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)

From the regression results, we can infer the following effects of the categorical variables on the dependent variable, `cnt` (which likely represents bike rentals or similar counts):

1. Seasonal Effects:
    - Spring: The coefficient for `season_spring` is negative (-0.0551), indicating that bike rentals are slightly lower in spring compared to other seasons, holding other factors constant. This effect is statistically significant at the 1% level (p-value = 0.009).
    - Summer: The coefficient for `season_summer` is positive (0.0610), suggesting that bike rentals are higher in summer, with a statistically significant positive effect (p-value = 0.000).
    - Winter: The coefficient for `season_winter` is also positive (0.0959), showing an increase in bike rentals during winter, with statistical significance (p-value = 0.000).
    - Inference: Summer and winter have a positive effect on bike rentals, while spring shows a slight decrease, likely due to weather conditions and outdoor activity preferences.
2. Month (September):
    - The coefficient for `mnth_Sep` is positive (0.0909), suggesting that bike rentals tend to be higher in September compared to other months, with strong statistical significance (p-value = 0.000).
    - Inference: September has a positive effect on bike rentals, possibly due to favorable weather conditions and an increase in outdoor activities post-summer.
3. Weather Situation:
    - Cloudy: The coefficient for `weathersit_Cloudy` is negative (-0.0801), showing that bike rentals decrease on cloudy days, with statistical significance (p-value = 0.000).
    - Rainy: The coefficient for `weathersit_Rainy` is much more negative (-0.2860), indicating a significant reduction in bike rentals on rainy days, also statistically significant (p-value = 0.000).
    - Inference: Both cloudy and rainy weather significantly decrease bike rentals, with rainy weather having a larger negative impact.

Conclusion: Seasonal changes, month of the year, and weather conditions (especially rain) all have significant effects on the dependent variable, with seasons like summer and winter promoting higher bike rentals, while cloudy and rainy conditions decrease the count significantly.

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` when creating dummy variables is important because it helps avoid the "dummy variable trap" and ensures the model is properly identified. Following are the reasons:

1. Avoiding Multicollinearity:
   When we create dummy variables for a categorical feature with more than two categories (e.g., "Season" with "Spring", "Summer", "Fall", "Winter"), you would normally get one dummy variable for each category. For example:
   - `Season_Spring` = 1 if Spring, 0 otherwise
   - `Season_Summer` = 1 if Summer, 0 otherwise
   - `Season_Fall` = 1 if Fall, 0 otherwise
   - `Season_Winter` = 1 if Winter, 0 otherwise
2. If we include all these dummy variables in the model, it will create perfect multicollinearity, meaning that one variable can be perfectly predicted by the others. For example, if `Season_Spring` = 0, `Season_Summer` = 0, `Season_Fall` = 0, then it must be that `Season_Winter` = 1. This redundancy leads to multicollinearity, which can distort the model's coefficients and inflate standard errors.
3. Model Identification:
   To avoid this problem, you typically drop one of the dummy variables (usually the first category or a reference category) when fitting the model. This "reference" category is the baseline, and the coefficients of the other categories are interpreted in relation to this reference group.
   For example, if we drop `Season_Spring`, the remaining variables (`Season_Summer`, `Season_Fall`, `Season_Winter`) will tell us how much more (or less) bike rentals are in those seasons relative to Spring. By dropping the first category, you prevent the variables from being linearly dependent, and the model remains identifiable.
4. Interpretability:
   Dropping the first category makes the model coefficients easier to interpret. Instead of comparing all categories to each other, you can focus on how each category differs from the baseline category.

Example:

If `Season_Spring` is dropped:

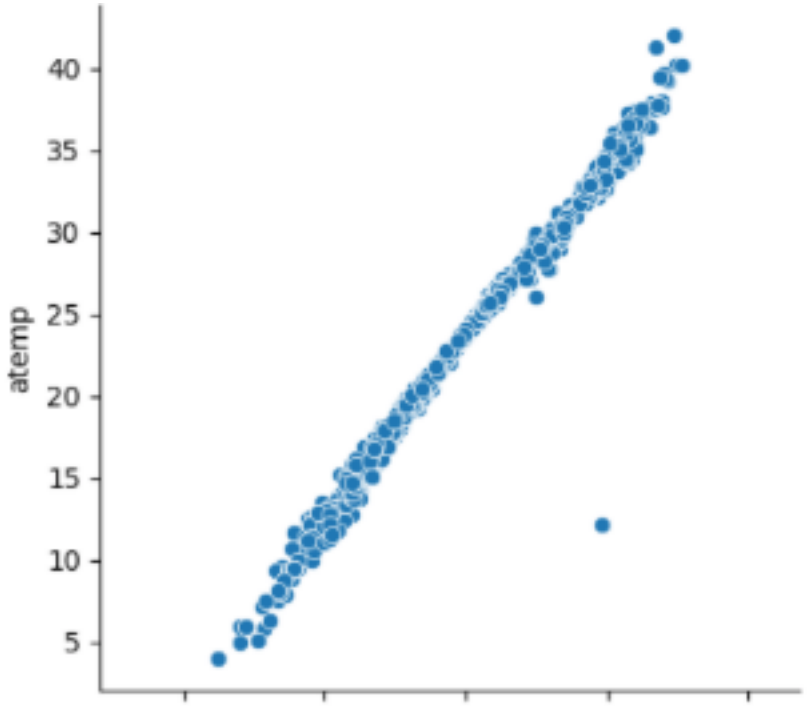
- `Season_Summer` might have a coefficient of 0.0610, meaning bike rentals in summer are 0.0610 higher than in spring (the baseline).
- `Season_Winter` might have a coefficient of 0.0959, meaning bike rentals in winter are 0.0959 higher than in spring.

If you didn't drop a category (i.e., included all the dummy variables), this would lead to multicollinearity and make the model's coefficients unreliable and harder to interpret.

In summary:
Using `drop_first=True` ensures the model avoids multicollinearity, makes it identifiable, and improves the interpretability of the categorical variable coefficients by using one category as a baseline.

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)

Among all the numerical variables 'temp' and 'atemp' has highest correlation with the target variable and they are also highly correlated with each other thus both can be assumed too similar.

