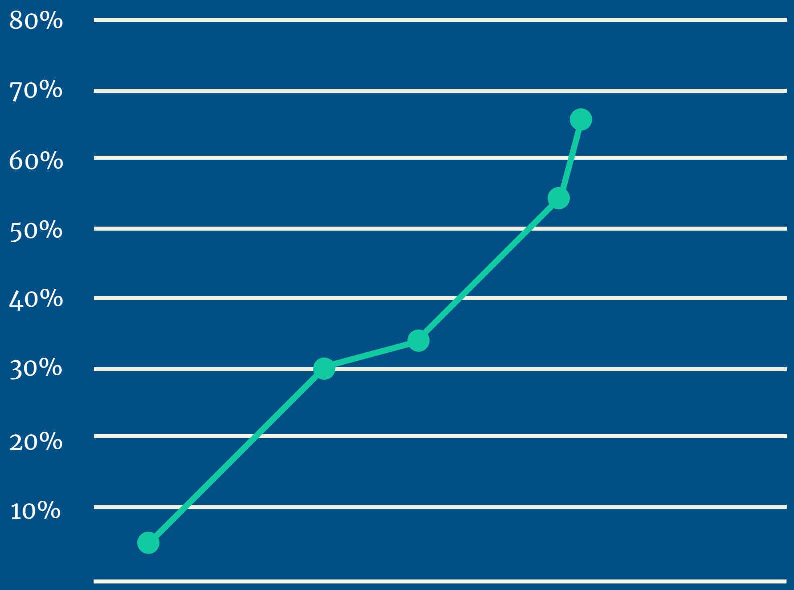
- Date :The date of the observation.
- Season : The season of the year when the observation was taken.
- Yr : The year of the observation.
- Month : The month of the observation.
- Holiday : Whether the day is a holiday or not while recording the observation.
- Weekday : Day of the week while recording the observation.
- Workingday : Whether the day is working day or not while recording the observation
- Weathesit : The weather conditions while recording the observation, whether it is heavily snowing, raining heavily, or experiencing a thunderstorm.
- Temp : The temperature in Celsius while recording the observation.
- Humidity: The Humidity while recording the observation.
- Windspeed : The speed of the wind while recording the observation.
- Casual : Count of the casual user.
- Registered : Count of the registered user.
- Cnt : Count of total rental bikes including both casual and registered.

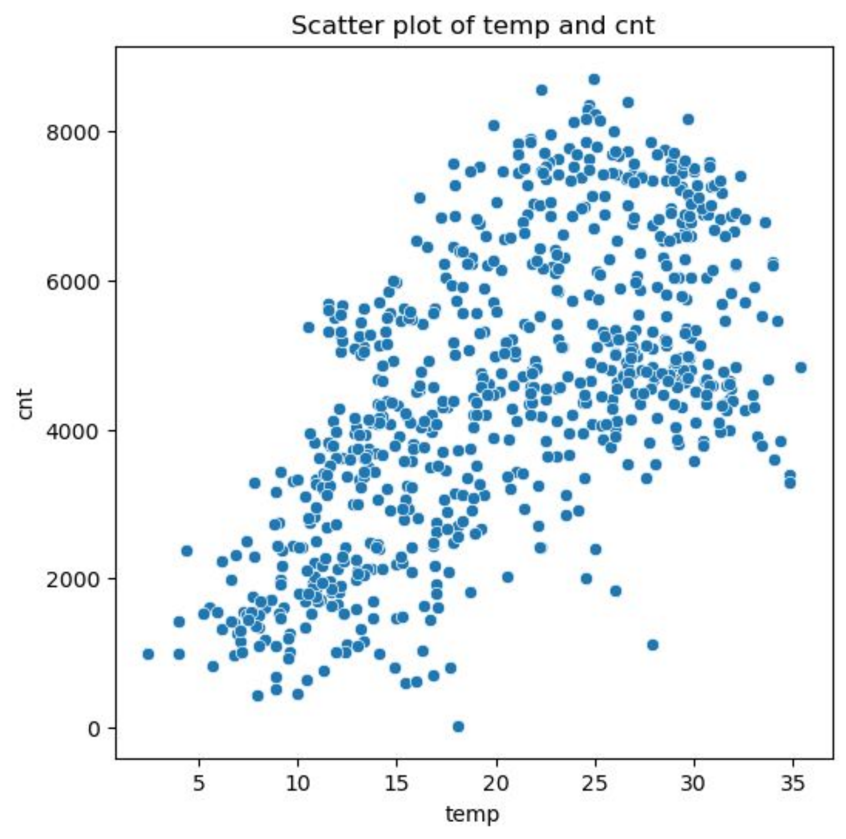
Data Cleaning

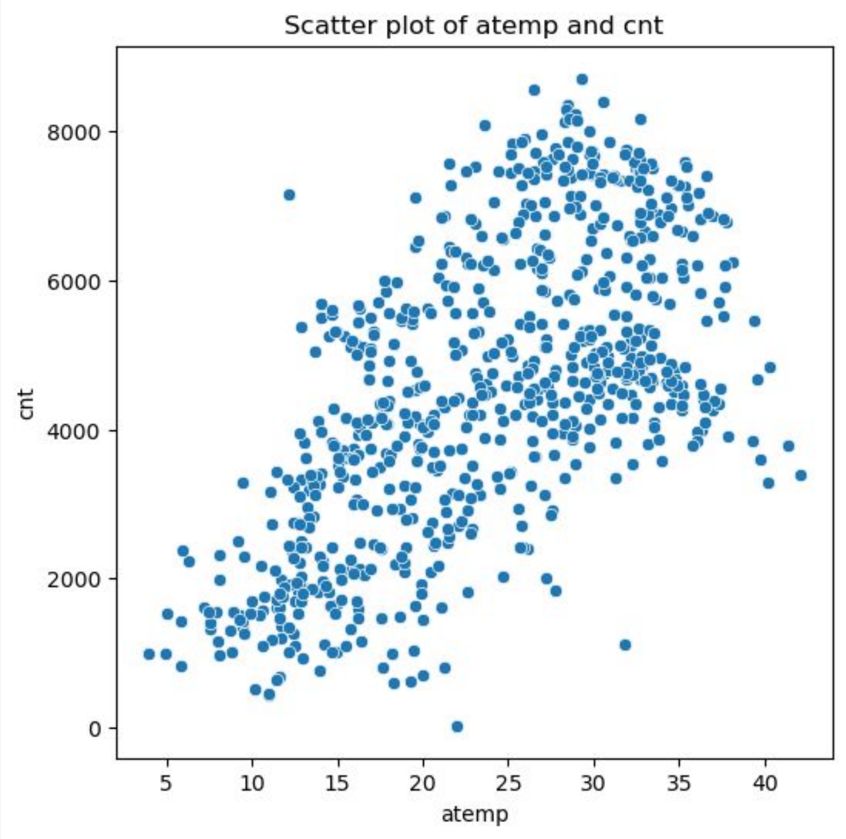
- There are no duplicate rows in the dataset.
- There are no missing values or Null values in the dataset.

