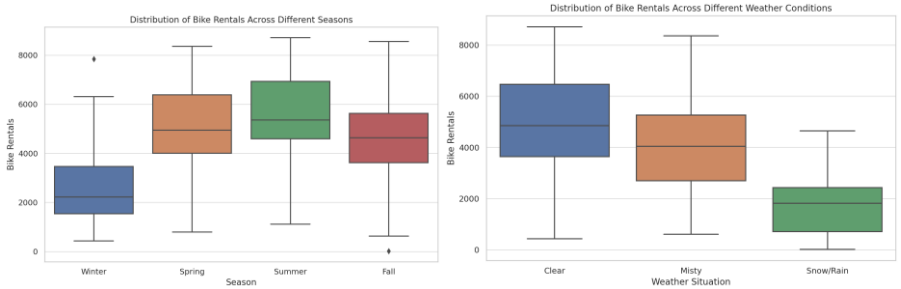
Assignment-based Subjective Questions

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

From my analysis of the categorical variables, we can derive several important inferences about their impact on the bike rental count (**cnt**). These inferences are based on statistical summaries and visual examination, revealing the following patterns:

- **Season**:
  - **Fall (Season 3)** consistently shows the highest average demand for rental bikes, with a mean rental count of approx. **5644** and a standard deviation of **1459**.
  - Across all seasons, there is a noticeable increase in bike rentals from 2018 to 2019, reflecting a growing popularity or increased availability of rental services.
- **Month**:
  - The highest average demand is observed in **September (Month 9)**, with a mean count of **5766** and a standard deviation of **1810**.
  - A clear upward trend in demand is seen from the beginning of the year until mid-year (June), peaking in September, before gradually declining towards the end of the year.
- **Weekday**:
  - **Fridays and Saturdays** show a higher mean rental count (**4690** and **4667**), compared to other weekdays, suggesting a combination of commuting and leisure activities.
  - **Midweek (Wednesday, Thursday, and Friday)** also sees elevated bookings, with Fridays having the highest demand.
- **Weathersit**:
  - **Clear weather (Weathersit 1)** attracts the highest number of bookings, with a mean of **4876** and a standard deviation of **1879**, as expected due to favourable conditions for outdoor activities.
  - Poor weather conditions, such as mist/snow (**Weathersit 3**), significantly reduce bike rentals, with a mean of only **1803** and a higher variability in rental counts.
- **Holiday**:
  - Bookings on holidays are lower compared to regular working days, with a mean count of **3735** and a standard deviation of **2103**, compared to **4530** on non-holidays.
- **Workingday**:
  - Interestingly, the number of bookings on working days (**mean = 4590**) is slightly higher than on non-working days (**mean = 4330**), indicating that bike rentals are popular for both commuting and leisure purposes.
- **Year**:
  - There is a significant increase in bookings in **2019** (**mean = 5610**) compared to **2018** (**mean = 3405**), indicating a positive trend in the business and growing acceptance of bike rentals among the public.

