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)

Sol: Year, Season, Weather Situation, Holiday, Month, Working day, and Weekday were the categorical variables in the dataset. A boxplot was used to visualise these. These variables influenced our dependent variable in the following ways:

Variable Name	Inferences
Year	count increased significantly in 2019 compared to 2018.
Season	count is higher for Fall (Autumn) and then followed by Summer.
Weather Situation	count is higher when the weather forecast was 'Clear, Partly Cloudy.' There is no data for Heavy rain/Snow.
Holiday	count is lower during holidays.
Month	count is highest in September and lowest in December.
Working day	count is unaffected whether it's a holiday or a working day.
Week day	count is similar for all weekdays

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

Sol: When creating dummy variables, the parameter drop_first=True is used to eliminate one of the categorical levels, also known as the reference category or baseline category, from the resulting dummy variables. It is important because:

a. **Multicollinearity reduction**: Including dummy variables for all levels of a categorical variable can lead to multicollinearity in the model. This can cause problems in regression models, such as unstable coefficient estimates, inflated standard errors, and difficulty in interpreting the impact of individual predictors. By dropping one dummy variable, its ensured that the remaining dummy variables are not perfectly correlated, reducing multicollinearity.

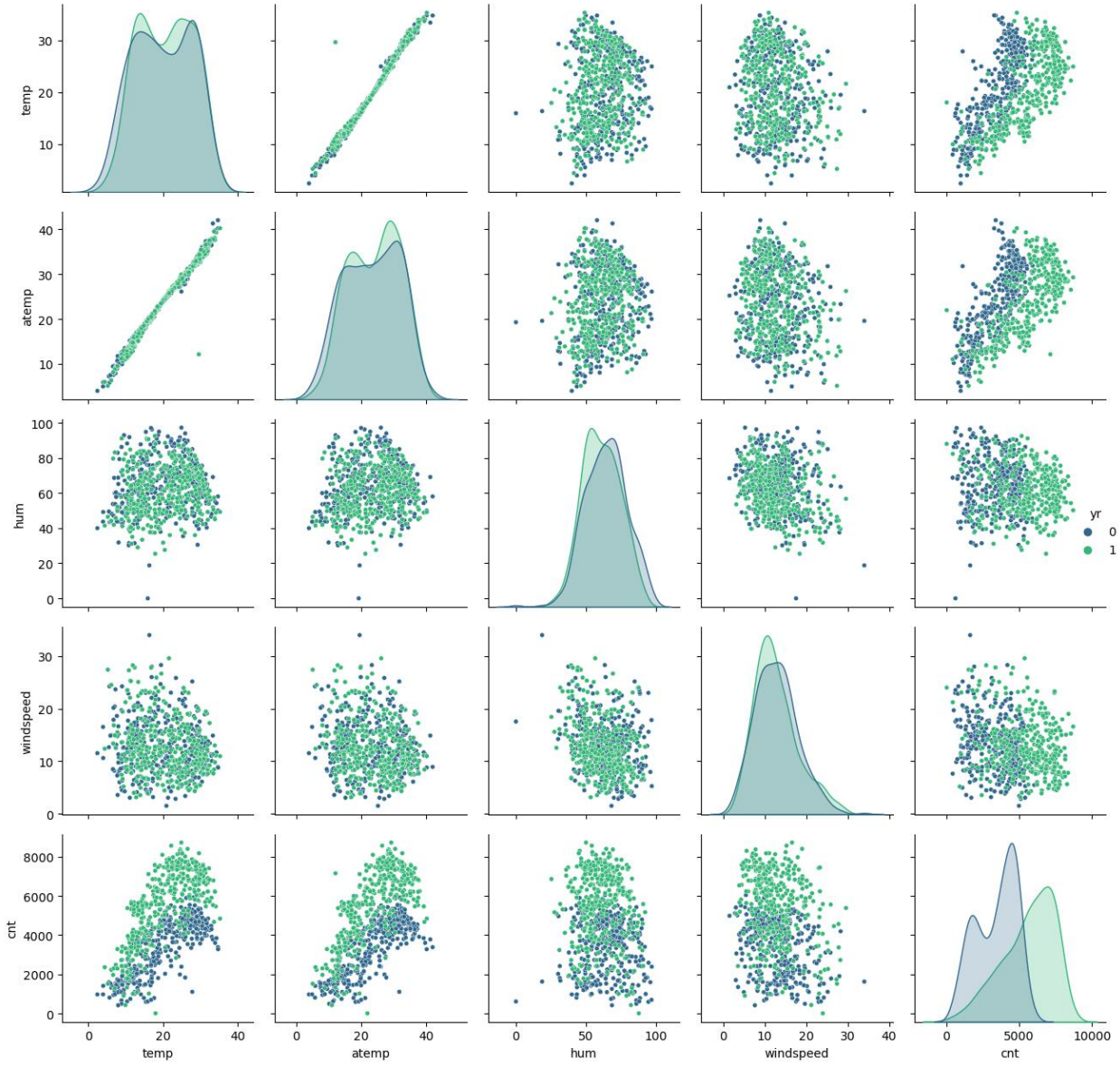
b. **Interpretability**: When interpreting regression coefficients, it's essential to have a reference category for categorical variables. By dropping the first dummy variable, it becomes the reference category, and the coefficients of the remaining dummy variables represent the difference between their respective categories and the reference category. This makes the interpretation more straightforward and meaningful.