Exploratory Data Analysis

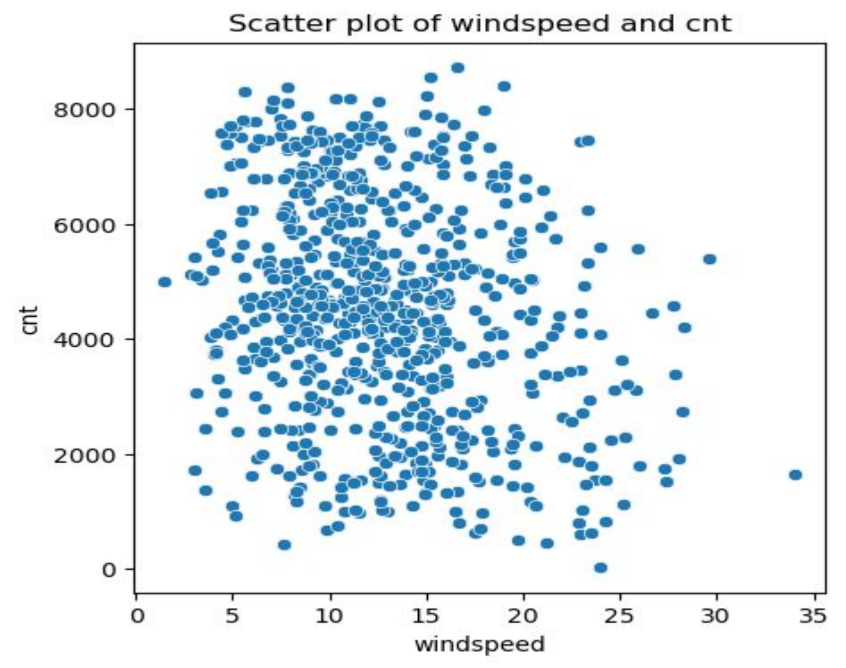
Analysing each column plotting the distribution of each columns





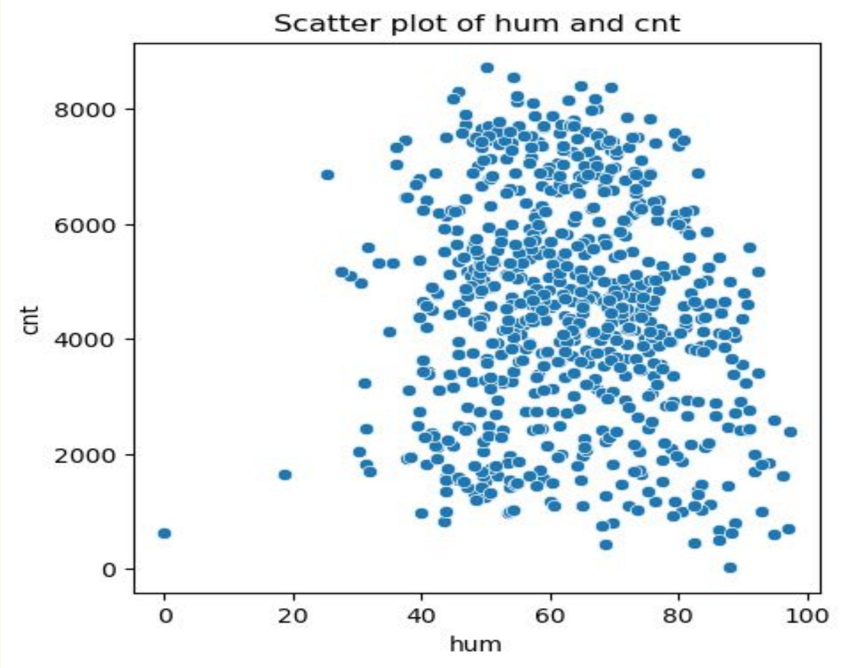


The Bike rental demand **increases** as the **temperature increases**.temp and atemp is almost having the similar distribution so we can drop either temp or atemp.



Wind Speed vs Count

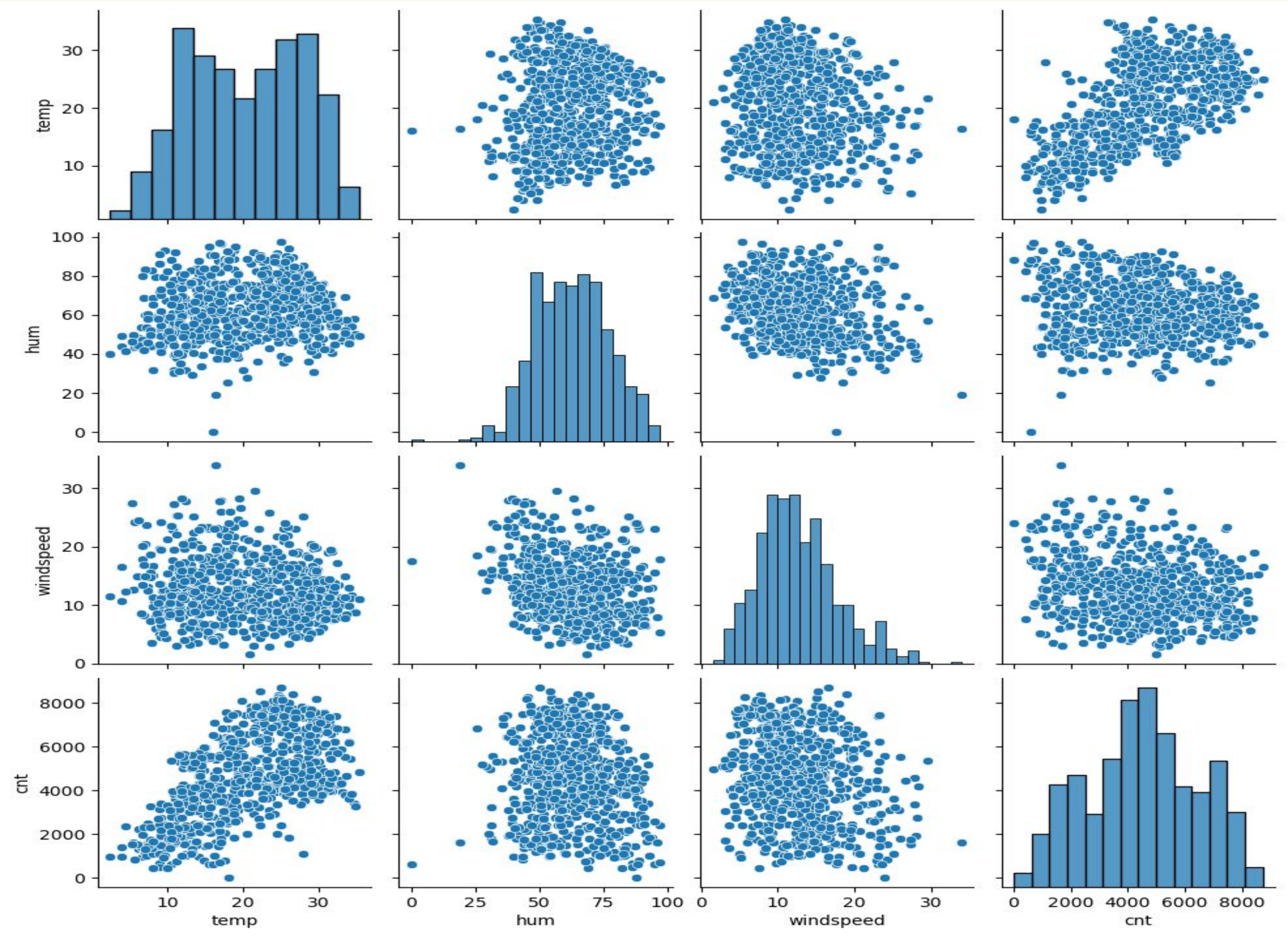
We can observe that wind speed has a positive influence on bike rental demand.

