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)

There are 7 categorical variables present in the dataset: 'season', 'yr', 'mnth', 'holiday', 'weekday', 'workingday' and 'weathersit'.

1. **Season (season):** The bikes count rented was the most during Fall season followed by Summer season. Spring season recorded least number of bike rentals.
2. **Year (yr):** The demand for rental bikes has increased a lot from 2018 to 2019.
3. **Month (mnth):** July month was pretty good with high median of bikes count and September witnessed the maximum count of rental bikes. January witnessed low bikes count in both whole years.
4. **Holiday (holiday):** Bikes were rented lesser in number during holidays.
5. **Weekday (weekday):** The rented bikes count followed similar pattern for all 7 days in a week.
6. **Working Days (workingday):** The bikes were slightly on higher demand during business working days
7. **Weather Type (weathersit):** The bikes were rented the maximum at the times of Clear weather. The count was very low when it is Light Rain weather type.

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)

We create dummy variables for non-binary categorical variables to separate multivalued data into a binary data which is either 0 or 1. If a categorical variable has ‘n’ levels of values, we need to create only ‘n-1’ dummy variables to reduce the correlations among them.

We generally make use of the method pd.get_dummies() to generate the dummy variables and this function by default creates ‘n’ dummy variables for ‘n’ levels respectively. While creating dummy variables, it is recommended to use **drop_first=True** as it would help in eliminating the extra column of the dummy variable data generated. This **drop_first** flag deletes the first column of dummy table.

Example: Lets take Feature Selection as a categorical variable with 3 levels: ‘Manual’, ‘Automated’ and ‘Mixed’ for data = [‘Manual’, ‘Automated’, ‘Mixed’, ‘Automated’].

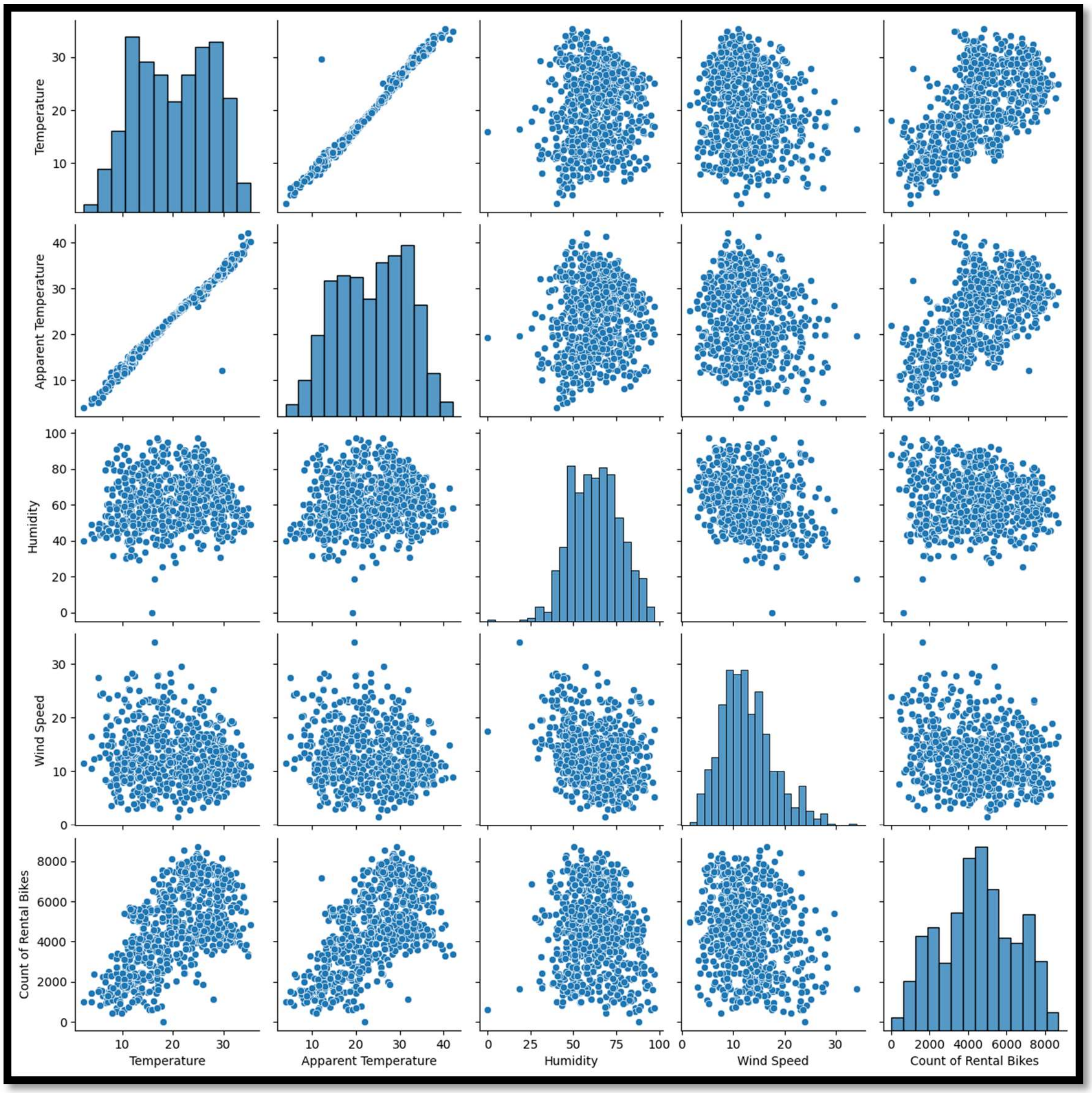
pd.get_dummies(data) generates a table like this:

