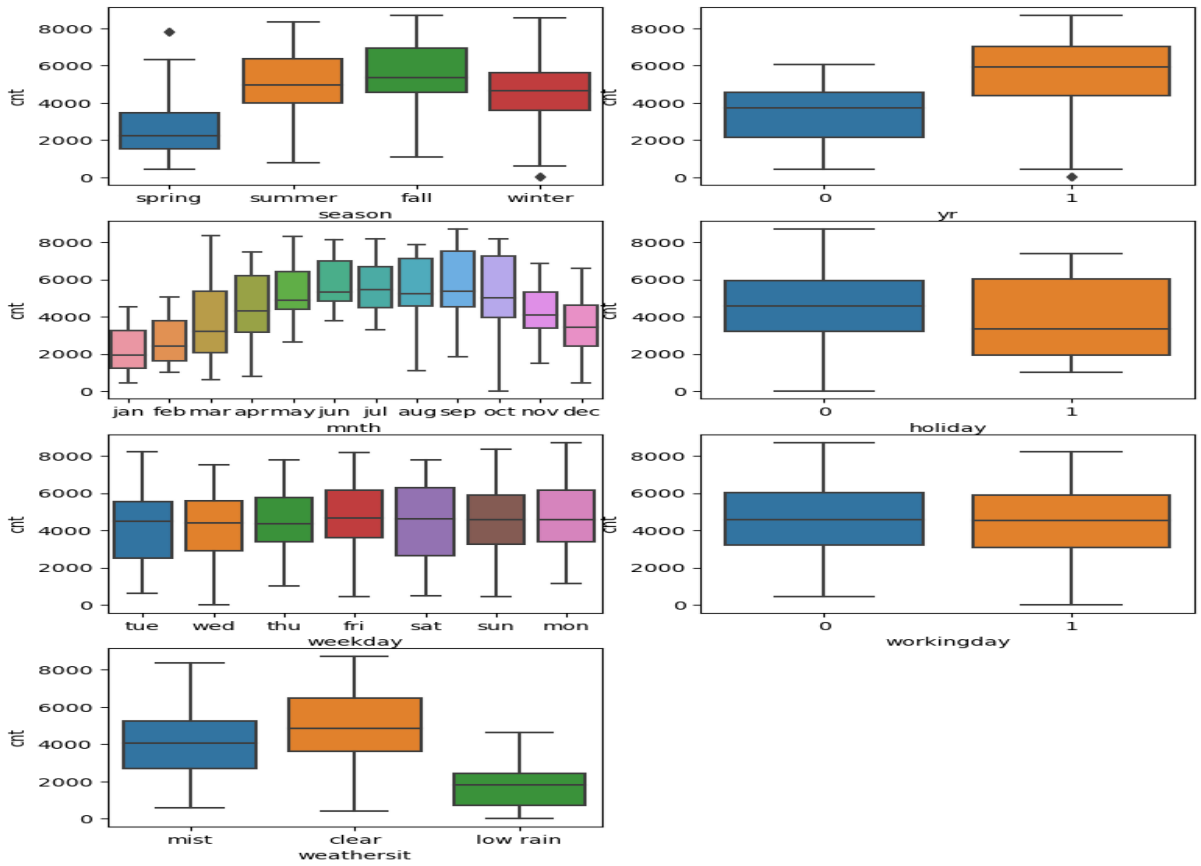
Assignment-based Subjective Questions

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

ANS:



The categorical variables are season, yr, mnth, holiday, weekday, workingday, weathersit and the insights getting from boxplot for categorical variables are:

- **Season:** The Fall season highest number of counts, The Spring season has lowest number of counts and summer and winter comes in middle as to have almost same amount of count.
- **Yr:** Between 2018 and 2019 ,2019 has highest number of counts.
- **mnth:** - march and September has maximum no of counts whereas Jan and oct has minimum no of counts. From June to September all have almost same maximum and medium and these months have higher no of counts among all months.
- **holiday:** Number of counts are higher when there is no holiday.
- **Weekday**: Almost all weekdays have same number of counts as they have almost same medium.
- **workingday:** working day has almost same number of counts.
- **weathersit:** Most of the users like to use the bike whenever the weather is clear, then number of count decreases in mist weather, very a smaller number of user use bike in low rain and No users uses bike in heavy rain weather according to the data.

