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)

1. `yr`:
    a. 2019 had a higher number of bookings compared to the previous year ,2018, and this indicates a strong and positive correlation for business growth for the next couple of years.
2. `season`:
    a. Fall season, i.e. september experienced an increase in bike sharing sales indicating that this month and season is more favorable to many customers
3. `holiday`:
    a. During holidays the bike sharing sales are lower due to many customers being busy with other factors which is understandable.
4. `weathersit`:
    a. Good weather conditions play a significant role in attracting more bookings due to it having Clear, Few clouds, Partly cloudy, Partly cloudy.
5. `weekday`:
    a. Bookings are more prevalent on Thursdays, Fridays, Saturdays and Sundays compared to earlier days of the week.
6. `windspeed`:
    a. A higher wind speed indicated that bike rentals decreased due to bad weather.

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)

drop_first = True is important because it helps in reducing the extra column created during the dummy variable creation. Therefore, it reduces the correlation created among dummy variables.

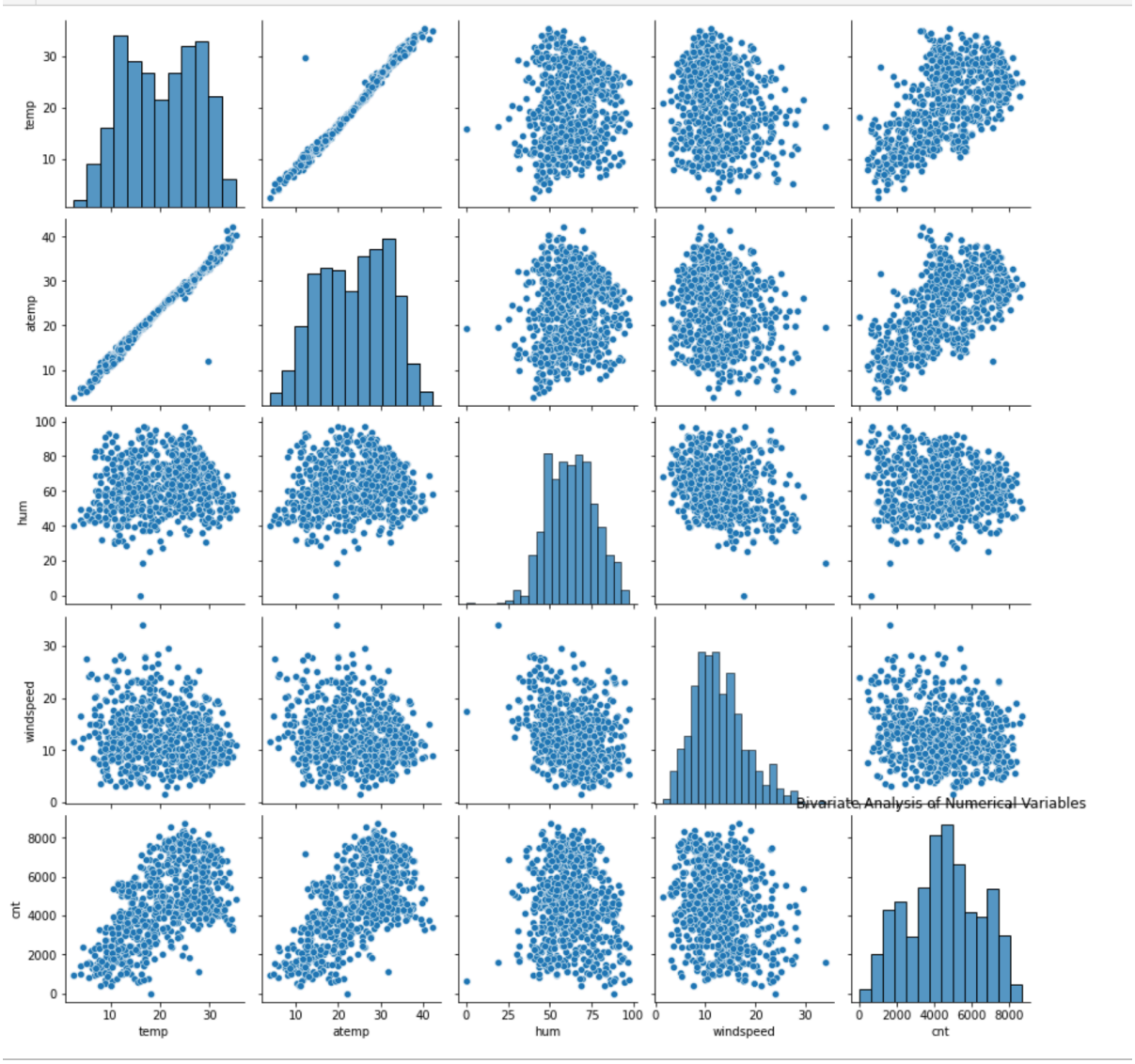
If we do not include **drop_first = True** then `n` dummy variables will get created and these predictors are themselves correlated which is known as *multicollinearity* and it leads to a *Dummy Variable Trap*.

The *Dummy variable trap* occurs when two more dummy variables are created by one-hot-encoding that are highly correlated and when one variable can be predicted using another variable. As a result, this makes it difficult to interpret the predicted coefficients and understand what variables have a positive or negative impact on the model.

Three main advantages of using **drop_first = True**

1. Prevents multicollinearity
2. Improves interpretability
3. Ensures model stability

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)



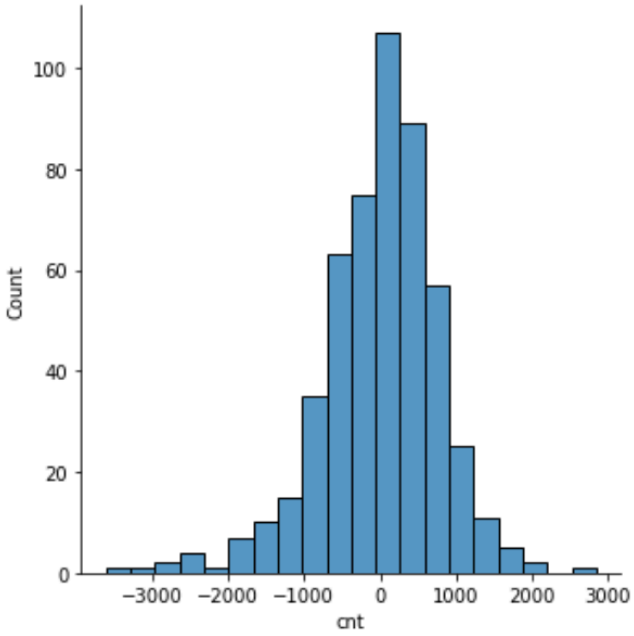
`temp` has the highest correlation with the target variable `cnt` out of all of the numerical variables.

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)

I validated the assumptions of linear regression after building the model on the training set through the following steps:

1. Residual Analysis:

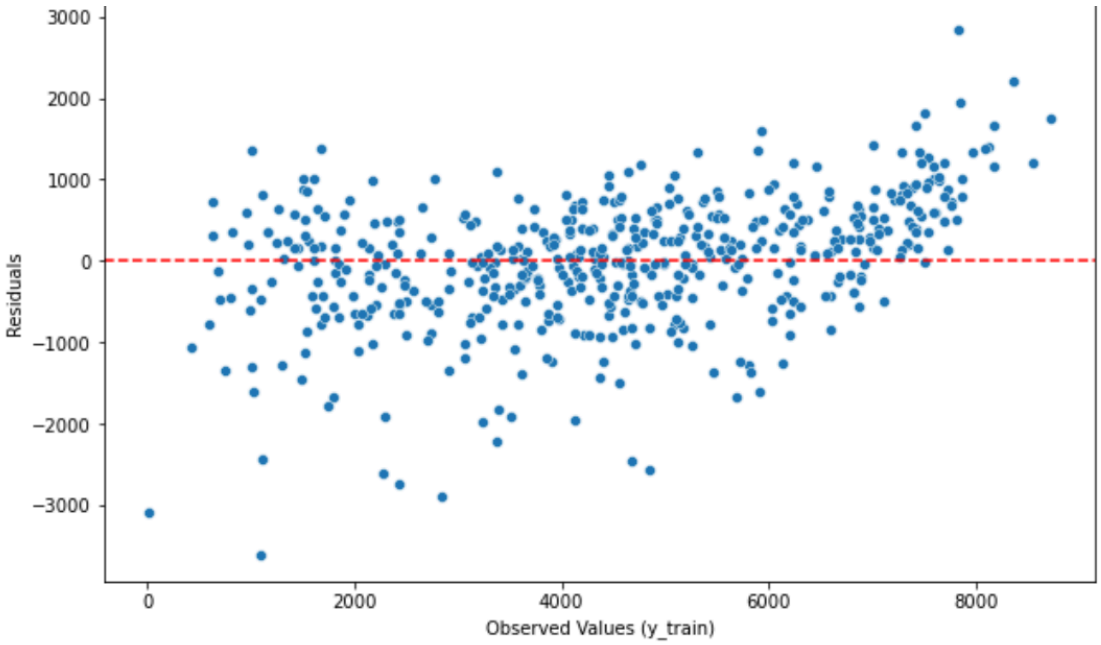
```
ut[83]: <seaborn.axisgrid.FacetGrid at 0x7f8b5567e1f0>
```



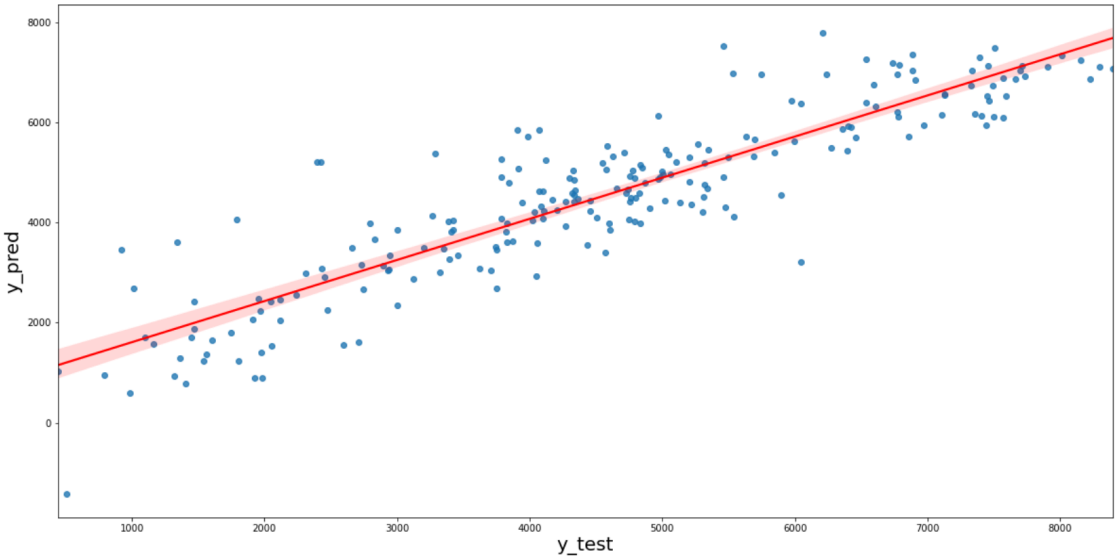
Based on the above we can see that the residuals are normally distributed and there is no patterns in the plot.

2. Homoscedasticity:

By plotting the scatter plots of residuals, I was able to validate that there is no significance pattern in them.



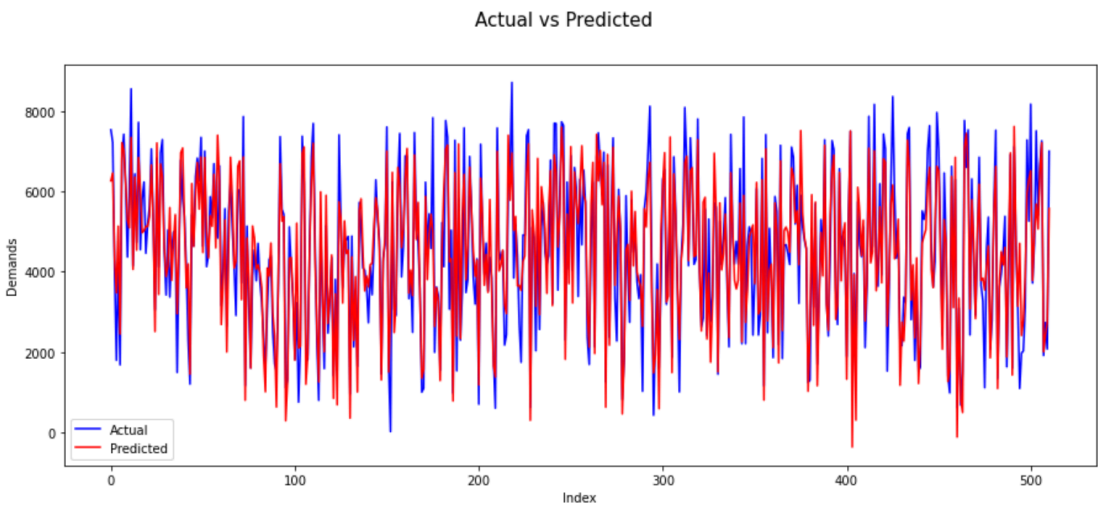
3. Linearity:



I created a scatter plot of observed vs predicted values and ensured that there is a linear relationship between the y_test and y_pred.

4. Independence of Residuals:

Examined the actual vs predicted residuals and ensured there is no pattern when plotted.



5. Multicollinearity:

By ensuring that VIF (Variance Inflation Factor) values of all the predictors in the final model are below (>) 5. After validating, there is no multicollinearity.

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)

The top three features are:

1. temp
2. yr
3. season (september)

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)

<Your answer for Question 6 goes here>

Linear regression is a supervised learning algorithm that is used to predict a continuous variable based on 1 or more independent variables (inputs). The goal of linear regression is to find the **best fit** line between the dependent and independent variables that minimizes the sum of squared differences between the observed and predicted values of the dependent variables. The relationship can be represented in a mathematical format: y = m*x + b where m is the slope of the regression line, x is the independent feature, and b is the y-intercept. A **positive linear relationship** will be called +ve if both independent and dependent variables increase. A **negative linear relationship** will be called -ve if both the dependent and independent variables decrease. There are two times of linear regression: **1. Simple Linear Regression** and **2. Multiple Linear Regression.**

Linear regression algorithm:
1. Model Representation:

Simple Linear Regression: Consists of a single independent variable:

$$y = b_0 + b_1 \cdot x + \varepsilon$$

where,
- y is the dependent variable
- x is the independent variable
- b0 is the y-intercept (constant term)
- b1 is the slope of the line, and
- E represents the error term

Multiple Linear Regression: Where there are multiple independent variables, the model is extended to:

$$y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + \cdots + b_n \cdot x_n$$

where:
- $(x_1, x_2, \ldots, x_n)$ are the independent variables, and
- $(b_0, b_1, b_2, \ldots, b_n)$ are the coefficients.

2. Objective Function:

Goal is to find the values that minimize the sum of the squared differences between the observed and the predicted value. This is often expressed as MSE:

$$\text{MSE} = \frac{1}{2m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

Where,
- m is the number of data points
- yi is the observed value
- y hat is the predicted value

3. Minimization:
Gradient descent is used to find the optimal values and is used for optimization.

$$\text{repeat until convergence}\{$$

$$w_j = w_j - \alpha\frac{\partial}{\partial w_j}J(\vec{w}, b)$$

$$b = b - \alpha\frac{\partial}{\partial b}J(\vec{w}, b)$$

$$\}$$

4. Training the Model:
Model is trained on a dataset where the algorithm learns the values to best fit the data, and this involves feeding the algorithm to various features.

5. Prediction:
After the model is trained, it can be used to make predictions on new and unseen data and new values can be obtained from non-seen input!

6. Evaluation:
Various metrics such as R^2, adjusted R^2, etc are used.

7. Assumptions:
1. Multicollinearity:
   a. Linear regression model assumes that there is very little to no multicollinearity in the data.
2. Auto-correlation:
   a. Linear regression assumes that there is very little or no auto-correlation in the data.

Occurs when there is no dependency between residual errors.

3. Relationship between:

a. Linear regression assumes that the relationship between response and feature variables must be linear.

4. Normality of error terms:

a. Error terms should be normally distributed

5. Homoscedasticity:

a. There should be no visible pattern in the residual values.

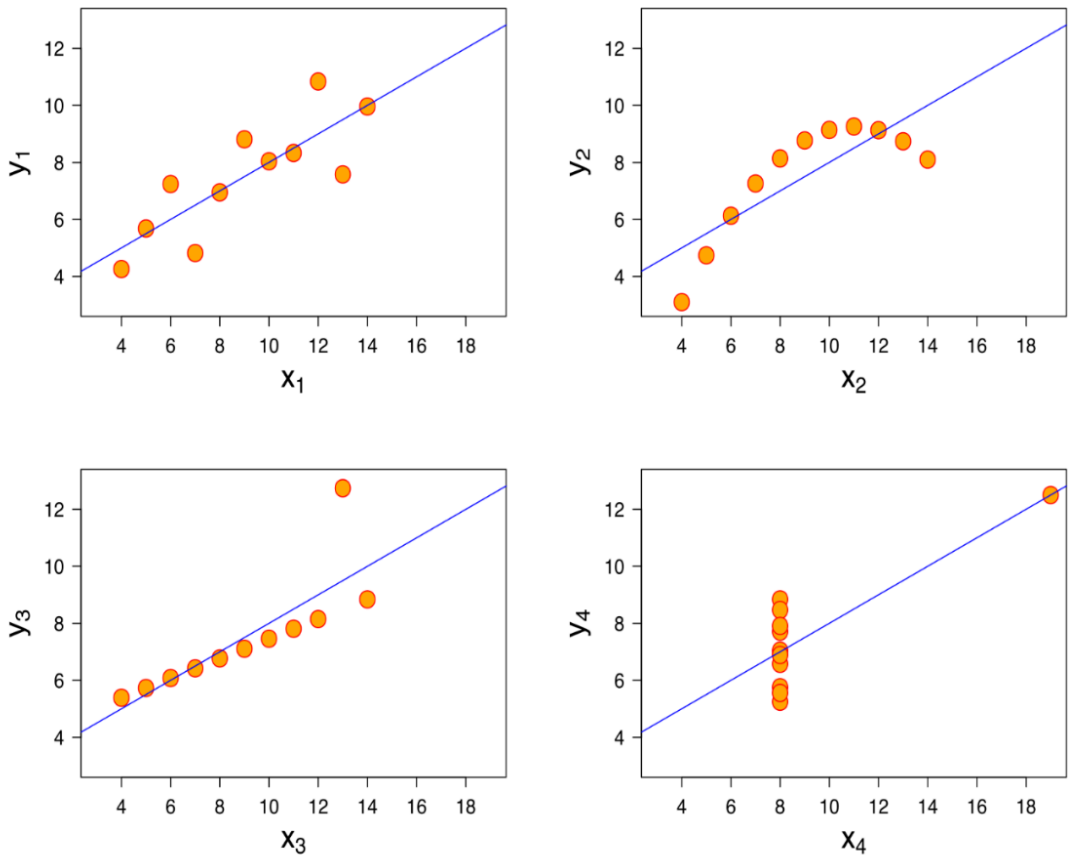
Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

Anscombe's quartet is a set of four datasets that have nearly identical statistical properties yet they differ when graphed. It was created by statistician **Francis Anscombe** in 1973. The importance is to visualize the data rather than relying solely on summary statistics. It demonstrates both the importance of graphing data when analyzing it and the effect of the outliers and other influential observations on statistical properties.

Each Dataset is:

1. Dataset 1:
    a. Linear relationship
    b. The data points are close to the regression line showing a strong and positive relationship
2. Dataset 2:
    a. Nonlinear relationship
    b. The data points form a curve, deviating significantly from the straight regression line
3. Dataset 3:
    a. Influenced by an outlier
    b. Most pints align well with the regression line but a single outlier heavily influences the summary statistics.
4. Dataset 4:
    a. Vertical alignment with an influential outlier
    b. Almost all x values are identical except for one outlier, which dominates the regression analysis.



Graphical representation of Anscombe's quartet

Importance of Anscombe's quartet:

1. Visualization is essential
2. Outliers and influence
3. Stats tools are context-sensitive
4. Avoid over reliance on correlation

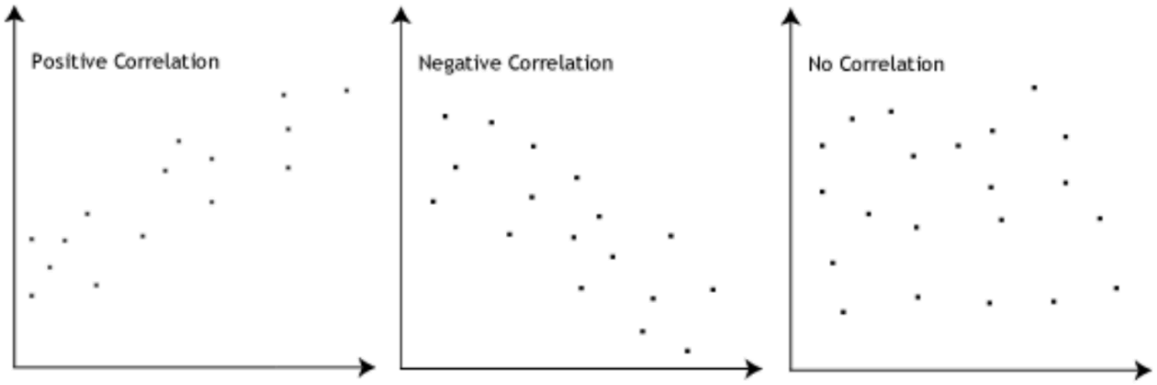
Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>

Pearson's R, also called Pearson's correlation coefficient, is a stats measure that quantifies the strength and direction of a linear relationship between 2 variables. It is one of the most commonly used correlation metrics in stats. It indicates both the direction and strength of the relationship between the variables.

The coefficient takes values between -1 and +1 where:
r = 1 => Perfect positive linear correlation
r = -1 => Perfect negative linear correlation
r = 0 => No linear correlation

A value greater than 0 indicates a positive association and vice versa for a value less than 0.



Correlation Graphs

Formula:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Where,

Xi,Yi: Individual data points.
X(bar), Y(bar): Mean of X and Y respectively.

Guidelines:

- 0.0 to 0.3 (or -0.3): Weak linear relationship
- 0.3 to 0.7 (or -0.7): Moderate linear relationship
- 0.7 to 1.0 (or -1.0): Strong linear relationship

Uses:

1. Correlation analysis
2. Hypothesis testing
3. predictive modeling

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)

<Your answer for Question 9 goes here>

Scaling is the process of transforming the features of a dataset so that they are on a similar scale or range, and it is a part of feature scaling in data preprocessing for machine learning algorithm models. The two ways you can scale is by **standardized scaling** or **normalized scaling.**

1 . Standardized Scaling:

Standardized scaling rescales the features so they have a mean of 0 and a standard deviation of 1.
- Formula:

$$X_{\text{standardized}} = \frac{X - \text{mean}(X)}{\text{std}(X)}$$

- Advantages: Less sensitive to outliers, and preserves the shape of the distribution.
- Disadvantages: Assumes that the variable follows a Gaussian Distribution.

2. Normalized scaling (min max scaling):

Min max scaling rescales the feature values to a fixed range of [0,1].
- Formula:

$$X_{\text{normalized}} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

- Advantages: Useful when the distribution of the variables is unknown or not gaussian.
- Disadvantages: sensitive to outliers.

Scaling is performed for the following reasons:

1. Ensures algorithm performance:
    a. Allows for values that have larger values to compress into distance-based calculations that are easier for many machine learning algorithms.
2. Improves convergence in optimization:
    a. Scaling helps the model converge faster by ensuring all features contribute equally for gradient based optimization algorithms (logistic regression,etc)
3. Interpretability and consistency:
    a. Scaling ensures that no feature's scale disproportionately influences the model or its interpretation.

Feature	Min-Max Scaling (normalization)	Standard Scaling (standardization)
Definition	Scales data between a fixed range of 0 and 1 using minimum and maximum values of the feature	Scales the data by subtracting the mean and dividing by the Standard Dev resulting in a distribution with a mean of 0 and a std of 1
Formula	(x - min) / (max - min)	(x - mean) / std_dev
Impact on outliers	Highly sensitive to outliers. A single extreme value can significantly affect the scaling	Less affected by outliers as the mean and std_dev are calculated across the whole dataset
When to use?	When the relative range of values within a feature is important and scale is known	When you want to normalize the data into a standard normal dist and are less concerned about outliers
Example applications	Image processing where pixel values need to be within a specific range	clustering algos, PCA where feature comparisons based on distances are important.

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)

<Your answer for Question 10 goes here>

VIF or Variance Inflation factor measures the degree of multicollinearity among the independent variables in a regression model. It tells us the impact of multicollinearity and improves model stability by addressing or removing to ensure a more reliable regression model is obtained.

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

Where Ri^2 is the R^2 value obtained by Xi against all other dependent variables.

The value of the VIF becomes infinite when the predictor variable under consideration has a perfect linear relationship with one or more predictor variables in the dataset. This is called **perfect multicollinearity.**

This occurs when the variable is completely predictable from the others.

For example:

X1	**X2**	**X3**
10	20	30
15	30	45
20	40	60

Notice that:

X3 = X1 + X2 for every row.

- X3 = 10 + 20 = 30
- X3 = 15 + 30 = 45
- X3 = 20 + 40 = 60

This means this is a **perfect linear combination**

And since R^2 = 1, we get VIF as infinite.

<u>The consequences of infinite VIF:</u>

1. regression model fails
2. unstable coefficients

<u>To handle them:</u>

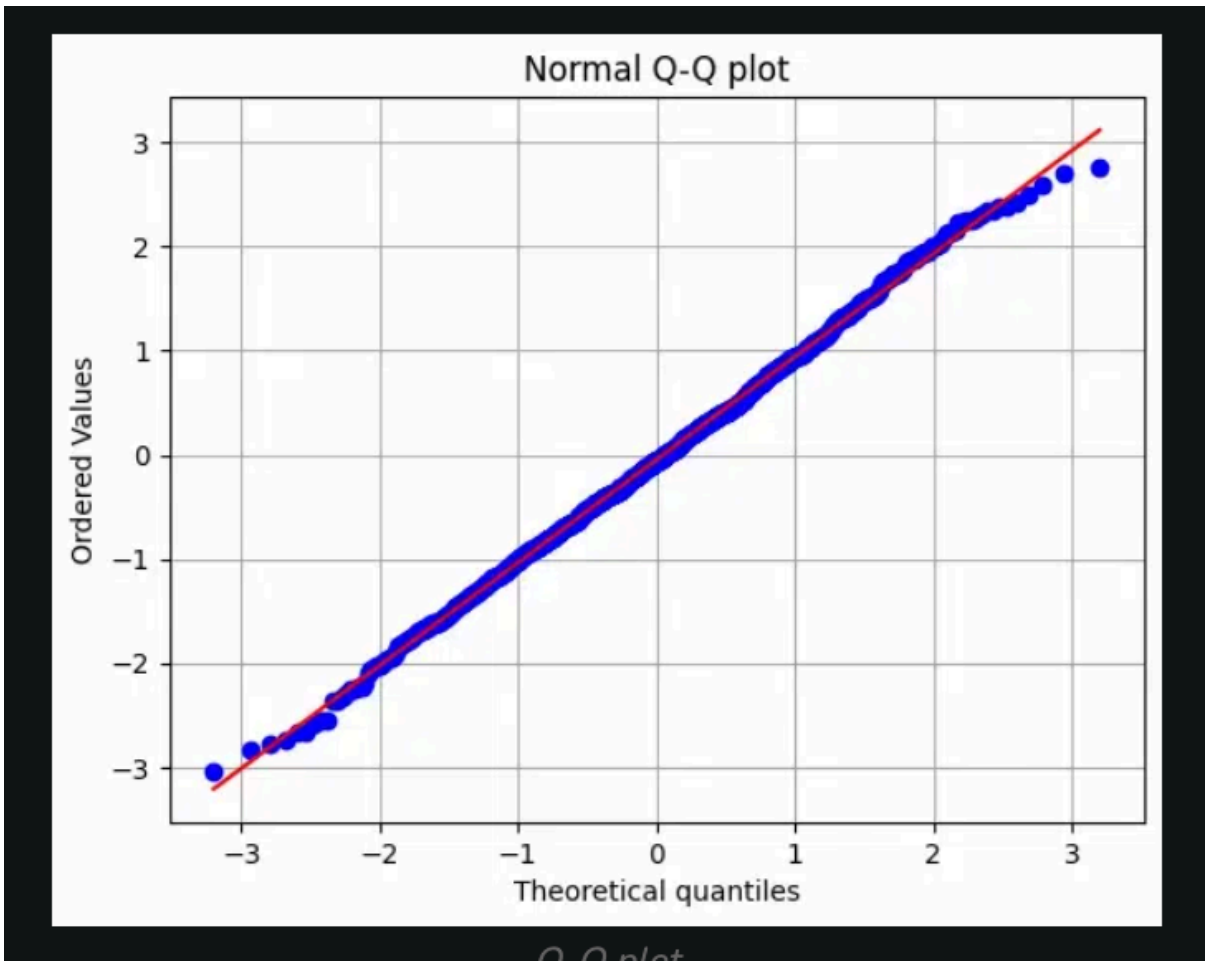
1. Remove redundant variables
2. combine variables
3. Regularization
4. dimensionality reduction

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)

<Your answer for Question 11 goes here>

A **Q-Q plot** is a graphical tool used to compare the distribution of a dataset to a theoretical distribution (normal distribution). It plots the quantiles of the data against the quantiles of a specified theoretical distribution to assess how well the data matches the distribution. If the points in the **Q-Q** plot approx fall along a straight line, it suggests the data is modeled well by the chosen theoretical distribution.

Example of a Q-Q plot:



The importance of a **Q-Q plot** in Linear Regression:

1. Checking normality of residuals
    a. Helps visually assess whether residuals are normally distributed and if they follow a straight line in a Q-Q plot.
2. Detecting Deviations from normality:
    a. Fat tails or heavy tails
    b. skewness

3. Identifying outliers and influential points
4. model fit assessment
5. validity of statistical tests

The **Q-Q** plot can provide more insights into the nature of the different types of analytical methods and understand how they are measured against each other.