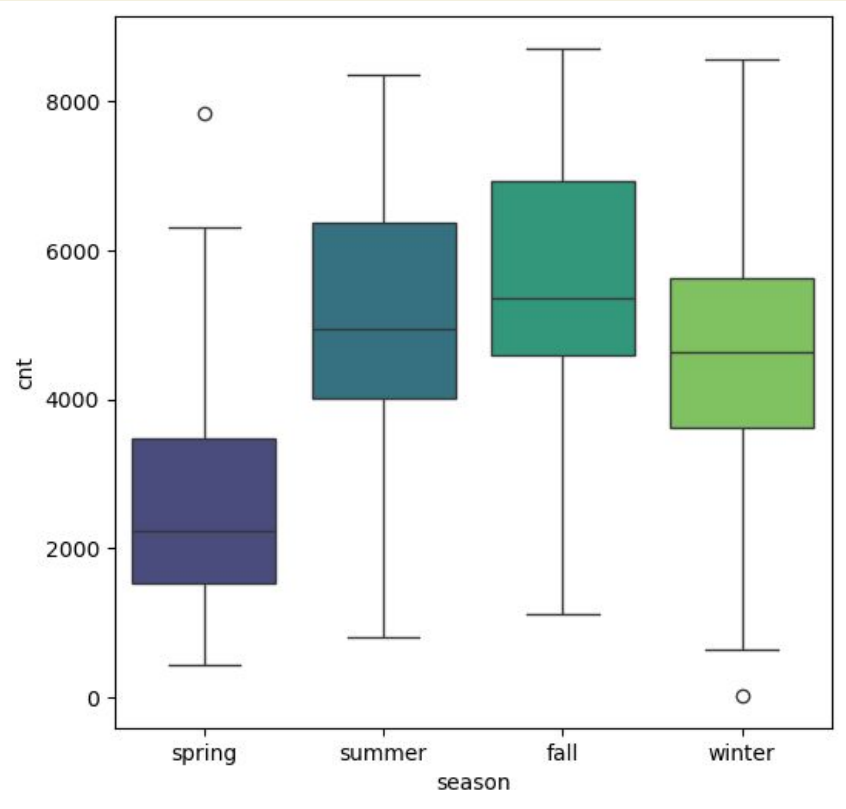


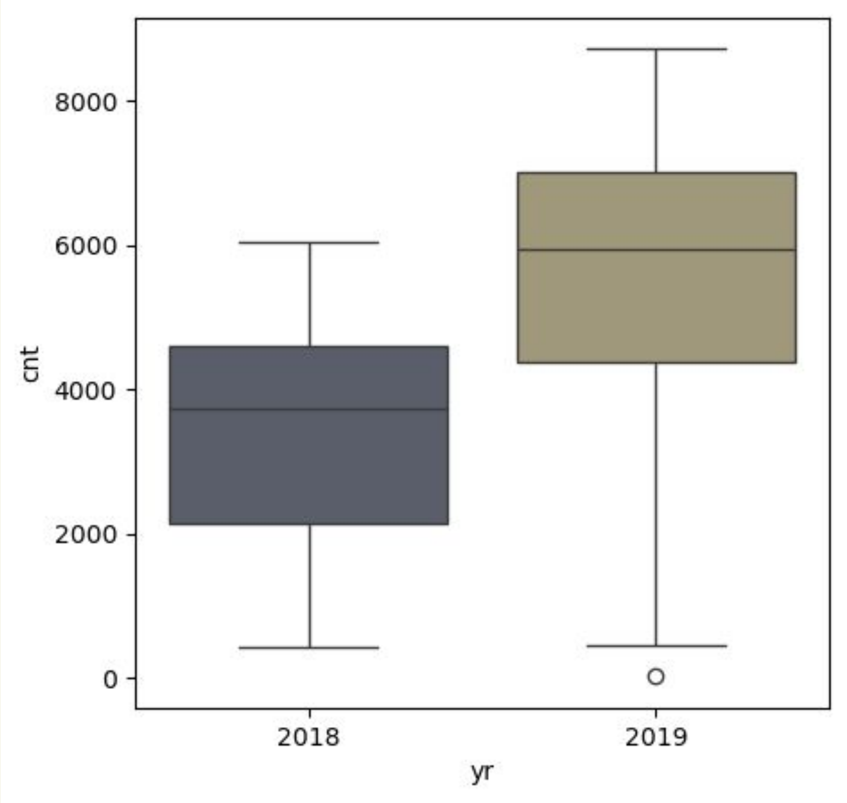
Humidity vs Count

We can observe that as humidity increases, there is a corresponding rise in bike rental demand.