c. **Dimensionality reduction**: Including dummy variables for all levels of a categorical variable increases the dimensionality of the dataset, which can lead to computational overhead and reduced model performance. Dropping the first

dummy variable reduces the number of dummy variables created by one, thus mitigating this issue.

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

Sol:

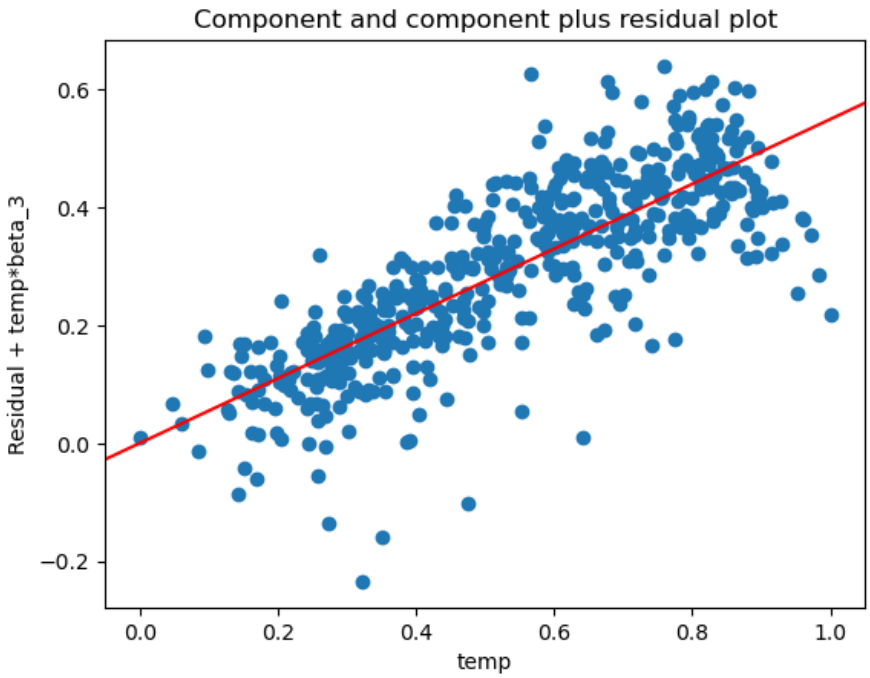


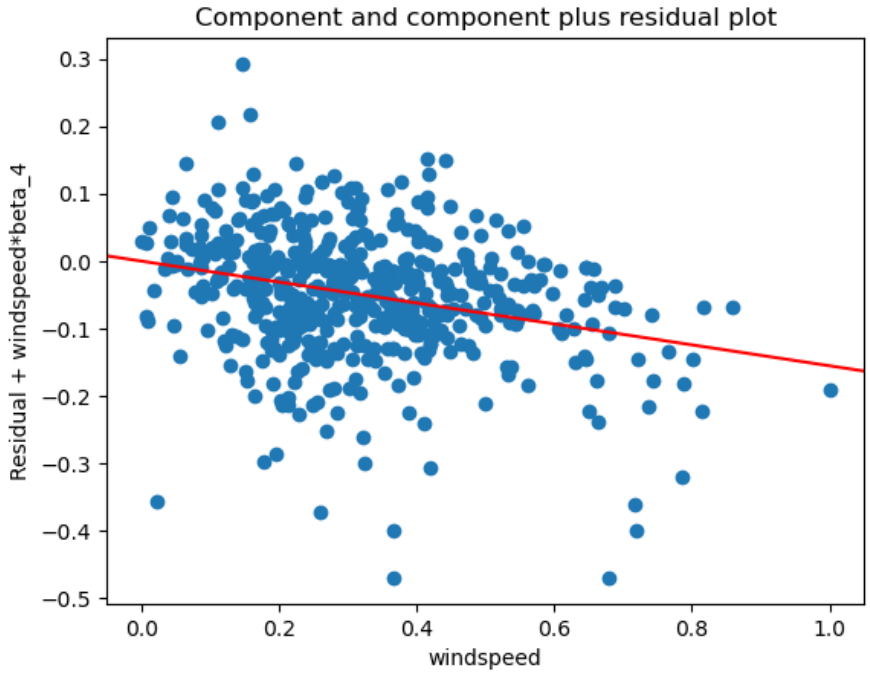
