Assignment-based Subjective Questions

Question 1 From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer: In our final model we have 4 categorical variables as explained below with their impacts. Each number in brackets (like -0.1413) represents the coefficient value from the regression model.
A positive value means the variable increases bike demand, while a negative value means it reduces demand.

a. season_name_spring (-0.1413)
    - Spring season slightly reduces bike demand compared to winter.
    - The coefficient -0.1413 means, keeping other factors constant, the average bike count decreases by about 0.14 units during spring.

b. mnth_name_Jul (-0.0715)
    - July also shows fewer rides than the reference month (likely January).
    - The -0.07 coefficient means there's a small drop in demand during this month.

c. weather_light_rain (-0.2413)
    - Rainy weather has the strongest negative effect on bike usage.
    - The -0.24 coefficient shows a clear fall in demand when it rains.

d. yr (0.2360)
    - The year variable has a positive coefficient (0.236), meaning demand increased in the later year.
    - It suggests more people started using bikes over time.

Refer to the Model Summary:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.793
Model:                            OLS   Adj. R-squared:                  0.791
Method:                 Least Squares   F-statistic:                     321.7
Date:                Mon, 10 Nov 2025   Prob (F-statistic):          1.37e-168
Time:                        23:58:55   Log-Likelihood:                440.46
No. Observations:                 510   AIC:                            -866.9
Df Residuals:                     503   BIC:                            -837.3
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const                0.2596      0.020     12.986      0.000       0.220       0.299
season_name_spring  -0.1413      0.014    -10.427      0.000      -0.168      -0.115
mnth_name_Jul       -0.0715      0.019     -3.769      0.000      -0.109      -0.034
weather_light_rain  -0.2413      0.027     -8.873      0.000      -0.295      -0.188
yr                   0.2360      0.009     25.713      0.000       0.218       0.254
temp                 0.4279      0.028     15.277      0.000       0.373       0.483
windspeed           -0.1524      0.028     -5.536      0.000      -0.206      -0.098
==============================================================================
Omnibus:                       54.941   Durbin-Watson:                   1.895
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              101.150
Skew:                          -0.657   Prob(JB):                     1.09e-22
Kurtosis:                       4.742   Cond. No.                         10.5
==============================================================================