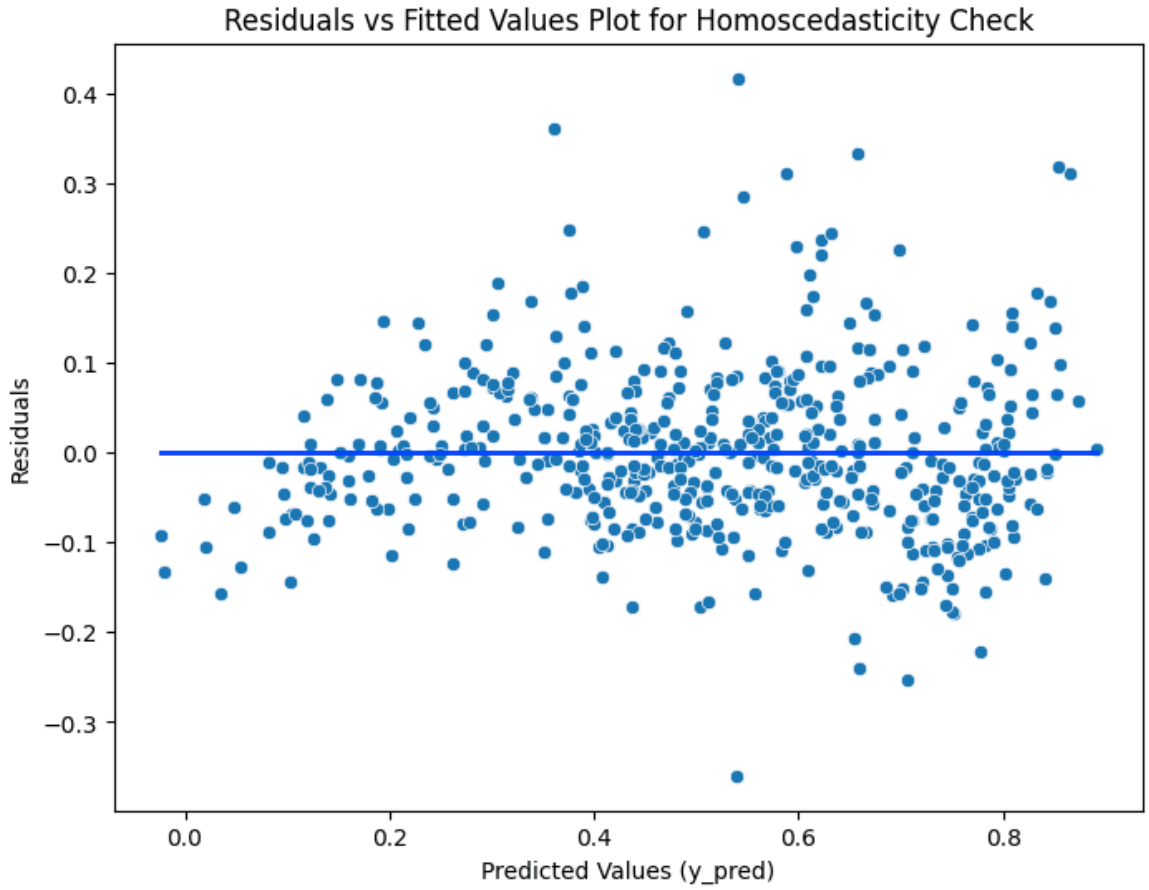


Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

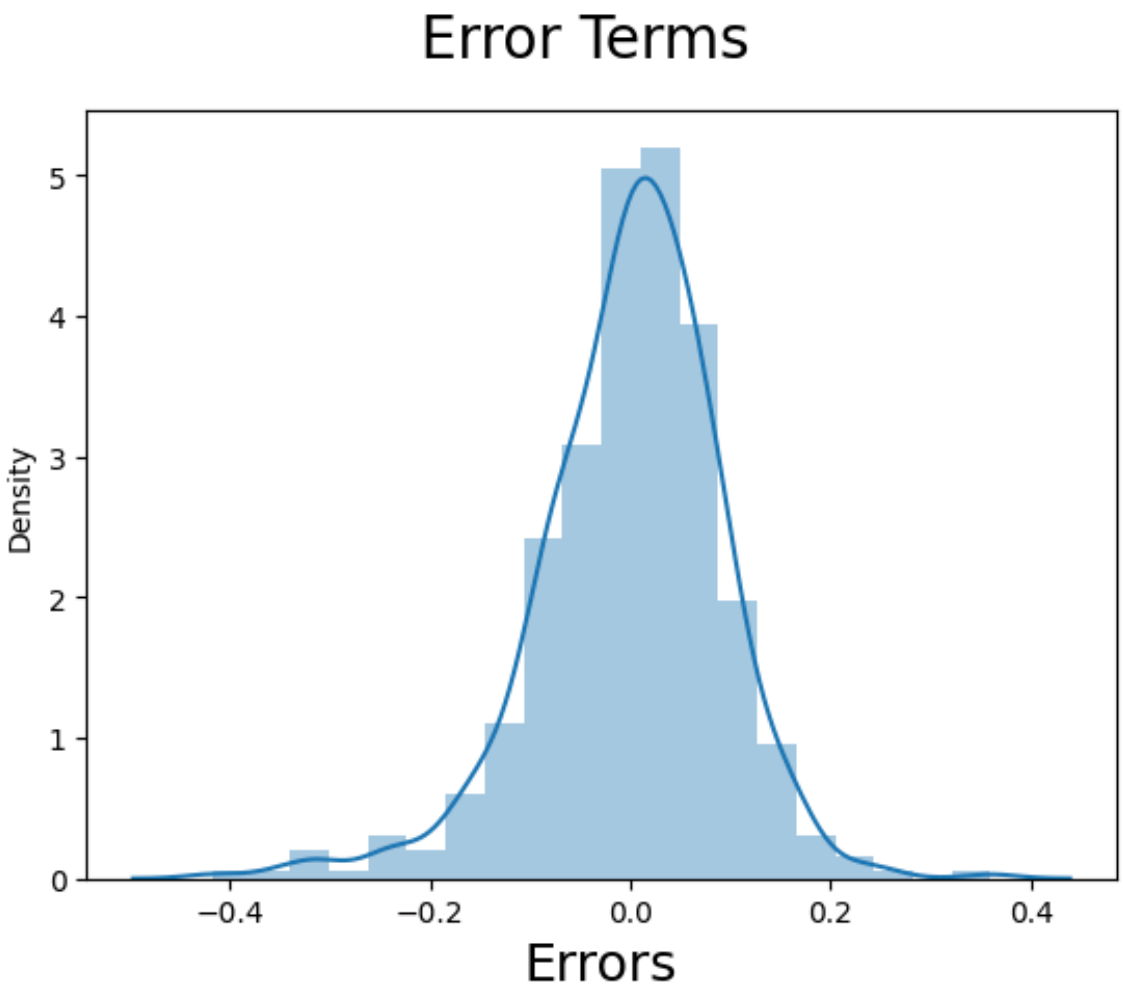
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

The assumptions of Linear Regression are as follows:
• Linear Relationship
• Homoscedasticity
• Absence of Multicollinearity
• Independence of residuals (absence of auto-correlation)
• Residuals are normally distributed

Homoscedasticity was tested by plotting residual vs predicted values and it shows no pattern in scatterplot thus verifying Homoscedasticity



The distribution of residual was checked using histogram which is normally distributed.



