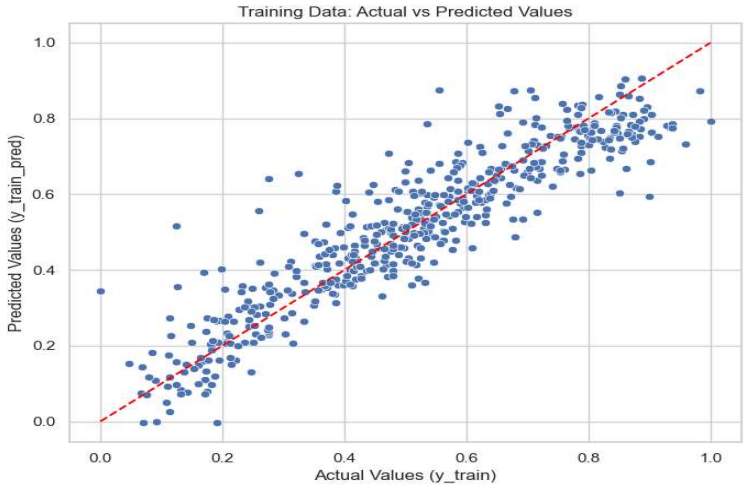
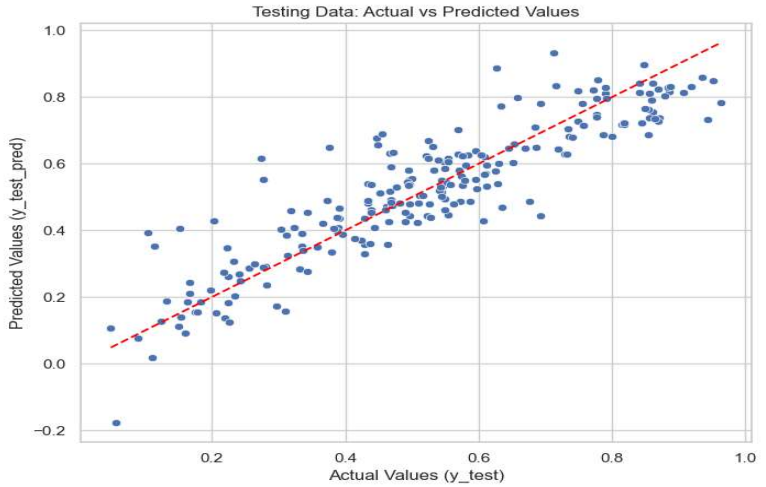
# LINEAR REGRESSION ASSIGNMENT SUBJECTIVE QUESTIONS


SUBMITTED BY -

ARNAB BERA ( ML C63 )

# 1. ASSIGNMENT BASED SUBJECTIVE QUESTIONS



# Q 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**. The categorical variables in the dataset were “season”, “year”, “month”, “holiday”, “weekday”, “workingday” and “weathersit”. These were visualized using a boxplot. These variables had the following effect on our dependent variable.

- **Season**:
  - **Fall**: Bike demand is highest in Fall.
  - **Summer and Winter**: Summer and winter has intermediate value of count with summer having greater count among the two.
  - **Spring**: The demand for bikes is lowest in spring, possibly due to less favourable weather.
- **Year**:
  - **2019 vs. 2018**: There is a clear increase in bike demand from the year 2018 to 2019. This trend suggests that the bike-sharing program gained popularity over this period.
- **Month:**
  - **High-Demand Months:** June, July, August and September are the months with the highest bike demand. Out of all these months, September has seen the highest no of rentals. This could be due to the warm summer weather, which is ideal for biking.

- **Low-Demand Months**: December has seen the lowest no of rentals. January, February, and December see the less bike demand, likely due to colder winter weather, which discourages biking.

