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)

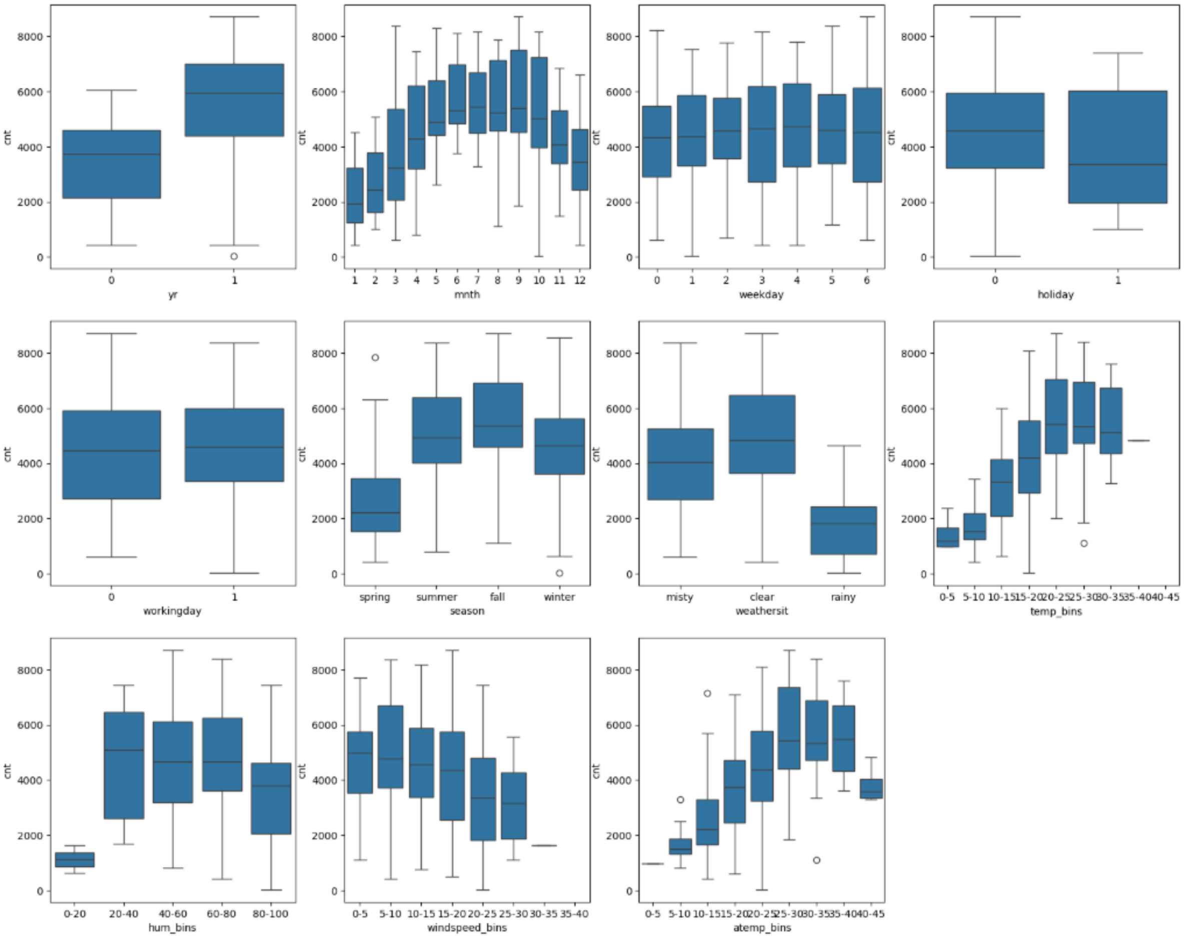
- Year variable seems to have a significant impact on dependent variable – can be associated with brand growth Year on Year.

- Demand increases across months from January to September and then demand seems to drop from October to December

- Less demand observed during holidays as median demand on holidays are lower than non-holidays.

- Higher demand observed during Fall and Summer seasons and least demand observed during spring season.

- Significantly lower demand observed during rainy weather situations.