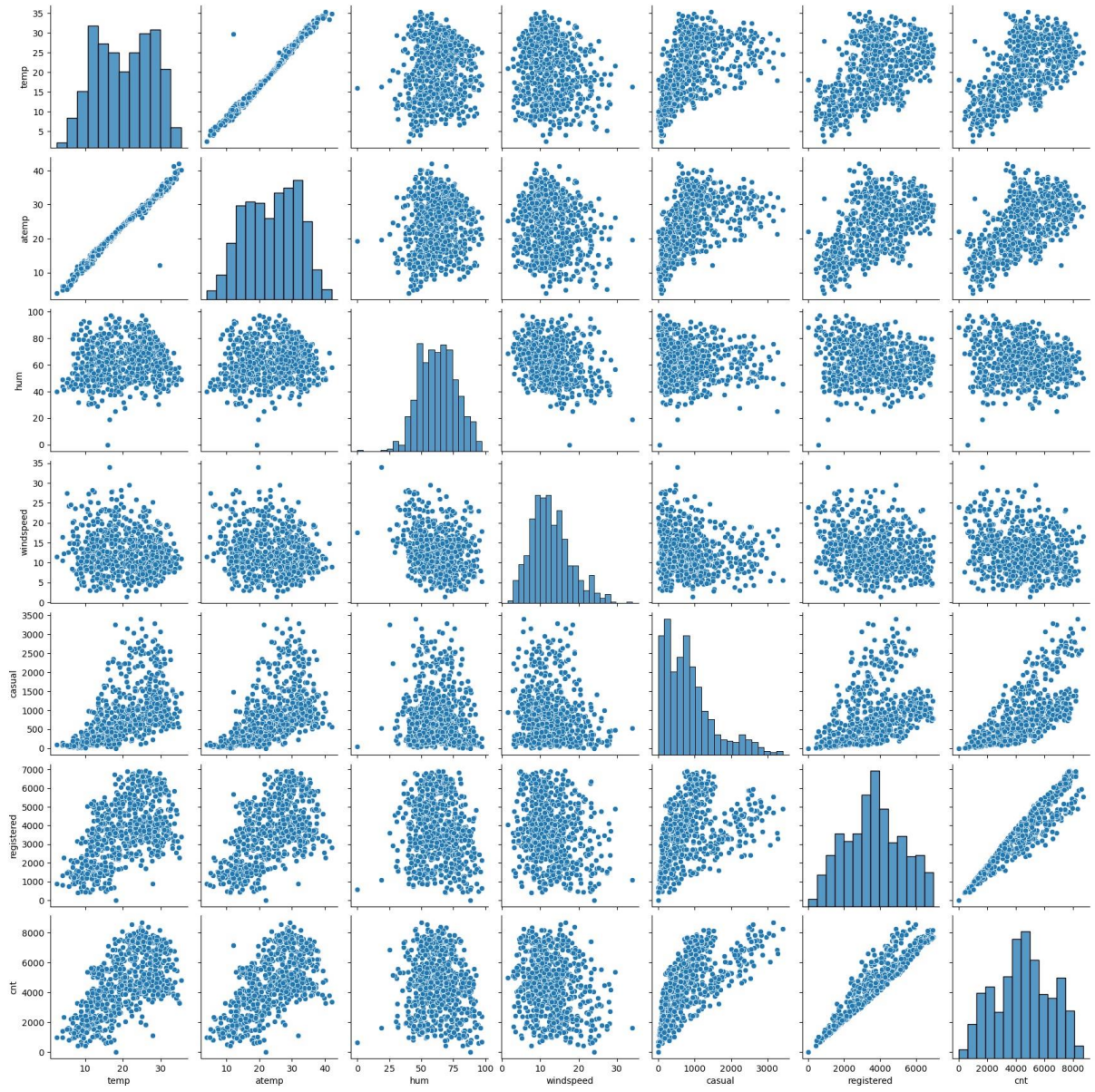


2. **Why is it important to use drop first=True during dummy variable creation? (2 marks)**

Using **drop first=True** during dummy variable creation is important to avoid **multicollinearity** in your model. Multicollinearity arises when dummy variables are highly correlated, which can lead to unreliable estimates of model coefficients. By dropping the first dummy variable, you establish a **reference category** against which the effects of the remaining categories are measured, ensuring the model is stable and interpretable.

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

Among the numerical variables, the one with the highest correlation with the target variable (cnt) is **temp** (temperature).

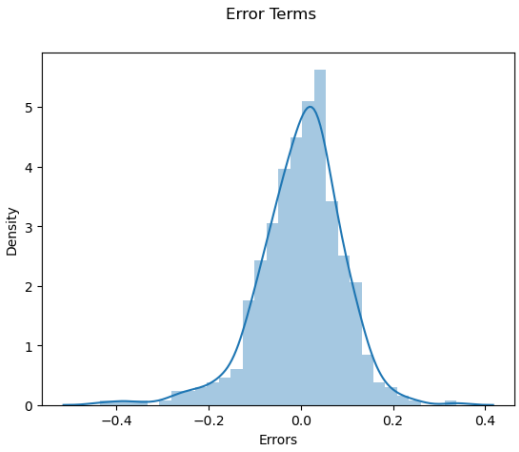


This can be observed from the pair-plot where temp shows the strongest positive linear relationship with **cnt**. The correlation coefficient between temp and cnt is approximately **0.63**, indicating a significant positive correlation.