Notes:
```

Question 2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

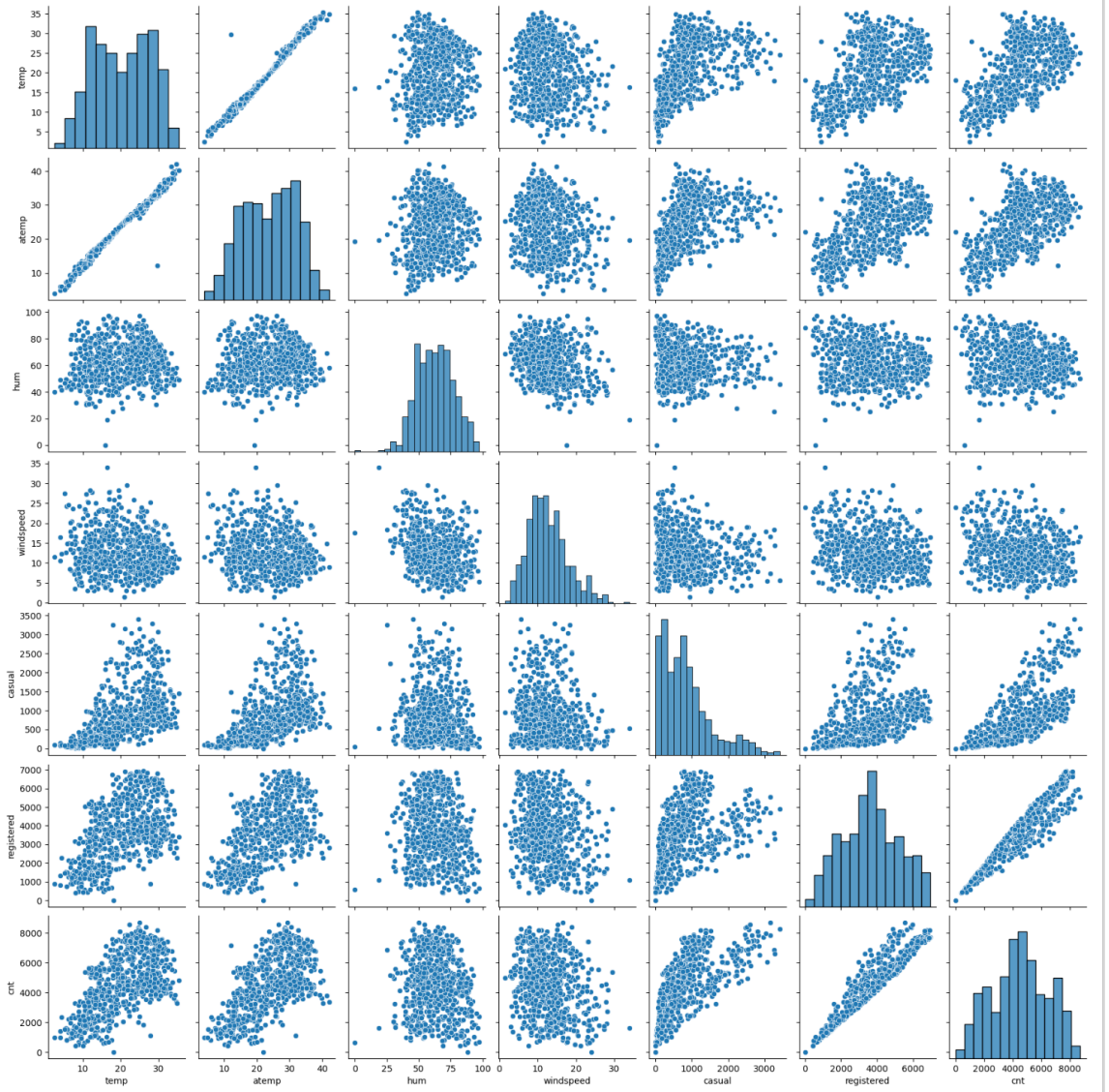
Answer: When we create dummy variables, keeping all categories causes duplicate information (multicollinearity).
Using drop_first=True removes one category, avoids confusion for the model, and makes results more stable and easy to interpret.

Example: if we have 2 seasons Winter and Summer, dummy creation makes two columns. If one is 1, the other will always be 0, creating duplicate information. Using drop_first=True drops one column so the model avoids confusion.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer: temp variables has the highest correlation

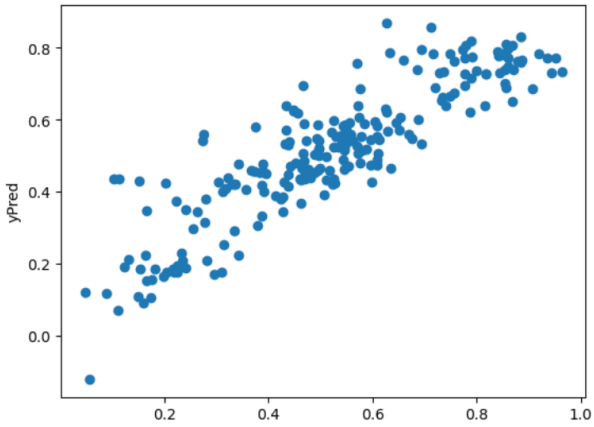
```
[890]: sns.pairplot(bike_df[numerical_columns])
       plt.show()
```



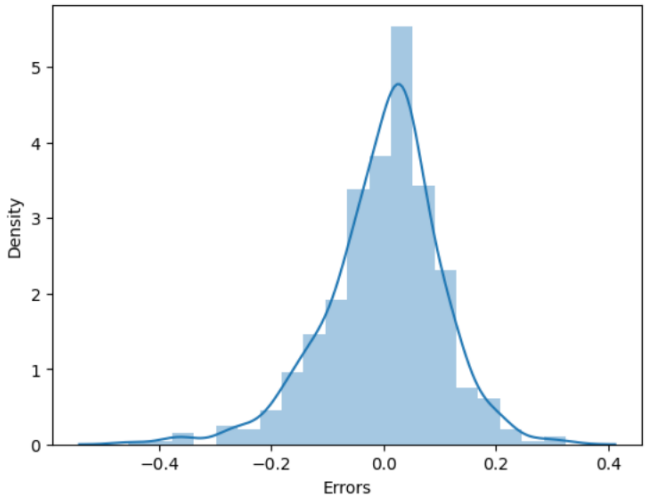
Question 4: How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer: After building the model, I checked the main assumptions of Linear R egression:

1. Linearity: Verified that the relationship between predicted and actual values is roughly a straight line.



2. Normality of errors: Checked that the residuals (errors) follow a normal distribution



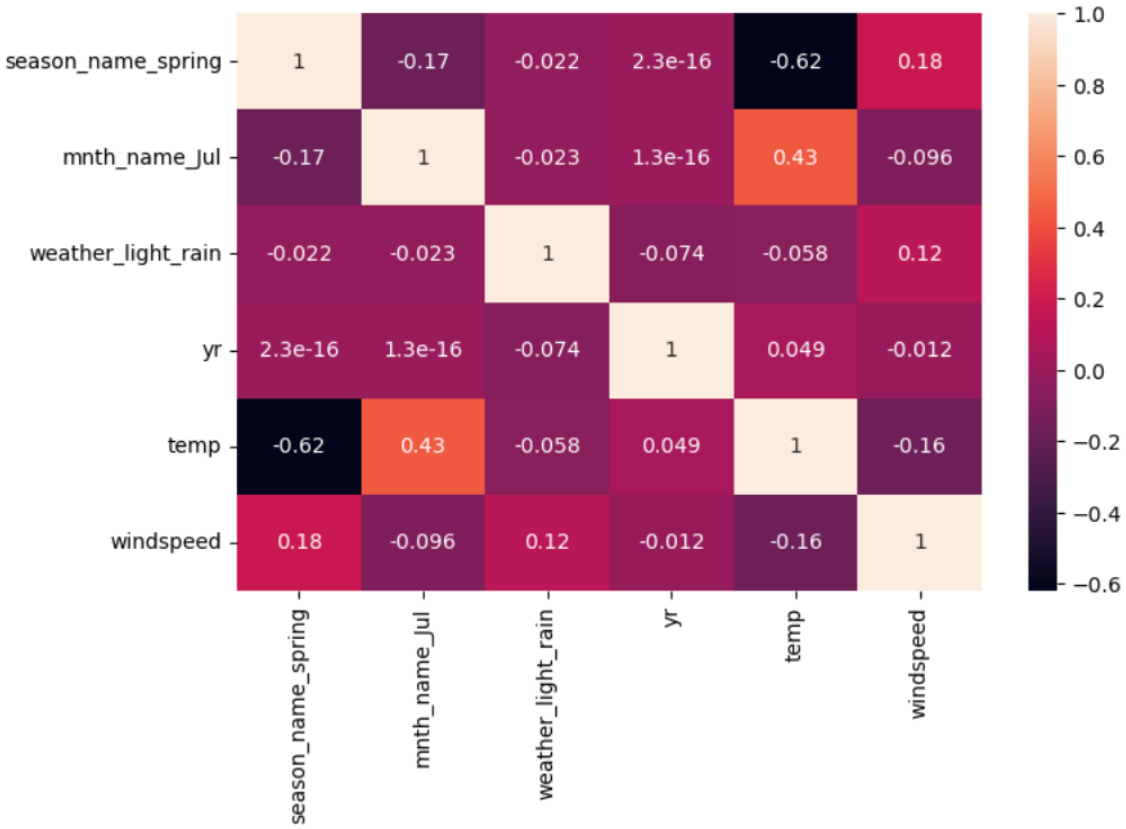
Analysis: The error graph is Normal Distributed!

3. Homoscedasticity: Made sure the residuals have constant spread (no pattern) across all predicted values.

4. No multicollinearity: Used VIF (Variance Inflation Factor) to confirm that independent variables are not highly correlated.

]:

	Features	VIF
0	season_name_spring	1.494392
1	mnth_name_Jul	1.256583
2	weather_light_rain	1.044848
3	yr	2.026288
4	temp	3.945895
5	windspeed	3.719729



QUESTION 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer:

- Temperature (temp) - has the highest positive impact; as temperature increases, more people use bikes.
- Year (yr) - demand increased in the next year, showing overall growth in bike usage over time.
- Weather (weather_light_rain) - has a strong negative impact; rainy weather reduces bike demand significantly.

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**