“temp” and “temp” are the two numerical variables which are highly correlated with the target variable “cnt”

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

Sol: Steps followed to you validate the assumptions of Linear Regression:

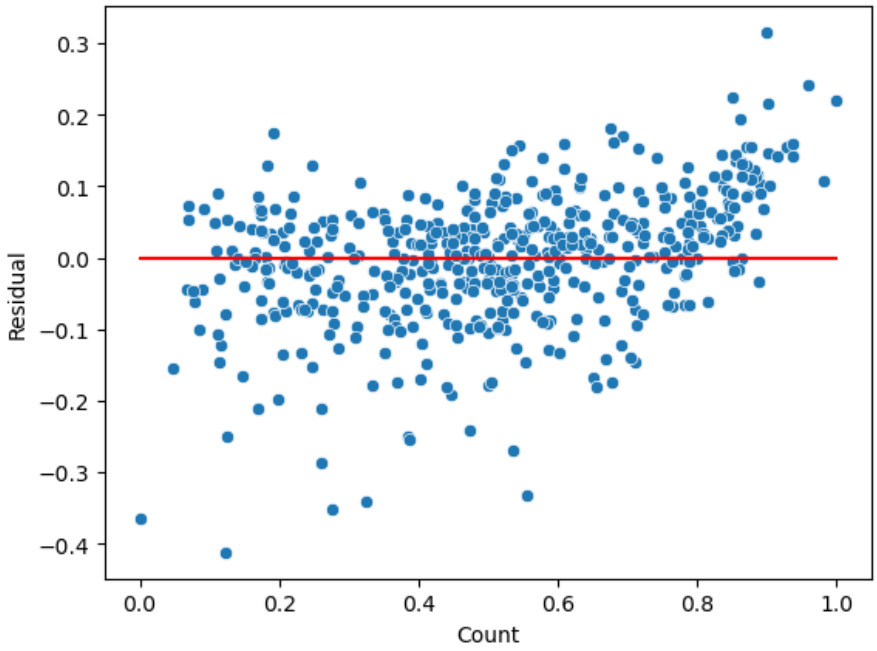
a) Linear Relationship: Made a CCPR plot with 2 predictor variables ‘temp’ and ‘windspeed’ to judge the effect of one regressor on the response variable by taking into account the effects of the other independent variables.