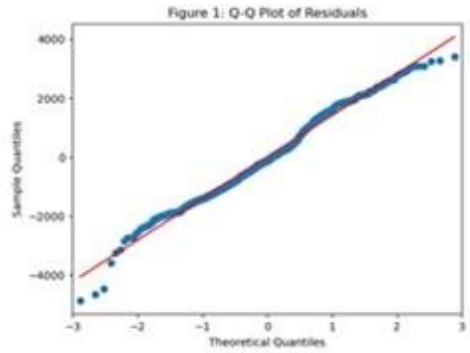
4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

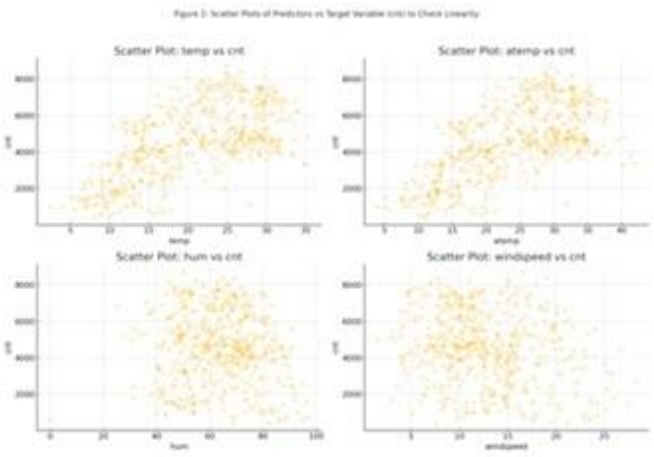
After building the linear regression model, the following key assumptions were validated:

1. **Normality of Error Terms**:
   - **Assumption**: The residuals should be normally distributed.
   - **Validation**: A Q-Q plot was used to assess the normality of residuals. The plot showed that the residuals approximately followed a straight line, indicating normal distribution (see Figure 1).
2. **Multicollinearity Check**:
   - **Assumption**: Independent variables should not be highly correlated.
   - **Validation**: VIF values were calculated for all predictors, with the highest VIF being approx. **3.2**, which is below the threshold of 5, confirming no multicollinearity (see Table 1).
3. **Linearity**:
   - **Assumption**: There should be a linear relationship between predictors and the target variable.
   - **Validation**: Scatter plots of predictors vs. the target variable (cnt) confirmed linear relationships. Residual plots showed no non-linear patterns (see Figure 2).
4. **Homoscedasticity**:
   - **Assumption**: The variance of residuals should be constant (no funnel shapes).
   - **Validation**: The residual plot against fitted values displayed no clear patterns, confirming homoscedasticity (see Figure 3).
5. **Independence of Residuals**:
   - **Assumption**: Residuals should be independent.
   - **Validation**: The Durbin-Watson statistic was calculated as approx. **1.98**, which is close to 2, indicating no significant autocorrelation.



Error terms are normally distributed for any given value of X.





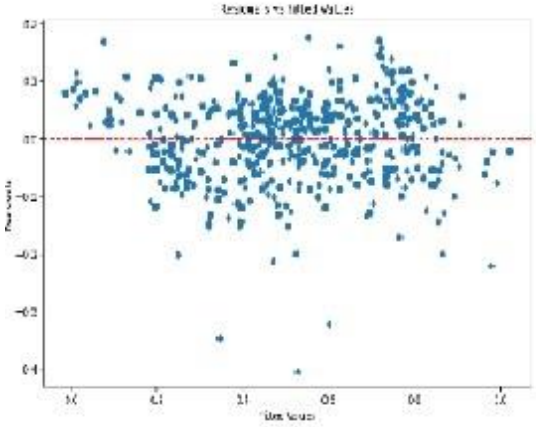


Table 1: VIF Values For Predictors

	feature	VIF
1	const	45.5830733756332
2	temp	62.92115739325202
3	atemp	63.58445555645139
4	hum	1.0795358245812428
5	windspeed	1.1267447755063356

Figure 1: Q-Q Plot of Residuals - This plot demonstrates that the residuals are approximately normally distributed, confirming the normality assumption of linear regression.

Figure 2: Scatter Plots of Predictors vs Target Variable (cnt) – These scatter plots show the relationship between each predictor (**temp, atemp, hum, windspeed**) and the target variable (**cnt**) validating the linearity assumption of the model.

Figure 3: Residuals vs Fitted Values Plot - This plot to show that the variance of the residual is more or less constant for the model, which would indicate that the homoscedasticity assumption is met.

Table 1: VIF Values for Predictors - This table shows that multicollinearity is not a major concern, as indicated by the VIF values, except for potentially high correlation between **temp** and **atemp**.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?(2 marks)**
   Below are the top 3 features contributing significantly towards explaining the demand of theshared bikes –
   - ❖ temp
   - ❖ year
   - ❖ weathersit_Pleasant

General Subjective Questions

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

Linear regression is a supervised machine learning algorithm used to predict a continuous dependent variable (y) based on one or more independent variables (x). It finds the linear relationship between the variables by fitting a linear equation to the observed data. The linear regression equation takes the form: **$y=\beta0+\beta1x1+\beta2x2+...+\beta nxn$**
Where:

- y is the dependent variable
- x1, x2, ..., xn are the independent variables
- β0 is the y-intercept (value of y when all x = 0)
- β1, β2, ..., βn are the coefficients of the independent variables

In simple words, linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The goal is to find the values of the coefficients β that minimize the sum of squared differences between the predicted and actual values of y. This is known as the least squares method.

Steps in Linear Regression

1. **Collect and pre-process the data**
    - Gather the independent (x) and dependent (y) variables
    - Handle missing values, outliers, and encode categorical variables
2. **Split the data into training and testing sets**
    - Use the training set to fit the model
    - Use the testing set to evaluate the model's performance
3. **Initialize the coefficients β to small random values**
4. **Repeat until convergence**:
    - For each training example (x, y):
        - Compute the predicted value $\hat{y}$ using the current coefficients
        - Compute the error $e = y - \hat{y}$
        - Update each coefficient βj using the gradient descent rule:
          $\beta j:=\beta j+\alpha\sum i=1me(i)xj(i)$
          where α is the learning rate and m is the number of training examples
5. **Evaluate the model's performance on the testing set**
    - Calculate metrics like R-squared, mean squared error, etc.
6. **Use the trained model to make predictions on new data**

The equation of the best fit regression line $Y = \beta_o + \beta_1 X$ can be found by the following two methods:

- Differentiation
- Gradient descent

We can use statsmodels or SKLearn libraries in python for the linear regression.

Types of Linear Regression

1. **Simple Linear Regression**:
    - **One** independent variable (x) and one dependent variable (y)
    - Equation: $y=\beta0+\beta1x$

2. **Multiple Linear Regression**:
    - **Multiple** independent variables (x1, x2, ..., xn) and one dependent variable (y)
    - Equation: $y=\beta0+\beta1x1+\beta2x2+...+\beta nxn$

3. **Polynomial Regression**:
    - Fits a curved line by adding polynomial terms to the linear regression equation
    - Equation: $y=\beta0+\beta1x+\beta2x2+...+\beta nxn$

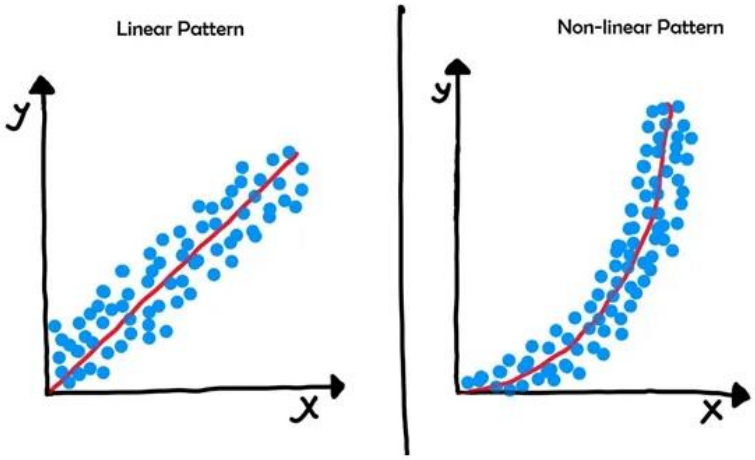
4. **Logistic Regression**:
    - Used for binary classification problems
    - Predicts the probability of the dependent variable being 1 or 0