Data Distribution





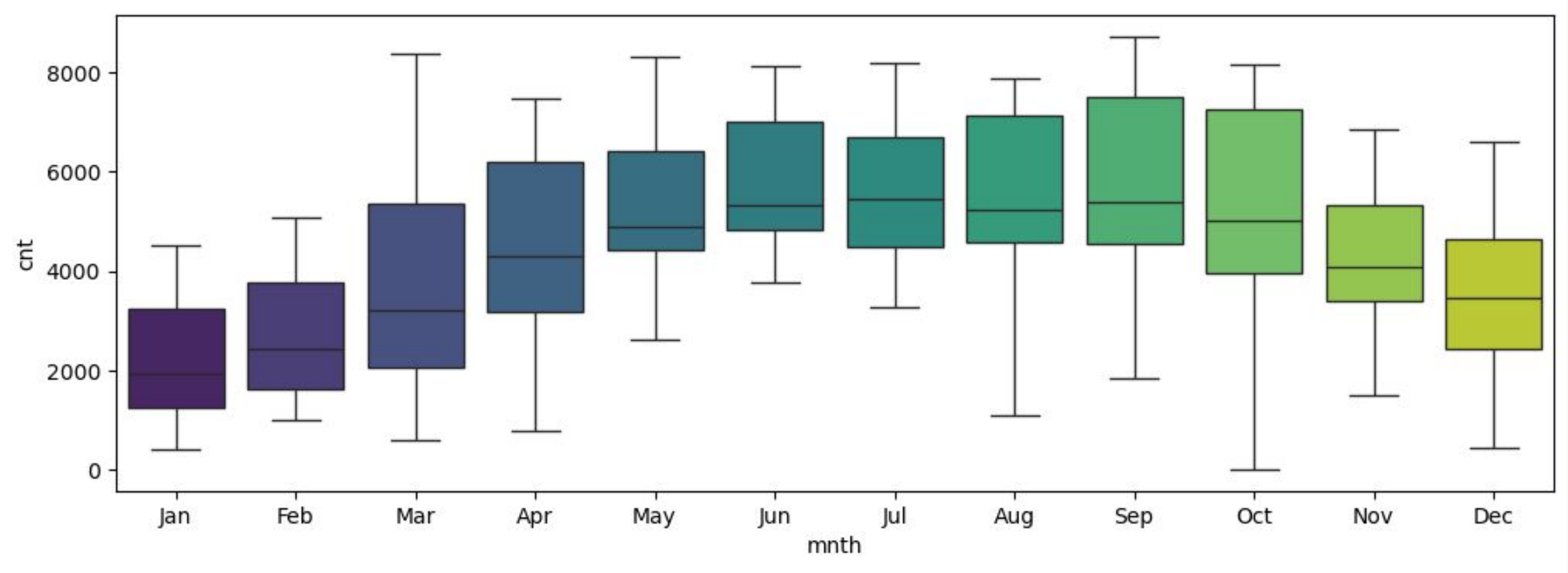


Wind Speed vs Count

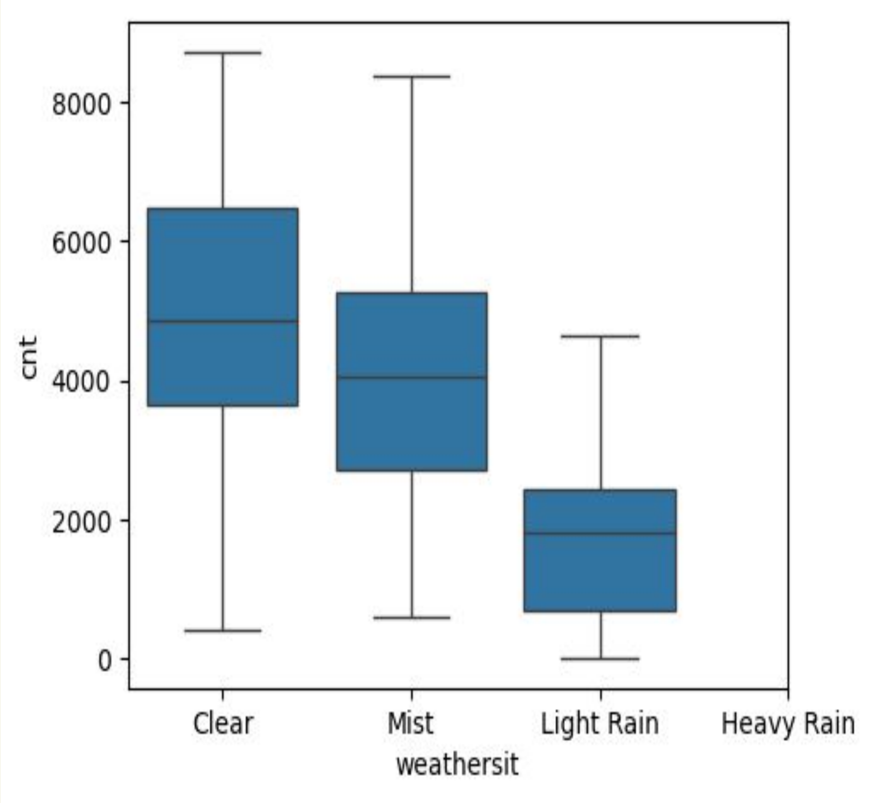
We can observe that during summer and fall there positive influence on bike rental demand.

Humidity vs Count

We can observe that there is increases in bike rental demand during the year 2019.

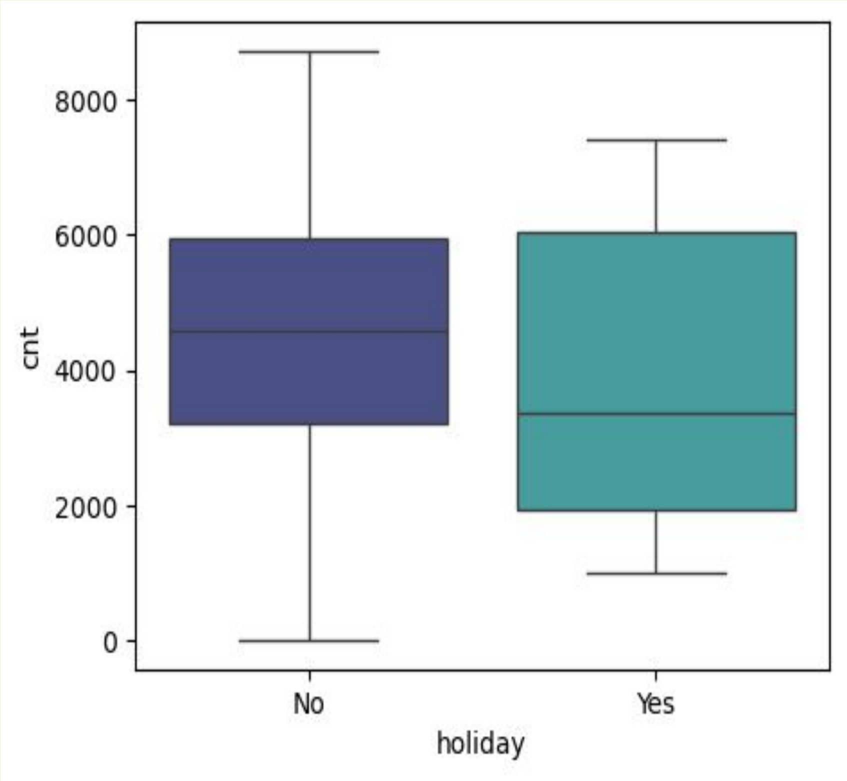


We can observe an increase in bike rental demand during the second and third quarters of the year.



Weather Condition vs Count

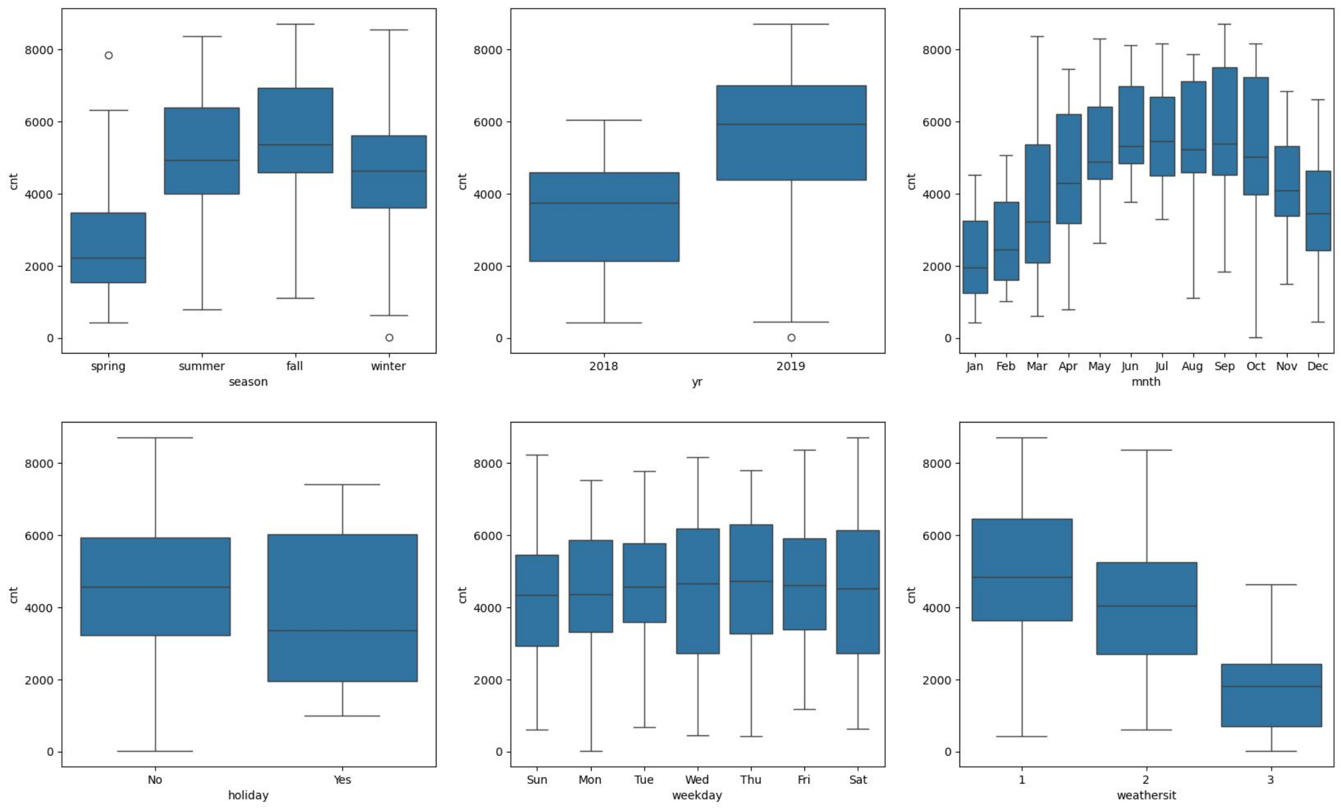
We can observe that when weather condition is clear bike rental count is higher



Holiday vs Count

We can observe that the mean bike rental count on non-holidays is higher than on holidays, indicating that bike rentals are more frequent on working days compared to non-working days.

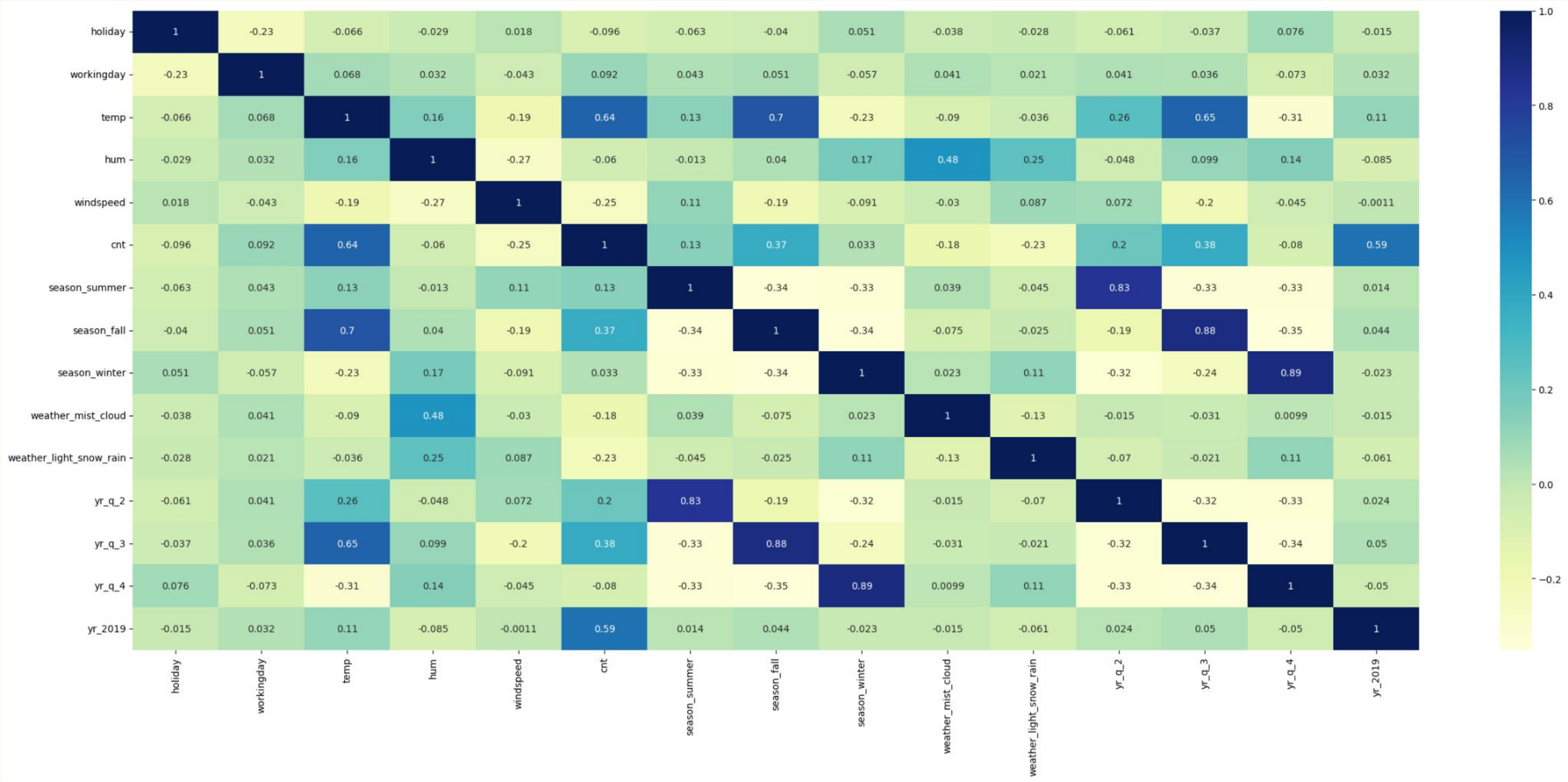
Data Distribution



Data Preparation

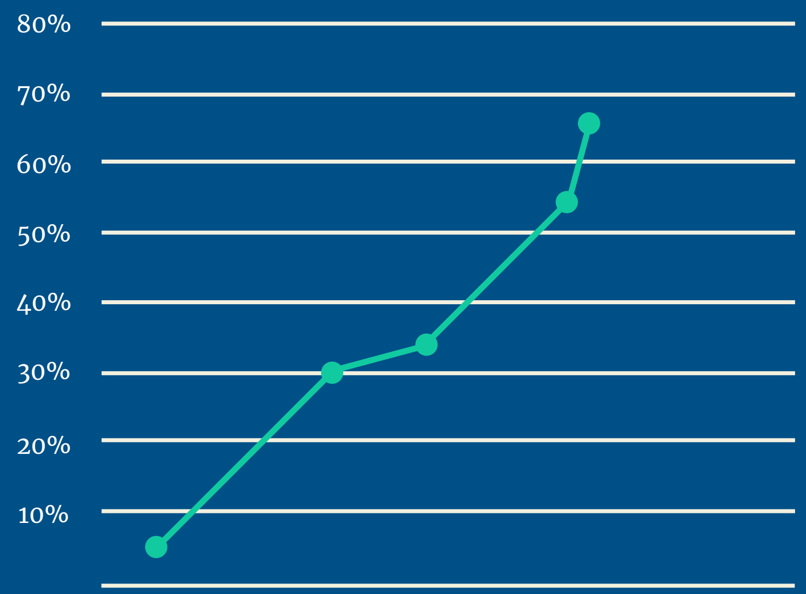
- Drops the columns instant,dteday,casual ,registered,atemp and weekday as these variables are not helping much for the model
- Apply the **min max scaling** for the feature set
- Create dummy variable for **Year**
  - **0** will correspond to yr_2018
  - **1** will correspond to yr_2019
- Create dummy variable for **Month**
  - **000** will correspond to Q1
  - **100** will correspond to Q1
  - **010** will correspond to Q3
  - **001** will correspond to Q4
- Create a dummy variable for **weathersit**
  - **00** will correspond to Clear,Few cloud,Partly Cloudy
  - **10** will correspond to Misty+Cloud,Mist+Broken Cloud,Mist+ Few Cloud,Mist
  - **01** will correspond to Light Snow,Light Rain + Thunderstorm + Scattered Cloud,Light Rain + Scattered Cloud
- Create a dummy variable for **Season**
  - **000** will correspond to spring
  - **100** will correspond to summer
  - **010** will correspond to fall
  - **001** will correspond to winter

Correlation of feature



Modelling

Analyzing the feature variables to predict the data.



