BIKE SHARING ASSIGNMENT

- ANAGHA RAVISHANKAR

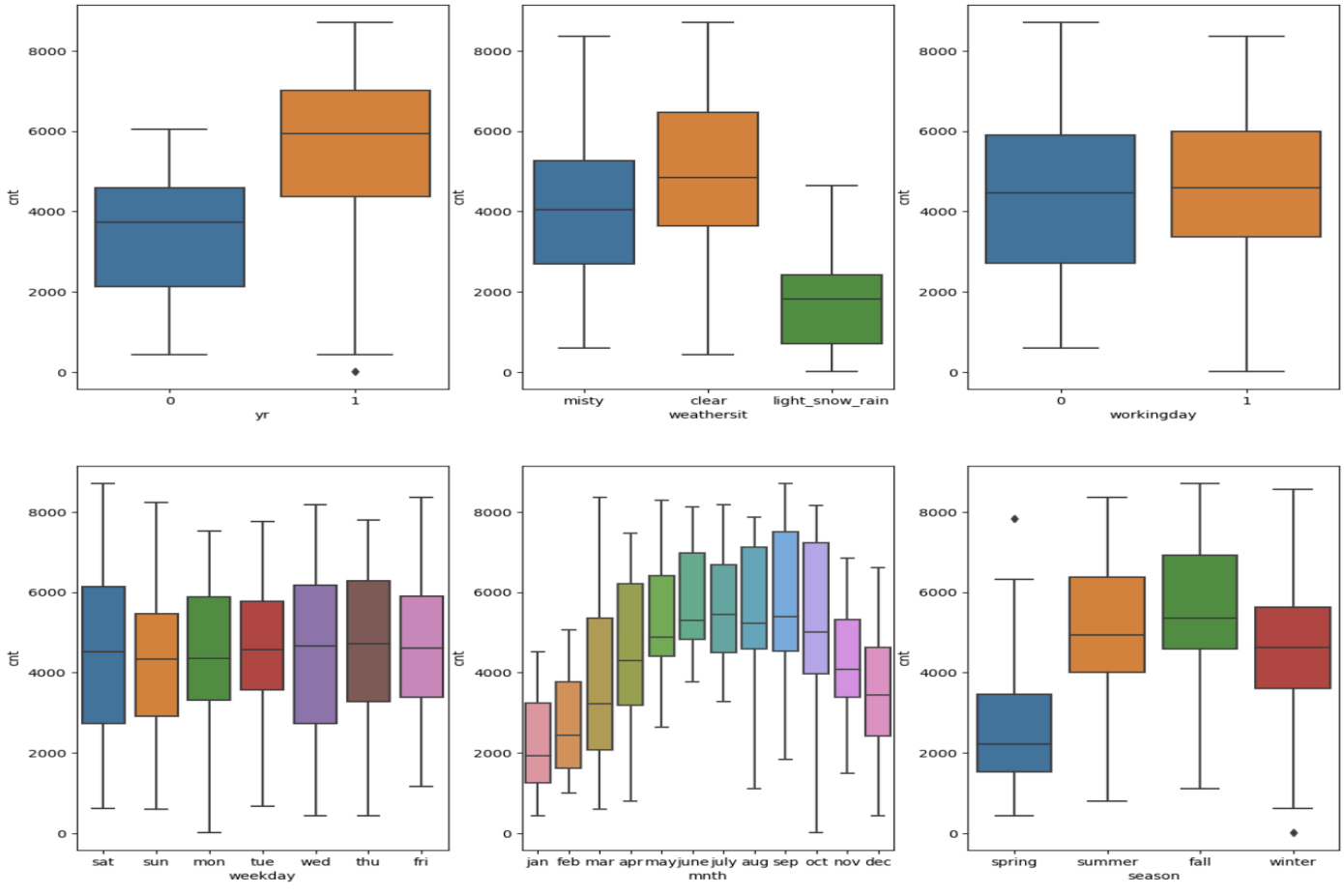
Assignment-based Subjective Questions

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

Answer: The dataset contains 6 categorical variables, which were analyzed using box plots to understand their impact on the dependent variable 'cnt'.

