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)

Answer :-

1. Season vs total_count

- Observation: Rentals are highest in the fall and summer seasons. Winter has the lowest rentals, and spring shows moderate usage.
- Summary: Fall and summer seasons see a peak in bike rentals, indicating favorable weather conditions. Winter sees the least bike usage.

2. Month vs total_count

- Observation: Rentals increase from January, peaking around May and October, and then decline towards the end of the year.
- Summary: Bike rentals are higher during the warmer months (March to October), with May and October being peak months. Rentals are lowest in December and January.

3. Year vs total_count

- Observation: There is a significant increase in rentals from 2018 to 2019.
- Summary: Bike rentals grew notably from 2018 to 2019, indicating increasing popularity or improved service.

4. is_holiday vs total_count

- Observation: Rentals on is_holidays are slightly higher than on non-is_holidays, but the difference is minimal.
- Summary: Bike rentals remain fairly consistent regardless of is_holidays, with a slight increase on is_holidays.

5. Weekday vs total_count

- Observation: Rentals are relatively even across the weekdays, with no significant drop during weekends.
- Summary: Bike rentals are steady throughout the week, suggesting consistent usage patterns on both weekdays and weekends.

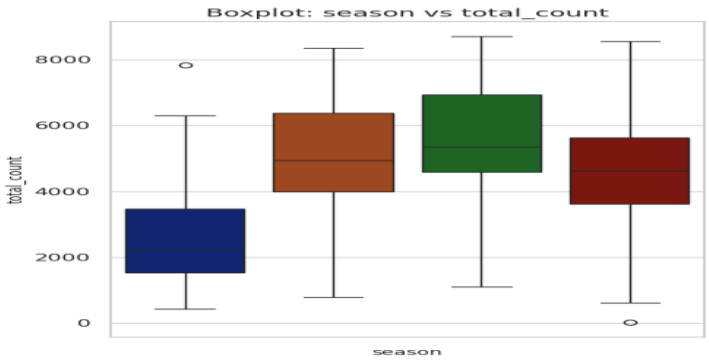
6. is_workingday vs total_count

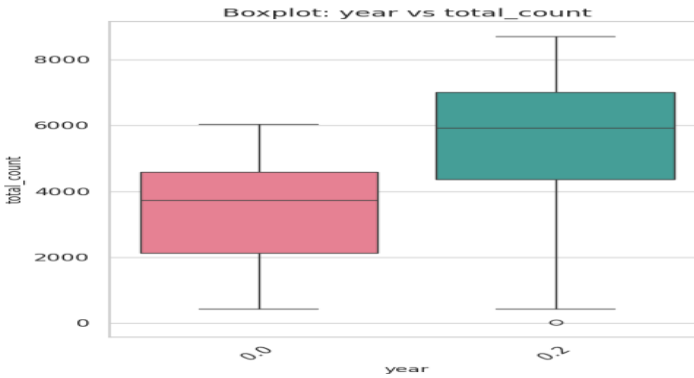
- Observation: Rentals on working days are slightly higher compared to non-working days.
- Summary: There is a marginal increase in bike rentals on working days, possibly indicating commuting usage.