- **Holiday:**
  - **Holidays vs Non-holidays**: Bike demand is higher on holidays compared to non-holidays. This increase can be attributed to people having more leisure time and choosing to bike for recreation or errands on holidays.
- **Weekday:**
  - **Even Distribution**: Bike demand is relatively evenly distributed across all weekdays, indicating consistent usage throughout the week.
  - **Slightly Higher on Fridays and Saturdays**: There is a slight increase in bike usage on Fridays and Saturdays, though the difference is not very pronounced.
- **Weathersit:**
  - **Clear Weather**: The highest bike demand occurs during clear weather conditions, due to favourable weather.
  - **Adverse Weather:** Bike demand decreases significantly during misty conditions, light snow/rain, and heavy snow/rain. There are no users when there is heavy snow/rain . The least demand is observed during light snow/rain, as adverse weather conditions make biking less appealing and potentially hazardous.

## Q2. Why is it important to use drop_first=True during dummy variable creation? (2 marks)

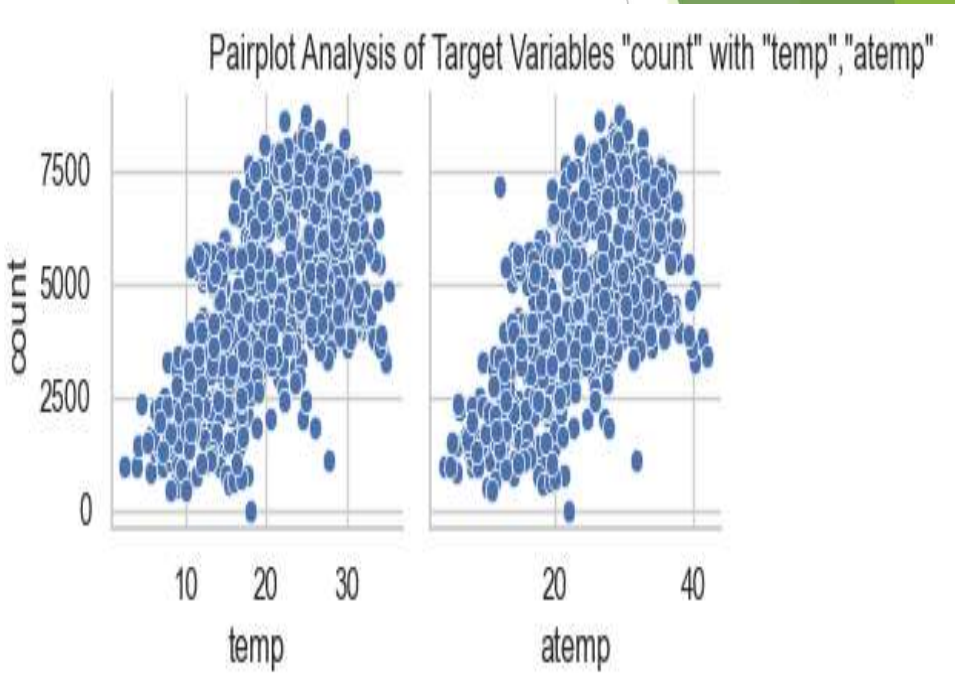
**Answer**. Using 'drop_first=True' during dummy variable creation is important for the following reasons :

- **Preventing Multicollinearity:** Including all dummy variables for a categorical feature can lead to multicollinearity, which occurs when predictor variables are highly correlated. This can make it difficult to determine the individual effect of each variable on the target variable. By dropping the first dummy variable, we avoid this issue and ensure that the remaining dummies can provide the necessary information without redundancy.
- **Reducing Redundancy:** By dropping the first category, the total number of dummy variables is reduced, which simplifies the model and improves efficiency.
- **Model Interpretability:** This practice ensures that the model remains interpretable and free from redundant variables, making it easier to understand and analyze the effects of other predictors.

**Syntax** - drop_first: bool, default False, which implies whether to get n-1 dummies out of n categorical levels by removing the first level.
**Example** - Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not B and C, then it is obvious A. So we do not need any extra variable or column to identify the A. Hence , we can drop first column A as it is redundant. So , for two dummy columns B and C the combination will be
A will be denoted by 00 , B will be denoted by 10 and C will be represented by 01.

## Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

**Answer.**
**Highest Correlation:**
In the pair-plot analysis, the two temperature variables , "temp" and "atemp" , show the highest correlation with the target variable "count" or "cnt" . This strong positive correlation indicates that higher temperatures are associated with an increase in bike bookings.



## Q4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

**Answer**. I have validated the assumption of Linear Regression Model based on below 5 assumptions -

- **Normality of Error Terms**: If the residuals follow a normal distribution, the assumption is met.

**Histogram**: Plotted a histogram of the residuals. If the residuals are normally distributed, the histogram should resemble a bell curve.

**Q-Q Plot**: Plotted a Q-Q plot of the residuals. If the residuals are normally distributed, the points should lie along the 45-degree line.

- **Multicollinearity Check:**

**Variance Inflation Factor (VIF):** Calculated the VIF for each predictor variable. VIF values less than 10 indicate that multicollinearity is not a concern.

- **Linear Relationship Validation:**

**Residual Plot:** Plotted residuals against the predicted values. If the residuals are randomly scattered around zero, it suggests that there is a linear relationship between the predictors and the response variable.

- **Homoscedasticity:**

**Residuals vs. Predicted Plot:** Plotted residuals against the predicted values to check for constant variance. The absence of a clear pattern indicates homoscedasticity.

- **Independence of residuals:**

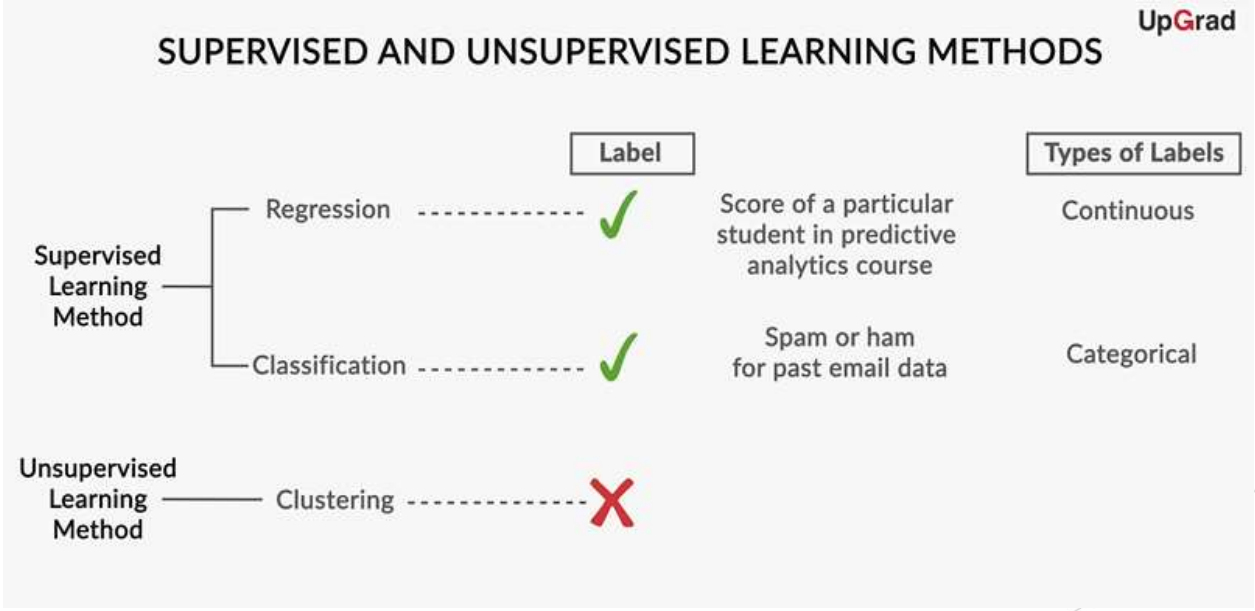
**Durbin-Watson Test:** Have calculated and checked the Durbin-Watson statistic to detect autocorrelation in the residuals.

## Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

**Answer.**
Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –

- **Temperature or "temp":** Higher temperatures are associated with increased bike usage.
- **year_2019**: The year 2019 seems to be a strong predictor , indicates an increasing trend in bike usage over time.
- **Light Snowy Rain Weather or weathersit_light_snow_rain**: Adverse weather conditions like light snow or rain discourage bike usage, impacting the demand for shared bikes negatively.

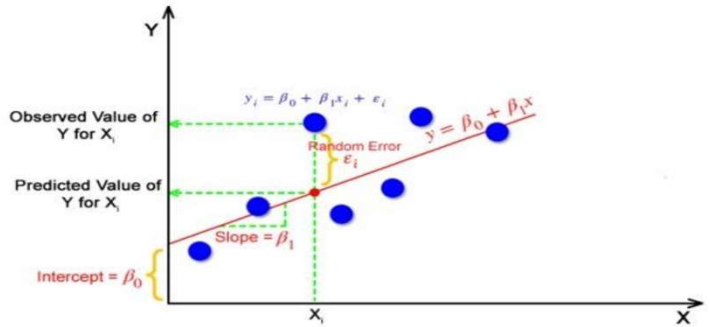
# 2. GENERAL SUBJECTIVE QUESTIONS



# Q1.Explain the linear regression algorithm in detail. (4 marks)

**Answer**. Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).
Dependent Variable (Y): The outcome or the variable we are trying to predict.
Independent Variables (X): The predictors or factors that influence the dependent variable.

Model Representation:
Simple Linear Regression:
In simple linear regression, the relationship between the dependent and independent variables is modeled as a straight line. The simplest form of linear regression is the simple linear regression, which involves one independent variable:
$Y=\beta 0+\beta 1X+\epsilon$



where:
Y is the dependent variable.
X is the independent variable.
$\beta_0$ is the intercept (the value of Y when X=0).
$\beta_1$ is the slope of the line (how much Y changes for a unit change in X).
$\epsilon$ is the error term (the difference between the observed and predicted values of Y).

Multiple Linear Regression:
When there are multiple independent variables, the model is extended to:
$Y=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_nX_n+\epsilon$
where
$X_1, X_2, ..., X_n$ are the independent variables.
$\beta_1, \beta_2, ..., \beta_n$ are the coefficients for each independent variable.

Assumptions of Linear Regression
Linearity: The relationship between the dependent and independent variables is linear.
Independence: The residuals (errors) are independent.
Homoscedasticity: The residuals have constant variance at every level of X.
Normality: The residuals of the model are normally distributed.

Steps in Linear Regression

Hypothesis:

The model starts with a hypothesis that there is a linear relationship between the dependent and independent variables.

Estimating Coefficients:

The coefficients (ß) are estimated using the least squares method. This method minimizes the sum of the squared differences between the observed and predicted values of Y.

Fitting the Model:

The linear equation is fitted to the data by adjusting the coefficients to minimize the residual sum of squares (RSS):

RSS=$\sum_{i=1}^{n}$ ( Yi $- Yi\_pred$)^2

where Yi_pred is the predicted value of Yi for the i-th observation.

Example Scenario

Imagine you are predicting house prices based on features like the size of the house, number of bedrooms, and age of the house. In a multiple linear regression model, the price of the house (dependent variable) would be predicted based on these features (independent variables):

Price=ß0+ß1(Size)+ß2(Bedrooms)+ß3(Age)+ϵ

Here, ß0 is the intercept,

ß1 is the coefficient for the size of the house,

ß2 is the coefficient for the number of bedrooms, and

ß3 is the coefficient for the age of the house.