Assumptions of Linear Regression

1. **Linearity**: The relationship between x and y is linear
2. **Independence**: The residuals are independent
3. **Homoscedasticity**: The residuals have constant variance
4. **Normality**: The residuals are normally distributed
5. **No multicollinearity**: Independent variables are not highly correlated

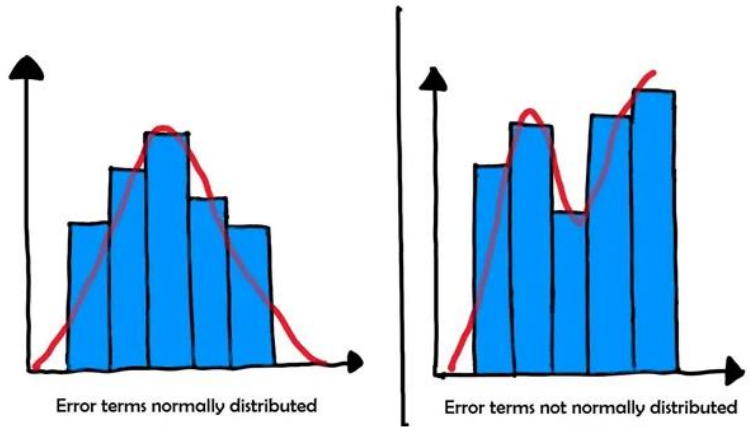
At the time of building a linear model, we assume that the target variable and predictor variables are linearly dependent. But, apart from these, below are few assumptions in linear regression model:

1. **Linear relationship between X and y:** X and Y should always display some sort of a linear relationship; otherwise, there will not be any use of fitting a linear model between them.

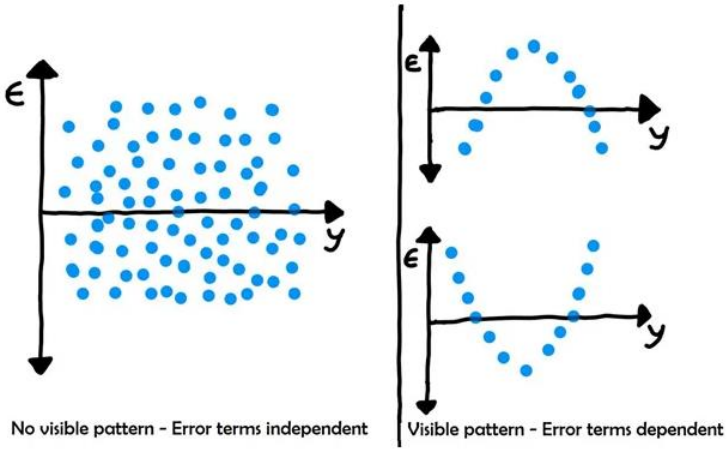


2. **Normal distribution of error terms:** It represents the assumption of normality. Which

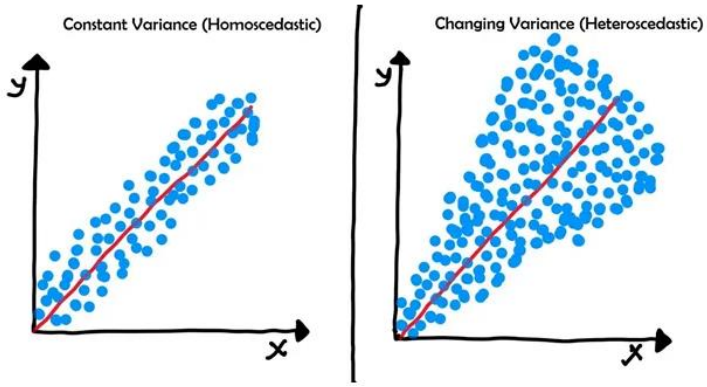
exhibits that error terms generally follow a **normal distribution with mean equal to zero** in most cases.



3. **Independence of error terms:** It explains that the error terms should not be dependent on one another. It means, there should not be any meaningful distribution between independent variable and error term.



4. **Constant variance of error terms:** This assumption says that the variance should not increase or decrease as the error values change. Also, the variance should not follow any pattern as the error terms change.



Evaluation Metrics

1. **R-squared (R^2)**:
    - Measures the proportion of variance in y that is predictable from x
    - Ranges from 0 to 1, higher is better
2. **Mean Squared Error (MSE)**:
    - Average squared difference between predicted and actual values
    - Lower is better
3. **Root Mean Squared Error (RMSE)**:
    - Square root of MSE
    - Interpretable in the same units as y
4. **Mean Absolute Error (MAE)**:
    - Average absolute difference between predicted and actual values
    - Lower is better

Linear regression is a powerful and widely used algorithm for predictive modelling and understanding relationships between variables.

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

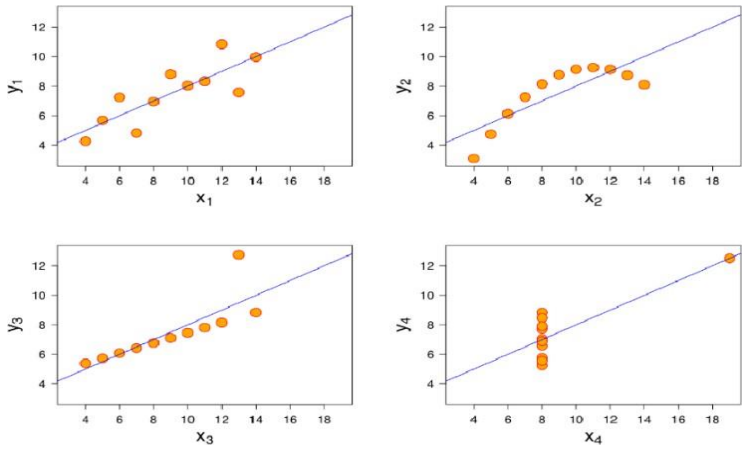
Anscombe's Quartet can be defined as **a group of four data sets which are nearly identical in simpledescriptive statistics**, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize completely, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	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
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

The summary statistics show that the means and the variances were identical for x and y across the groups:

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) betweenx and y is 0.816 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they showthe same regression lines as well but each dataset is telling a different story:



- Dataset I appear to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by anoutlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.
- This quartet emphasizes the importance of visualization in Data Analysis. Looking at thedata reveals a lot of the structure and a clear picture of the dataset.

3. What is Pearson's R? (3 marks)

Pearson's R, or the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is denoted as r and ranges from -1 to 1.

Definition and Interpretation

- **Positive Correlation**: A value of r*r* close to +1 indicates a strong positive relationship, meaning that as one variable increases, the other variable tends to also increase. For example, height and weight often exhibit a positive correlation.
- **Negative Correlation**: A value close to -1 indicates a strong negative relationship, where an increase in one variable corresponds to a decrease in the other. An example is the relationship between the speed of a vehicle and the time taken to reach a destination; as speed increases, time decreases.
- **No Correlation**: An r*r* value around 0 suggests no linear relationship between the variables, indicating that changes in one variable do not predict changes in the other.

Pearson's R Calculation

The Pearson correlation coefficient is calculated using the formula:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Where:

- n = number of pairs of scores
- ∑xy = sum of the product of paired scores
- ∑x and ∑y = sums of the individual scores
- ∑x2 and ∑y2 = sums of the squares of the individual scores

This formula allows researchers to determine how closely two variables are related and to assess the strength and direction of their relationship.

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

Scaling is a crucial pre-processing step in data analysis and machine learning that involves adjusting the range and distribution of data features. This is performed to ensure that different features contribute equally to the analysis, especially when they are measured on different scales. Without scaling, algorithms that compute distances, such as k-nearest neighbors or gradient descent-based methods, may be biased towards features with larger numerical values, leading to suboptimal performance.