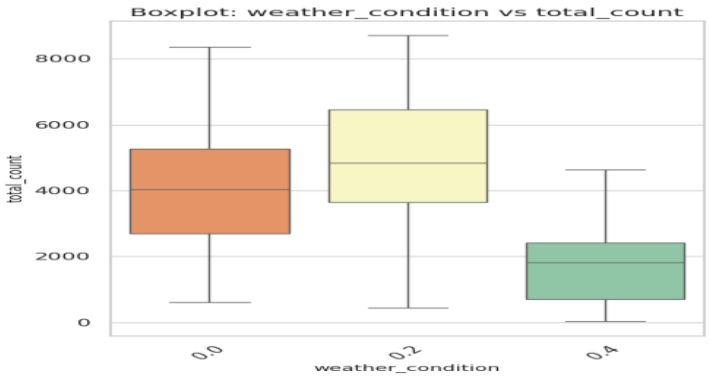
7. weather_condition vs total_count

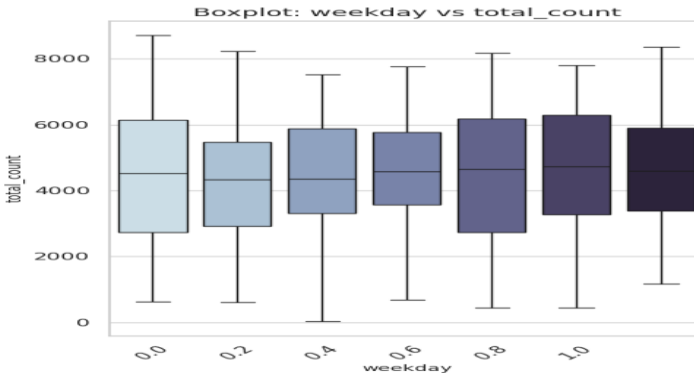
- Observation: Clear weather conditions (good) have the highest rentals, while bad and severe weather conditions see a significant drop.

- Summary: Favorable weather conditions (clear) lead to higher bike rentals. Poor weather conditions (bad and severe) result in lower usage.









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

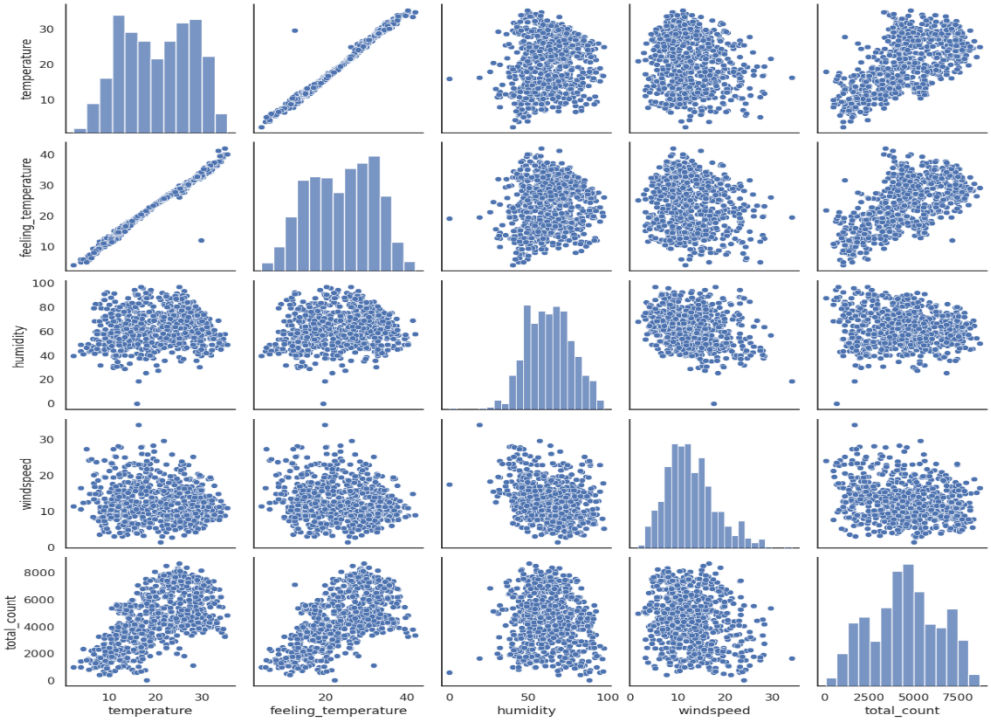
Answer :-

If we do not use drop_first = True, then n dummy variables will be created, and these predictors(n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap. Also If we have three labels for a variable (summer,winter and spring) we can use only two labels summer and winter with Boolean values to represent the third label spring

Summer	Winter	
1	0	-- Represents Summer
0	1	-- Represents Winter
0	0	-- Represents Spring

