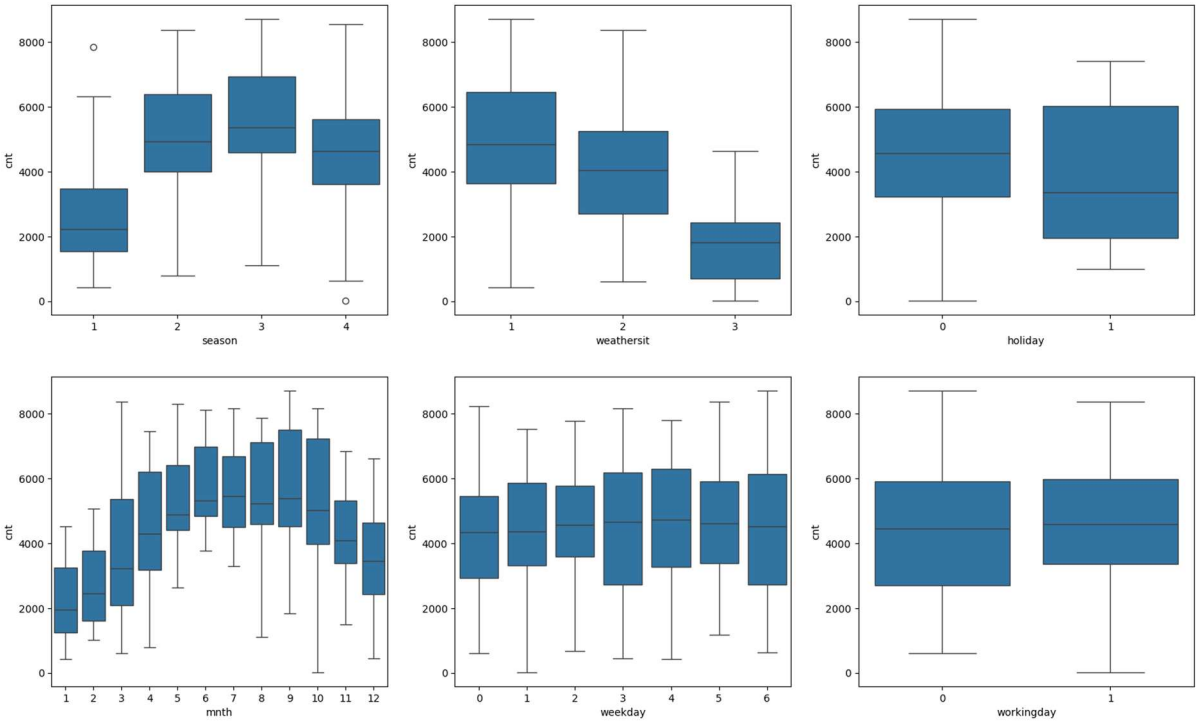
1**. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

Cnt is our dependent variable, and we are having season, weathersit, month and weekdays are categorical variables in the dataset is having significance impact. It has been shown in below plots.



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

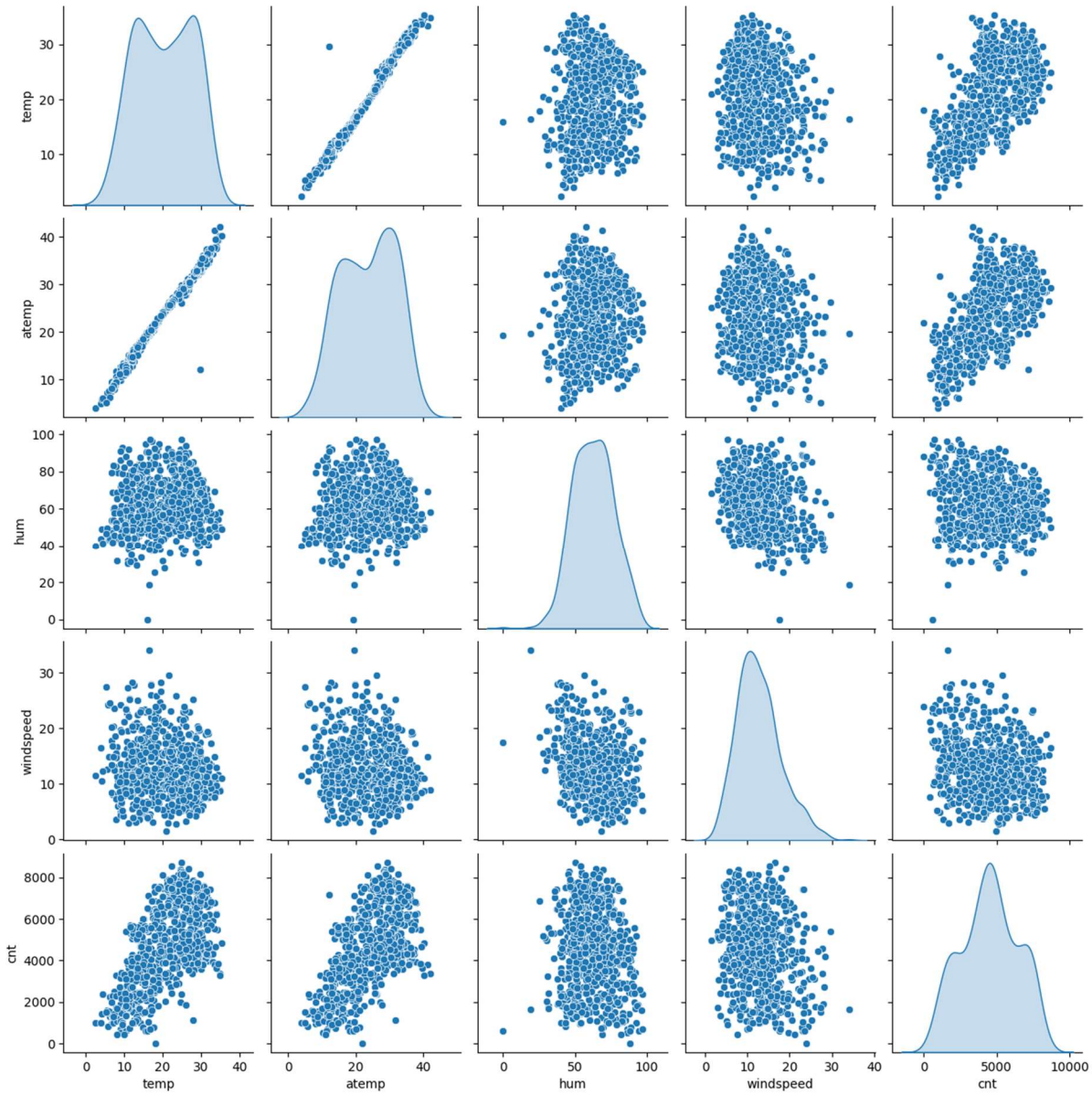
It is to avoid multicollinearity and ensure efficient model.

When creating dummy variables, each category of a categorical variable is converted into a separate binary variable. If all categories are included, it can lead to perfect multicollinearity, where one variable can be perfectly predicted from the others.

Using drop_first=True helps maintain the integrity of the regression model and ensures that the coefficients are interpretable and the model is not overfitted due to multicollinearity.

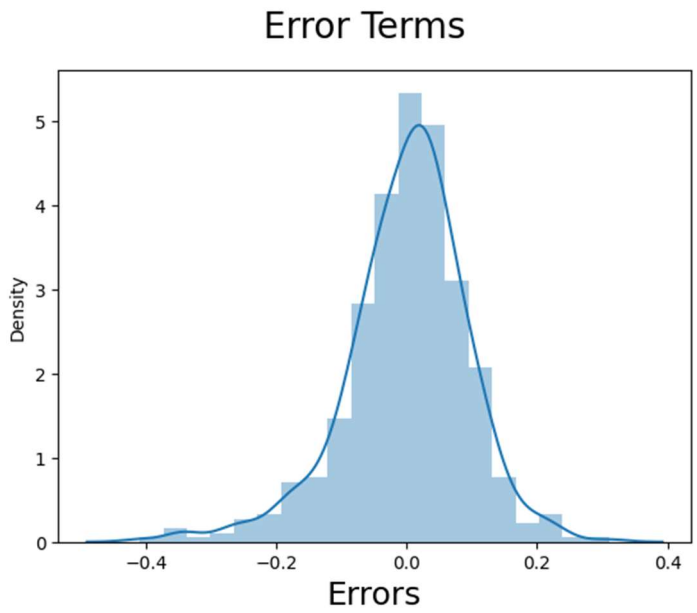
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

Temp is having highest correlation with target variable.