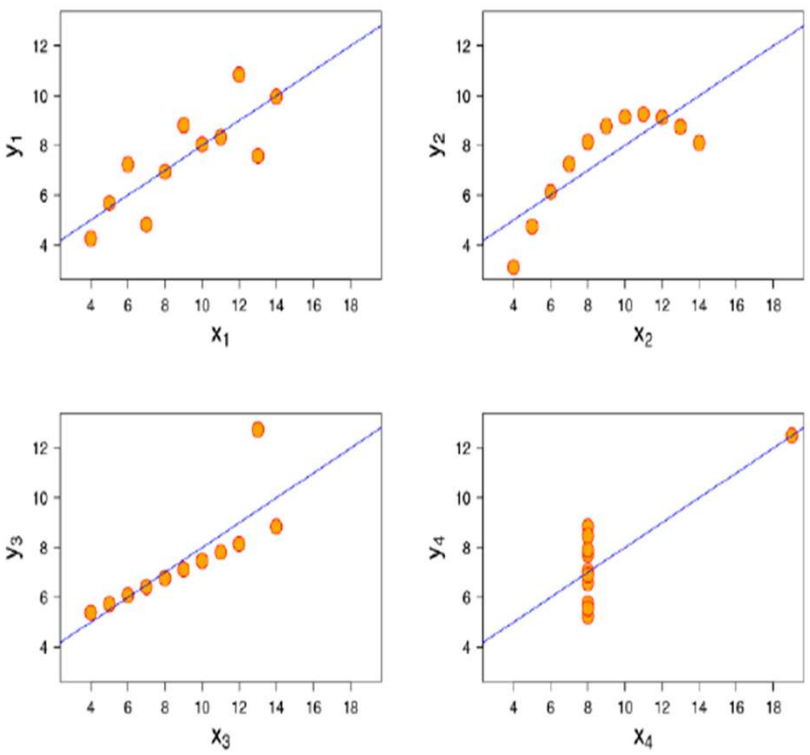
## Q2. Explain Anscombe's quartet in detail.(3 marks)

**Answer**. 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. Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple
descriptive 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. 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) between x and y is 0.816 for each dataset

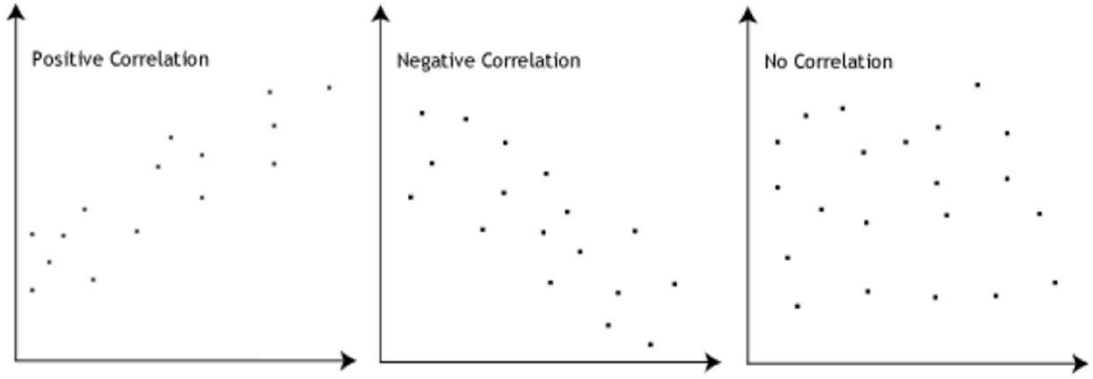



- Graph 1 (Dataset 1) appears to have clean and well-fitting linear models.
- Graph 2 (Dataset 2) is not distributed normally.
- In Graph 3 (Dataset 3) the distribution is linear, but the calculated regression is thrown off by an outlier.
- Graph 4 (Dataset 4) shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.
Despite the similarities in statistical data, the datasets differ drastically in their graphical representations.

## Q3.What is Pearson's R? (3 marks)

**Answer**. Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.
The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:




r = 1 means the data is perfectly linear with a positive slope
r = -1 means the data is perfectly linear with a negative slope
r = 0 means there is no linear association

## Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

**Answer.**
Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give
wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

## Difference between normalized scaling and standardized scaling

| Sl No. | Normalized scaling | Standardized scaling |
|---|---|---|
| 1. | Minimum and maximum value of features are used for scaling. | Mean and standard deviation is used for scaling. |
| 2. | It is used when features are of different scales. | It is used when we want to ensure zero mean and unit standard deviation. |
| 3. | Scales values between [0, 1] or [-1, 1]. | It is not bounded to a certain range. |
| 4. | It is really affected by outliers. | It is much less affected by outliers. |
| 5. | Scikit-Learn provides a transformer called MinMaxScaler for Normalization. | Scikit-Learn provides a transformer called StandardScaler for standardization. |

## Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

**Answer. VIF - the variance inflation factor -** The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. (VIF) =1/(1-R^2 ).

$$\mathrm{VIF}(X_i) = \frac{1}{1-R_i^2}$$

If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

**Perfect Multicollinearity:** When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity.

$$\mathrm{VIF}(X_i) = \frac{1}{1-1} = \frac{1}{0} = \infty$$

**Impact on Model:** High VIF values can lead to unstable coefficient estimates and inflated standard errors, compromising model reliability. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

## Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

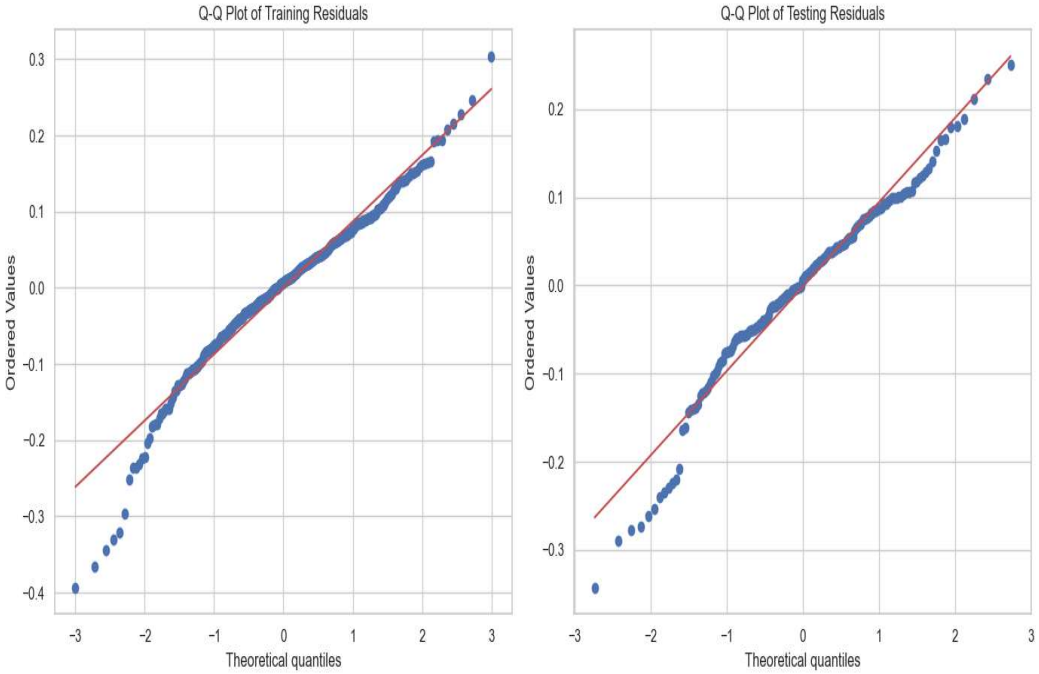
**Answer.**

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions. If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line $y = x$

If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line $y = x$. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.




**Use in Linear Regression**

In the context of linear regression, a Q-Q plot is primarily used to assess whether the residuals (the differences between observed and predicted values) follow a normal distribution. This is important because one of the key assumptions of linear regression is that the residuals are normally distributed. Normality of residuals ensures the validity of hypothesis tests and confidence intervals.

**Importance of a Q-Q Plot in Linear Regression**

- **Checking Normality**: Assesses if residuals are normally distributed by comparing them to a theoretical normal distribution.
- **Identifying Deviations**: Detects skewness, kurtosis, and other deviations from normality, indicating potential data issues.
- **Model Diagnostics**: Evaluates the fit and appropriateness of the regression model, highlighting violations of assumptions for refinement.

# -- THE END --