Scaling is performed for several reasons:

1. **Equal Weighting**: It ensures that all features contribute equally to the distance calculations and model training, preventing features with larger ranges from dominating the results.
2. **Improved Convergence**: In optimization algorithms, particularly those used in machine learning, scaling can lead to faster convergence during training, as it helps in navigating the loss landscape more effectively.
3. **Handling Different Units**: When features are measured in different units (e.g., height in centimetres and weight in kilograms), scaling allows for a uniform comparison.

Difference between Normalized Scaling and Standardized Scaling

Normalized scaling and standardized scaling are two common techniques used to scale data, each with distinct methodologies and applications:

Normalized Scaling (Min-Max Scaling):

Definition: This technique rescales the feature values to a fixed range, typically [0, 1]. The formula used is:

$$X_{\text{new}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

Use Case: It is most effective when the data does not contain outliers, as it can skew the results. Normalization is often used in algorithms that do not assume any distribution of the data, such as neural networks and k-nearest neighbors.

Standardized Scaling (Z-Score Normalization):

Definition: This method transforms the data to have a mean of 0 and a standard deviation of 1. The formula is:

$$X_{\text{new}} = \frac{X - \text{mean}}{\text{std}}$$

Use Case: Standardization is preferred when the data follows a Gaussian distribution. It is less affected by outliers compared to normalization, making it suitable for datasets where outliers are present.

In summary, while both techniques aim to prepare data for analysis, the choice between normalization and standardization depends on the specific characteristics of the dataset and the requirements of the machine learning algorithms being employed.

5.You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

The value of the Variance Inflation Factor (VIF) can become infinite primarily due to perfect multicollinearity among the independent variables in a regression model. This occurs when one independent variable is an exact linear combination of one or more other independent variables. In such cases, the R2*R*2 value obtained from regressing one variable against the others becomes 1, leading to the VIF formula:

$$VIF_j = \frac{1}{1 - R_j^2}$$

When *Rj*2=1, the denominator becomes zero, resulting in VIFj tending towards infinity.

In practical terms, infinite VIF indicates that the model cannot uniquely estimate the coefficients of the involved variables due to their perfect correlation. This situation can severely impair the interpretability of the regression results, as it becomes impossible to determine the individual effect of each variable on the dependent variable. To address this issue, one might consider removing or combining the collinear variables to ensure that each independent variable contributes unique information to the model

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.(3 marks)

A Q-Q plot, or Quantile-Quantile plot, is a graphical tool used in statistics to compare the quantiles of two probability distributions. It plots the quantiles of one dataset against the quantiles of another, allowing researchers to visually assess whether the two distributions are similar. If the points on the plot fall approximately along the 45-degree line (y = x), it suggests that the two distributions are similar; deviations from this line indicate differences in distribution characteristics such as location, scale, and skewness.

Use of Q-Q Plots in Linear Regression

In the context of linear regression, Q-Q plots are primarily used to assess the normality of the residuals, which is a key assumption for many statistical tests and models. Here are the main uses and importance of Q-Q plots in linear regression:

1. **Normality Check**: Q-Q plots help determine if the residuals from a regression model are normally distributed. This is crucial because many inferential statistics, such as confidence intervals and hypothesis tests, rely on the assumption of normality. If the residuals are not normally distributed, the validity of these statistical inferences may be compromised.
2. **Identifying Deviations**: By examining the Q-Q plot, analysts can identify specific deviations from normality, such as skewness or the presence of outliers. For instance, if points deviate significantly from the line, it may indicate that the model is not capturing all the variability in the data, or that the data contains outliers that need further investigation.
3. **Model Diagnostics**: Q-Q plots serve as a diagnostic tool for validating the assumptions of the regression model. If the residuals appear to follow the line closely, it suggests that the model is appropriate. Conversely, significant deviations can prompt the analyst to consider transformations of the dependent variable or alternative modelling approaches.

In summary, Q-Q plots are essential for validating the assumptions of linear regression models, particularly the normality of residuals, thereby ensuring the reliability of statistical conclusions drawn from the analysis.