- **Year (yr) :** There is a significant increase in the number of bike bookings from 2017 to 2018. Therefore, as the yr progresses, the dependent variable cnt would also increase..
- **Weather Situation (weathersit):** Most bike bookings occurred during 'clear' weather, Indicating that cnt increases when the weather is clear.
- **Working Day:** A slightly higher percentage of bike bookings occurred on working days. This suggests that working days would not have much impact on cnt.
- **Weekday:** Bike bookings showed a consistent trend across weekdays. This variable may have some or no influence on cnt.
- **Month (mnth):** majority of the bike bookings took place from May to September. This indicates a trend in bookings by month and cnt would be dependent on it.
- **Season:** Most of the bike bookings occurred in fall. This suggests that season could have an impact on cnt.

Overall, yr, weathersit, mnth and season appear to have an effect on the dependent variable cnt, while workingday and weekday may have limited influence.



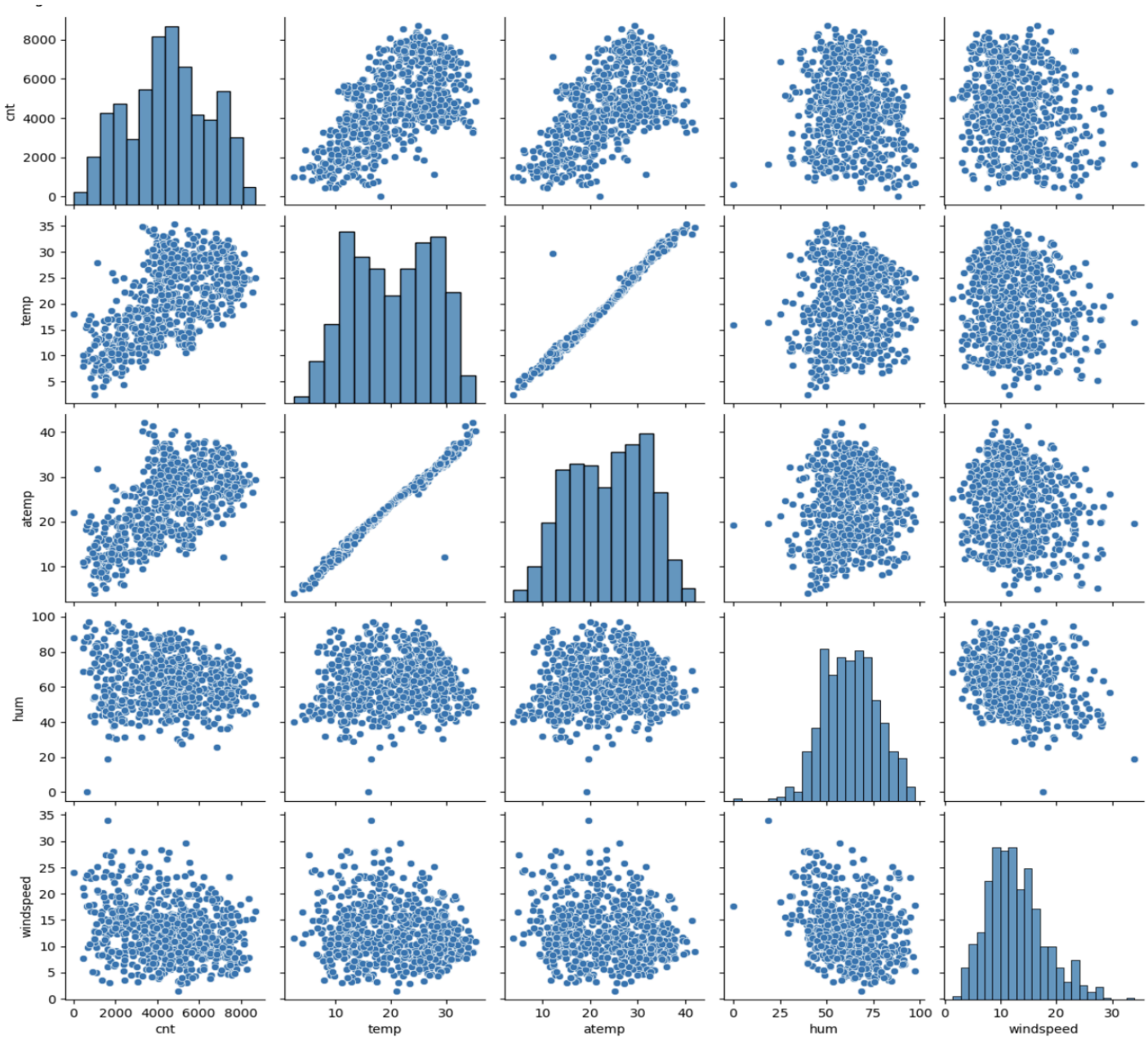
2. Why is it important to use drop_first=True during dummy variable creation?