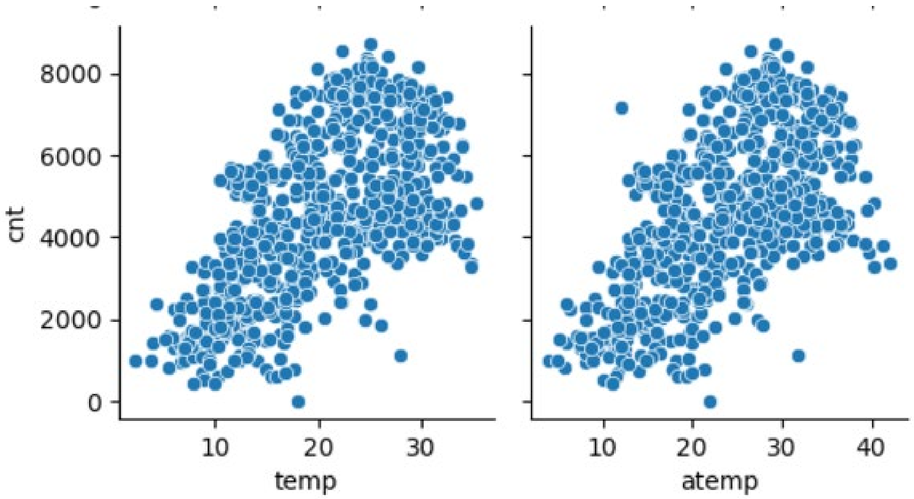
Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)

Using drop_first=True during dummy variable creation prevents the dummy variable trap, which occurs when all dummy variables for a categorical variable are included, leading to perfect multicollinearity.
When a categorical variable with k categories is represented by k dummy variables, one of them can be inferred from the others. This redundancy makes it impossible for models like linear regression to estimate unique coefficients for all variables.
By setting drop_first=True, one dummy variable is removed. This avoids redundancy, improves model stability, and ensures coefficients represent contrasts with the reference category. It also enhances interpretability, as the intercept reflects the outcome for the reference category.

In summary, drop_first=True eliminates multicollinearity, avoids the dummy variable trap, and simplifies model interpretation.

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)

Based on the pairplot, Temp and atemp variables seem to have positive correlation with the dependent variable.

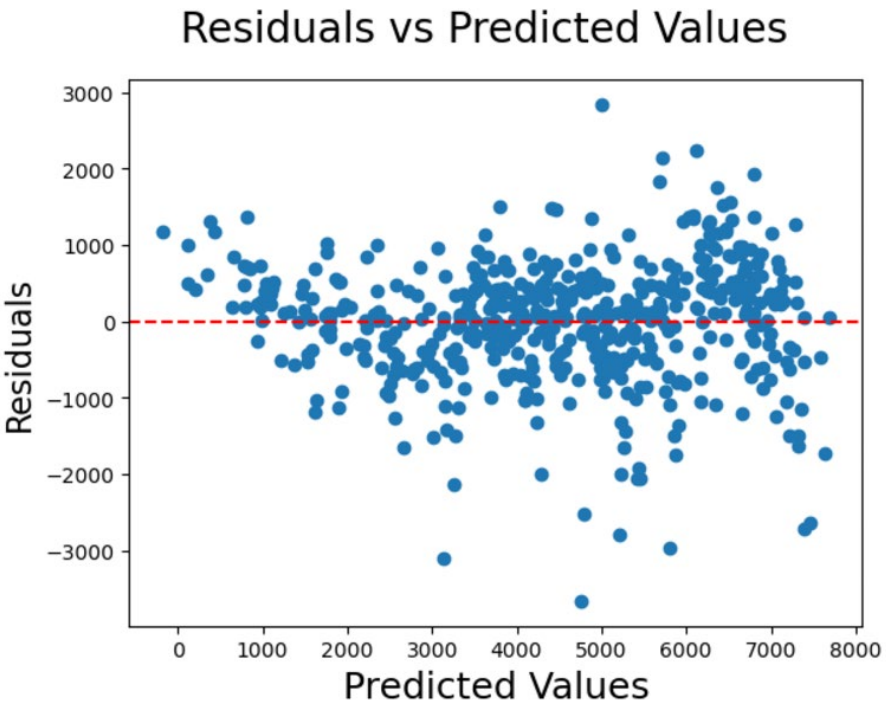


Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

Major Assumptions of Linear Regression:

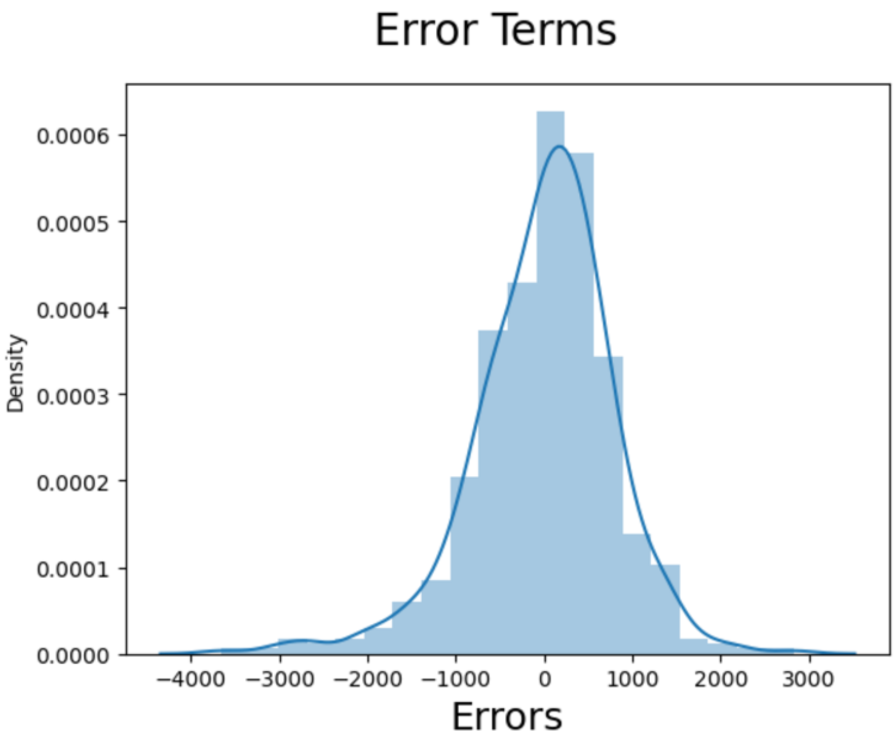
1. Linearity - The relationship between independent variables and the dependent variable is linear.
2. Independence of Errors - Residuals (errors) are independent of each other.
3. Homoscedasticity - The variance of residuals remains constant across all levels of predicted values.
4. Normality of Residuals - Residuals are normally distributed.
5. Multicollinearity - Independent variables are not highly correlated with each other.

- To validate linearity assumption, we can visualize the plot of residuals versus the predicted values. The residuals are randomly spread around 0, indicating there is no systemic pattern which validates the linearity assumption. The spread remains consistent across residuals which validates the homoscedasticity assumption as well.



- To validate independence of errors assumption, Durbin-Watson test should be performed. A value of close to 2 indicates no significant correlation. The final model had a Durbin Watson score of 2.050 which validates the assumption of independence of errors.

- Histogram of residuals are plotted to visualize the distribution. The visualization clearly shows the residuals are normally distributed as it is in the shape of the bell curve.



- VIF for all the variables in the model is less than 5, which validates the multicollinearity assumption.

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)

Based on our final model, the top three predictor variables that significantly influence bike bookings are:

Temperature (temp)
A coefficient value of 3888 indicates that a one-unit increase in the scaled temperature variable results in an increase of 3888 bike hires. This highlights the strong positive impact of temperature on demand.

Weather Situation 3 (weathersit_rainy)
A coefficient value of -2494 indicates that bike bookings decrease by 2494 units on heavily rainy days. This demonstrates the adverse effect of bad weather on demand.

Year (yr)
A coefficient value of 2039 indicates that a one-unit increase in the year variable leads to an increase of 2039 bike hires. This can be attributed to year-on-year brand growth and increased popularity.

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 machine learning algorithm used to find the relationship between one or more independent variables and a continuous dependent variable. It works by fitting a straight line to the data, represented by the equation $y=\beta0+\beta1x1+\beta2x2+\cdots+\beta pxp+\epsilon y$.

Here, $\beta0$ is the intercept, $\beta1,\beta2,\ldots,\beta p$ are the coefficients, and ϵ is the error term. The goal is to estimate the coefficients so the difference between the actual and predicted values is as small as possible. This is done using the least squares method, which minimizes the sum of the squared differences (errors) between actual and predicted values.

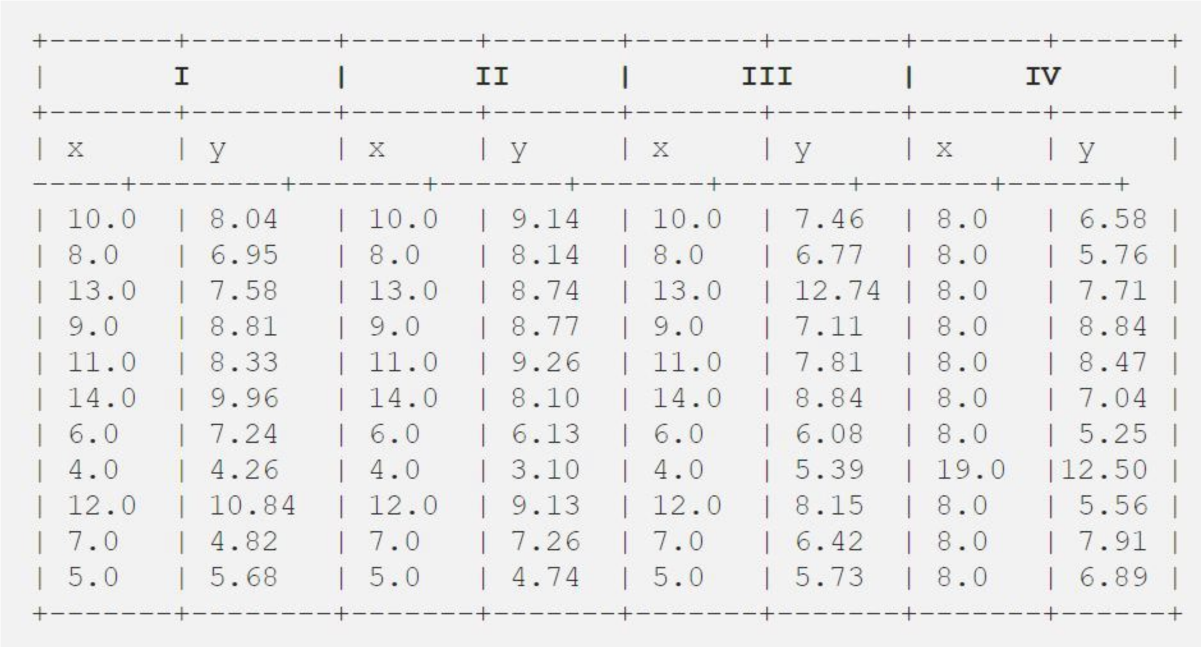
Once the coefficients are calculated, the model can predict new values of y for given X. Linear regression assumes that the relationship between X and y is linear, the errors are independent and normally distributed, the variance of the errors is constant, and the predictors are not highly correlated. It is easy to use and interpret but may not work well if these assumptions are not met or if the data has outliers or non-linear patterns. Despite its simplicity, linear regression is widely used in predictive modeling and data 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 group of four datasets that share nearly identical descriptive statistical characteristics, such as mean, variance, correlation, and linear regression equations. However, when plotted on a graph, these datasets display strikingly different patterns. This highlights the limitations of relying purely on numerical metrics and emphasizes the critical role of data visualization in analysis.

The four datasets of Anscombe's quartet:

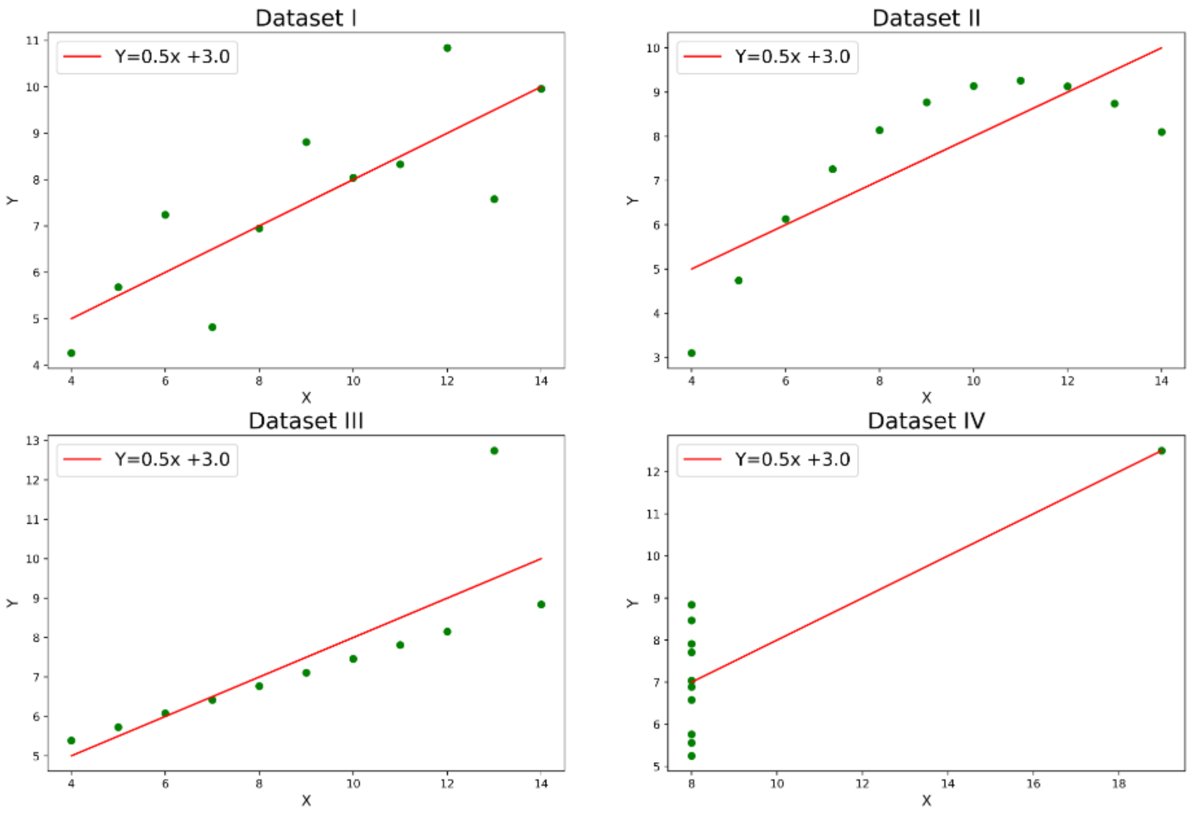


```
+-------+--------+-------+-------+-------+-------+-------+------+
|       I        |      II       |      III      |      IV      |
+-------+--------+-------+-------+-------+-------+-------+------+
| x     | y      | x     | y     | x     | y     | x     | y    |
-----+--------+-------+-------+-------+-------+-------+------+
| 10.0  | 8.04   | 10.0  | 9.14  | 10.0  | 7.46  | 8.0   | 6.58 |
| 8.0   | 6.95   | 8.0   | 8.14  | 8.0   | 6.77  | 8.0   | 5.76 |
| 13.0  | 7.58   | 13.0  | 8.74  | 13.0  | 12.74 | 8.0   | 7.71 |
| 9.0   | 8.81   | 9.0   | 8.77  | 9.0   | 7.11  | 8.0   | 8.84 |
| 11.0  | 8.33   | 11.0  | 9.26  | 11.0  | 7.81  | 8.0   | 8.47 |
| 14.0  | 9.96   | 14.0  | 8.10  | 14.0  | 8.84  | 8.0   | 7.04 |
| 6.0   | 7.24   | 6.0   | 6.13  | 6.0   | 6.08  | 8.0   | 5.25 |
| 4.0   | 4.26   | 4.0   | 3.10  | 4.0   | 5.39  | 19.0  |12.50 |
| 12.0  | 10.84  | 12.0  | 9.13  | 12.0  | 8.15  | 8.0   | 5.56 |
| 7.0   | 4.82   | 7.0   | 7.26  | 7.0   | 6.42  | 8.0   | 7.91 |
| 5.0   | 5.68   | 5.0   | 4.74  | 5.0   | 5.73  | 8.0   | 6.89 |
+-------+--------+-------+-------+-------+-------+-------+------+
```

The summary stats of these four datasets:

	I	II	III	IV
Mean_x	9.00	9.00	9.00	9.00
Variance_x	11.00	11.00	11.00	11.00
Mean_y	7.50	7.50	7.50	7.50
Variance_y	4.13	4.13	4.12	4.12
Correlation	0.82	0.82	0.82	0.82
Linear_Regression_Slope	0.50	0.50	0.50	0.50
Linear_Regression_Intercept	3.00	3.00	3.00	3.00

Each of the four datasets has the same key statistical summaries, creating the illusion that they are similar. However, visualizing these datasets through scatterplots unveils their distinct characteristics.



- Dataset 1 (Top Left): The scatterplot reveals a clear linear relationship between x and y, fitting well with the regression line.
- Dataset 2 (Top Right): The data exhibits a non-linear relationship, which is not captured accurately by the regression line.
- Dataset 3 (Bottom Left): Most points follow a horizontal pattern, but a single outlier significantly influences the regression results.
- Dataset 4 (Bottom Right): A single high-leverage point heavily affects the correlation, creating a misleadingly high value despite the rest of the data having no meaningful trend.

Key Takeaway:
Although the statistical summaries for all four datasets are identical, their scatterplots reveal significant differences. This demonstrates the necessity of complementing numerical analysis with visual tools to uncover underlying data patterns and ensure accurate interpretation.

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>

The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

- Between 0 and -1 indicates negative correlation, When one variable changes, the other variable changes in the opposite direction.
- 0 indicates no correlation. There is no relationship between the variables.

- Between 0 and +1 indicates positive correlation, When one variable changes, the other variable changes in the same direction.

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 adjusting the values of numerical features to a common range or scale, without distorting differences in the data. It ensures all variables contribute equally to the model. Scaling improves model performance and speeds up convergence during training.

Normalized vs. Standardized Scaling

- Normalization rescales data to a fixed range, usually [0, 1] by using formula x-min(x) / max(x) – min(x).
- Standardization rescales data to follow standard normal distribution have a mean of 0 and a standard deviation of 1, by using formula x-mean / SD

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)

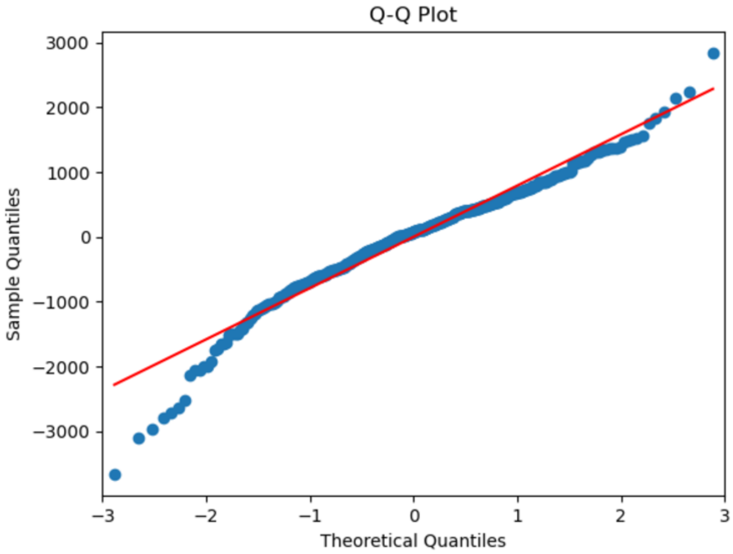
<Your answer for Question 10 goes here>

VIF becomes infinite when there is perfect multicollinearity among the independent variables in a dataset. This means one or more independent variables can be expressed as an exact linear combination of others.
The formula for VIF is 1 / (1-R^2). When there is a perfect correlation, R^2 becomes 1, hence 1/(1-1) = infinite.

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 (Quantile-Quantile) plot is a tool used to visually assess whether the residuals of a regression model follow a normal distribution. It does this by plotting the quantiles of the residuals against the expected quantiles of a normal distribution. Points falling close to a straight diagonal line indicate that the residuals are normal. Significant deviations from the line suggest potential issues like skewness or outliers. Ensuring residuals are normally distributed helps validate the model's assumptions, making statistical tests and predictions more reliable.


Q-Q Plot
Sample Quantiles
Theoretical Quantiles