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

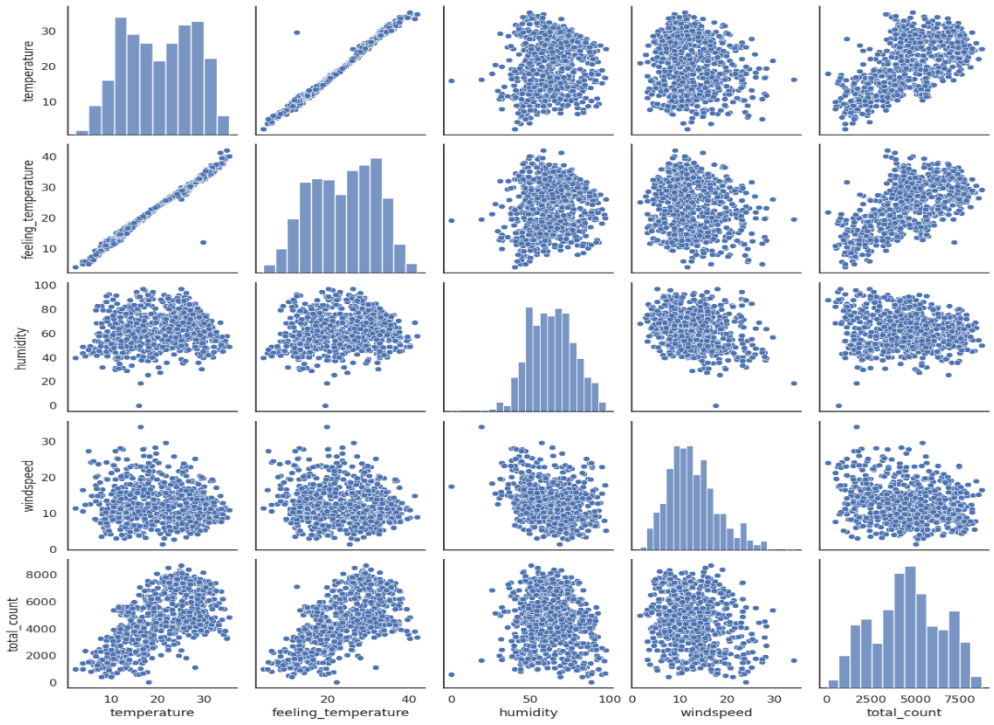
Answer :- Feeling Temperature has the highest correlation with the target variable total_count. The same was found when we plotted correlation matrix between the two variables which came to around (0.65)



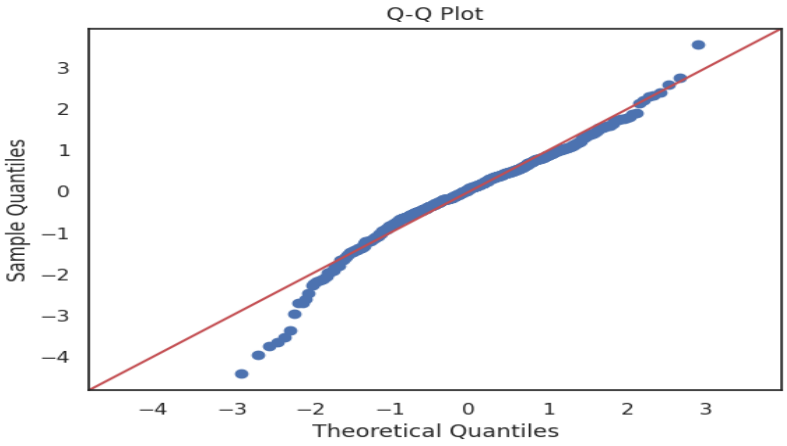
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

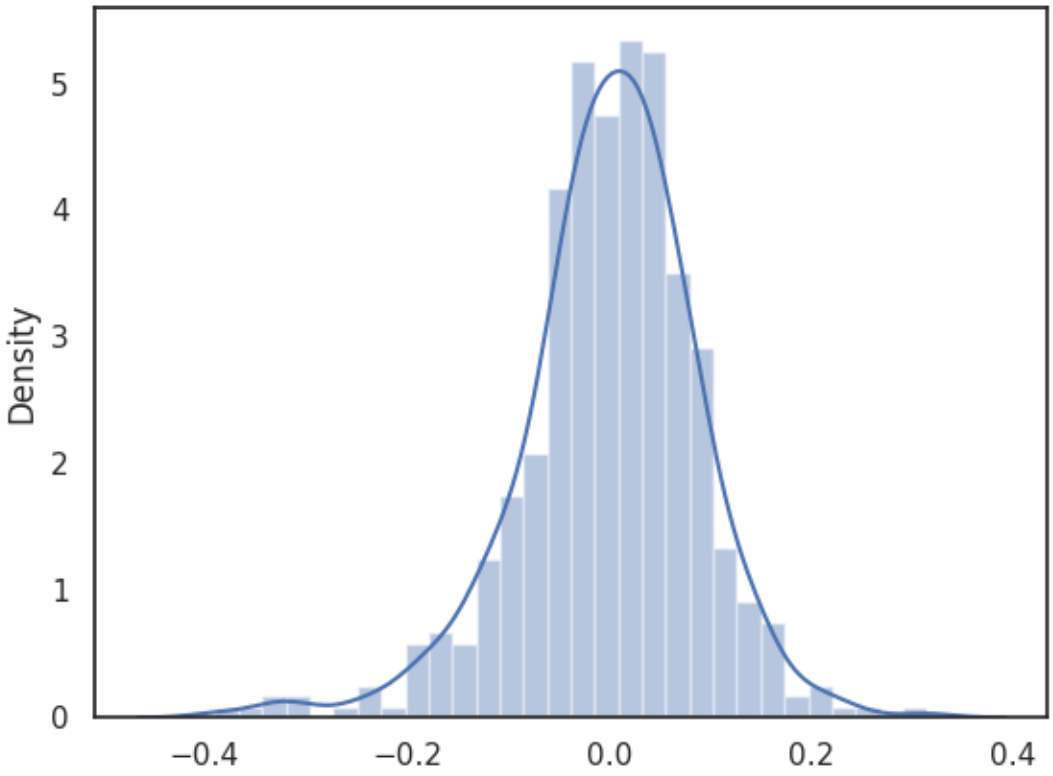
We validated the assumptions of Linear Regression by doing the below checks

1. Linear Relationship :- There is a linear relationship between the dependent and the independent variables, and the assumption holds.

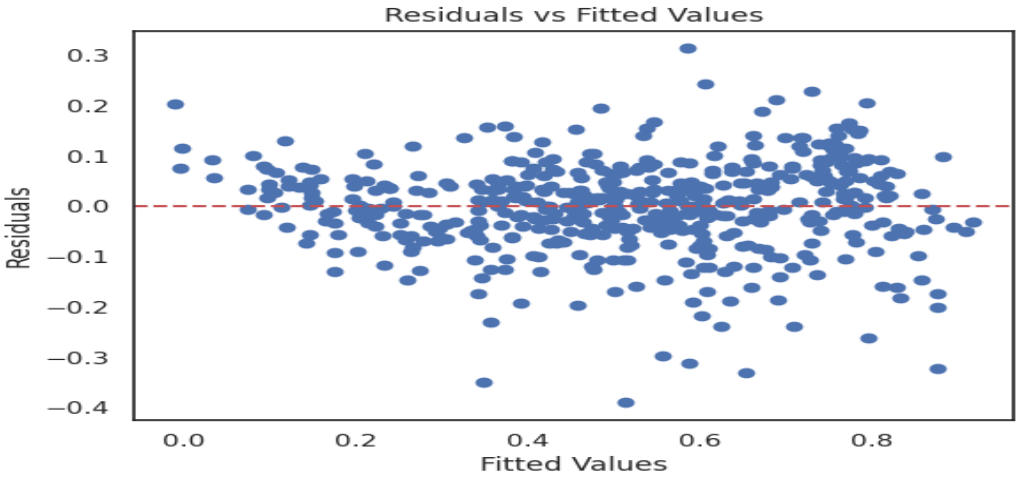


2. Checking normality of error terms using qq plot and distplot





3. Homoscedasticity:- Homoscedasticity means the residuals have constant variance at every level of x. The absence of this phenomenon is known as heteroscedasticity. Heteroscedasticity generally arises in the presence of outliers and extreme values.



