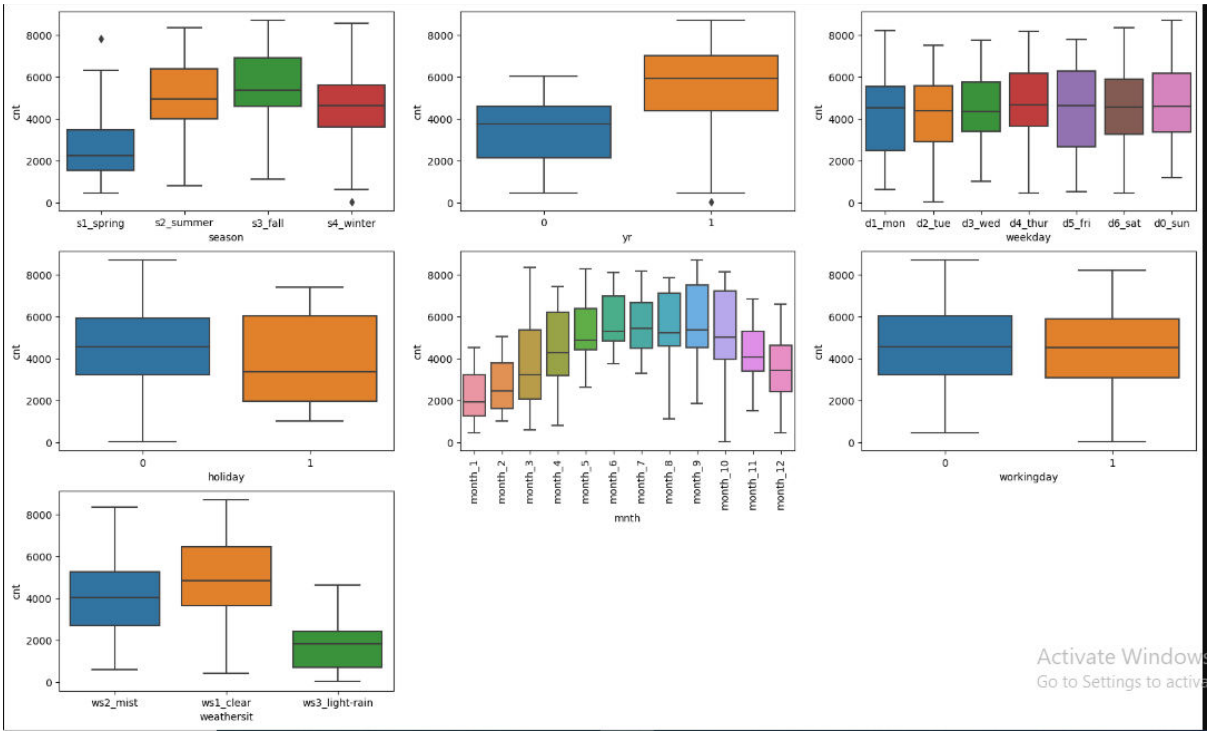
Assignment-based Subjective Questions

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



The categorical variables in the dataset are ‘season’, ’yr’, ’mnth’, ‘weekday’, ‘holiday’, ‘workingday’ and ‘weathersit’.
These categorical variables have following effect on dependent variable (cnt) :