Linear Regression Model

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.831
Model:                            OLS   Adj. R-squared:                  0.827
Method:                 Least Squares   F-statistic:                     174.3
Date:                Mon, 24 Feb 2025   Prob (F-statistic):          5.40e-181
Time:                        15:34:57   Log-Likelihood:                 492.41
No. Observations:                 510   AIC:                            -954.8
Df Residuals:                     495   BIC:                            -891.3
Df Model:                          14
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                      0.2140      0.029      7.477      0.000       0.158       0.270
holiday                   -0.0726      0.027     -2.673      0.008      -0.126      -0.019
workingday                 0.0178      0.009      1.946      0.052      -0.000       0.036
temp                       0.4725      0.037     12.832      0.000       0.400       0.545
hum                       -0.1372      0.039     -3.504      0.001      -0.214      -0.060
windspeed                 -0.1772      0.027     -6.628      0.000      -0.230      -0.125
season_summer              0.1062      0.022      4.853      0.000       0.063       0.149
season_fall                0.0352      0.030      1.189      0.235      -0.023       0.093
season_winter              0.1859      0.024      7.604      0.000       0.138       0.234
weather_mist_cloud        -0.0564      0.011     -5.162      0.000      -0.078      -0.035
weather_light_snow_rain   -0.2418      0.027     -8.837      0.000      -0.296      -0.188
yr_q_2                     0.0163      0.023      0.712      0.477      -0.029       0.061
yr_q_3                     0.0516      0.028      1.824      0.069      -0.004       0.107
yr_q_4                    -0.0272      0.024     -1.132      0.258      -0.074       0.020
yr_2019                    0.2297      0.008     27.263      0.000       0.213       0.246
==============================================================================
```

	Features	VIF
0	const	47.74
7	season_fall	9.90
12	yr_q_3	8.82
8	season_winter	6.51
13	yr_q_4	6.41
11	yr_q_2	5.54
6	season_summer	5.17
3	temp	4.02
4	hum	1.90
9	weather_mist_cloud	1.57
10	weather_light_snow_rain	1.25
5	windspeed	1.20
1	holiday	1.07
2	workingday	1.07
14	yr_2019	1.03

Running the linear regression model with all variables resulted in an R-squared value of 0.831 and an Adjusted R-squared value of 0.827. However, the p-values for **yr_q_2**, **yr_q_4**, and **season_fall** are relatively high.

The VIF values for **season_fall, yr_q_3, season_winter, yr_q_4, yr_q_2,** and **season_summer** are relatively high.

Linear Regression Model

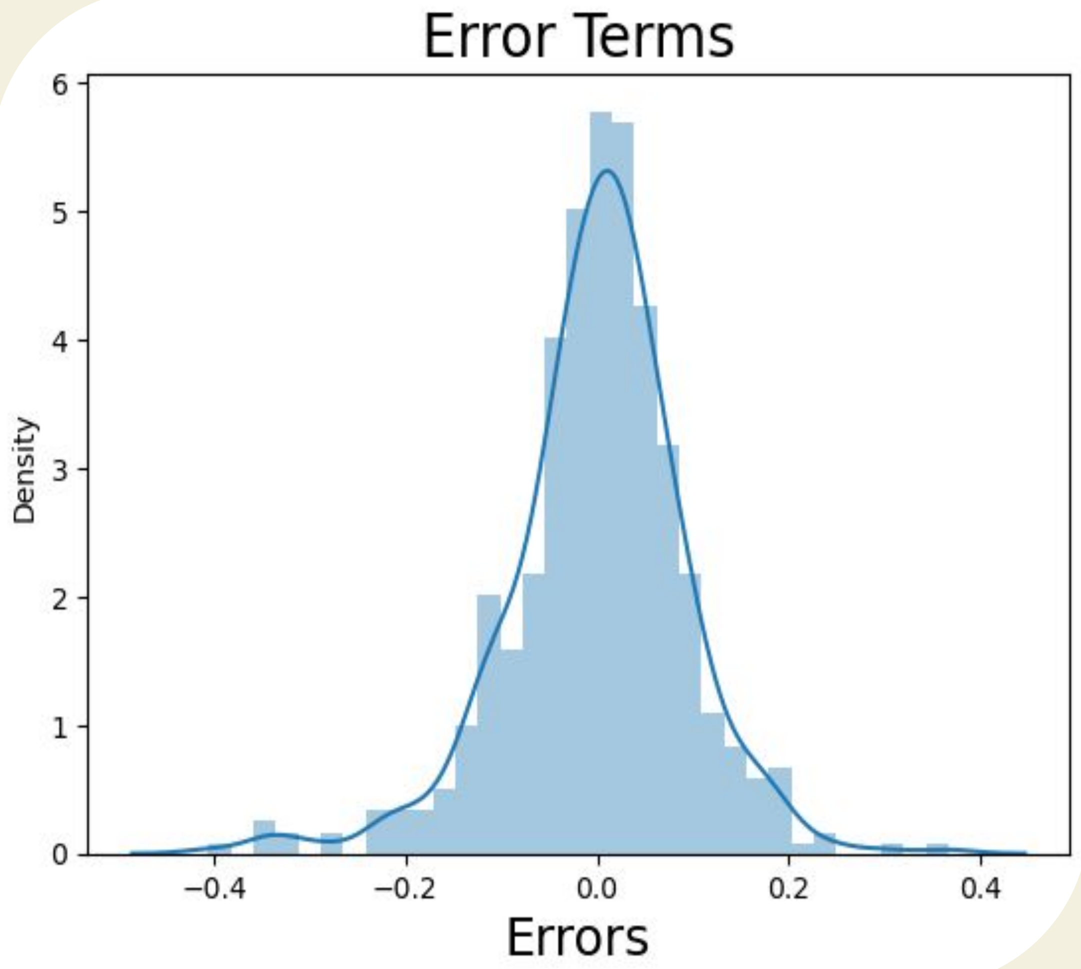
After removing features such as **season_fall, yr_q_1, yr_q_2, and yr_q_1**, the p-values and VIF values of the remaining features fall within the expected range, with minimal change in the **R-squared** value and an increase in the **Adjusted R-squared** value. This indicates that the selected variables are the best fit for model building.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.830
Model:                            OLS   Adj. R-squared:                  0.826
Method:                 Least Squares   F-statistic:                     221.1
Date:                Mon, 24 Feb 2025   Prob (F-statistic):          1.19e-183
Time:                        15:34:57   Log-Likelihood:                 490.43
No. Observations:                 510   AIC:                            -956.9
Df Residuals:                     498   BIC:                            -906.0
Df Model:                          11
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
const                       0.2119      0.029      7.417      0.000       0.156       0.268
temp                        0.5102      0.029     17.829      0.000       0.454       0.566
hum                        -0.1496      0.039     -3.868      0.000      -0.226      -0.074
windspeed                  -0.1820      0.027     -6.834      0.000      -0.234      -0.130
season_summer               0.1092      0.013      8.355      0.000       0.084       0.135
season_winter               0.1540      0.012     13.389      0.000       0.131       0.177
weather_mist_cloud         -0.0548      0.011     -5.027      0.000      -0.076      -0.033
weather_light_snow_rain    -0.2389      0.027     -8.759      0.000      -0.292      -0.185
yr_q_3                      0.0680      0.016      4.220      0.000       0.036       0.100
holiday                    -0.0745      0.027     -2.742      0.006      -0.128      -0.021
workingday                  0.0182      0.009      1.990      0.047       0.000       0.036
...
==============================================================================
```

	Features	VIF
0	const	47.51
8	yr_q_3	2.86
1	temp	2.43
2	hum	1.85
4	season_summer	1.84
6	weather_mist_cloud	1.56
5	season_winter	1.44
7	weather_light_snow_rain	1.24
3	windspeed	1.19
9	holiday	1.07
10	workingday	1.07
11	yr_2019	1.03

Residual Analysis

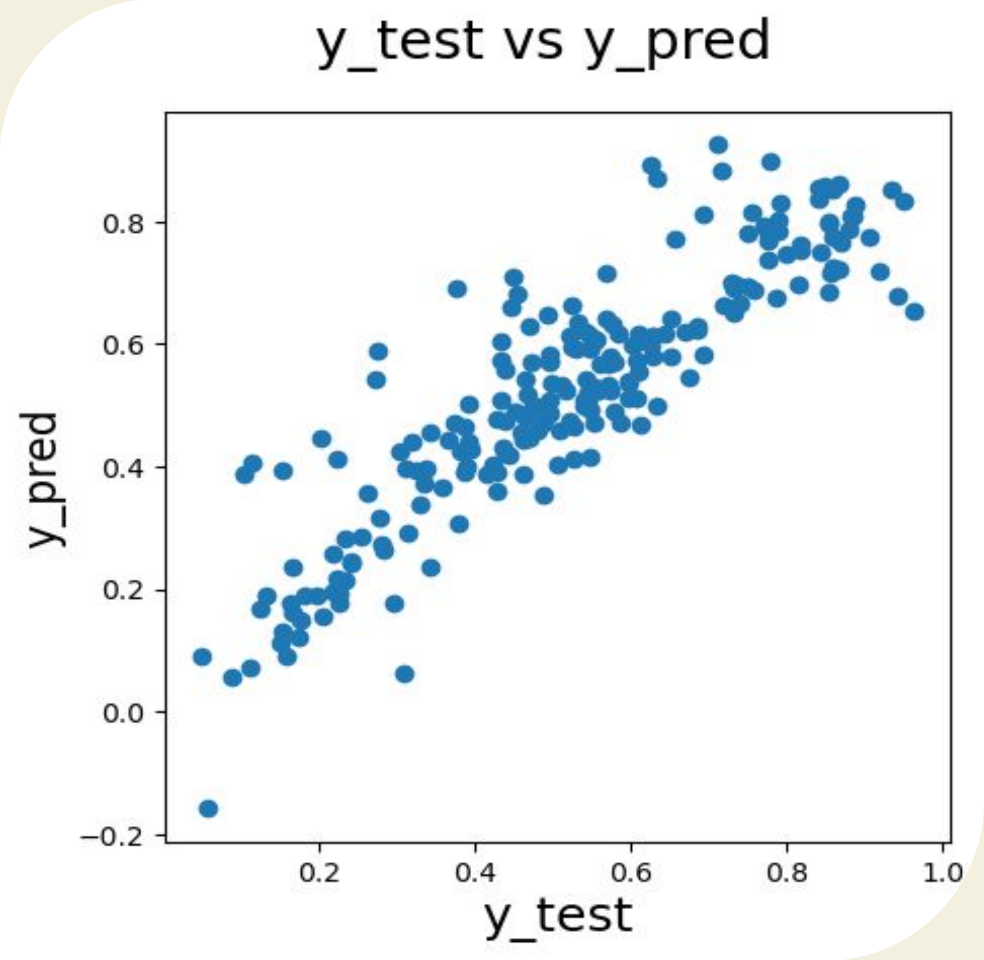
Residual errors being normally distributed means that the differences between the actual and predicted values (residuals) follow a normal distribution.



Model Evaluation

Mean Squared Error = 0.009686748493429775

R2 score = 0.7960504543310527



Linear Regression Model (RFE)