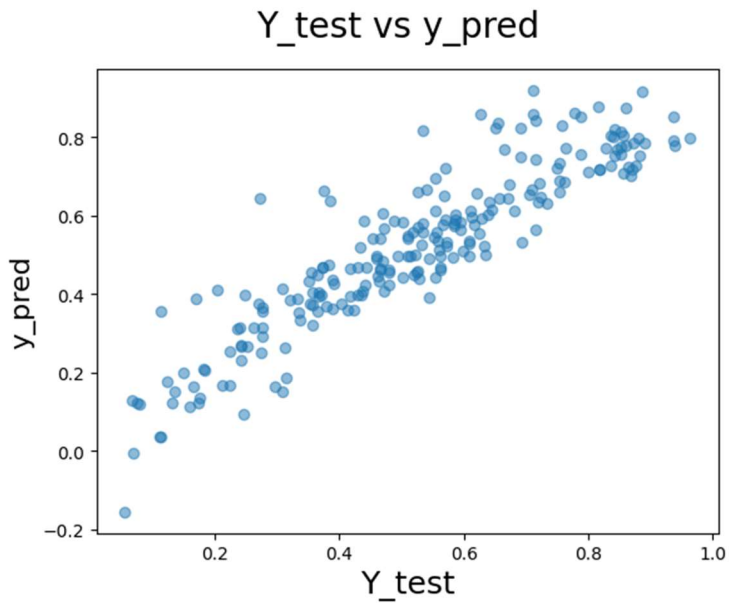


4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

Residuals followed normal distribution



Linearity verified through model prediction



There is No Multicollinearity between the predictor variables

Features	VIF	
2	temp	4.71
1	workingday	4.03
3	windspeed	4.03
0	yr	2.00
7	weekday_6	1.65
4	season_2	1.55
8	weathersit_2	1.53
5	season_4	1.38
6	mnth_9	1.20
9	weathersit_3	1.07

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

There are 3 predictor variables we are having from this dataset

1. Temperature (temp) –> Coefficient 0.5644
2. Weather Situation (Weather_sit) –> Coefficient (-0.3071)
3. Year (yr) -> Coefficient 0.2303

General Subjective Questions

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

It is a statistical method used to model between target variable and one or more independent variables.

Final goal to find the best fitting straight line through data points that can be used to predict the target variable.

Equation of the Line: The linear regression model is represented by the equation: $y=b0+b1x+\epsilon$

(y) is the dependent variable.

(x) is the independent variable.

(b0) is the y-intercept.

(b1) is the slope of the line.

(ϵ) is the error term (the difference between the observed and predicted values).

We are having some assumptions for linear regression. Namely Linearity, Independence, Homoscedasticity, Normality.

We need to go through below steps to find the linear regression on given dataset.

1. Data Collection
2. Data Cleansing
3. Model training
4. Model Evaluation
5. Prediction

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

Anscombe's quartet is a collection of four datasets that have nearly identical simple descriptive statistics, such as mean, variance, and correlation, but appear very different when graphed. It was created by statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data before analysing it.

Anscombe quartet illustrates relying on statistical summary will mislead and visualisation is very important to find the anomalies and outliers.

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

Pearson correlation coefficient is a measure of the linear relationship between two variables. It quantifies the strength and direction of this relationship and is denoted by (r).

Range and Interpretation:

- The value of (r) ranges from -1 to 1.
- (r = 1): Perfect positive linear relationship.
- (r = -1): Perfect negative linear relationship.
- (r = 0): No linear relationship.

Calculation: The formula for Pearson's R is:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

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

Scaling is a preprocessing technique used to adjust the range of data features. It ensures that all features contribute equally to the model's performance, preventing any single feature from disproportionately influencing the results due to its scale.

Necessity of scaling

Improves Model Performance

Speeds Up Convergence

Ensures Consistency

Normalization scales the data to a specific range, typically [0, 1] or [-1, 1].

Normalization: Preferred when the data does not follow a normal distribution and when you need to bound the data within a specific range.

Standardization transforms the data to have a mean of 0 and a standard deviation of 1.

Standardization: Preferred when the data is normally distributed and when we need to center the data around the mean with unit variance.

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

The Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in regression analysis. It quantifies how much the variance of a regression coefficient is inflated due to collinearity with other predictors.

- VIF becomes infinite when there is perfect multicollinearity, meaning one predictor variable is a perfect linear combination of one or more other predictor variables.

- Mathematically, this occurs when the correlation matrix of the predictors is singular, i.e., it does not have an inverse.

VIF for a predictor (X_i) is calculated as:

$$\text{VIF}(X_i) = \frac{1}{1 - R_i^2}$$

where (R_i^2) is the coefficient of determination of the regression of (X_i) on all other predictors.

If (R_i^2 = 1) (indicating perfect multicollinearity), the denominator becomes zero, making VIF infinite.

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

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, typically the normal distribution. It helps to assess whether the data follows a specified distribution.

Use and Importance in Linear Regression

1. Assessing Normality of Residuals:
    - In linear regression, one of the key assumptions is that the residuals are normally distributed. A Q-Q plot helps to visually check this assumption.
    - The quantiles of the residuals are plotted against the quantiles of a normal distribution. If the residuals are normally distributed, the points should lie approximately along a straight diagonal line.
2. Detecting Deviations from Normality:
    - Heavy Tails: If the points deviate from the line at the ends, it indicates heavy tails (more extreme values than expected).
    - Skewness: If the points form a curve, it suggests skewness in the data.
    - Outliers: Points that are far from the line indicate potential outliers.
3. Model Validation:
    - By using a Q-Q plot, we can validate whether the linear regression model's residuals meet the normality assumption. If they do not, it may indicate that the model is not appropriate, and transformations or different modelling techniques might be needed.