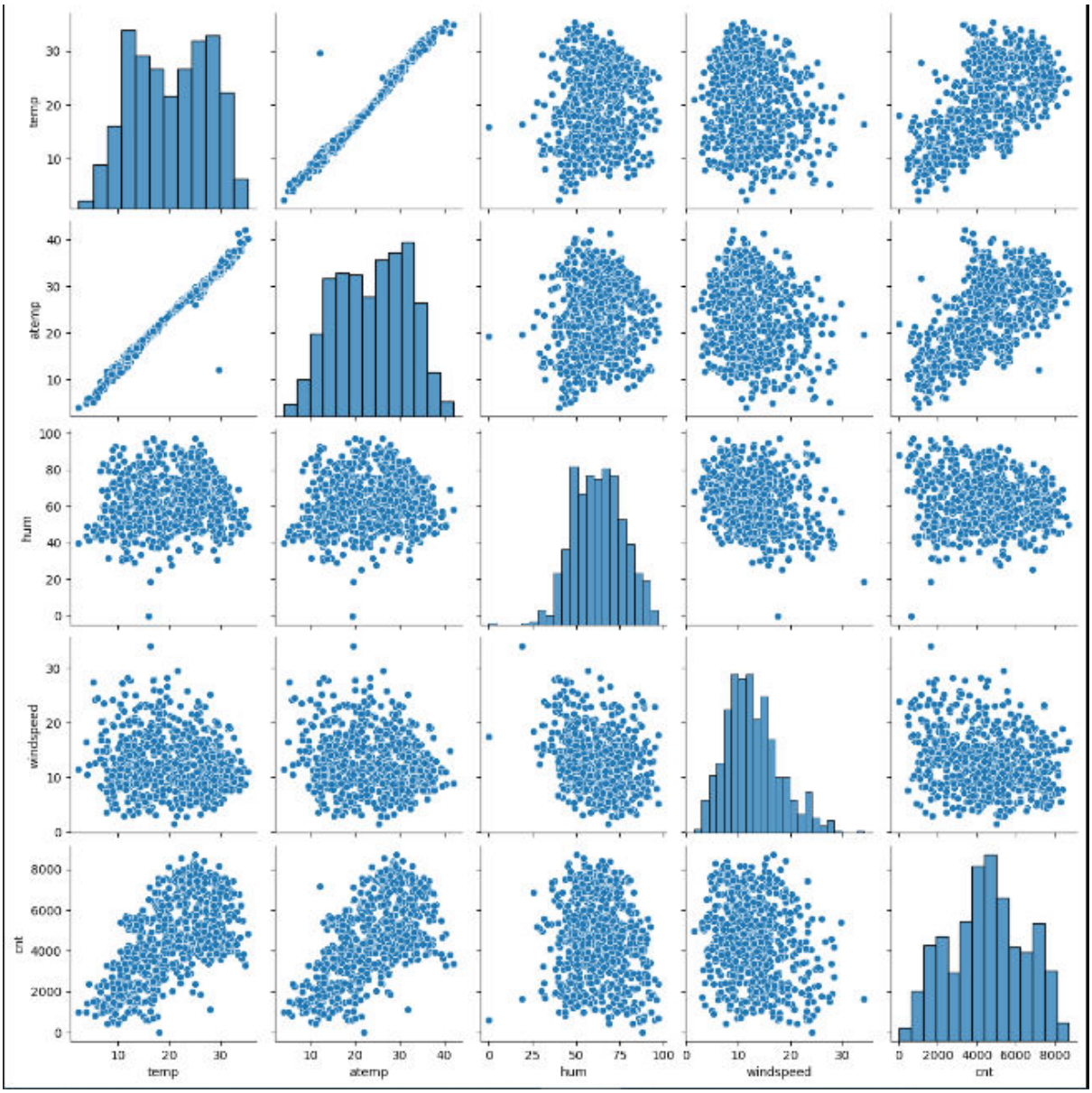
- Season: ‘fall season’ has the highest count of total rental bikes and ‘spring season’ has the lowest count of rental bikes.
- Year (yr): yr_1 (2019) has more count of total rental bikes in comparison to yr_0 (2018).
- Month (mnth): Box plot shows that month_9 (sept) has highest count of rental bikes and month_1 (jan) has lowest count.

- Holiday: Holiday_0 (no holiday) has more count of rental bikes in comparison to holiday_1 (a holiday).
- Weekday: Saturday, Sunday and Monday has more count of rental bikes.
- Working day: workingday_0 (non-working day) has more count of total rental bikes in comparison to workingday_1 (yes a working day).
- Weather sit: Light-rain weather situation has low count of total rental bikes in comparison to clear and mist.

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

- Dummy variables are used to represent categorical data in a numerical format, which is required in many machine learning algorithms.
- Using ‘drop_first = True’, one of dummy variables is omitted because for n categories we need only n-1 dummy variables. The omitted category becomes the reference category.
- Thus ‘drop_first = True’ helps to avoid multicollinearity because if we include all dummy variables, they would be perfectly multicollinear. This can lead to problems in regression analysis.
- Also, interpretation of coefficients becomes easier on dropping one variable, as the omitted category serves as the baseline for comparison.

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

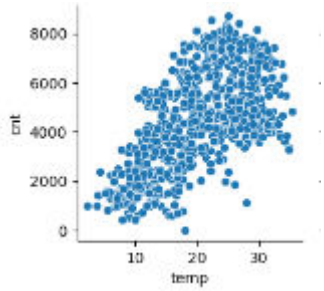


'temp' and 'atemp' are the numerical variables that have the highest correlation with the target variable('cnt').