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

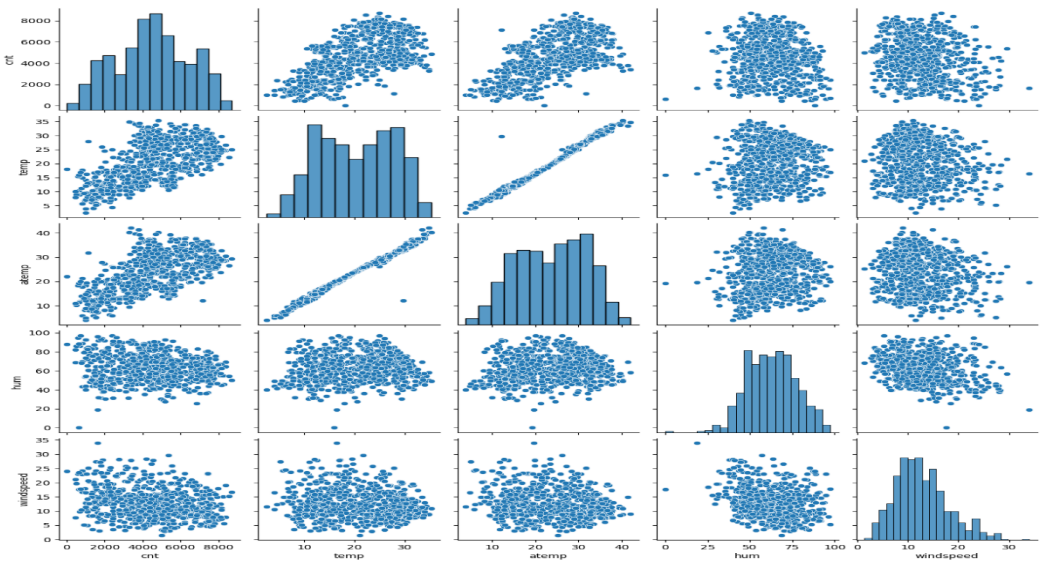
ANS: It is important to drop first variable or column getting after creating dummy variables because –

- There will no use of the column in the future.
- We can get the answer from the others columns too.
- If we don not drop the column, there are chances that we will get the VIF value infinity, which can affect the model and we will not get the perfect rank of the variables for use.
- For ex. If we do not remove the first dummy column created from weathersit, we get the inf value in the VIF and they are on the top of the VIF although they have negative impact on the mode.

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

ANS: - temp and atemp has highest correlation with the target variable(cnt)

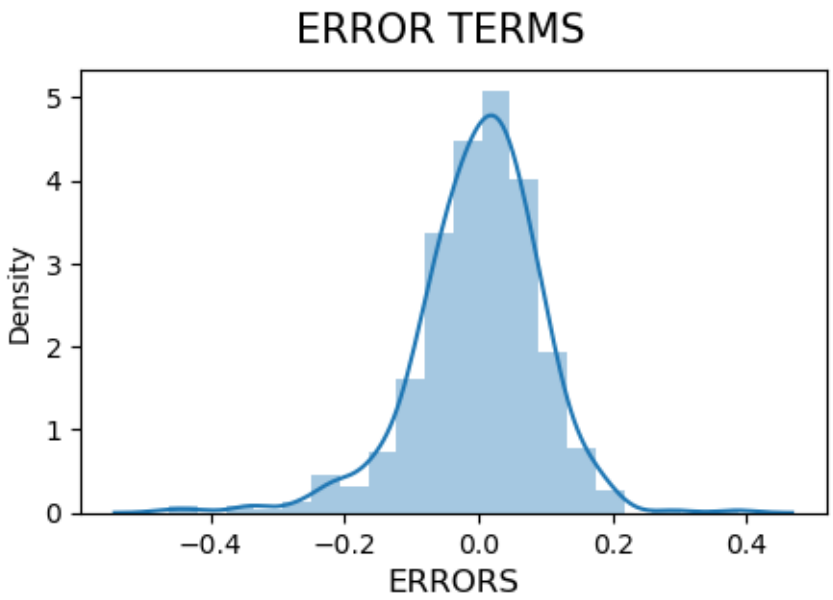
- temp and atemp strongly linearly related to each other as both variables are temprature in celcius



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

ANS: Assumptions:

- In Linear Regression, there should be linear correlation between the dependent variable and independent variable and we have made a pair plot which tells the correlation between variables
- In Linear Regression, there should be normal distributed curve for the residuals getting after actual value – predicted value and its follows the



assumptions as shown in fig.

- In Linear Regression, there should be minimum ,no collinearity or constant which is caused when the independent variables are highly correlated with each other's. So, we have checked the multicollinearity using VIF in the notebook and the final model's VIF shows the vVIFs under acceptable range or minimum.

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

ANS: The feature which can affect the count are:

- **Temp variable** with coefficient of 0.433 which tells that a unit increase in temp increases the count by 0.433 units
- **yr variable** with coefficient of 0.233 tells that a unit increase in yr increases the count by 0.233 units
- **low rain** (weatersit) variable with coefficient of (-0.307) tells that unit increase in low rain decreases the count of by .0307 units

General Subjective Questions

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

ANS: Linear regression is a widely used supervised machine learning algorithm used for predicting a continuous output variable (also known as the dependent variable) based on one or more input features (also known as independent variables). It establishes a linear relationship between the input features and the output variable, assuming that the relationship can be approximated by a straight line.

Linear regression algorithm:

- Assumptions:
  - Linearity: Linear regression assumes that the relationship between the input features and the output variable is linear.
  - Independence: The input features should be independent of each other, meaning they do not have strong correlations.
  - Homoscedasticity: The variance of the errors (residuals) should be constant across all levels of the input features.
  - Normality: The errors should follow a normal distribution with a mean of zero.
- Model Representation:
  - Simple linear regression, we have one input feature (X) and one output variable (Y), and the relationship can be represented as:

    $Y = \beta0 + \beta1 * X + \varepsilon$

    Y: The predicted output variable.

    X: The input feature.

    β0: The y-intercept (the value of Y when X is 0).

    β1: The slope of the line (the change in Y corresponding to a unit change in X).

    ε: The error term (the difference between the predicted value and the actual value).
  - Multiple linear regression, we have more than one input feature (X1, X2, ..., Xn), and the relationship can be represented as:

    $Y = \beta0 + \beta1 * X1 + \beta2 * X2 + ... + \beta n * Xn + \varepsilon$
- Fitting the Model:

  Once the optimal values of β0, β1, ..., βn are determined (either through gradient descent or analytical solutions), the linear regression model is considered "trained" and can be used for making predictions on new data.

- Making Predictions:

  To make predictions on new data, simply plug the input feature values into the regression equation:

  $Y\_predicted = \beta0 + \beta1 * X1 + \beta2 * X2 + ... + \beta n * Xn$

- Evaluation:

  The performance of the linear regression model is typically evaluated using metrics such as R-squared, Mean Absolute Error (MAE), Mean Squared Error (MSE), or Root Mean Squared Error (RMSE).

Linear regression is a simple yet powerful algorithm, especially when the relationship between the input features and the output variable is approximately linear. However, it may not perform well when dealing with complex, nonlinear relationships in the data.

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

ANS:
Anscombe's quartet is a collection of four small datasets that have nearly identical statistical properties, yet they exhibit significantly different patterns when graphed. This is to emphasize the importance of data visualization and to illustrate that relying solely on summary statistics can be misleading.

The four datasets in Anscombe's quartet :

- 1. Linear Relationship
  The dataset forms a relatively clear linear relationship, where the points roughly follow a straight line. A linear regression on this data would yield a good fit.
- 2. Non-linear Relationship
  The dataset looks somewhat like a linear relationship but with a slightly curved pattern. However, it is not a perfect fit for a straight line, indicating that a non-linear model might be a better fit.
- 3.Outlier
  This dataset has a linear relationship, but with outlier. This outlier significantly affects the regression line and shows the importance of detecting and handling outliers in data analysis.
- 4. Discontinuous Relationship
  This dataset exhibits a clear discontinuous relationship. The data points can be divided into two groups, with one group forming a straight line and the other group having a different slope.
- Summary statistics (mean, variance, correlation, etc.) can be identical across different datasets, but the underlying patterns and relationships can vary significantly. It underscores the importance of data visualization and exploring data graphically to gain deeper insights before drawing conclusions or making predictions.

3. What is Pearson's R? (3 marks)
ANS: Pearson's R, also known as Pearson correlation coefficient or simply correlation coefficient, is a statistical measure that quantifies the linear relationship between two continuous variables. It is denoted by the symbol "r."

The Pearson correlation coefficient ranges from -1 to +1, where:

- r = +1 indicates a perfect positive linear relationship between the two variables. This means that as one variable increases, the other also increases proportionally.
- r = -1 indicates a perfect negative linear relationship between the two variables. This means that as one variable increases, the other decreases proportionally.
- r = 0 indicates no linear relationship between the two variables. In this case, the two variables are considered independent of each other.

Pearson's correlation is widely used in various fields, including statistics, social sciences, economics, and many other areas where relationships between variables need to be assessed. It helps researchers and analysts understand how two variables are related and to what extent they tend to move together or in opposite directions.

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

ANS: Scaling refers to the process of transforming the numerical values of different variables to a specific range or distribution. The purpose of scaling is to ensure that all the variables are on a comparable scale, which can be crucial for certain algorithms and analyses. Scaling is performed to avoid issues caused by variables with vastly different ranges, magnitudes, or units, as these differences can lead to biased results and incorrect model behavior.
There are two common methods of scaling:

- **Normalized Scaling (also known as Min-Max scaling):**
  Normalized scaling transforms the data so that it falls within a specified range, typically between 0 and 1. The formula for normalizing a variable x is as follows:
  $x\_normalized = (x - min(x)) / (max(x) - min(x))$
- **Standardized Scaling (also known as Z-score scaling or standardization):**
  Standardized scaling transforms the data in such a way that it has a mean of 0 and a standard deviation of 1. The formula for standardizing a variable x is as follows:
  $x\_standardized = (x - mean(x)) / std(x)$

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

(3 marks)

ANS:

- Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in a regression analysis. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated with each other. VIF helps assess how much the variance of an estimated regression coefficient is increased due to multicollinearity.
- VIF is calculated for each independent variable in the regression model.
- When the VIF is equal to or greater than infinity, it indicates that there is a perfect or near-perfect linear relationship between that independent variable and the combination of other independent variables in the model.
- For ex. If we do not remove the first dummy column created from weathersit , we get the infinity(inf) value in the VIF and they are on the top of the VIF although they have negative impact on the mode, It shows that dummies have perfectly correlated with each other.

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

(3 marks)

ANS:

- Q-Q plot, short for quantile-quantile plot, is a graphical tool used to assess the similarity between the distribution of a dataset and a theoretical distribution, such as the normal distribution. It is a powerful diagnostic tool to identify whether a dataset follows a particular distribution or to detect deviations from the expected distribution. In the context of linear regression, Q-Q plots are essential for evaluating the assumption of normality of the residuals
- The use and importance of a Q-Q plot in linear regression are as follows:
  - **Checking normality assumption**: Q-Q plots of residuals are used to assess this assumption. If the residuals are approximately normally distributed, the points in the Q-Q plot will fall close to the straight line.
  - **Identifying outliers and heavy tails:** Q-Q plots can reveal outliers and heavy tails in the data, which might not be apparent in a histogram or other graphical representations. Deviations from the straight line in the Q-Q plot indicate departures from the expected distribution, suggesting the presence of outliers or extreme values.
  - **Validating model assumptions**: Q-Q plots are a valuable tool to validate the assumptions, especially the normality assumption.
  - **Model improvement:** If the Q-Q plot shows significant deviations from the straight line, it suggests that the model might be mis specified or that transformations of the data might be necessary to improve the model's performance.