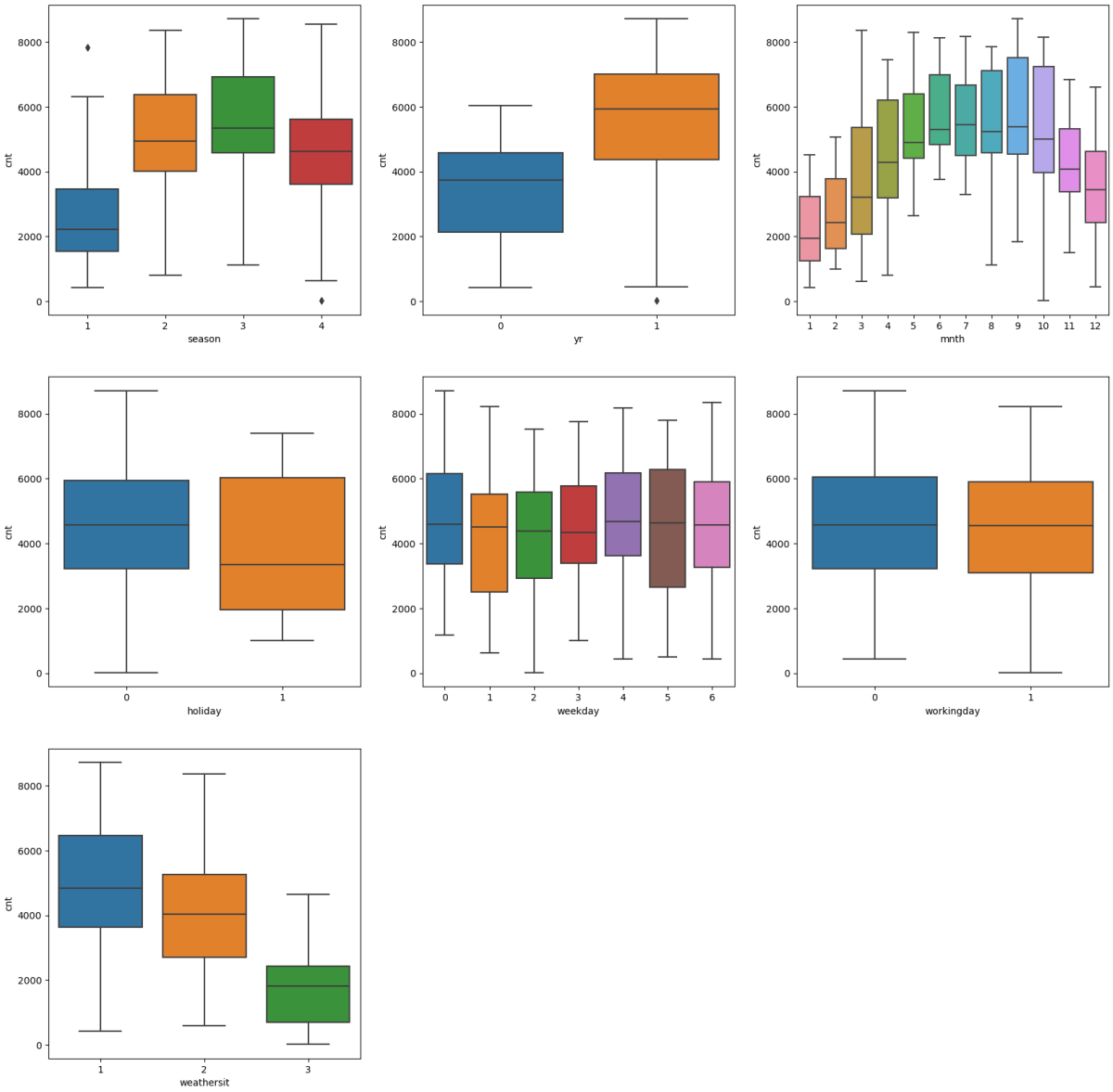
Assignment-based Subjective Questions

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



- Categorical variables in the dataset, including season, year, holiday, weekday, working day, weathersit, and month, were analyzed using boxplots (see attached Fig.). The impact of these variables on our dependent variable is summarized as follows:

- Season: The boxplot revealed that the spring season had the lowest value of "cnt," while fall had the maximum. Summer and winter exhibited intermediate values.

- Weathersit: Users were notably absent during heavy rain or snow, indicating unfavorable weather conditions. The highest count occurred when the weather situation was 'Clear, Partly Cloudy.'

- Yr: The number of rentals in 2019 surpassed that of 2018.

- Holiday: Rentals decreased during holidays.

- Mnth: September witnessed the highest number of rentals, whereas December had the least. This aligns with the weathersit observation, suggesting heavy snowfall in December impacting rentals.

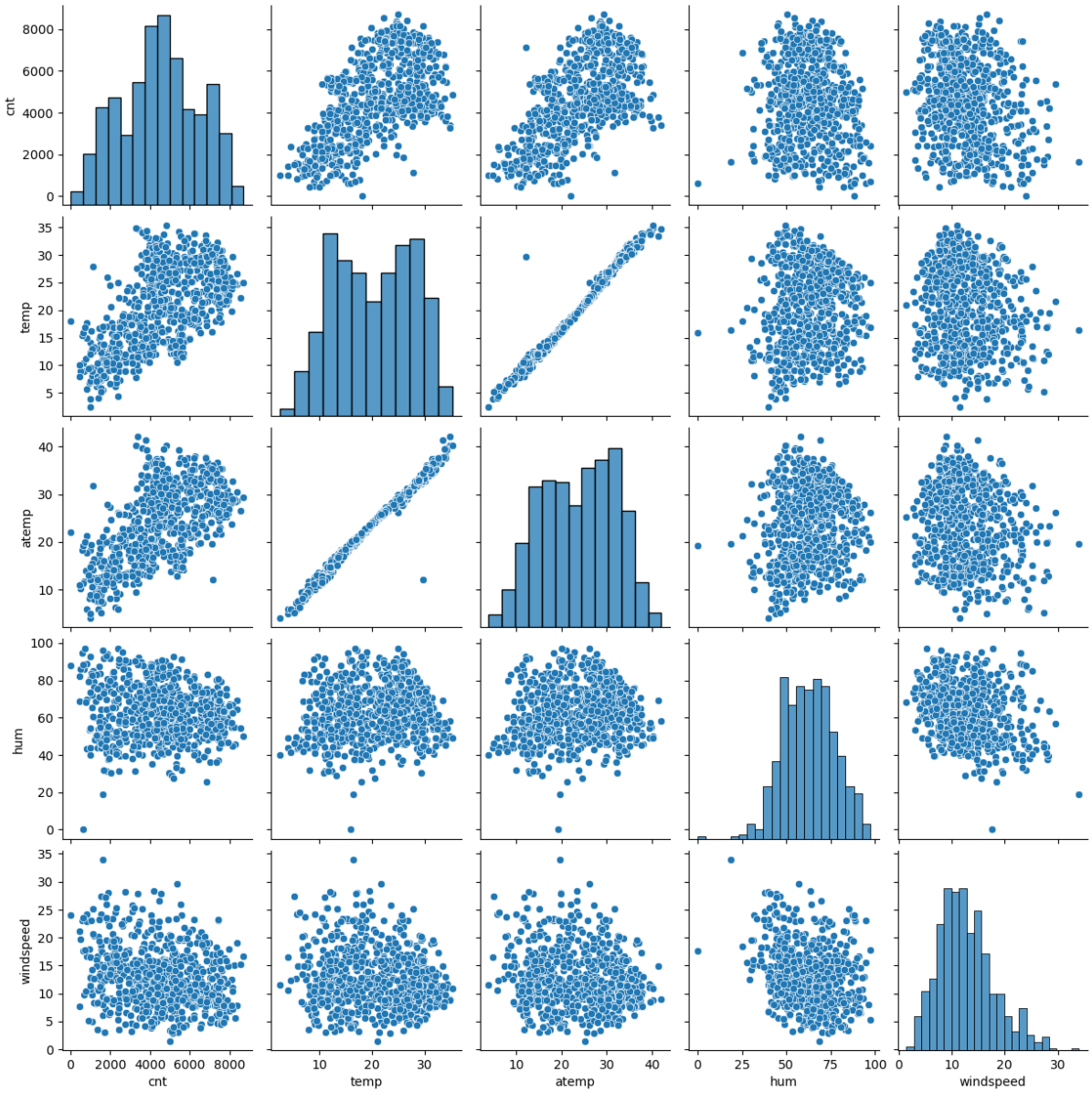
- Weekday: Rental counts remained relatively even throughout the week.

- Workingday: The median count of users remained consistent throughout the week.

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

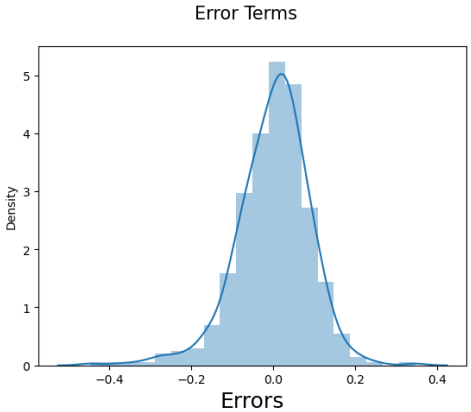
   Not dropping the first column when creating dummy variables can result in correlation and redundancy issues, especially with smaller cardinalities. This can adversely affect model convergence, distort variable importance lists, and impact performance. By using 'drop_first=True' during dummy variable creation, we mitigate these issues, leading to improved model performance by reducing the number of columns and training time.

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



By observing above pairplot it can be seen that, “temp” and “atemp” are the two numericalvariables which are highly correlated with the target variable (cnt)

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



- Initially, we assessed the linearity assumption by visualizing the numeric variables through a pairplot. The aim was to determine whether there exists a linear relationship between the independent and dependent variables. Further details can be found in the notebook.

- The second test involved checking the normal distribution of residuals with a mean centered around 0. This assumption was verified by plotting a distplot of residuals and confirming their adherence to a normal distribution. The diagram below illustrates that residuals are distributed around a mean of 0.

- Addressing the assumption of minimal multicollinearity, we employed the Variance Inflation Factor (VIF). This quantitative measure helped assess how strongly the feature variables are correlated with each other in the new model. Additional information can be found in the notebook.

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

    - According to the final model, the top three features contributing significantly to explaining demand are as follows:

    1. Temperature (Coefficient: 0.437655)
    2. Weather Situation (weathersit): Light Snow, Light Rain + Mist & Cloudy (Coefficient: -0.292892)
    3. Year (Coefficient: 0.234287)

General Subjective Questions

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

Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values. Linear Regression is the most basic form of regression analysis. Regression is the most commonly used predictive analysis model.

Linear regression is based on the popular equation “y = mx + c”.

It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x).

In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable. Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc. Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error.

In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term.

Regression is broadly divided into simple linear regression and multiple linear regression.

- **Simple Linear Regression**: SLR is used when the dependent variable is predicted using onlyone independent variable.

- Best Fitting Line Equation

$$\hat{y}_i = b_0 + b_1 x_i + \varepsilon_0$$

Response Variable — Predictor Variable — Residual Error

- **Multiple Linear Regression**: MLR is used when the dependent variable is predicted usingmultiple independent variables.

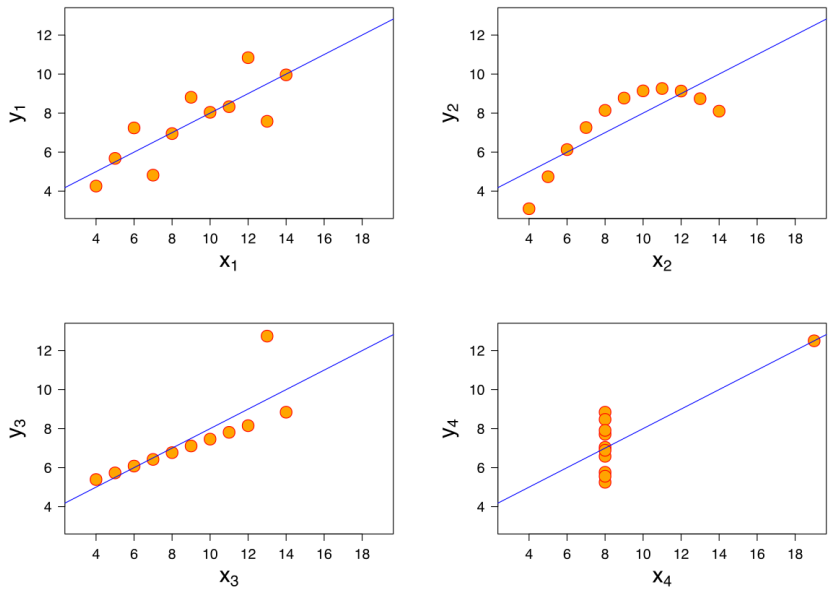
$$\text{observed data} \rightarrow \quad y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p + \varepsilon$$

$$\text{predicted data} \rightarrow \quad y' = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

$$\text{error} \rightarrow \quad \varepsilon = y - y'$$

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

Francis Anscombe created Anscombe's Quartet, consisting of four datasets with nearly identical statistical characteristics. Despite their similar statistics, these datasets exhibit distinct distributions and display significantly different patterns when graphed. The purpose of Anscombe's Quartet is to underscore the significance of visually representing data before conducting analysis and to highlight the impact of outliers and influential observations on statistical properties.



● The first scatter plot (top left) appears to be a simple linear relationship.

● The second graph (top right) is not distributed normally; while there is a relation between them, it's not linear.

● In the third graph (bottom left), the distribution is linear, but should have a different regression line The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

● Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

3. What is Pearson's R? (3 marks)
   Pearson's r is a numerical summary of the strength of the linear association between the variables.It value ranges between -1 to +1. It shows the linear relationship between two sets of data

Formula

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

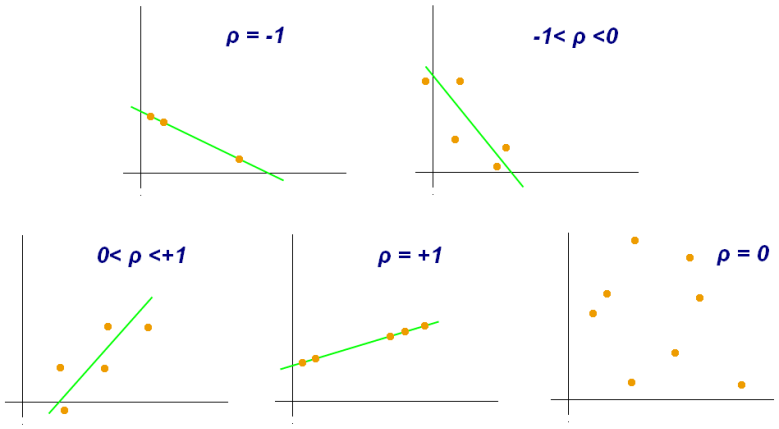
r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable



- r = 1 means the data is perfectly linear with a positive slope
- r = -1 means the data is linear with negative slope.
- R = 0 means there is no linear association.

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

- **Feature scaling** is a method used to normalize or standardize the range of independent variablesor features of data. It is performed during the data preprocessing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.
- Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbours and Neural Networks.

- Standardization, on the other hand, can be helpful in cases where the data follows a Gaussiandistribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, theywill not be affected by standardization.

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

The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. If there is perfect correlation, then VIF = infinity. It gives a basic quantitative idea about how much thefeature variables are correlated with each other. It is an extremely important parameter to test our linear model.

$$VIF = \frac{1}{1 - R^2}$$

Where R-1 is the R-square value of that independent variable which we want to check how well thisindependent variable is explained well by other independent variables. If that independent variablecan be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1. So, VIF = 1/(1-1) which gives VIF = 1/0 which results in
"infinity"

The numerical value for VIF tells you (in decimal form) what percentage the variance (i.e. the standard error squared) is inflated for each coefficient. For example, a VIF of 1.9 tells you thatthe variance of a particular coefficient is 90% bigger than what you would expect if there was no multicollinearity — if there was no correlation with other predictors.

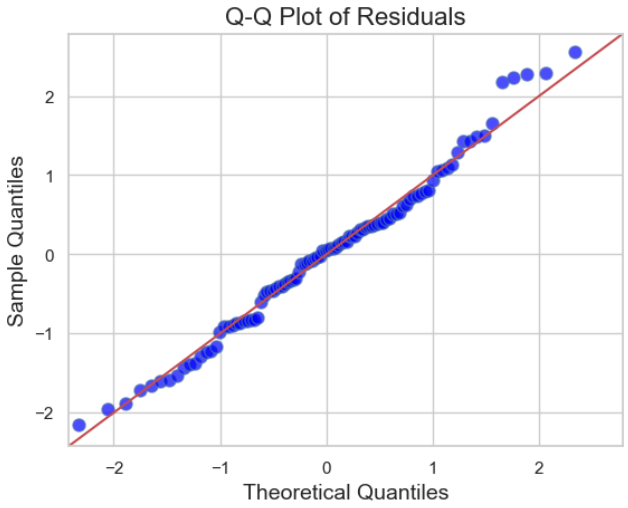
A **rule of thumb** for interpreting the variance inflation factor:

- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.

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

   Q-Q plots in linear regression assess the normality of residuals by comparing the quantiles of observed data to those of a theoretical normal distribution. A 45-degree diagonal line indicates normal distribution. Departures from linearity suggest non-normality, aiding in assumption checking for valid statistical inferences. Q-Q plots identify outliers and skewness, crucial for model validity, serving as a diagnostic tool alongside residual plots. Interpretation involves checking for consistency with the straight line, identifying curvature indicating non-normality, and spotting outliers. In summary, Q-Q plots are vital for ensuring the normality of residuals in linear regression, contributing to model validity and interpretation.

Below is the example of Q-Q plot



The Q-Q plot compares the quantiles of the residuals to the quantiles of a theoretical normal distribution. If the residuals follow a normal distribution, the points should fall along the 45-degree reference line. Departures from the line may suggest non-normality.