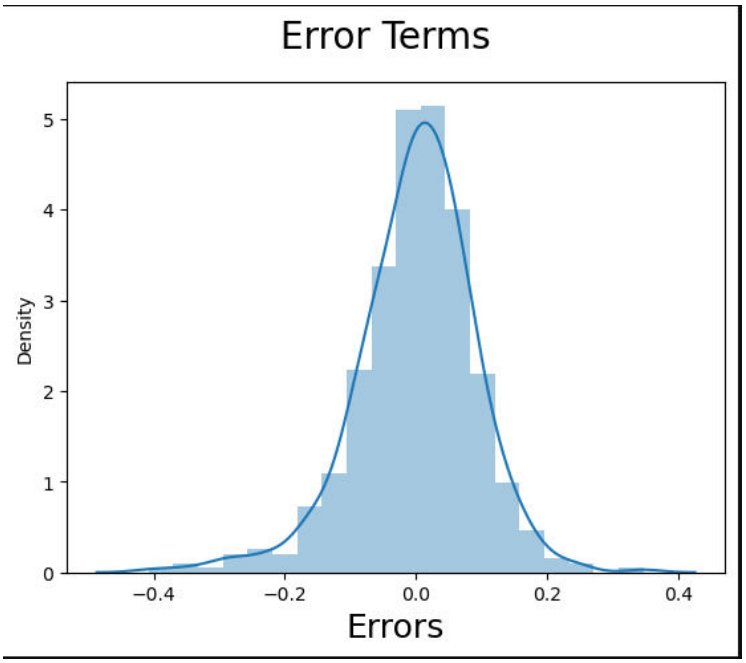
4. How did you validate the assumptions of the Linear Regression after building the model on the training set?

Assumptions of linear regression are validated in following ways:

- Linear relationship between X and y: The relationship between the dependent variable (cnt) and independent variable (temp) is linear.



- Validating assumption of Normal distribution of error terms – From below graph we can see that distribution of residual is normal distribution with mean equal to zero.



- There is no multicollinearity as the values of VIFs for final model are within the acceptable range. It shows that independent variables are not highly correlated with each other.

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

Top 3 features contributing significantly towards the demand of shared bikes, according to final model are-

- Temperature(temp) : It has a coefficient of '0.5174', which means a unit increases in temp variable increases the count of total rental bikes by 0.5174 units.

- Weathersit_3(ws3_light-rain): It has a coefficient of '-0.2828', which means a unit increase in ws3_light-rain variable decreases the count of total rental bikes by 0.2828 units.

  (where, weathersit_3 is: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)

- Year(yr): It has a coefficient of '0.2325', which means a unit increase in yr variable increases the count of total rental bikes by 0.2325 units.

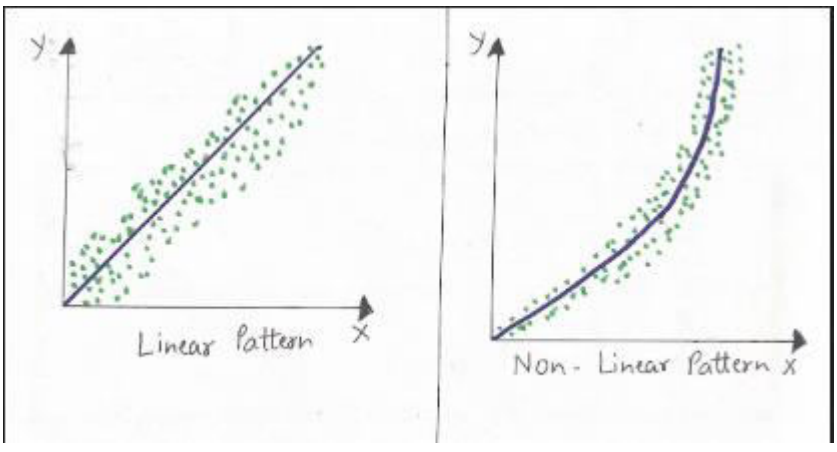
General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

- Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent variables.
- Linear regression is one the easiest and most popular Machine Learning algorithms. It is a statistical method used for predictive analysis.
- Linear regression makes predictions for continuous or numeric variables.
- It creates the best straight-line fitting to the provided data to find the best linear relationship between the independent and dependent variables.

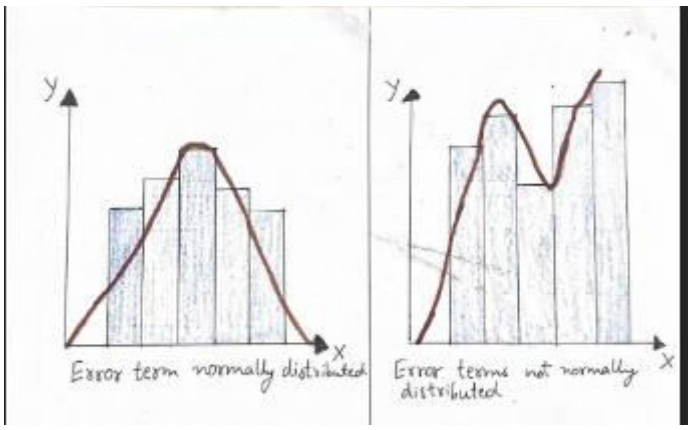
Assumptions in linear regression model are –

1) Linear relationship between X and y: X and y should display some sort of a linear relationship, otherwise there is no use of fitting a linear model between them.

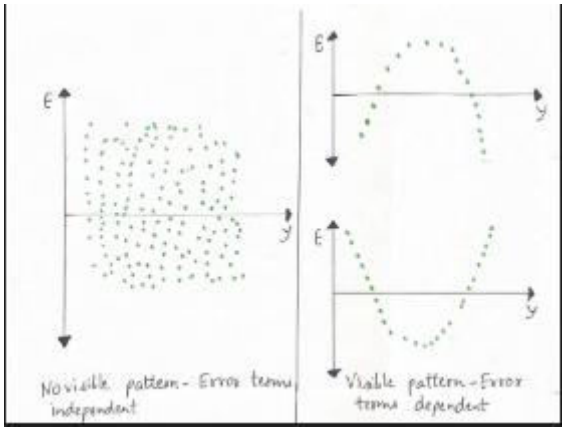


2) Normal distribution of error terms: It represents the assumption of normality. As it has been seen that error

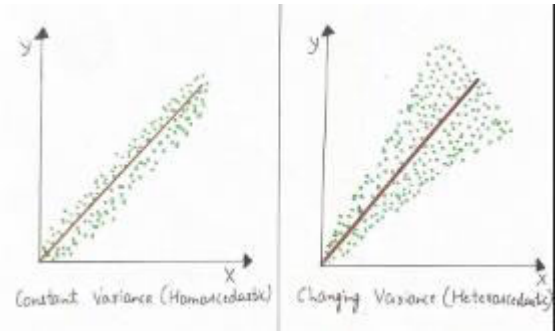
terms generally follow a normal distribution with mean equal to zero in most cases.



3) Independence of error terms: The error terms should not be dependent on one another.



4) Constant variance of error terms: According to this assumption, the variance should not increase or decrease as the error value changes. And variance should not follow any pattern as the error term changes.



Linear Regression is of 2 types:

1) **Simple Linear Regression (SLR)** - It explains the relationship between a dependent variable and only one independent variable using a straight line.
   Equation for SLR: $y = \beta_0 + \beta_1 X$
   Where, y is the dependent variable, X is the independent variable, β_0 is intercept (or constant), β_1 is the slope.

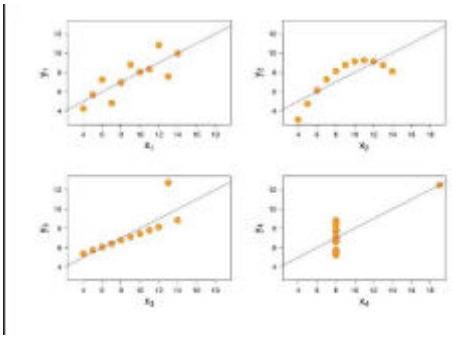
2) **Multiple Linear Regression (MLR)** - It shows the relationship between one dependent variable and multiple independent variables. It fits a 'hyperplane' instead of a straight line.
   Equation for MLR: $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ldots + \beta_p X_p$
   Where, y is the dependent variable, (X_1, X_2, till X_p) are the independent variables, β_0 is constant term, β_1 is coefficient of X_1, β_2 is coefficient of X_2 and so on.

2. Explain the Anscombe's quartet in detail.

- Anscombe's quartet is basically a set of four dataset, which have identical descriptive statistical properties in terms of mean, variance, R-squared, correlations, and linear regression line but when we scatter plot on graph they have different representations.
- Statistician Francis Anscombe in 1973 created these datasets in order to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.
- The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data.
- Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics.
- Anscombe's quartet also emphasizes on the importance of using data visualization to spot trends, outliers, and other

crucial details that might not be obvious from summary statistics alone.