Manual	Automated	Mixed
1	0	0
0	1	0
0	0	1
0	1	0

pd.get_dummies(data, drop_first=True) generates a table like this:

Automated	Mixed
0	0
1	0
0	1
1	0

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)



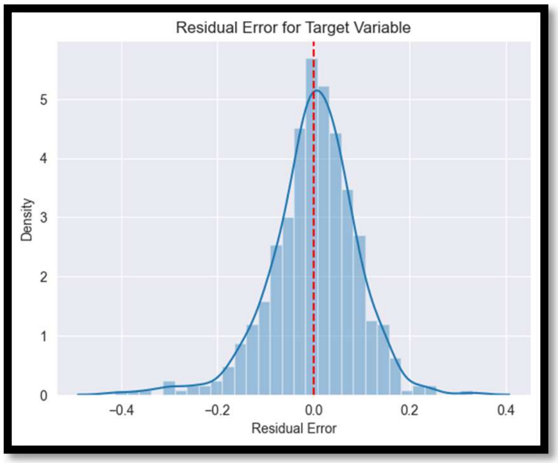
Out of the 5 numerical variables, Temperature (temp) and Apparent Temperature (atemp) have high correlation with the target variable, Count of Rental Bikes (cnt)

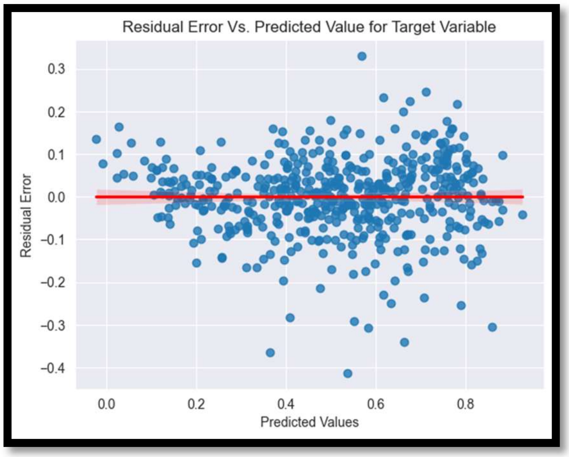
Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

After the training of linear regression model, prediction was performed on the target variable for training data. With respect to target variable, there are 2 values now: Actual value and Model Predicted value. The Residual error for target variable is calculated by the difference between its actual value and predicted value. And we have the mathematical expression derived for the linear regression model developed:

cnt = 0.09 + 0.52 * temp + 0.23 * yr + 0.14 * winterSeason + 0.11 * sep + 0.1 * summerSeason + 0.06 * sat + 0.05 * workingday + 0.05 * aug - 0.06 * holiday - 0.08 * mistWeather - 0.15 * windspeed - 0.29 * lightRainWeather

Following are visualizations of density distribution plot of Residual error and a distribution plot of Residual Error Vs. Predicted values of Target variable :





Linearity: The Target variable varies linearly with the dependent variables either positively or negatively as seen in the above mathematical expression of the model developed. The changes in predictor variables are associated with proportional changes in the target variable collectively.