   Linear Regression is one of the most basic and widely used methods in data analysis. It helps us understand how the target variable changes based on one or more other variables. The main goal is to draw a straight line that best fits the data points. This line shows the relationship between the dependent variable (like bike demand) and the independent variables (like temperature, season, or windspeed).

   The equation for this line looks like:

   $\mathbf{y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n}$

   Here,

   y is what we're trying to predict

   $\mathbf{x_1, x_2 \ldots}$ are the factors affecting it

   $\mathbf{b_0}$ is the intercept

   $\mathbf{b_1, b_2 \ldots}$ show how much each factor influences y.

   The algorithm finds the best values for these coefficients by reducing the difference between the predicted and actual values - this is done using a method called **Ordinary Least Squares (OLS)**.

   After building the model, we check if it fits the data well using metrics like $\mathbf{R^2}$ and by looking at the error patterns.

2. **Explain the Anscombe's quartet in detail. (3 marks)**

   Answer: Anscombe's Quartet is a famous example in statistics created by Francis Anscombe in 1973.
   It consists of four different datasets that look very different when plotted, but all have almost identical statistical properties such as the same mean, variance, correlation, and regression line.
   Each dataset contains 11 (x, y) pairs, and when you calculate simple statistics like average, correlation, or regression coefficients, all four datasets seem identical.
   However, when you plot them on a graph, they look completely different one shows a linear trend, another is curved, one has an outlier, and one has most points the same except one extreme point.
   The main lesson from Anscombe's Quartet is that:
   - Summary statistics alone can be misleading.

- Data visualization is very important before making conclusions or building models.

3. **What is Pearson's R? (3 marks)**

   Answer: Pearson's R is the Pearson correlation coefficient, which is a number that measures the strength and direction of the linear relationship between two variables. It always lies between -1 and +1:

   - +1 means a perfect positive relationship - as one variable increases, the other increases too.
   - -1 means a perfect negative relationship - as one increases, the other decreases.
   - 0 means no linear relationship between the two variables.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**

   Answer: Scaling means changing the range of data so that all features are on a similar scale. Some columns may have very large values (like employee salary in thousands, lakhs) and others small (like employee age or experience).
   Large numbers can dominate the model and make results less accurate, hence we scale the variables to a common range
   Types of Scaling
   1. Normalized Scaling or Min-Max Scaling
      - Rescales data between 0 and 1.
      - Used when you want all features in a fixed range.
   2. Standardized Scaling or Z-score Scaling
      - Converts data so it has mean = 0 and standard deviation = 1.
      - Used when data follows a normal distribution or when you don't want to limit values between 0 and 1.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

   Answer: VIF becomes infinite when two or more independent variables are perfectly correlated, i.e. one column can be exactly predicted from another.
   In that case, the model can't separate their effects, so the denominator in the VIF formula ($1 - R^2$) becomes zero, making the VIF infinite.

Example:
If we have both temperature_in_celsius and temperature_in_fahrenheit as features.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

Answer:

A Q-Q plot (Quantile-Quantile plot) is a graph that helps check whether a variable (mostly the residuals) follows a normal distribution. On the plot, if all points lie roughly along a straight diagonal line, residuals are normally distributed. If points curve away, residuals are not normal, which violates one of the key assumptions of linear regression.

Normality of residuals ensures that statistical tests (like p-values and confidence intervals) are valid. If the Q-Q plot shows major deviation, it means the model's predictions or significance tests may not be reliable.
Example from the assignment: