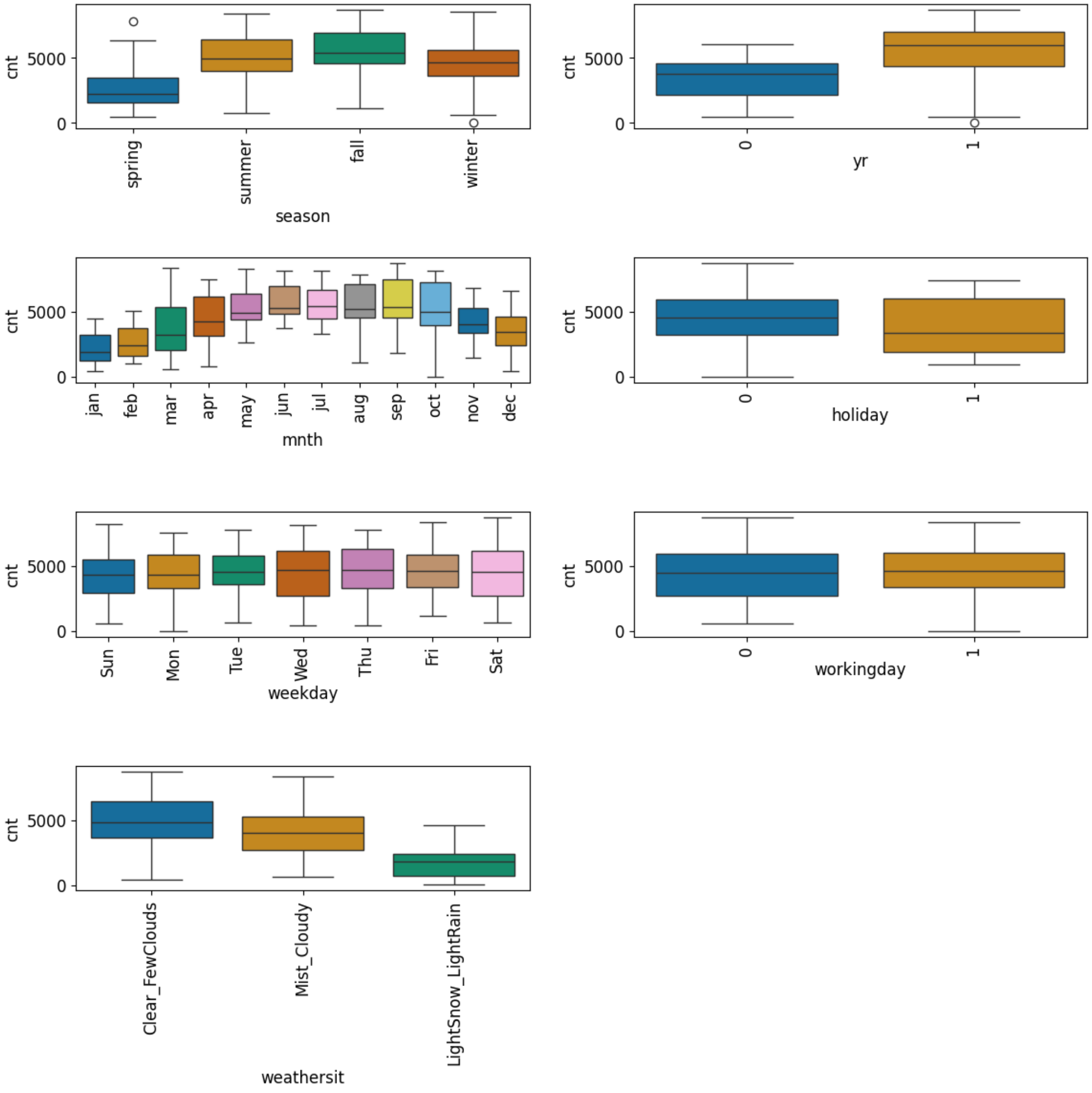
Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)



Insights -

1. The highest number of bike-sharing activities occur during the fall season, while the lowest occurs in spring.
2. The bike-sharing count is higher in 2019 compared to 2018.
3. Bike-sharing activity increases during the mid-year months and decreases towards the end of the year.
4. The bike-sharing count is highest when the weather is clear or partly cloudy and lowest during light snow or light rain conditions.

5. There is no significant difference in bike-sharing counts between working days and non-working days.
6. Bike-sharing counts remain consistent across all weekdays.
7. On holidays, there is a noticeable decline in the bike-sharing count.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)

Let's first understand dummy variable creation process, what it is-

1. Dummy encoding creates separate binary columns for each category in a categorical variable.
2. If there are n categories, it creates n columns.

Now understand what happen if we use all generated columns –

1. Including all n columns causes multicollinearity because the sum of all dummy columns equals 1 for any row (linear dependency).
2. Multicollinearity inflates standard errors of regression coefficients, making them unreliable.

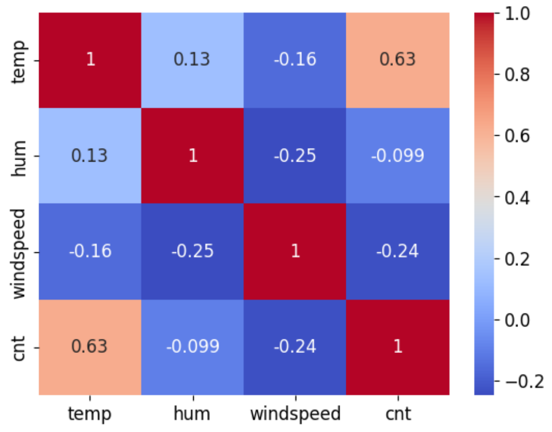
Work of drop_first = True

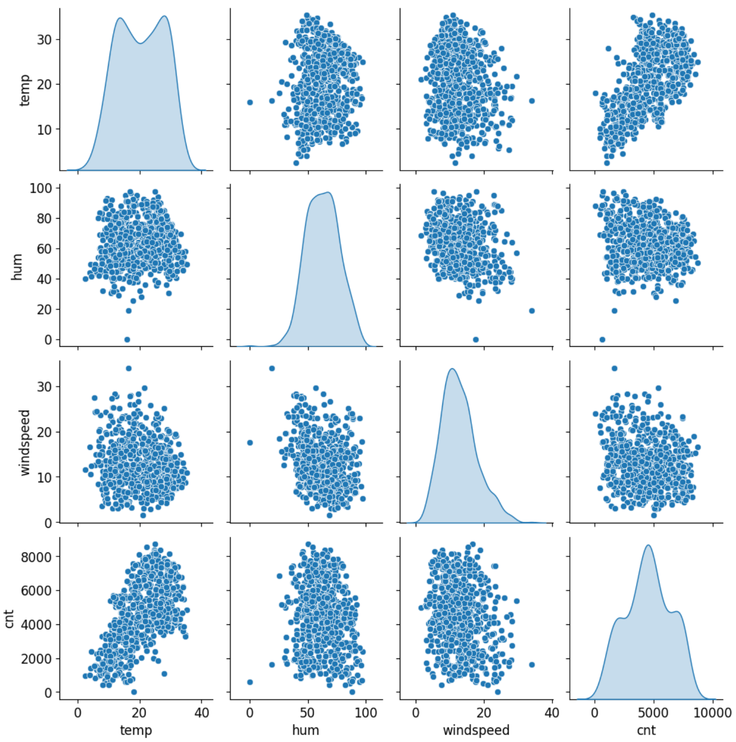
1. When you use drop_first=True, one category (usually the first) is dropped and treated as the reference category.
2. The dropped category is the baseline. The coefficients of the remaining categories are interpreted relative to the reference category.

Example: For categories A, B, C:
Without drop_first: A, B, C (3 columns, redundancy exists).
With drop_first: B, C (2 columns, A is reference).

Conclusion is that if we can get the required output from n-1 columns then we should drop it.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)



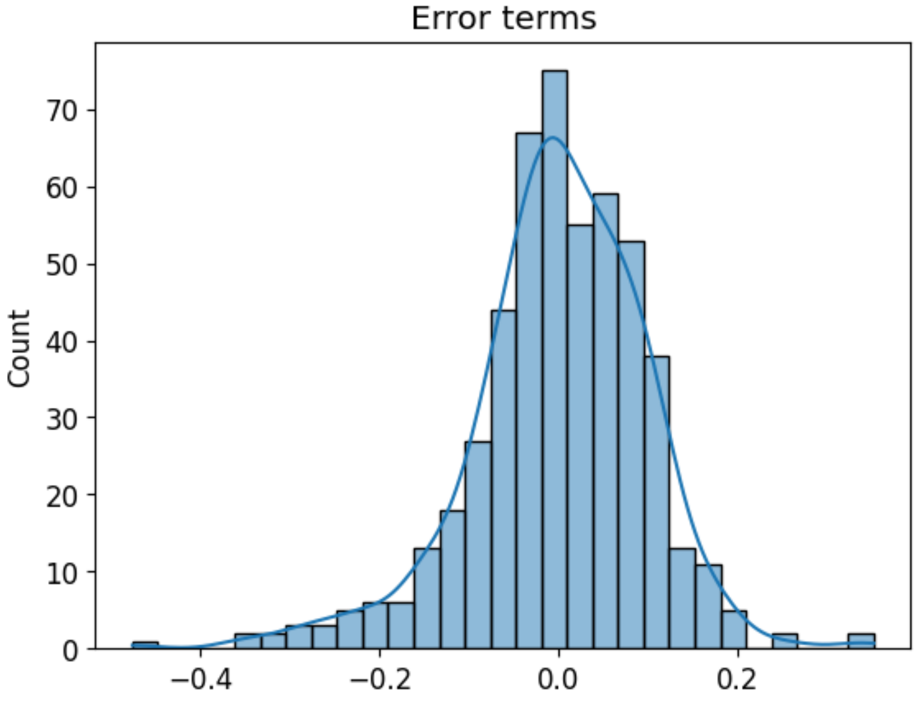


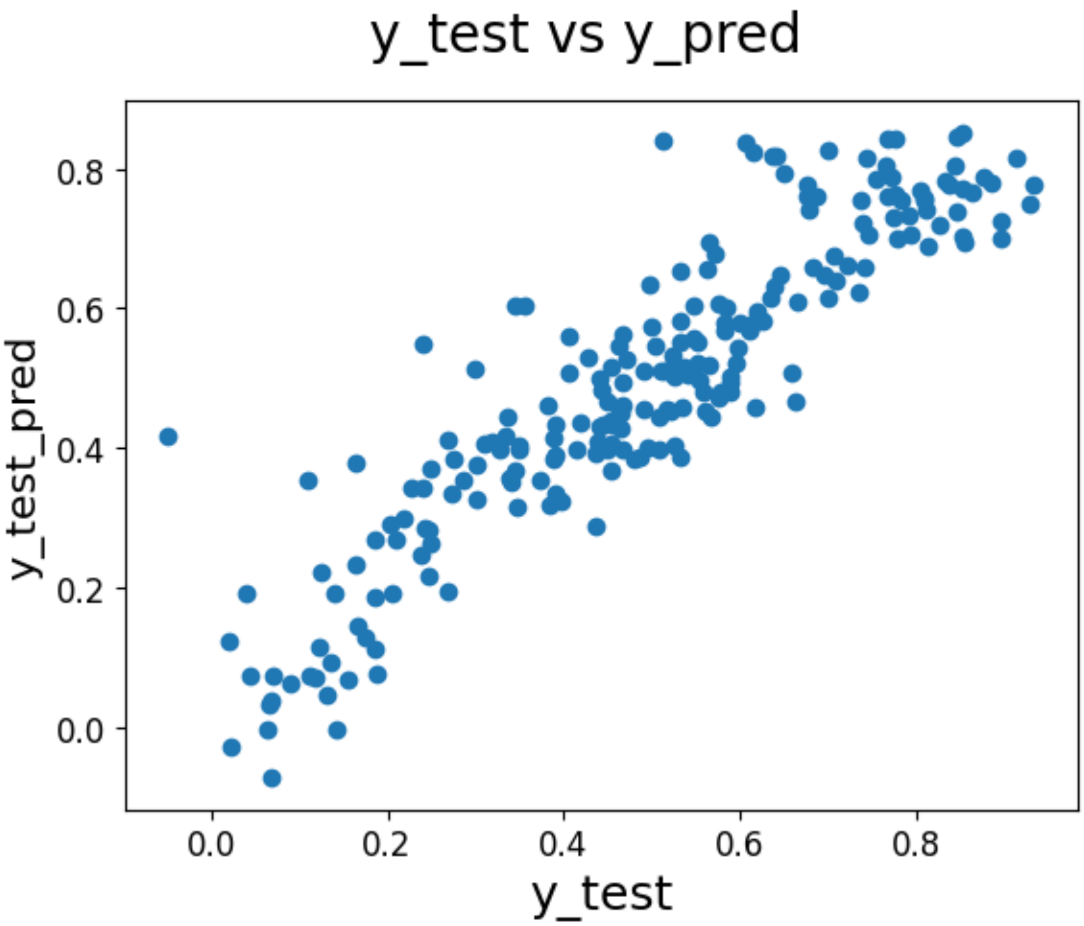
From above Heatmap and pair plot we can see highest correlation between ‘Temp’ and ‘cnt’.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)





Most of the data points align closely with the straight line, indicating that the error terms follow a normal distribution.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

Based on final model top three features contributing significantly towards explaining the demand are:

* Temperature: (0.583)
* Windspeed: (-0.150)
* Year: (0.251)

Hence, it can be clearly concluded that the variables temperature, windspeed and month are significant in predicting the demand for shared bikes.

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Linear Regression is a fundamental algorithm in machine learning and statistics used to model the relationship between a dependent variable (target/output) and one or more independent variables (features/input). Below is a detailed explanation:

1. Objective of Linear Regression

Linear Regression aims to find the best-fitting line (or hyperplane in multiple dimensions) that minimizes the difference between the predicted and actual values of the dependent variable.

The goal is to model the relationship as:
$y=\beta0+\beta1x1+\beta2x2+\cdots+\beta nxn+\epsilon y$

Where:

- y: Dependent variable (target).
- $x1,x2,\ldots xn$: Independent variables (features).
- $\beta0$: Intercept (constant term).
- $\beta1,\beta2,\ldots,\beta n$: Coefficients (weights for features).
- ϵ: Error term (difference between actual and predicted values).

2. Types of Linear Regression

1. **Simple Linear Regression**:
   - Models the relationship between one independent variable (x) and the dependent variable (y).
   - Equation:
     $y =\beta0+\beta1x$
2. **Multiple Linear Regression**:
   - Models the relationship between two or more independent variables ($x1,x2,\ldots,xn$) and the dependent variable (y).
   - Equation:
     $y=\beta0+\beta1x1+\beta2x2+\cdots+\beta nxn$

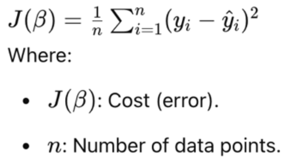
3. Key Assumptions of Linear Regression

For accurate modelling, Linear Regression relies on these assumptions:

1. **Linearity**: The relationship between independent variables and the target variable is linear.
2. **Independence of Errors**: Error terms (ϵ) are independent.
3. **Homoscedasticity**: The variance of the error terms remains constant across all levels of the independent variables.
4. **Normality of Errors**: The error terms follow a normal distribution.
5. **No Multicollinearity**: Independent variables are not highly correlated with each other.

4. Working of Linear Regression

1. **Hypothesis Function**:
   The model predicts y^ (predicted value) using the equation:
   y^= β0+β1x1+β2x2+⋯+βnxn
2. **Cost Function**:
   The cost function quantifies the error between predicted (y^) and actual (y) values. For Linear Regression, the most common cost function is the **Mean Squared Error (MSE)**:



$J(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Where:

- $J(\beta)$: Cost (error).
- n: Number of data points.

3. **Optimization**:
   Linear Regression minimizes the cost function J(β) to find the best coefficients (β). This is typically done using:
   - **Gradient Descent**: Iteratively updates coefficients to minimize the cost function.
   - **Normal Equation**: Directly calculates optimal coefficients without iteration:

$$\beta = (X^T X)^{-1} X^T y$$

5. Performance Metrics

To evaluate the model, the following metrics are commonly used:

1. **R-squared (R2)**: Measures how much variance in the dependent variable is explained by the independent variables.

$$R^2 = 1 - \frac{\text{Sum of Squared Errors (SSE)}}{\text{Total Sum of Squares (TSS)}}$$

   Values range from 0 to 1; higher is better.

2. **Adjusted R-squared**: Adjusts R2 for the number of predictors in the model, preventing overfitting.
3. **Mean Absolute Error (MAE)**: Average of absolute errors.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

4. **Root Mean Squared Error (RMSE)**: Square root of MSE, penalizing large errors more than MAE.

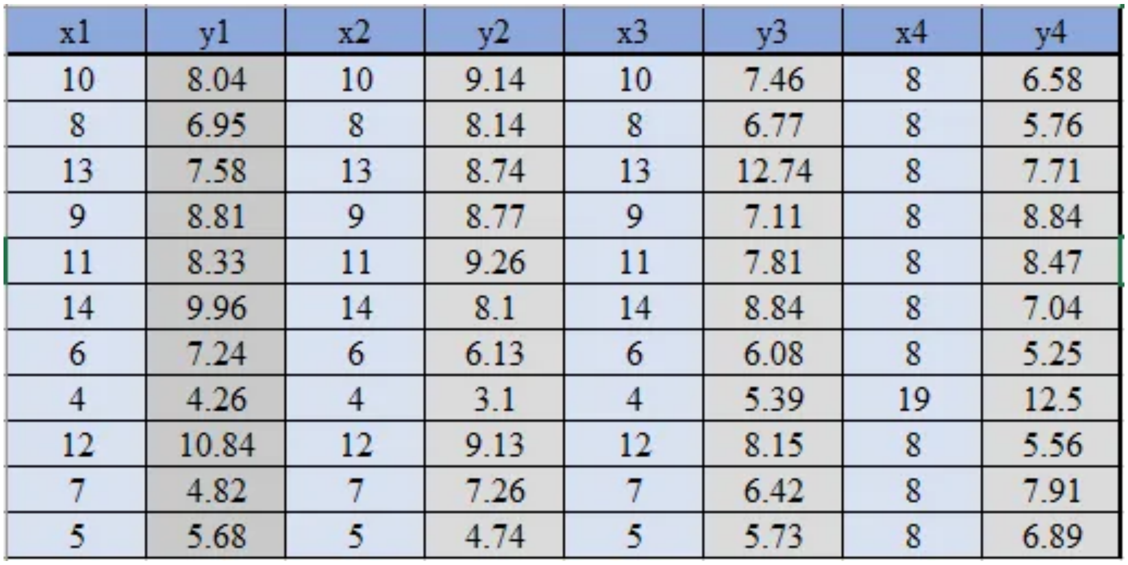
Limitations of Linear Regression

1. Assumes a linear relationship, which may not hold for complex data.
2. Sensitive to outliers, as they can significantly affect the regression line.

3. Struggles with multicollinearity (correlated features) and heteroscedasticity (non-constant error variance).
4. Requires feature scaling when coefficients vary in magnitude.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Anscombe's Quartet is the modal example to demonstrate the importance of data visualization which was developed by the statistician Francis Anscombe in 1973 to signify both the importance of plotting data before analysing it with statistical properties. It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyse about these data-sets is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behaviour irrespective of statistical analysis.



x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Apply the statistical formula on the above data-set,

Average Value of x = 9

Average Value of y = 7.50

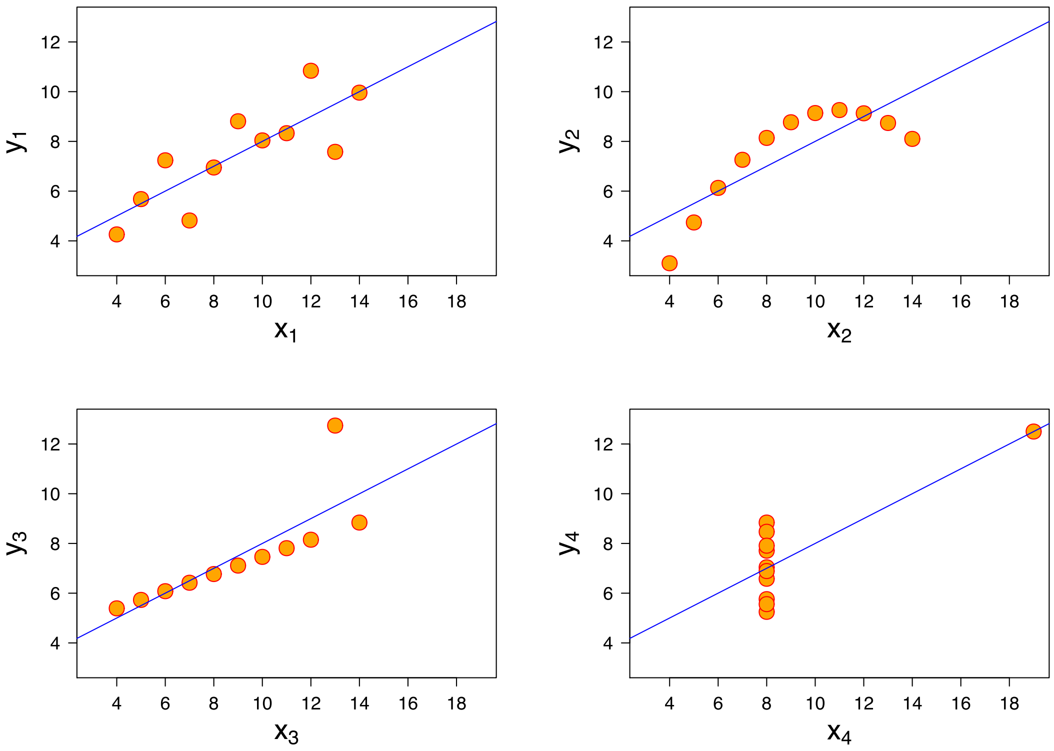
Variance of x = 11

Variance of y =4.12

Correlation Coefficient = 0.816

Linear Regression Equation : y = 0.5 x + 3

However, the statistical analysis of these four data-sets are pretty much similar. But when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represent the different behaviour.



1. Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.
2. Data-set II — shows a curve shape but doesn't show a linear relationship (might be quadratic?).
3. Data-set III — looks like a tight linear relationship between x and y, except for one large outlier.
4. Data-set IV — looks like the value of x remains constant, except for one outlier as well.

Data-sets which are identical over a number of statistical properties, yet produce dissimilar graphs, are frequently used to illustrate the importance of graphical representations when exploring data. This isn't to say that summary statistics are useless. They're just misleading on their own. It's important to use these as just one tool in a larger data analysis process. Visualizing our data allows us to revisit our summary statistics and re-contextualize them as needed.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson's correlation coefficient, often denoted as r, is a measure of the linear correlation between two variables. It quantifies the strength and direction of the linear relationship between paired continuous data.

Pearson's correlation coefficient r is defined as the covariance of two variables X and Y divided by the product of their standard deviations. Mathematically, it is expressed as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where:

- x and y are the individual data points of variables X and Y,
- x¯ and y¯ are the mean values of X and Y respectively,
- n is the number of data points.

Properties:

1. Range of r: Pearson's r ranges from -1 to 1.

- r = 1: Perfect positive linear correlation (as X increases, Y also increases proportionally).
- r = −1: Perfect negative linear correlation (as X increases, Y decreases proportionally).
- r = 0: No linear correlation between X and Y.

2. Strength of Correlation:

- The closer r is to 1 or -1, the stronger the linear relationship between X and Y.
- The closer r is to 0, the weaker the linear relationship (or no linear relationship).

3. Direction:
   - If r is positive, X and Y have a positive linear relationship (both variables increase or decrease together).
   - If r is negative, X and Y have a negative linear relationship (as one variable increases, the other decreases).

4. Assumption: Pearson's r assumes that the relationship between the variables is linear and that both variables are normally distributed.
5. Sensitive to Outliers: Pearson's r can be sensitive to outliers, especially when they significantly influence the covariance between X and Y.

Limitations:

- Pearson's r measures only linear relationships. It may not capture nonlinear relationships.
- It assumes that both variables are normally distributed and can be influenced by outliers.
- It does not imply causation; a strong correlation does not necessarily mean that changes in one variable cause changes in the other.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

Scaling is a preprocessing step in data analysis and machine learning where numerical features are transformed to a standard scale, typically to make the data comparable and improve the performance and stability of certain algorithms. Here's a detailed explanation addressing your questions:

What is Scaling?

Scaling refers to the process of transforming the range of variables (features) to a similar scale. It is essential because many machine learning algorithms perform better or converge faster when features are on a relatively similar scale and close to normally distributed.

Why is Scaling Performed?

1. Algorithm Performance: Many machine learning algorithms, such as linear and logistic regression, K-nearest neighbors, support vector machines (SVMs), and neural networks, are sensitive to the scale of input features. Features with larger scales can dominate those with smaller scales, affecting the algorithm's ability to learn effectively.
2. Convergence: Gradient descent-based optimization algorithms, used in many machine learning models, converge faster when features are scaled. This is because the steps taken in each iteration are more consistent when the scales are similar.
3. Distance-based Algorithms: Algorithms that rely on distances between data points, such as K-nearest neighbors and clustering algorithms (e.g., K-means), are sensitive to the scale of features. Features with larger scales will have a greater impact on the distance calculations.

Types of Scaling:

1. **Normalized Scaling (Min-Max Scaling):**

- Formula: Xnorm = (X−Xmin) / (Xmax−Xmin)
- Range: Transforms data to a scale between 0 and 1.
- Purpose: Ensures all features have the same scale, preserving the shape of the original distribution. Useful when the data needs to be bound within a specific range.

2. **Standardized Scaling (Z-score Normalization):**

- Formula: Xstd = (X−μ)/σ
- Properties: Transforms data to have a mean of 0 and a standard deviation of 1.
- Purpose: Centres the data around zero and adjusts the variance, making it suitable for algorithms that assume normally distributed data, such as linear regression and linear discriminant analysis.

Differences between Normalized Scaling and Standardized Scaling:

- **Range of Values**:

I. Normalized Scaling: Values are scaled to a fixed range (e.g., 0 to 1).

II. Standardized Scaling: Values are cantered around 0 with a standard deviation of 1.

- **Impact on Distribution**:

I. Normalized Scaling: Preserves the shape of the original distribution, but may not handle outliers well.

II. Standardized Scaling: Does not bound values to a specific range, but is less sensitive to outliers and makes the data more suitable for algorithms that assume normally distributed features.

Algorithm Suitability:

I. Normalized Scaling: Often used when the algorithm (e.g., neural networks) requires inputs to be within a specific range.

II. Standardized Scaling: Generally preferred for algorithms that assume normally distributed data, or when interpreting coefficients in linear models is important.

Considerations:

1. Choice of Scaling: The choice between normalized and standardized scaling depends on the specific requirements of the machine learning algorithm and the characteristics of the data.
2. Impact on Interpretation: Standardized scaling is often preferred when the interpretation of coefficients or feature importance is important, as the scaling does not change the relationship between variables, only their scales.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

Variance Inflation Factor (VIF) is a measure used to quantify how much the variance of a regression coefficient is inflated due to collinearity with other predictor variables in a linear regression model. It assesses how much the variance of the estimated regression coefficients increases if your predictors are correlated.

Calculation of VIF:

For each predictor variable X_j, the VIF is calculated as:

$$\mathrm{VIF}_j = \frac{1}{1-R_j^2}$$

Where:

- R_j^2 is the R^2 value obtained by regressing X_j against all other predictor variables X_{-j}.

Occurrence of Infinite VIF:

1. **Perfect Multicollinearity**: If a predictor variable X_j is perfectly linearly dependent on other predictor variables (i.e., X_j can be expressed as a perfect linear combination of other predictors), then R_j^2 will be 1.
2. **Calculation of VIF**: When R_j^2=1, the denominator 1−R_j^2 becomes 0. This leads to:

   VIF_j = 1/0 = ∞

   Hence, the VIF for the predictor variable X_j becomes infinite.

Practical Implications:

- **Model Fitting**: Infinite VIF values indicate that one or more predictor variables are perfectly collinear with others. In such cases, the regression coefficients cannot be estimated uniquely because the model matrix is rank-deficient.
- **Interpretation**: Infinite VIFs make it impossible to interpret the regression coefficients properly because the model is overparameterized due to perfect multicollinearity.

Handling Infinite VIF:

To address infinite VIF values, consider the following steps:

1. **Identify and Remove Variables**: Identify variables that are causing perfect multicollinearity and consider removing one of them from the model.
2. **Combine Variables**: Sometimes, variables may be transformed or combined to remove multicollinearity. For example, instead of using both height in centimeters and height in inches, use only one.
3. **Regularization Techniques**: Techniques like Ridge Regression or Lasso Regression can help in reducing multicollinearity by penalizing large coefficients.
4. **Principal Component Analysis (PCA)**: PCA can be used to reduce dimensionality and address multicollinearity by creating new orthogonal variables (principal components) that are uncorrelated.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
(Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

A Q-Q plot (Quantile-Quantile plot) is a graphical technique used to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. It compares the quantiles of the dataset against the quantiles of a theoretical distribution.

Understanding Q-Q Plot:

1. **Purpose**:
   - **Distribution Assessment**: Q-Q plots are used to visually inspect if a dataset follows a specific distribution (e.g., normal distribution).

   - **Residual Analysis**: In linear regression, Q-Q plots of residuals (the differences between observed and predicted values) are used to check if the residuals are normally distributed.
2. **Construction**:
   - **The Q-Q plot is constructed by**:
     - Sorting the data points in ascending order.
     - Computing the quantiles for both the dataset and the theoretical distribution (e.g., normal distribution).
     - Plotting the quantiles of the dataset against the quantiles of the theoretical distribution.
3. **Interpretation**:
   - If the dataset follows the theoretical distribution (e.g., normal distribution), the points in the Q-Q plot should roughly follow a straight line.
   - Deviations from the straight line indicate departures from the assumed distribution.

Importance of Q-Q Plot in Linear Regression:

1. **Assumption Checking**:
   - **Normality of Residuals**: In linear regression, one of the key assumptions is that the residuals (errors) are normally distributed with mean 0 and constant variance (homoscedasticity). Q-Q plots of residuals help assess this assumption.
   - **Diagnostic Tool**: Q-Q plots provide a visual check to see if residuals exhibit patterns or deviations from normality, which can indicate potential issues with the regression model.
2. **Model Validity**:
   - A linear regression model assumes that residuals are normally distributed. Deviations from normality can affect the validity of statistical inference, such as hypothesis testing and confidence intervals.
   - Q-Q plots help identify situations where non-normality of residuals might affect the reliability of model predictions and interpretations.
3. **Decision Making**:
   - **Based on the Q-Q plot**:
     - If residuals closely follow a straight line, it suggests that the assumption of normality is reasonable, and the linear regression model is appropriate for inference.
     - If residuals deviate significantly from the straight line, it may indicate that the linear regression assumptions are violated, and further investigation or model refinement is necessary.

Practical Use:

- **Model Refinement**: If Q-Q plots reveal non-normality in residuals, transformations (e.g., logarithmic transformation) or alternative modeling approaches (e.g., generalized linear models) may be considered to address the issue.
- **Diagnostic Tool**: Q-Q plots complement other diagnostic tools in regression analysis, such as residual plots and tests for homoscedasticity, providing a comprehensive assessment of the model's assumptions.