- On plotting Anscombe's quartet four dataset, each dataset seems to have a unique variability patterns and distinctive correlation strength.



- First scatter plot represents linear relationship with some variance.
- Second scatter plot doesn't show linear relationship, instead show a curve.
- Third scatter plot represents a tight linear relationship between x and y, expect one large outlier.
- Forth scatter plot represents that value of x remain constant, except for one outlier.

3. What is Pearson's R?

- Pearson's R also known as Pearson correlation coefficient, is a statistical test that measures the strength between the different variables and their relationships. It is denoted by 'r'.
- In simple words, Pearson's R calculates the effect of change in one variable when the other variable changes. It is the most common way of measuring a linear correlation.

- It basically estimates the relationship strength between the two continuous variables.
- Pearson's R varies between -1 and +1.
- Where, r = +1 indicates perfect positive correlation between the variables.(i.e., both variables tend to change in the same direction)
- Where, r = -1 indicates perfect negative correlation between the variables. (i.e., both variables tend to change in the different directions)
- Where, r = 0 indicates that No correlation exists between the variables.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

- Scaling is the process of standardizing or transforming the features or variables in a dataset so that they have a similar scale or range.
- Many times, it has been seen that in dataset multiple variables are in different ranges. In such case scaling is necessary to bring them all in a single range.
- Scaling helps to ensure that no single feature has an undue influence on the learning algorithm just because it has a larger magnitude.(It is a important step in data preprocessing before training the model)
- Some models like LR which interpret coefficients as feature importance. So, scaling ensures that the coefficients represent the actual impact of each feature on the target variable.

Difference between Normalized scaling and standardized scaling:

1) **Normalized scaling**- It is also called as Min-Max scaling and used to transform features to be on a similar range. Minimum and maximum values of features are used for scaling. It is used when features are of different scales. This scaling makes the features bounded to a specific range. This scales the range to [0,1] or sometimes [-1,1]. It is useful when there are no outliers, as it cannot cope up with them.

Formula of Normalized scaling:

$$X_{new} = (X - X_{min})/(X_{max} – X_{min})$$

2) **Standardized scaling**- It is also called as Z-score normalization. Mean and standard deviation is used for this scaling. It scales the data to have a mean of 0 and standard deviation (sd) of 1. This scaling does not have a bounding range. It does not get affected by outliers as there is no predefined range of transformed features. It is useful when feature distribution is Gaussian or normal.

Formula of Standardized scaling:

$$X_{new} = (x – mean(x))/sd(x)$$

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

- A Variance Inflation Factor (VIF) is the measure of the amount of multicollinearity in regression analysis.

- Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model.
- The value of VIF is calculated by the formula:

  $VIF_i = 1/ (1-R_i^2)$
- In perfect correlation case, value of R-squared is equal to 1, so denominator will be zero and overall value become infinite.
- When value of VIF is infinite, it is the case of perfect multicollinearity.
- Value of VIF is infinite strongly indicates that in the regression model there's a severe multicollinearity issue.
- So we need to address this situation, as this can lead to unreliable coefficients estimation and difficulty in model interpretation.

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

- A Q-Q plots are also known as Quantile-Quantile plot. It is a graphical method for determining whether two samples of data came from the same population or not.
- A Q-Q plot is a plot of the quantiles of the first data set against the quantiles of second data set. (By quantile it means the fraction or percent of points below the given value.)

Use of Q-Q plot in Linear Regression:

- Q-Q plots are also used to check if a dataset follows a particular distribution like exponential, normal or uniform.
- Particularly in Linear regression model, they are used in model validation, to assess if the residuals follow a normal distribution (which is one of the assumption of many regression model).

Importance of Q-Q plot are-

- Distribution Assessment: Q-Q plots are important for assessing whether a dataset follows a particular probability distribution, (such as normal, exponential, or others). So, this helps in selecting appropriate statistical methods for analysis.
- Outlier detection: Q-Q plots can help identify outliers or deviations from the expected distribution. As outliers can have a significant impact on statistical analysis, so detecting them is important.
- Model validation: In domain like regression analysis, Q-Q plot are used to validate assumptions about the distribution of residuals. If the residuals don't follow the expected distribution, it suggests the model may not be the best fit for the data.

  In statistics, Q-Q plots serve as an important diagnostic tool, aiding in the interpretation of data and in selecting appropriate statistical techniques.