Normality: Error terms are normally distributed with mean approximately at 0 as seen in the density distribution plot of Residual error.

Independence of Residuals: Total residual sum of error always lies at 0 level but the data points are scattered independently which indicates the observations are independent of each other

Homoscedasticity: Error terms follow a constant variance in nature as the scattered data points are approximately linear in nature. The residual error terms always vary between -0.2 to +0.2 throughout the predicted values of target variable from 0 to 1. Few points which are out of this zone, indicate outliers or influential points.

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)

Here is the mathematical expression for the linear regression model developed:

cnt = 0.09 + 0.52 * temp + 0.23 * yr + 0.14 * winterSeason + 0.11 * sep + 0.1 * summerSeason + 0.06 * sat + 0.05 * workingday + 0.05 * aug - 0.06 * holiday - 0.08 * mistWeather - 0.15 * windspeed - 0.29 * lightRainWeather

The top 3 features which significantly impact the count of rental bikes:
Temperature (temp) – Having coefficient of +0.52. For every unit increase in temperature, the bikes count is predicted to go up by 0.52.
Light Rain Weather type (lightRainWeather) – Category level of weathersit variable having coefficient of -0.29. For every unit change in weathersit towards lightRainWeather, the bikes count is predicted to come down by 0.29
Year (yr) – Having coefficient of +0.23. For every unit increase in year, the bikes count is predicted to go up by 0.23.

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 machine learning algorithm which is used to predict the value of a given target variable in terms of one or more predictor variables.
It has a generic expression in this format:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n$$

y is the target variable
$x_1, x_2, x_3, \dots, x_n$ are independent predictor variables
β_0 is the y intercept (predicted y value when all x values are 0
$\beta_1, \beta_2, \beta_3, \dots, \beta_n$ are respective coefficients of the predictor variables which represent the change in value of y for a unit change in x

Based on the number of predictor variables, there 2 types of linear regression:

1. Simple Linear Regression: When there is only 1 independent predictor variable impacting the target variable. $y = \beta_0 + \beta_1 x$
2. Multiple Linear Regression: When there are more than 1 independent predictor variables impacting the target variable. $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n$

As this is a predictive statistical model, this doesn't always predict the exact value of target variable. So, an error always exists between the actual value and predicted value of the target variable which is termed as Residual Error. Lower the Residual error, better is the predictive nature of the model.

There are 4 assumptions for Linear Regression:

1. Linearity: The relationship between the independent and dependent variable is linear.
2. Independence: The residual errors are independent of each other.
3. Homoscedasticity: The residual errors should have constant variance at all levels of the independent variables.
4. Normality: The residuals should be approximately normally distributed with residual mean at 0.

The cost function used in linear regression is the Mean Squared Error (MSE):

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

n is number of observations
y_i is actual value
$\hat{y}_i$ is predicted value

The square root of MSE gives Root Mean Squared Error (RMSE) which has the same units as target variable. This cost function should be minimized by using optimization techniques like Gradient Descent to achieve a best fitting regression model.

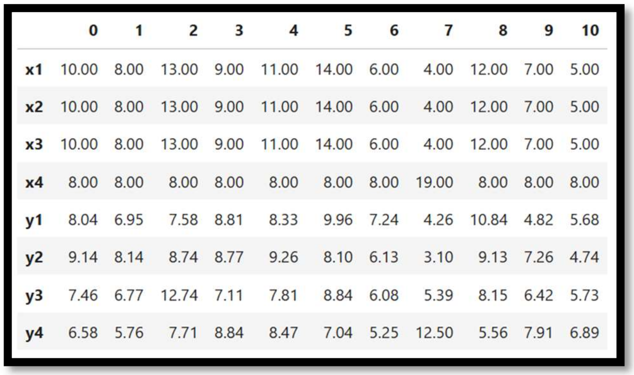
The model should be trained and developed in such a way that it does not memorize the whole training data leading to overfitting of model and not too simplified causing underfitting. These kinds of models fail to predict for new data(or test data) thereby deteriorating the primary objective of predictive analysis.