Answer: It is important to use `drop_first=True` during dummy variable creation **to avoid multicollinearity** issues in the regression model. When creating dummy variables, if you include all levels of a categorical variable, it introduces multicollinearity because one level of the categorical variable can be perfectly predicted from the other levels.

By setting `drop_first=True`, one level of the categorical variable is dropped. This ensures that the model coefficients are properly estimated. Additionally, it **reduces the number of features** in the dataset to **n-1**, making the model efficient.

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

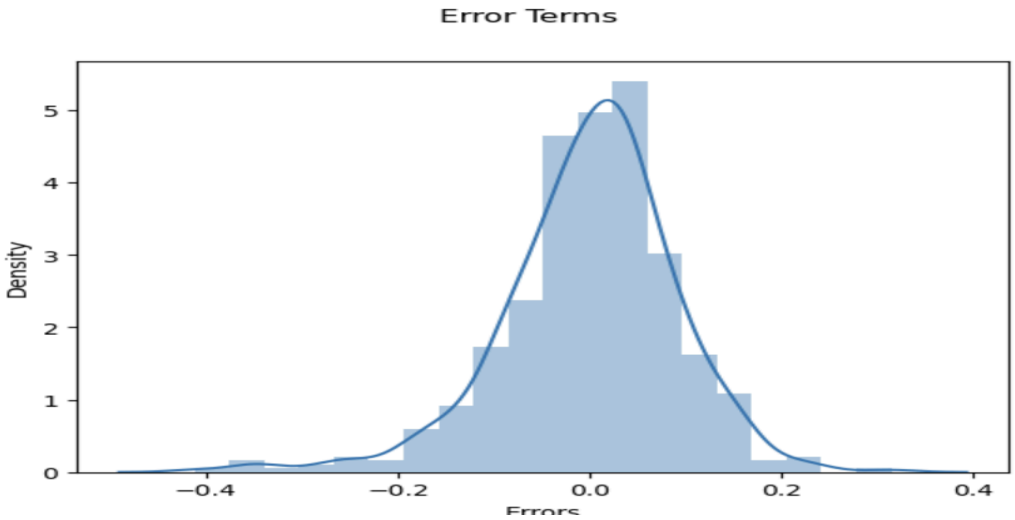
Answer: The variables **'temp' and 'atemp'** have the strongest correlation with the target variable 'cnt' compared to the other variables.



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

Answer: I have assessed the assumptions of the Linear Regression Model using the following five criteria:

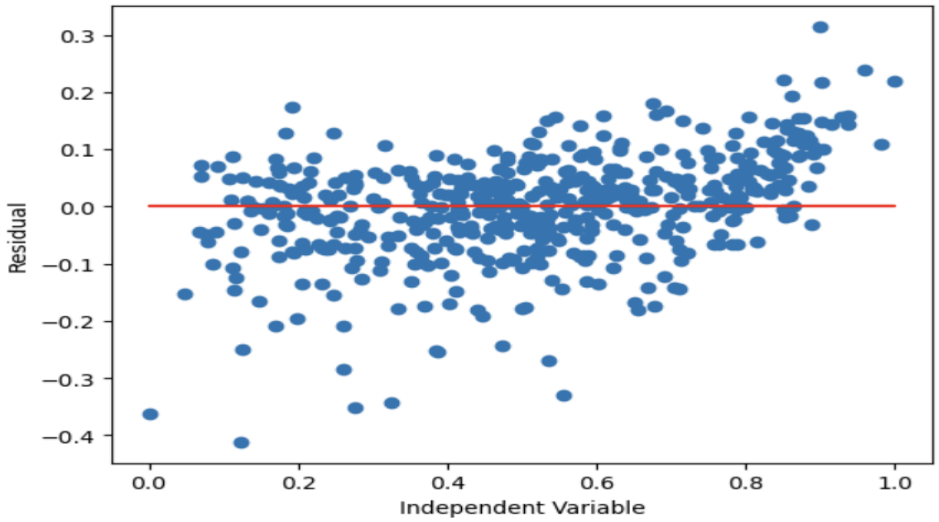
- **Assumption of Normality of Error Terms**: Error terms are expected to follow a normal distribution with mean at 0, which they are following.



- **Multicollinearity Assessment**: There should be no significant multicollinearity observed among the predictor variables.

- **Validation of Linear Relationship**: The relationship between the predictor variables and the target variable should exhibit linearity.

- **Verification of Homoscedasticity**: Residual values should not display any visible pattern, indicating homoscedasticity.



- **Independence of Residuals**: Residuals should be independent of each other, implying no auto-correlation.

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

Answer: The top 3 features contributing significantly towards explaining the demand of the shared bikes are **temp(temperature), weathersit_light_snow_rain(weather) and yr(year)**

- **Temperature**: When the temperature goes up by one unit, bike hires increase by about 0.5499 units.
- **Weather Condition**: If there's light snow or rain, bike hires drop by around 0.2871 units for each unit increase in this condition.
- **Year**: Every year, bike hires go up by approximately 0.2331 units.

It is evident from the regression equation obtained for the model:
$cnt = 0.075 + 0.2331 * yr + 0.0561 * workingday + 0.5499 * temp - 0.1552 * windspeed + 0.0886 * season\_summer + 0.1307 * season\_winter + 0.0974 * mnth\_sep + 0.0675 * weekday\_sat - 0.2871 * weathersit\_light\_snow\_rain - 0.0800 * weathersit\_misty$

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer: Linear regression is a statistical model used to analyze the linear relationship between a dependent variable and a set of independent variables. It helps us understand how changes in one or more independent variables affect the dependent variable. It aims to find the best-fitting linear equation that describes the relationship between the variables.

The relationship between variables is represented mathematically by the equation:
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon$

Here,

- Y : is the dependent variable we want to predict.
- $X_1, X_2 \ldots X_n$: are the independent variable we use for making predictions.
- β_0: is the intercept
- $\beta_1, \beta_2, ..., \beta_n$: are the coefficients representing the change in Y for a one-unit change in each respective independent variable.
- ϵ : is the error term

The linear relationship can be positive or negative:

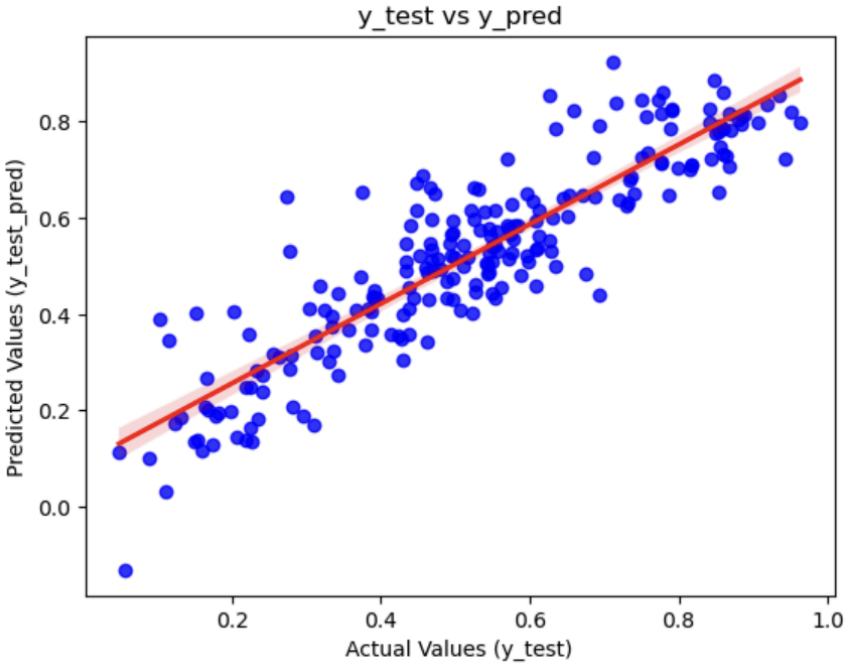
- **Positive Linear Relationship**: Both independent and dependent variables increase together.
- **Negative Linear Relationship**: Independent variable increases while the dependent variable decreases.

Linear regression can be of two types:

- **Simple Linear Regression**: Involves one independent variable.
- **Multiple Linear Regression**: Involves more than one independent variable.

Algorithm:

The goal of linear regression is to estimate the coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ that best fit the data. This is done using the method of least squares, which minimizes the sum of the squared differences between the observed values of Y and the values predicted by the linear equation. Once the coefficients are estimated, they can be used to make predictions on new data by plugging the values of the independent variables into the equation.



Assumptions:

- Linear regression relies on several assumptions, including:
- Linearity: The relationship between the independent and dependent variables is linear.
- Independence: The observations are independent of each other.
- Homoscedasticity: The variance of the error term is constant across all levels of the independent variables.
- Normality: The error term follows a normal distribution.

Conclusion:

Linear regression is a powerful and widely used tool for modeling the relationship between variables.

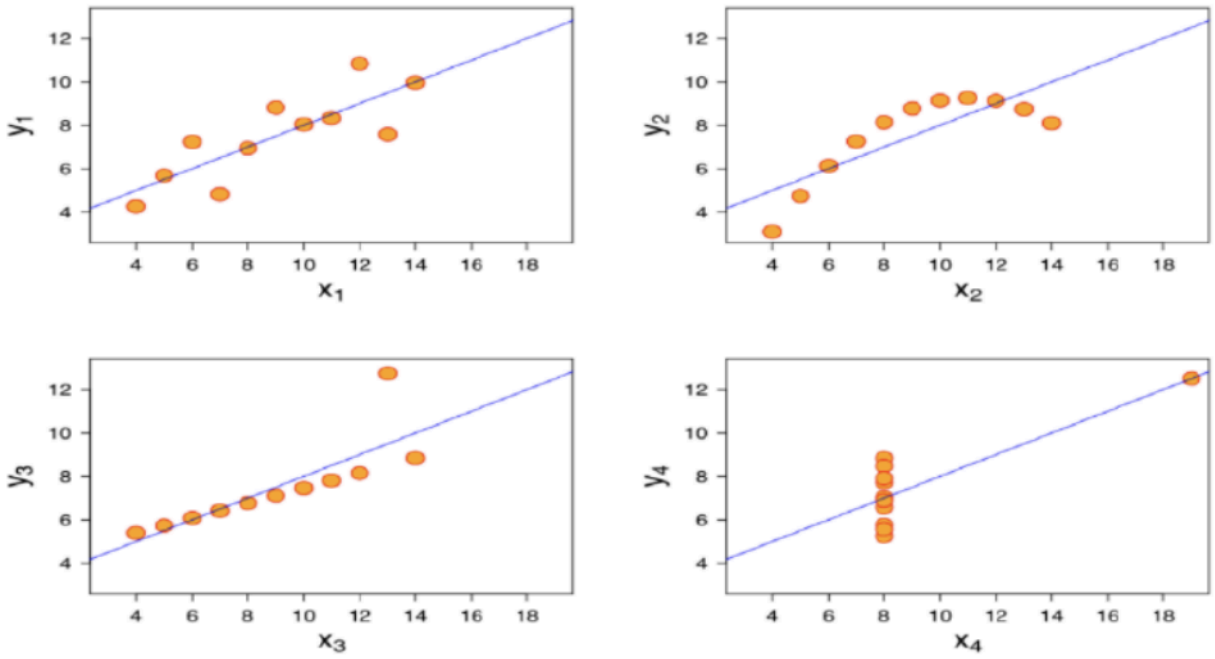
2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet is a famous example in statistics that consists of four datasets, each containing eleven (x, y) data points. What makes Anscombe's special is that despite having very different distributions and relationships between the variables, they all have nearly identical statistical properties when analyzed using statistical measures such as mean, variance, correlation coefficient, and linear regression coefficients.

Background:

Anscombe's quartet was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data before applying statistical analysis. At first glance, the four datasets appear to have vastly different patterns and relationships between the variables. However, when analyzed using traditional summary statistics, they appear almost identical.

Characteristics of Anscombe's Quartet:
Let's see each dataset in Anscombe's quartet:



- **Dataset I**:
  - Linear relationship between x_1 and y_1.
  - Well-behaved, no outliers.

- **Dataset II**:
  - Non-linear relationship between x_2 and y_2.
  - One outlier that significantly influences the regression line.

- **Dataset III**:
  - Strong linear relationship between x_3 and y_3 , except for one outlier.
  - The regression line is heavily influenced by the outlier.

- **Dataset IV**:
  - No apparent relationship between x_4 and y_4, except for one outlier.
  - The outlier significantly affects the regression line.

Significance:
Anscombe's quartet illustrates the limitations of relying solely on statistics to understand data and emphasizes the importance of data visualization. While statistics provide useful insights, they may not capture the full complexity of the data. Visual examination of the data, through techniques such as scatter plots, can reveal patterns, trends, and outliers that may not be apparent from summary statistics alone.

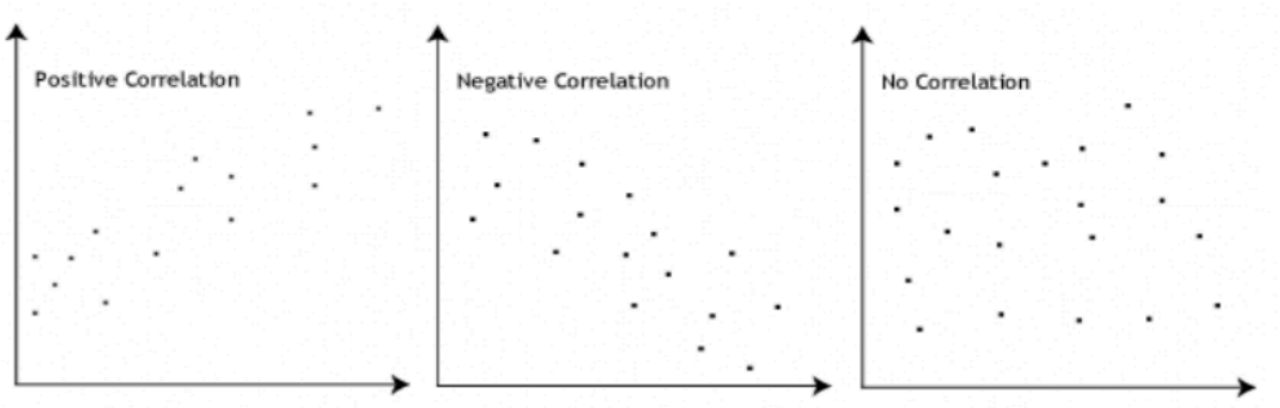
Implications:
The quartet highlights the importance of graphical analysis in exploratory data analysis (EDA). It serves as a caution against over-reliance on statistics and emphasizes the need to visually inspect data before drawing conclusions.