4. No Collinearity :- The independent variables shouldn't be correlated. If multicollinearity exists between the independent variables, it is challenging to predict the outcome of the model. In essence, it is difficult to explain the relationship between the dependent and the independent variables. In other words, it is unclear which independent variables explain the dependent variable.

Feature	VIF
year	1.030918
is_holiday	1.148865
is_workingday	1.751623
temperature	1.604383
humidity	1.882878
windspeed	1.183010
season_Summer	1.330257
season_Winter	1.288511
month_July	1.433382
month_September	1.190843
weekday_Sunday	1.669023
weather_condition_Light Snow & Rain	1.243118
weather_condition_Mist & Cloudy	1.558397

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

Answer :-

Based on the table below the top 3 features contributing significantly towards explaining the demand of the shared bikes are

1) Temperature People like biking when it's warm. (Positive Trend)

2) Year :- People are picking up biking as a trend.

3) Weather Condition (Light Snow & Rain) :- People don't like biking when its rainy or snowing outside.

const	0.235
year	0.229
is_holiday	-0.114
is_workingday	-0.010
temperature	0.595
humidity	-0.173
windspeed	-0.188
season_Summer	0.082
season_Winter	0.137
month_July	-0.046
month_September	0.094
weekday_Sunday	-0.053
weather_condition_Light Snow & Rain	-0.240
weather_condition_Mist & Cloudy	-0.054

General Subjective Questions

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

Answer :-

Linear regression is a statistical method that is used to model the relationship between a dependent variable (target) and independent variables (predictors) by fitting a linear equation to the data points.

Steps of Linear Regression:

1. Define the Problem: Identify the dependent variable (the variable we want to predict or explain) and the independent variables (variables that may influence the dependent variable also known as predictors).

2. Collect Data: Gather the dataset containing observations of the dependent and independent variables. Ensure that the data is clean and does not contain any missing values or outliers that could skew the results.

3. Explore the Data: Perform exploratory data analysis (EDA) to understand the distribution, relationships, and summary statistics of the data. Visualize relationships between variables using scatter plots or correlation metrices.

4. Split Data: Split the dataset into training and testing sets. The training set is then used to train the model, while the testing set is used to evaluate the model's performance.

5. Choose a Model: Select the linear regression as the modeling technique, assuming there is a linear relationship between the independent and dependent variables.

6. Formulate the Model: The linear regression model is represented as: $y = \beta 0 + \beta 1x 1 + \beta 2x 2 + \cdots + \beta_n x_n + \epsilon$, where: ▪ y is the dependent variable (target). ▪ $x1, x2, \ldots, xn$ are the independent variables (predictors). ▪ $\beta 0$ is the intercept (where the line crosses the y-axis). ▪ $\beta 1, \beta 2, \ldots, \beta n$ are the coefficients (slopes) of the independent variables. ▪ ϵ is the error term (residuals), representing the difference between observed and predicted values.

7. Fit the Model: Use the training data to estimate the coefficients $\beta 0, \beta 1, \ldots, \beta n$ that minimize prediction errors using methods like Ordinary Least Squares (OLS).

8. Evaluate the Model: Assess the model's performance using evaluation metrics such as R-squared, on the testing dataset. R-squared measures how well the model explains the variance in the dependent variable.

9. Make Predictions: Use the fitted model to make predictions on testing data by applying the learned coefficients to the independent variables.

10. Validate and Iterate: Validate the model by comparing predicted values with actual values in the testing set. Iterate by refining the model, adjusting variables.

Example: Suppose a bike sharing company wants to predict daily bike rentals based on weather conditions and time-related factors. Here's how they would apply linear regression:

Step 1: Define the problem—predict daily bike rentals based on variables like temperature, humidity, windspeed, and time-related factors (hour, day of week).

Step 2: Collect data—gather historical data on daily bike rentals, weather conditions (temperature, humidity, windspeed), time indicators (hour, day of week), and other relevant factors.

Step 3: Explore data—plot variables like temperature against bike rental count to see if there's a relationship.

Step 4: Split data—divide the dataset into training (e.g., 80%) and testing (e.g., 20%) sets.

Step 5: Choose model—select linear regression because it's effective for predicting a continuous outcome based on multiple predictors.

Step 6: Formulate model— $total count = \beta 0 + \beta 1 \times temperature + \beta 2 \times humidity + \beta 3 \times windspeed + \cdots + \epsilon$

Step 7: Fit model—use training data to estimate coefficients β0, β1, β2, ... that minimize prediction errors.

Step 8: Evaluate model—use metrics like R-squared and RMSE to measure how well the model predicts bike rental counts on the testing set.

Step 9: Make predictions—apply the model to new data (upcoming days or months) to forecast bike rental demand.

Step 10: Validate and iterate—compare predicted rentals with actual counts to refine the model, possibly adding new variables or adjusting parameters for better predictions.

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

Answer :-

Anscombe's quartet is a set of four small datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, etc.) but differ considerably when graphed or analyzed further. It was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analyzing it and to illustrate the effect of outliers and influential observations on statistical properties.

Description of the Quartet

Dataset Characteristics:

Each dataset consists of 11 points (pairs of x and y values).

Despite having identical summary statistics, they exhibit markedly different distributions and relationships when plotted.

Summary Statistics:

Mean of x: Approximately 9 for all four datasets.

Mean of y: Approximately 7.50 for all four datasets.

Variance of x: Approximately 11 for all datasets.

Variance of y: Approximately 4.12 for all datasets.

Correlation between x and y: Approximately 0.816 for all datasets.

Linear regression line: $y = 3 + 0.5x$ (almost identical for all datasets).

Graphical Representation:

When graphed, each dataset shows a different pattern:

Dataset I: Linear relationship.

Dataset II: Non-linear, possibly exponential.

Dataset III: Clustered around two distinct groups.

Dataset IV: Influenced by an outlier.

Importance and Implications

Graphical Insight: Anscombe's quartet highlights the importance of visualizing data. Despite having identical summary statistics, the datasets look very different when plotted, emphasizing that statistical summaries alone can be misleading.

Outliers and Influential Points: Dataset IV, for instance, demonstrates how a single outlier can significantly affect correlation and regression parameters, making it crucial to identify and handle outliers appropriately.

Statistical Assumptions: It challenges assumptions underlying statistical methods that rely solely on summary statistics, reinforcing the need for robust exploratory data analysis (EDA).

Teaching and Understanding: Anscombe's quartet is frequently used in statistics education to illustrate concepts such as the effect of outliers, the importance of visual inspection, and the limitations of relying solely on summary statistics.

Conclusion

Anscombe's quartet remains a powerful illustration of the principle "graphical exploration of data is an essential step in statistical analysis." It underscores that while summary statistics provide a convenient snapshot of data, they may mask underlying complexities and patterns that can only be revealed through visual inspection and deeper analysis. Therefore, when analyzing data, it's critical to complement numerical summaries with graphical representations to gain a comprehensive understanding of the dataset's characteristics and relationships.

3. What is Pearson’s R? (3 marks)

Pearson's correlation coefficient, often denoted as r, is a measure of the linear relationship between two variables. It quantifies the strength and direction of the association between two continuous variables. Here are key aspects of Pearson's r:

Key Properties

1. Range: Pearson's r ranges from -1 to +1.
    - r=1: Perfect positive linear relationship (as X increases, Y increases proportionally).
    - r=−1: Perfect negative linear relationship (as X increases, Y decreases proportionally).
    - r=0: No linear relationship between X and Y.
2. Strength of Relationship:
    - The closer |r| is to 1, the stronger the linear relationship.
    - |r| near 0 suggests a weak linear relationship.
3. Assumption: Pearson's r assumes a linear relationship between X and Y.
4. Sensitive to Outliers: Pearson's r can be sensitive to outliers because it's based on means and deviations.

Interpretation

- Positive r: Indicates that as X increases, Y tends to increase.
- Negative r: Indicates that as X increases, Y tends to decrease.
- Zero r: Indicates no linear relationship between X and Y.

Use Cases

- Data Analysis: Used extensively in data analysis to understand relationships between variables.
- Research: Commonly used in scientific research to analyze experimental data and observational studies.
- Assumption Checking: Often checked in statistical modeling to ensure the independence assumption between errors (in regression models).

Limitations

- Linear Relationship Assumption: Pearson's r measures only linear relationships; it may not detect nonlinear associations.
- Sensitive to Outliers: Outliers can distort r and affect its interpretability.

- Only Measures Strength and Direction: Does not provide information about causation or the slope of the relationship.

In summary, Pearson's correlation coefficient r is a widely used measure in statistics to quantify the degree and direction of the linear relationship between two variables, providing valuable insights into data relationships when its assumptions are met.

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

In the context of data analysis and statistics, "scaling" refers to the process of transforming variables so that they fit within a specific scale, making them comparable or improving their interpretability. Scaling is often used to standardize variables or adjust their ranges to facilitate easier comparison, visualization, or analysis. Here are the key aspects of scaling:

Types of Scaling

1. Standardization (Z-score normalization):
    - Purpose: Standardization transforms the data to have a mean of 0 and a standard deviation of 1.
    - Effect: Ensures all variables are on the same scale, which is useful for algorithms that rely on distance measures (e.g., clustering, principal component analysis).
2. Min-Max Scaling:
    - Purpose: Rescales variables to a fixed range, typically between 0 and 1.
    - Effect: Preserves the original distribution shape and is sensitive to outliers. Useful for algorithms that require data on a bounded interval.
3. Normalization:
    - Purpose: Scales data to have a unit norm (typically unit length).
    - Effect: Useful for algorithms that weight input variables equally and do not make assumptions about the distribution of the input data.

Importance of Scaling

- Comparability: Enables variables measured in different units or scales to be directly compared.
- Algorithm Performance: Improves the performance of many machine learning algorithms by ensuring that all variables contribute equally to the analysis.
- Interpretability: Makes the data more interpretable and easier to visualize, especially when dealing with multiple variables.

Considerations

- Effect on Interpretation: Scaling does not change the shape of the distribution or the relationships between variables but modifies the numerical values.
- Outlier Sensitivity: Some scaling methods are sensitive to outliers, which may need to be handled separately.

- Contextual Application: The choice of scaling method depends on the specific context, the nature of the data, and the requirements of the analysis or algorithm.

In summary, scaling is a preprocessing step in data analysis that transforms variables to a comparable scale, facilitating better analysis, visualization, and interpretation of the data. It is an essential technique in statistics and machine learning to ensure that variables contribute meaningfully and uniformly to the analysis process.

Difference between Normalized Scaling and Standardized Scaling

- **Normalization** focuses on scaling the values of a variable to a specific range (typically [0, 1]), ensuring they have similar magnitudes. It doesn't change the distribution shape but emphasizes the relative magnitude of values.
- **Standardization** adjusts the mean and standard deviation of a variable to a common scale (mean 0, standard deviation 1). It modifies the distribution shape to be centered around zero and is more suitable when the shape of the distribution and the spread of values are important.
- **Contextual Use**: The choice between normalization and standardization depends on the specific requirements of the data analysis or machine learning algorithm. Normalization is suitable when the magnitude of variables matters, while standardization is preferred when the distribution and spread of variables need adjustment.

In summary, while both normalized scaling and standardized scaling are methods for making variables comparable, they achieve this through different transformations that suit different analytical contexts and algorithmic requirements.

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

(3 marks)

Answer :-

VIF is opposite of R2 when the two variables are perfectly multicollinear VIF comes out as infinite.

The occurrence of an infinite value for Variance Inflation Factor (VIF) typically happens when there is perfect multicollinearity among the predictor variables in a regression model. Perfect multicollinearity means that one or more of the independent variables can be exactly predicted from the others, leading to a situation where the variance of the regression coefficients cannot be calculated or is undefined.

Reasons for Infinite VIF:

1. Perfect Linear Relationship: If one predictor variable can be expressed as a perfect linear combination of other predictor variables in the model, then the correlation matrix used to compute VIF will have one or more eigenvalues that are zero. This results in the inverse of the

correlation matrix not existing, and thus, VIF for that variable (or variables involved in the perfect multicollinearity) becomes infinite.

2. Mathematical Calculation: The formula for VIF involves computing the inverse of the correlation matrix of the predictors. If this matrix is not invertible due to perfect multicollinearity, the VIF calculation breaks down, resulting in an infinite value.
3. Implications: Infinite VIF indicates that the variance of the regression coefficient for the affected variable cannot be estimated due to the perfect multicollinearity. This situation can lead to unreliable regression coefficients and standard errors, making the interpretation of the model problematic.

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

(3 marks)

Answer :-

A Q-Q plot, short for Quantile-Quantile plot, is a graphical tool used to assess whether a dataset follows a particular distribution (often compared to a theoretical distribution such as normal distribution). It compares the quantiles of the dataset against the quantiles of a specified theoretical distribution. Here's a detailed explanation of its use and importance in the context of linear regression:

How Q-Q Plot Works:

1. Construction:
    - The Q-Q plot is constructed by plotting the quantiles of the dataset against the quantiles of a specified theoretical distribution (usually a normal distribution).
    - If the dataset perfectly follows the theoretical distribution, the points in the Q-Q plot will lie approximately on a straight line.
2. Interpretation:
    - Normal Distribution Check: For linear regression, it's often crucial to check if the residuals (errors) of the regression model are normally distributed. The Q-Q plot of residuals against the theoretical quantiles of a normal distribution helps in visually assessing this.
    - Pattern Analysis: Deviations from a straight line in the Q-Q plot indicate departures from normality. Specifically:
        - S-Shaped Curve: Indicates heavier tails or skewness compared to the normal distribution.
        - Non-Straight Linearity: Indicates significant departures from normality, which might affect the reliability of statistical tests and assumptions in linear regression.

Importance in Linear Regression:

1. Assumption Checking: Linear regression assumes that the residuals (errors) follow a normal distribution. Violations of this assumption can lead to biased parameter estimates, incorrect standard errors, and unreliable hypothesis tests.
2. Model Validity: A Q-Q plot provides a clear visual indication of whether the residuals conform to a normal distribution. If the residuals are not normally distributed, it suggests that the model may not adequately capture the variability in the data, or there may be underlying issues such as omitted variables, nonlinear relationships, or heteroscedasticity.
3. Decision Making: Based on the Q-Q plot:
    - If the residuals follow a straight line reasonably well, it supports the validity of using linear regression and interpreting its coefficients.
    - If the residuals deviate significantly from a straight line, corrective actions such as transforming variables, considering different modeling techniques, or addressing outliers and influential points may be necessary.

4. Statistical Inference: Normality of residuals is crucial for valid inference in linear regression, including confidence intervals, hypothesis testing (t-tests for coefficients), and prediction intervals. Q-Q plots help ensure these inferences are based on reliable assumptions.

Practical Use:

- Step in Model Diagnostics: Q-Q plots are typically included in the diagnostic toolkit alongside residual plots, leverage plots, and other diagnostic plots to assess the assumptions and performance of the regression model.
- Visualization Aid: They provide a straightforward way to communicate the distributional characteristics of residuals to stakeholders and decision-makers.

Conclusion:

In linear regression analysis, a Q-Q plot serves as a valuable diagnostic tool for assessing the normality assumption of residuals. By visually comparing observed quantiles against expected quantiles from a normal distribution, it helps analysts and researchers make informed decisions about the adequacy of their regression models and the reliability of their statistical inferences. Therefore, understanding and interpreting Q-Q plots are essential skills in ensuring the robustness and validity of linear regression analyses.