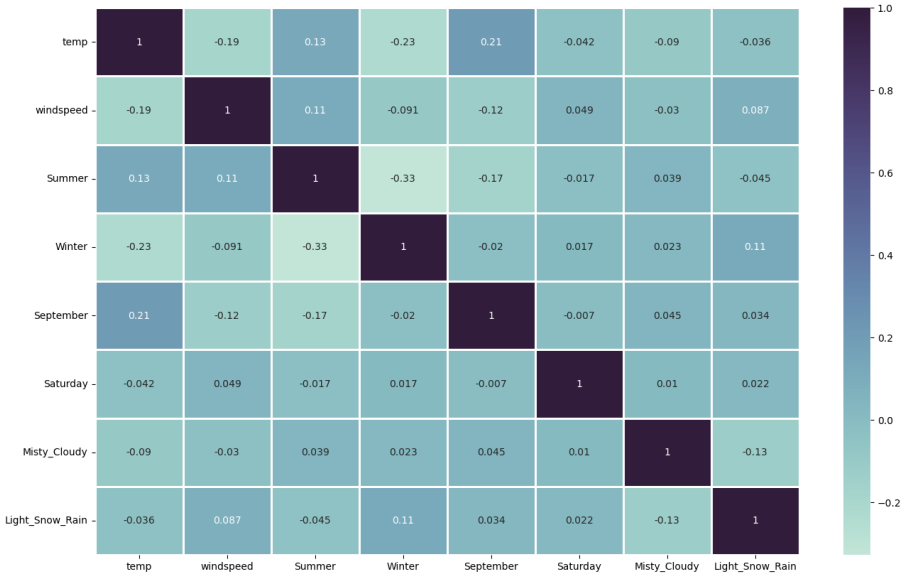
Linearity is preserved.

b) Homoscedasticity: Checked whether the spread of the residuals is consistent and independent of the magnitude of the predicted values.



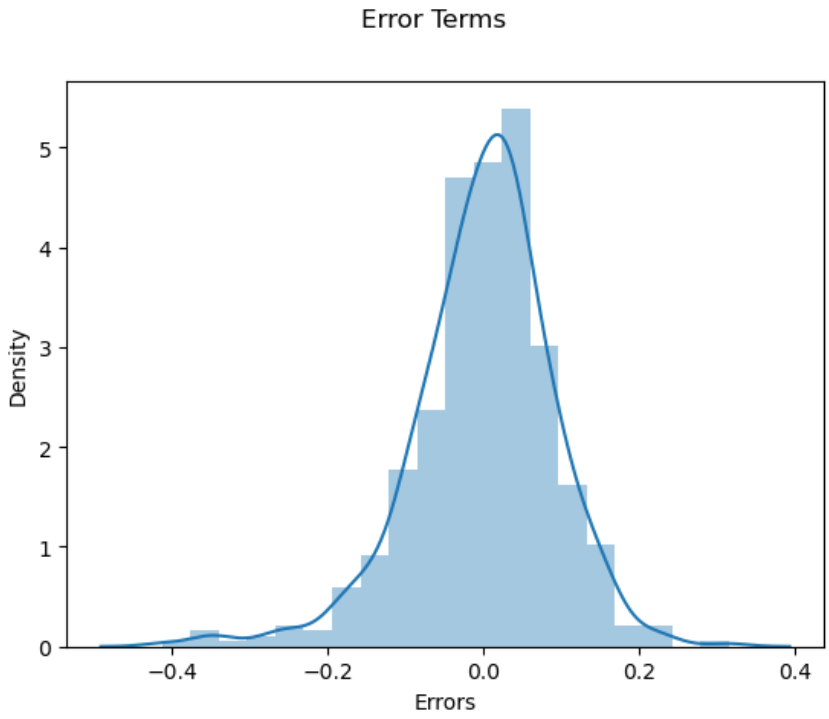
No visible pattern in residual values.

c) Absence of Multicollinearity: Checked if variables are highly correlated to each other.