Conclusion:
Anscombe's quartet is a classic example that stresses the importance of data visualization in statistical analysis. By combining statistical analysis with graphical exploration, we can gain a more understanding of the underlying patterns.

3. What is Pearson's R?

Answer: Pearson's correlation coefficient, denoted by r, is a measure of the linear relationship between two continuous variables. It states the strength and direction of the linear association between the variables. Pearson's r ranges from -1 to +1, where:

- r = +1 indicates a perfect positive linear relationship.
- r = -1 indicates a perfect negative linear relationship.
- r = 0 indicates no linear relationship between the variables.



Formula:

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

Assumptions:
Pearson's correlation coefficient assumes that the relationship between the variables is linear, that the data is normally distributed, and that there are no outliers.

Conclusion:
Pearson's correlation coefficient, r, is a valuable tool for indicating the strength and direction of linear relationships between continuous variables. It allows for easy interpretation.

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

Answer:

Scaling:
Scaling refers to the process of transforming the values of variables to a standard range. It involves adjusting the numerical values of the features in a dataset so that they fall within a specific range. Scaling does not change the shape of the data distribution but rather changes the scale or units of measurement.

Purpose of Scaling:
Scaling is performed for several reasons:

- Normalization: Scaling ensures that all variables have a similar scale, preventing features with larger scales from dominating those with smaller scales.
- Faster Convergence: Scaling can help algorithms converge faster during training.
- Improved Performance: Some machine learning algorithms perform better when features are scaled.
- Interpretability: Scaling makes it easier to interpret the coefficients because the coefficients represent the impact of each feature on the target variable in a standardized manner.

Difference:

Aspect	Normalized Scaling	Standardized Scaling
Range of Values	Typically scaled to range [0, 1]	Centered around mean 0, with standard deviation of 1
Purpose	Preserves original range of data, suitable when algorithm requires features to be within a specific range	Centers data around 0, scales to have unit variance, suitable for algorithms assuming normal distribution of features
Sensitivity to Outliers	More sensitive to outliers	Less sensitive to outliers
When to use?	used when we don't know about the distribution	used when distribution is normal.

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

Answer: Variance inflation factor (VIF) is a measure used to detect multicollinearity in regression. Multicollinearity occurs when independent variables in a regression model are highly correlated with each other.

When the VIF is infinite, it typically indicates perfect multicollinearity. Perfect multicollinearity occurs when one or more independent variables in the regression model can be exactly predicted from a linear combination of other variables. (i.e there is a perfect linear relationship among the independent variables)

This perfect multicollinearity leads to mathematical issues in estimating the regression coefficients. Iit becomes impossible for the regression model to estimate unique coefficients for each variable because one variable can be expressed as a perfect linear combination of others.

Perfect multicollinearity can arise due to various reasons such as:

- Including a variable that is a linear combination of other variables already in the model.
- Measurement errors that lead to redundant variables.
- Over-specification of the model, where too many variables are included relative to the sample size.

To address infinite VIF values, it's necessary to identify and resolve the multicollinearity issue by either removing redundant variables, transforming variables, or using techniques like principal component analysis (PCA).

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

Answer: A Q-Q plot (quantile-quantile plot), is a graphical tool used to assess whether a dataset follows a specific probability distribution. In a Q-Q plot, the quantiles of the dataset being analyzed are plotted against the quantiles of a known theoretical distribution, like normal distribution. If the points in the Q-Q plot fall approximately along a straight line, it indicates that the dataset follows the assumed distribution.

In linear regression, Q-Q plots are used to check the assumption of normality of the residuals. The normality of residuals is an important assumption in linear regression because many statistical tests and confidence rely on the assumption that the residuals are normally distributed.

The importance of Q-Q plots in linear regression lies in their ability to help identify potential problems with the regression model. By diagnosing issues such as non-normality of residuals early in the analysis, we can make informed decisions about the model and take appropriate steps to improve the regression results.