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>

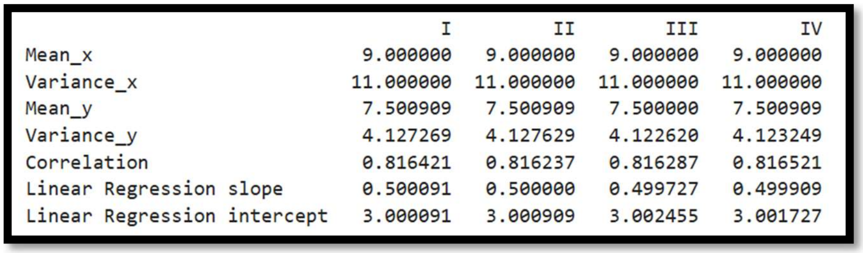
Anscombe's quartet is a set of four datasets that have identical summary statistics—such as mean, variance, and correlation coefficients—but display very different distributions and relationships when visualized. It was created to illustrate the importance of visualizing data before drawing conclusions from statistical analyses. Each dataset consists of paired values (x, y):

1. Dataset I: This follows a linear trend.
2. Dataset II: This also shows a linear trend, but with a slight deviation due to an outlier.
3. Dataset III: This dataset has a clear nonlinear relationship, resembling a curve.
4. Dataset IV: This dataset appears as a vertical line, indicating that all x-values are the same but y-values vary.



	0	1	2	3	4	5	6	7	8	9	10
x1	10.00	8.00	13.00	9.00	11.00	14.00	6.00	4.00	12.00	7.00	5.00
x2	10.00	8.00	13.00	9.00	11.00	14.00	6.00	4.00	12.00	7.00	5.00
x3	10.00	8.00	13.00	9.00	11.00	14.00	6.00	4.00	12.00	7.00	5.00
x4	8.00	8.00	8.00	8.00	8.00	8.00	8.00	19.00	8.00	8.00	8.00
y1	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68
y2	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74
y3	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73
y4	6.58	5.76	7.71	8.84	8.47	7.04	5.25	12.50	5.56	7.91	6.89

Summary Statistics:

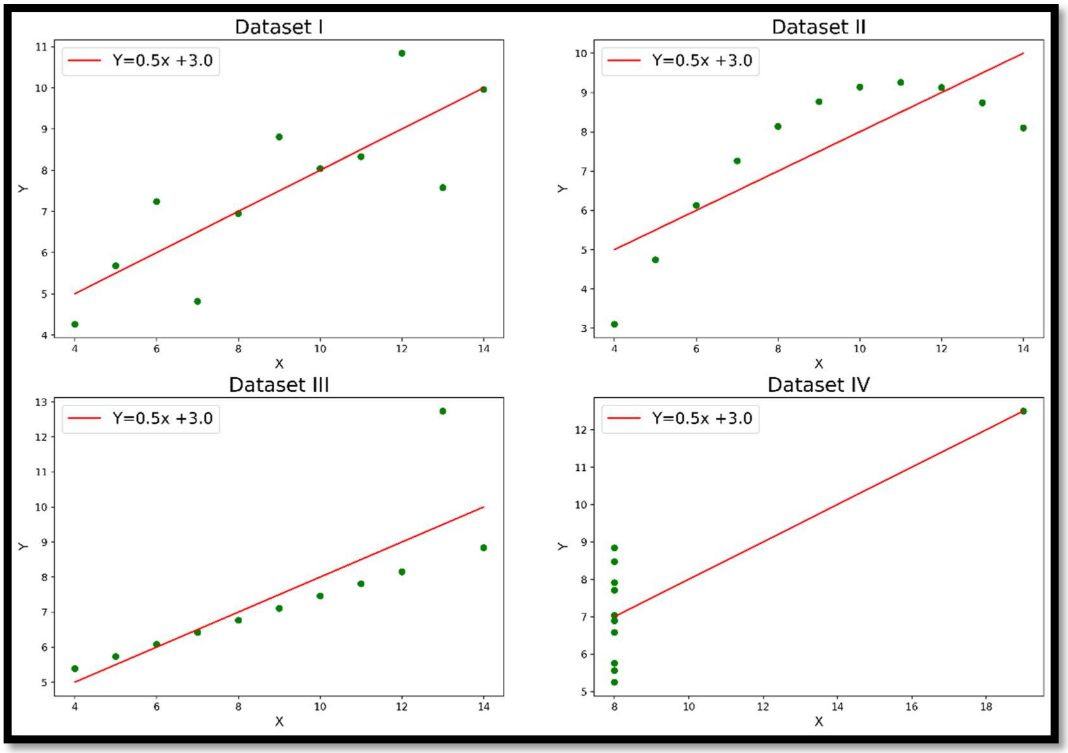


```
                                   I         II        III         IV
Mean_x                      9.000000   9.000000   9.000000   9.000000
Variance_x                 11.000000  11.000000  11.000000  11.000000
Mean_y                      7.500909   7.500909   7.500000   7.500909
Variance_y                  4.127269   4.127629   4.122620   4.123249
Correlation                 0.816421   0.816237   0.816287   0.816521
Linear Regression slope     0.500091   0.500000   0.499727   0.499909
Linear Regression intercept 3.000091   3.000909   3.002455   3.001727
```