No high correlation of any variables.

d) Normality of error: The distribution of residuals should be normal and centred around 0. (The mean is 0). Tested this residuals assumption by producing a distplot of residuals to see if they follow a normal distribution or not.



Residuals are scattered around mean = 0

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

Sol:

Feature	Correlation Coefficient	Inference
temp	0.549936	Change in temperature affects rentals
yr	0.233056	Increase in bike rentals per year
Light_Snow_Rain	-0.288021	Snow and rain inversely proportional to rentals

General Subjective Questions

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

Sol: Linear regression is a supervised learning algorithm used for predicting a continuous output variable (also called the dependent variable) based on one or more input features (also called independent variables). The goal of linear regression is to find the best-fitting linear relationship between the input features and the output variable. This relationship is represented by a straight-line equation of the form:

$$y = mx + c$$

Where:

- y is the predicted output (dependent variable)
- x is the input feature (independent variable)
- m is the slope of the line (coefficient), representing how much y changes when x increases by one unit
- c is the y-intercept, where the line intersects the y-axis when x is zero.

The linear regression algorithm finds the values of m and c that minimize the difference between the predicted output and the actual output in the training data. This minimization is typically achieved using the method of least squares.

Regression is performed when the dependent variable is of continuous data type and predictors or independent variables could be of any data type like continuous, nominal/categorical etc. Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error. In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term.
Regression is broadly divided into simple linear regression and multiple linear regression.

I. Simple Linear Regression: SLR is used when the dependent variable is predicted using only one independent variable.

$$E(y) = \beta_0 + \beta_1 x$$

β_1 is the change in E(y) corresponding to a unit increase in x.

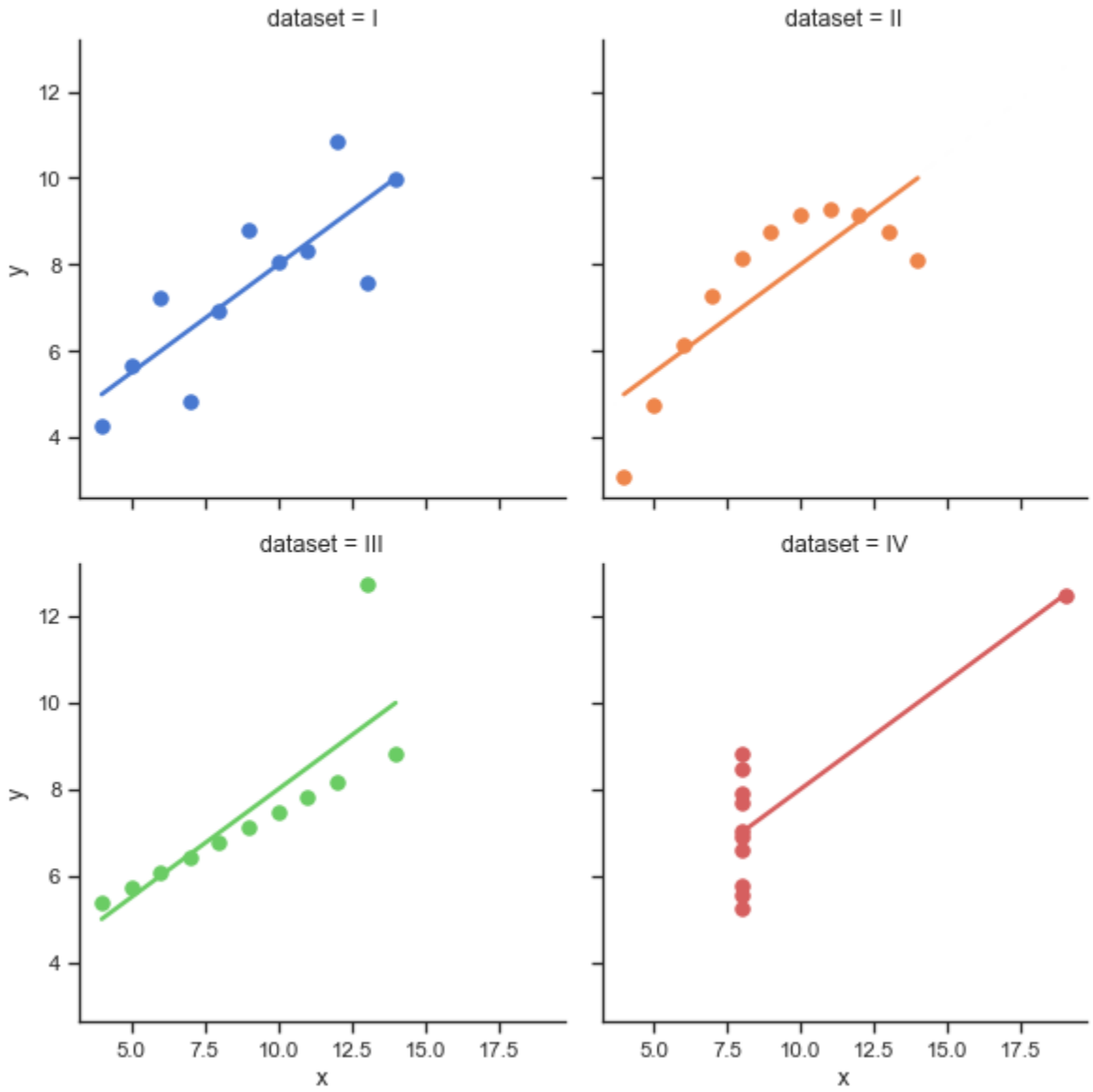
II. Multiple Linear Regression: ML is used when the dependent variable is predicted using multiple independent variables.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