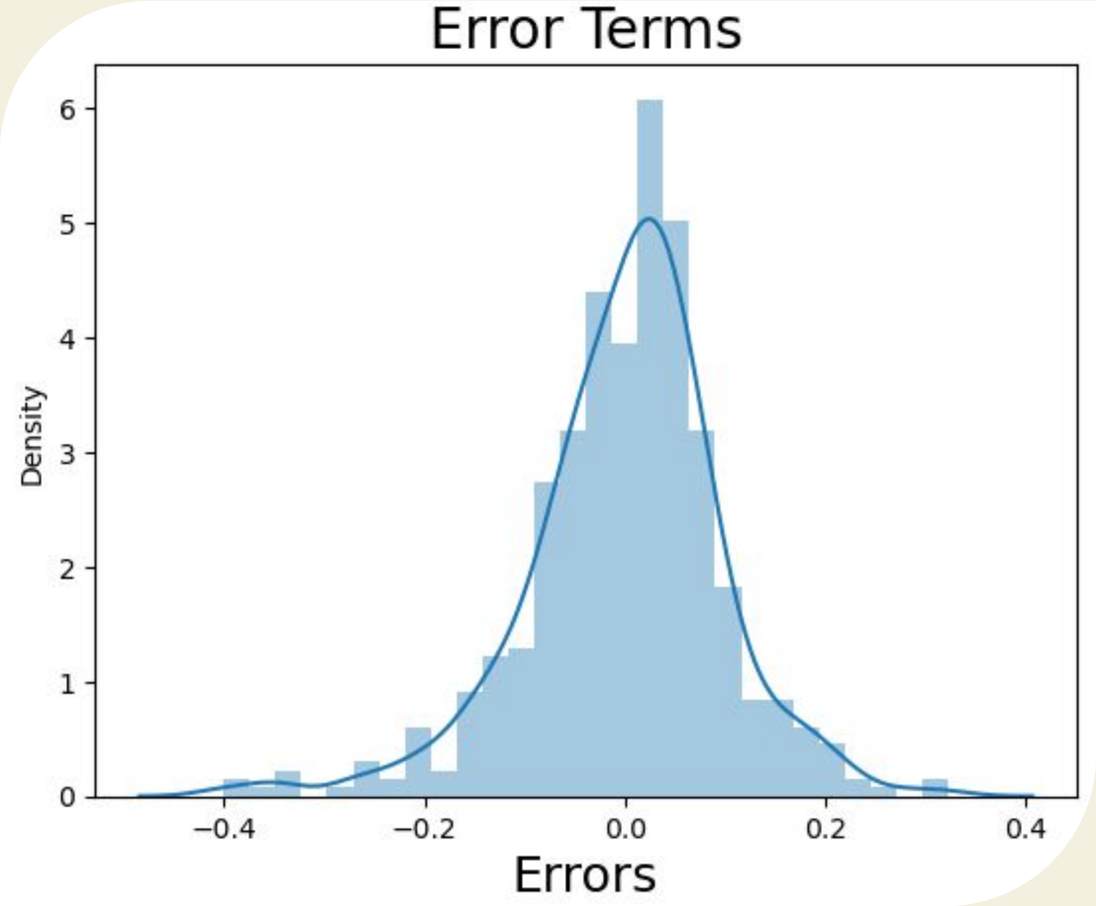
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.812
Model:                            OLS   Adj. R-squared:                  0.809
Method:                 Least Squares   F-statistic:                     309.9
Date:                Mon, 24 Feb 2025   Prob (F-statistic):          1.06e-177
Time:                        15:34:58   Log-Likelihood:                 464.81
No. Observations:                 510   AIC:                            -913.6
Df Residuals:                     502   BIC:                            -879.7
Df Model:                           7
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
const                       0.2510      0.027      9.164      0.000       0.197       0.305
temp                        0.6208      0.021     29.958      0.000       0.580       0.662
hum                        -0.2554      0.033     -7.719      0.000      -0.320      -0.190
windspeed                  -0.2114      0.028     -7.683      0.000      -0.265      -0.157
season_summer               0.0761      0.011      7.059      0.000       0.055       0.097
season_winter               0.1391      0.011     12.464      0.000       0.117       0.161
weather_light_snow_rain    -0.1898      0.027     -7.028      0.000      -0.243      -0.137
yr_2019                     0.2269      0.009     25.822      0.000       0.210       0.244
==============================================================================
Omnibus:                       58.844   Durbin-Watson:                   1.947
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              132.505
Skew:                          -0.628   Prob(JB):                     1.69e-29
Kurtosis:                       5.158   Cond. No.                         15.2
==============================================================================
```

	Features	VIF
0	const	39.82
5	season_winter	1.24
2	hum	1.23
1	temp	1.16
3	windspeed	1.16
4	season_summer	1.14
6	weather_light_snow_rain	1.10
7	yr_2019	1.02

The linear regression model using REF selected fewer features than the manually selected model. Additionally, the R-squared value of the REF model is slightly lower compared to the manually selected features. Moreover, all the features present in the REF model are also included in the manually selected model.

Residual Analysis

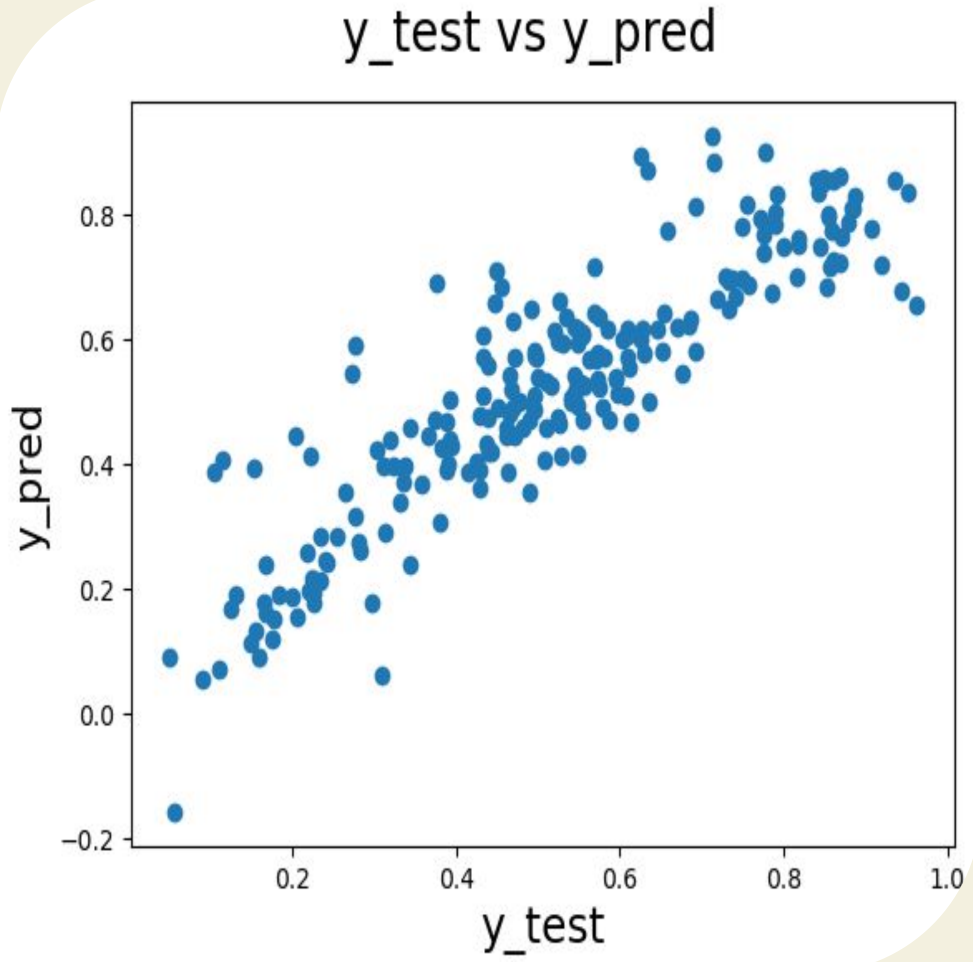
Residual errors being normally distributed means that the differences between the actual and predicted values (residuals) follow a normal distribution.



Model Evaluation

Mean Squared Error = 0.010273183192060383

R2 score = 0.7837033710521443



Business Goal and Key Insights

The objective is to analyze the factors influencing bike rental demand and make data-driven decisions to optimize the service. Based on the model results, the following insights have been identified:

Key Findings from the Model

1. **Temperature is the strongest positive predictor of bike rental demand** (coef = 0.5102, $p < 0.001$).
   - A **one-unit increase in temperature** leads to a **0.5102-unit increase in bike rental demand**, confirming that people prefer riding bikes in warmer conditions.
2. **Humidity and wind speed negatively impact bike rentals:**
   - **Humidity (hum) has a significant negative impact** (coef = -0.1496, $p < 0.001$), meaning that higher humidity leads to fewer bike rentals.
   - **Wind speed (windspeed) also reduces demand** (coef = -0.1820, $p < 0.001$), indicating that strong winds discourage people from renting bikes.
3. **Seasonal Influence on Demand:**
   - **Winter (season_winter) and Summer (season_summer) have a strong positive impact** on bike rentals, with winter (coef = 0.1540, $p < 0.001$) having the highest increase in demand.
4. **Impact of Weather Conditions:**
   - **Cloudy/Misty weather (weather_mist_cloud) reduces demand** (coef = -0.0548, $p < 0.001$), but the impact is relatively small.
   - **Snow or rain (weather_light_snow_rain) has a significant negative impact** (coef = -0.2389, $p < 0.001$), meaning that bike rentals drop sharply during poor weather conditions.
5. **Quarterly Trends:**
   - **Bike rentals increase significantly in Q3** (yr_q_3, coef = 0.0680, $p < 0.001$), suggesting that demand is higher in the **third quarter of the year**.
6. **Effect of Holidays and Working Days:**
   - **Bike rentals are lower on holidays** (holiday, coef = -0.0745, $p = 0.006$), indicating that people rent fewer bikes on non-working days.
   - **Working days show a positive impact** (workingday, coef = 0.0182, $p = 0.047$), though the effect is small, suggesting a slight increase in rentals on workdays.
7. **Yearly Growth in Demand:**
   - **Bike rentals increased significantly in 2019 compared to 2018** (yr_2019, coef = 0.2292, $p < 0.001$), indicating strong growth in demand over time.

Conclusion

- **Weather conditions, temperature, seasonality, and working schedules** are the most important factors affecting bike rental demand.
- Bike rental demand is **highest in warm weather, during working days, and in Q3.**
- Demand drops during extreme weather conditions such as **rain, snow, and high humidity.**
- The model has a strong fit (**R-squared value is high**), making it reliable for predicting bike rental demand trends.

These insights can help businesses **optimize bike availability, plan for seasonal demand changes, and develop targeted promotions** based on weather and time of year.

Best Fit Line can be derived from the below

$$cnt = 0.5102 \times temp + (-0.1496 \times hum) + (-0.1820 \times windspeed) + (0.1092 \times \text{season\_summer}) + (0.1540 \times \text{season\_winter}) + (-0.0548 \times \text{weather\_mist\_cloud}) + (-0.2389 \times \text{weather\_light\_snow\_rain}) + (0.0680 \times \text{yr\_q3}) + (-0.0745 \times holiday) + (0.0182 \times workingday) + (0.2292 \times \text{yr\_2019}) + (0.2119)$$