Multicollinearity was checked via heatmap and VIF where no column had high correlation or VIF after pruning.

Independence of residual was verified by Durbin-Watson statistic where value of final model is 2.076 which is close to 2 which indicates non-autocorrelation.

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 the final model, the top 3 features contributing significantly to explaining the demand for shared bikes are:

1. Temperature (`temp`):
   The coefficient for `temp` is 0.4782, which is the largest in magnitude among the variables, indicating that temperature has a strong positive effect on bike rentals. As temperature increases, bike rentals tend to rise significantly, which makes sense as warmer weather typically encourages outdoor activities like cycling.

2. Year (`yr`):
   The coefficient for `yr` is 0.2341, showing a positive trend over time. This suggests that bike rentals have increased as time has progressed, likely due to growing bike-sharing adoption or improved infrastructure over the years.

3. Rainy Weather (`weathersit_Rainy`):
   The coefficient for `weathersit_Rainy` is -0.2860, and it has one of the largest negative impacts on bike rentals. Rainy weather significantly reduces the demand for shared bikes, which is expected since people tend to avoid biking in poor weather conditions.

These features—temperature, year, and rainy weather—have the most significant and substantial effects on bike rental demand.

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)

Linear regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. In its simplest form, simple linear regression, the model assumes a linear relationship between the dependent variable y and a single independent variable x, expressed by the equation:

$$y = \beta 0 + \beta 1x + c$$

Where:

- y is the predicted value.
- $\beta 0$ is the intercept (constant).

- β1 is the coefficient (slope) that represents the effect of the independent variable x
- c is the error term or noise

In multiple linear regression, there are multiple independent variables, and the equation becomes:

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + \cdots + \beta n\, xn + c$$

Where:

- y is the predicted value.
- β0 is the intercept (constant).
- β1 is the coefficient (slope) that represents the effect of the independent variable x1
- β1 is the coefficient (slope) that represents the effect of the independent variable x2
- βn is the coefficient (slope) that represents the effect of the independent variable xn
- c is the error term or noise

The goal of linear regression is to find the values of the coefficients β0 ,β1 ,…,βn that minimize the sum of squared errors (SSE), which represents the difference between the actual values and predicted values.

The model is trained using methods like Ordinary Least Squares (OLS) or Gradient Descent to find the optimal coefficients. Once trained, the model can be used to predict new values of y based on new inputs x.

Key evaluation metrics for linear regression include:

- R-squared (R^2): A measure of how well the model explains the variability of the data.
- Mean Squared Error (MSE): A metric for assessing the average squared difference between actual and predicted values.

Linear regression is widely used due to its simplicity, interpretability, and efficiency in cases where there is a linear relationship between the variables

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)

Anscombe's quartet is a set of four datasets that have identical simple descriptive statistics (mean, variance, correlation, and regression line) but appear very different when graphed. Created by Francis Anscombe in 1973, the quartet is used to demonstrate the importance of visualizing data in addition to analyzing numerical summaries.

Each dataset in the quartet contains 11 data points and consists of two variables, x and y. The key features of all four datasets are:

- Identical Mean: The mean of x is 9, and the mean of y is approximately 7.5 for all datasets.
- Identical Variance: Both x and y have the same variance in all datasets.
- Identical Correlation: The correlation between x and y is 0.82 for all datasets.
- Identical Regression Line: The linear regression line (slope ~0.5, intercept ~3) fits all four datasets similarly.

However, the datasets differ greatly when plotted:

1. Dataset I shows a linear relationship with no outliers.
2. Dataset II is linear but contains one outlier.
3. Dataset III exhibits a non-linear relationship.
4. Dataset IV shows a vertical pattern with an outlier.

Anscombe's Quartet highlights that summary statistics can be misleading, and it's crucial to visualize data to understand its true structure.

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)

Pearson's r (also known as the Pearson correlation coefficient) is a statistic that measures the strength and direction of the linear relationship between two continuous variables. It tells us whether and how strongly two variables are related to each other.

The key thing to remember about Pearson's r is that it specifically measures linear relationships, meaning it detects if the variables move together in a straight-line pattern (either upwards or downwards) as one variable changes.