β_1 = coefficient for x_1 variable
β_2 = coefficient for x_2 variable
β_n = coefficient for x_n variable
β_0 is the intercept (constant term).

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

Sol:



I	x	10	8	13	9	11	14	6	4	12	7	5
	y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68
II	x	10	8	13	9	11	14	6	4	12	7	5
	y	9.14	8.14	8.74	8.77	9.26	8.1	6.13	3.1	9.13	7.26	4.74
III	x	10	8	13	9	11	14	6	4	12	7	5
	y	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73
IV	x	8	8	8	8	8	8	8	19	8	8	8
	y	6.58	5.76	7.71	8.84	8.47	7.04	5.25	12.5	5.56	7.91	6.89

Anscombe's quartet was introduced by the British statistician Francis Anscombe in 1973 to emphasize the importance of visualizing data and not relying solely on summary statistics. It comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

- The first scatter plot appears to be a simple linear relationship.
- The second graph is not distributed normally. While there is a relation between them it's not linear.
- In the third graph the distribution is linear, but should have a different regression line. The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

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

Sol: Pearson's correlation coefficient, often denoted as Pearson's R or simply R, is a statistical measure that quantifies the linear relationship between two continuous variables. It is a widely used method to assess the strength and direction of the association between two variables. The coefficient ranges from -1 to 1, where:

- If R is close to -1, it indicates a perfect negative linear relationship, meaning that as one variable increases, the other decreases in a straight-line pattern.
- If R is close to 1, it indicates a perfect positive linear relationship, meaning that as one variable increases, the other increases in a straight-line pattern.
- If R is close to 0, it indicates a weak or no linear relationship, suggesting that there is no linear trend between the two variables.

The formula to calculate Pearson's correlation coefficient between two variables X and Y is as follows:

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

where:

- x_i and y_i are the individual data points for variables x and y, respectively.
- $\bar{x}$ and $\bar{y}$ are the means of variables x and y, respectively.

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

Sol: Scaling refers to the process of transforming the numerical values of different variables to a common scale. The goal is to bring all variables to a similar range, typically between 0 and 1 or around a mean of 0 with a standard deviation of 1. Scaling is an essential step in data preprocessing before applying certain machine learning algorithms, as it can improve the model's performance and convergence.

Scaling is performed for the following reasons:

- **Feature Magnitude**: Variables with significantly different scales can dominate the learning process, especially in distance-based algorithms (e.g., k-nearest neighbors, support vector machines). Scaling helps ensure that all features contribute equally to the model.

- **Gradient Descent**: Many optimization algorithms used in machine learning, such as gradient descent, converge faster when the features are on a similar scale. It helps in more efficient and effective model training.

- **Regularization**: Regularization techniques (e.g., L1 and L2 regularization) penalize large coefficients. Scaling can prevent certain features from receiving disproportionate regularization penalties due to their larger magnitudes.

- **Distance Metrics**: Scaling is important for clustering and similarity-based algorithms (e.g., k-means) that rely on distance metrics. Variables with different scales can result in incorrect clustering.

Normalized Scaling	**Standardized Scaling**
Scales the model using minimum and maximum values.	Scales the model using the mean and standard deviation.
When features are on various scales, it is functional.	When a variable's mean and standard deviation are both set to 0, it is beneficial.
x = $\frac{x-\min(x)}{\max(x)-\min(x)}$	x = $\frac{x-\text{mean}(x)}{sd(x)}$
Values on the scale fall between [0, 1] and [-1, 1].	Values on a scale are not constrained to a particular range.
When the feature distribution is unclear, it is helpful.	When the feature distribution is consistent, it is helpful.
Outliers information is lost.	Outliers information is preserved.

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

Sol: VIF is a measure used to detect multicollinearity among predictor variables in regression analysis. It quantifies how much the variance of the estimated regression coefficient is inflated due to multicollinearity.

$$\text{VIF} = \frac{1}{1 - R_i^2}$$

where R_i^2 is the coefficient of determination (R-squared) when the predictor variable XiXi is regressed on all other predictor variables.

The VIF becomes infinite when there is perfect multicollinearity or nearly perfect multicollinearity among predictor variables. Perfect multicollinearity occurs when one predictor variable is a perfect linear combination of other predictor variables. If an independent variable can be completely described by other independent variables, it has perfect correlation and has an R-squared value of 1. This is because the denominator in the VIF formula is close to zero or zero (VIF = 1/(1-1).

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

Sol: A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. It is a visual comparison between the quantiles of the observed data and the quantiles of the theoretical distribution. The main purpose of the Q-Q plot is to check if the data deviates significantly from the expected distribution and to identify any departures from normality.

A Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

- **Similar Distribution:** If all point of quantiles lies on or close to straight line at an angle of 45 degrees from x -axis
- **Y-values < X-values:** If y-quantiles are lower than the x-quantiles.
- **Y-values < X-values:** If x-quantiles are lower than the y-quantiles.
- **Different distribution:** If all point of quantiles lies away from the straight line at an angle of 45 degrees from x -axis

Use and importance of Q-Q plot:

- **Checking Normality Assumption**: In linear regression, one of the crucial assumptions is that the residuals (the differences between actual and predicted values) are normally distributed. Q-Q plots are frequently used to assess whether the residuals of a regression model are approximately normally distributed. If the residuals follow a straight line on the Q-Q plot, it suggests that the normality assumption is reasonable. However, if the points deviate significantly from the line, it indicates departures from normality.

- **Identifying Skewness and Outliers**: Q-Q plots can also reveal skewness and the presence of outliers in the data. If the points in the Q-Q plot deviate from the line at the

tails, it suggests skewness in the data. Additionally, if there are outliers in the dataset, they may appear as points significantly far from the line.

- **Model Validity and Interpretability**: Normality of residuals is not only important for the validity of statistical tests and confidence intervals in linear regression but also for the interpretability of the regression coefficients. Departures from normality can lead to biased estimates and inaccurate inferences.