Visualizing the Data:

When plotted, the stark differences between the datasets become evident:

- Dataset I shows a clear linear trend.
- Dataset II shows a linear trend but is influenced by an outlier (a point that deviates significantly from the trend).
- Dataset III highlights a quadratic relationship, making it inappropriate to use linear regression.
- Dataset IV shows no correlation despite having similar summary statistics.



Insights after the Analysis:

1. Importance of Visualization: Anscombe's quartet emphasizes that summary statistics alone can be misleading. Visualizing data reveals underlying patterns and relationships that numbers may obscure.
2. Modeling Considerations: The quartet illustrates the importance of selecting appropriate models for data analysis. Different datasets may require different analytical approaches based on their distributions.
3. Outliers: The influence of outliers can dramatically change the interpretation of data, which underscores the need to identify and understand them in analysis.

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 known as Pearson correlation coefficient, is a statistical measure that assesses the strength and direction of a linear relationship between two continuous variables. Its value ranges between -1 to +1, indicating both the magnitude and direction of the correlation.. The mathematical formula is given by:

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

R is Pearson correlation coefficient
x_i is value of x in a sample
$\bar{x}$ is mean of x
y_i is value of y in a sample
$\bar{y}$ is mean of y
n is total number of observations in sample

Significance of R limits:

- $R = 1$: Perfect positive linear correlation. As one variable increases, the other variable increases proportionally.
- $R = -1$: Perfect negative linear correlation. As one variable increases, the other decreases proportionally.
- $R = 0$: No Linear correlation. Changes in one variable do not predict changes in the other.

Derived Parameters from R:

1. The Square of R - (R^2) represents the coefficient of determination. It indicates the proportion of the variance in the dependent variable that can be explained by the independent variable in a linear regression model. Since it is a square of coefficient, it varies from 0 to 1.
2. The Adjusted (R^2) is a modified version of the coefficient of determination (R^2) that adjusts for the number of predictors in a regression model. Unlike R^2, which can only increase or stay the same when additional predictors are added, Adjusted R^2 can decrease if the new predictors do not improve the model sufficiently.

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>

Feature scaling is a crucial technique in data preprocessing that adjusts the range of independent variables or features in a dataset. This process is essential to ensure that different features contribute equally to the performance of machine learning algorithms, especially when those features have varying units or scales. Without feature scaling, algorithms may disproportionately prioritize features with larger values, which can lead to biased results.

When features are on different scales, algorithms that rely on distance calculations (like K-Nearest Neighbors or Support Vector Machines) may be misled by the magnitude of the features rather than their actual relationships. For instance, a feature with values ranging from 1 to 1000 will dominate the distance calculations compared to a feature that ranges from 0 to 1, which could lead to incorrect model predictions.

There are two primary methods of feature scaling: Normalization and Standardization.

1. Normalization:

   Normalization typically rescales the data to a range between 0 and 1. This is especially useful when the data distribution is not Gaussian (i.e., it does not follow a bell curve). The normalization process can be represented by the formula:

   $$\acute{x} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

   By normalizing, all features are treated equally, ensuring that the algorithm learns from the relative magnitudes rather than absolute values. This scaling is also called as Min-Max scaling and is most suitable technique for normalizing data in developing linear regression models.

2. Standardization:

   Standardization, on the other hand, transforms the data to have a mean of 0 and a standard deviation of 1. This is achieved using the formula:

   $$\acute{x} = \frac{x - \mu}{\sigma}$$

   where μ is the mean and σ is the standard deviation of the feature. This method is particularly beneficial when the data approximates a Gaussian distribution. However, it can also be applied to data with outliers since standardization does not restrict values to a specific range. Outliers will retain their relative positions in the standardized data.

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>

The Variance Inflation Factor (VIF) is a metric used to quantify how much the variance of an estimated regression coefficient increases due to multicollinearity among the independent variables in a linear regression model. A high VIF indicates a high degree of multicollinearity, which can affect the stability and interpretability of the regression coefficients. The formula for VIF for an independent variable (X_i) is given by:

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

A VIF value becomes infinite when an independent variable is perfectly correlated with one or more other independent variables in the model. This perfect correlation means that the variable can be expressed as a linear combination of the others, leading to redundancy. This phenomenon of having an infinite VIF is indeed related to Multicollinearity. It indicates that the variable does not provide any unique information beyond what is already accounted for by the other variables in the model.

Example, Consider a dataset with three variables: X_1, X_2, and X_3. If X_3 is a perfect linear combination of X_1 and X_2 (e.g., $X_3 = 2\,X_1 + 3\,X_2$), then:
When you calculate the VIF for X_3, the R-squared value (R_3^2) from regressing X_3 on X_1 and X_2 will equal 1. This leads to:

$$VIF(X_3) = \frac{1}{1 - 1} = \infty$$

The numerical value of the VIF indicates, in decimal form, the percentage increase in variance (i.e., the square of the standard error) for each coefficient due to multicollinearity. For instance, a VIF of 1.9 implies that the variance of a specific coefficient is 90% greater than it would be if there were no multicollinearity—meaning there is no correlation with other predictors.

Interpretation Guidelines for VIF:

- 1: No correlation as R_i^2 will be equal to 0 which indicates no correlation
- 1 to 5: Low correlation as R_i^2 will value between 0 to 0.8.
- 5 to 10: Low correlation as R_i^2 will value between 0.8 to 0.9
- Greater than 10: High correlation as R_i^2 will be more than 0.9

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)

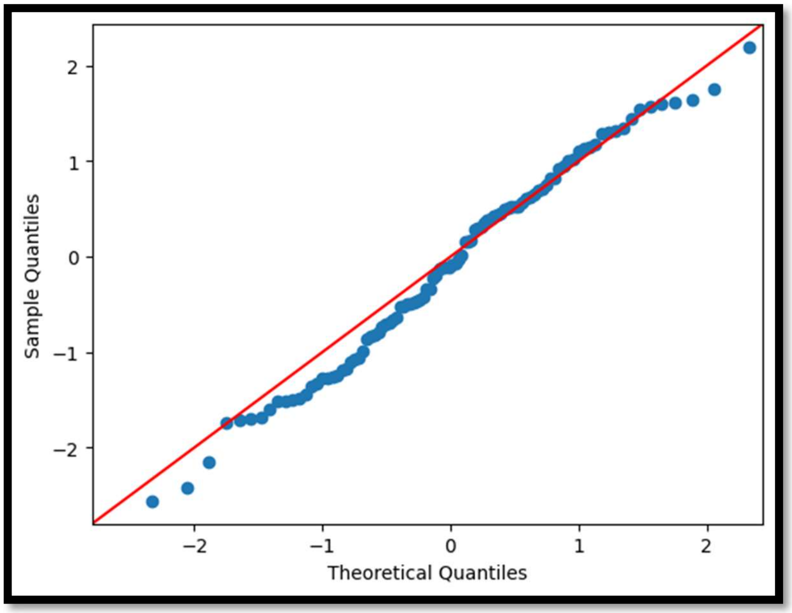
<Your answer for Question 11 goes here>

A Quantile-Quantile (Q-Q) plot is a graphical tool used to assess whether a dataset follows a specified theoretical distribution, typically the normal distribution. In a Q-Q plot, the quantiles of the sample data are plotted against the quantiles of the theoretical distribution. If the points on the plot fall approximately along a straight line, it suggests that the sample data follows the specified distribution closely.

Structure of a Q-Q Plot:

X-axis: Represents the quantiles of the theoretical distribution (e.g., a standard normal distribution).
Y-axis: Represents the quantiles of the sample data.
Reference Line: A 45-degree line (y = x) is often included to help visualize how closely the sample quantiles match the theoretical quantiles.

Sample plot from random generated data:



Interpretation of Q-Q Plots:

Points on the Line: If the points lie close to the diagonal reference line, the data are likely normally distributed.
S-shaped Pattern: If the plot has an S-shaped curve, it indicates that the data may have heavier tails than a normal distribution (indicating positive skewness) or lighter tails (indicating negative skewness).
Points Far from the Line: Points that are far from the line, especially at the ends, suggest departures from normality, indicating potential issues with the assumptions of linear regression.