Range of Pearson's r

The value of r can range from -1 to 1, and here's what different values mean:

- r=1: Perfect positive linear relationship. As one variable increases, the other increases in a perfectly straight line. For example, if the number of hours studied increases, the score on a test might increase in a perfectly predictable manner.

- r=−1: Perfect negative linear relationship. As one variable increases, the other decreases in a perfectly straight line. For instance, if the temperature decreases, the sale of hot chocolate might increase in a perfectly predictable inverse relationship.

- r=0: No linear relationship. The variables don't change in any discernible straight-line pattern. For example, the number of hours someone sleeps may have no linear relationship with their favorite color.

- Between 0 and 1 (positive correlation): A positive linear relationship. As one variable increases, the other tends to increase as well, but it's not a perfect linear relationship. For example, the more time you spend exercising, the higher your fitness level might be, but not in a perfectly straight line.

- Between 0 and -1 (negative correlation): A negative linear relationship. As one variable increases, the other tends to decrease, but it's not a perfect negative linear relationship. For example, the more junk food someone eats, the lower their energy might be, but not perfectly in a straight line.

Interpreting Pearson's r in Context

1. Strength of the Relationship: The closer r is to 1 or -1, the stronger the linear relationship. So, if r is 0.9, you have a strong positive relationship, while r=−0.8 indicates a strong negative relationship.

2. Direction of the Relationship: If r is positive, the relationship is direct: as one variable increases, the other increases. If r is negative, the relationship is inverse: as one variable increases, the other decreases.

Example: Real-World Applications

- Height and Weight: There's usually a positive correlation between height and weight. As height increases, weight tends to increase as well (though not perfectly). The Pearson r might be something like 0.8, indicating a fairly strong relationship, but not a perfect one.

- Hours Studied and Test Scores: Generally, there's a positive correlation here too. The more hours you study, the higher your score might be, but again, it's not a perfect correlation. So, the Pearson r could be around 0.7 or 0.8, showing a moderate positive relationship.

- Ice Cream Sales and Temperature: There's a positive correlation between temperature and ice cream sales. As temperature rises, ice cream sales also rise, but not perfectly —other factors could be at play.

- Exercise and Body Fat: Often, there's a negative correlation between exercise and body fat percentage. The more you exercise, the lower your body fat, but other factors like diet or genetics can also influence the outcome.

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)

Scaling refers to the process of transforming data so that different features have comparable magnitudes, making them suitable for machine learning algorithms that are sensitive to the scale of the data. It adjusts the range of the data or modifies the distribution of the features to ensure uniformity.

Scaling is performed for several reasons:

1. Ensure Equal Contribution: It prevents features with larger ranges from dominating the model's behavior.
2. Improve Model Performance: Algorithms like k-NN, SVM, and gradient descent-based methods rely on distances and will perform better when the features are on similar scales.
3. Faster Convergence: In optimization algorithms (e.g., linear regression or neural networks), scaling can speed up convergence by ensuring that the gradient descent steps are consistent across all features.

Difference Between Normalized Scaling and Standardized Scaling

- Normalized Scaling (Min-Max Scaling): Rescales the data to a fixed range, usually [0, 1] or [-1, 1]. It adjusts the data based on the minimum and maximum values of each feature.

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

- Standardized Scaling (Z-Score Scaling): Transforms the data so that it has a mean of 0 and a standard deviation of 1. This scaling method centers the data around the mean and scales it based on its variance.

$$X_{\text{std}} = \frac{X - \mu}{\sigma}$$

Key Difference:

- Normalization is sensitive to outliers, as it uses the min and max values of the data.
- Standardization is not sensitive to outliers and is better when the data follows a Gaussian distribution.

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)

Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in a regression model. It quantifies how much the variance of a regression coefficient is inflated due to collinearity with other predictors.

A VIF becomes infinite when there is perfect multicollinearity among the predictors, meaning one predictor is a perfect linear function of another predictor(s). This occurs when:

1. Perfect Linear Relationship: One of the independent variables in the model can be perfectly predicted by the other independent variables, leading to a deterministic relationship (e.g.,X1 =2X2).

2. Singular Matrix: In the calculation of VIF, the predictor variables are used to create a covariance matrix. If the predictors are perfectly collinear, the covariance matrix becomes singular (non-invertible), causing a mathematical breakdown, which results in infinite VIF.

3. Infinite VIF indicates that the model cannot separate the effects of predictors accurately, and such variables may need to be removed or combined.

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)

A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare the distribution of a dataset with a theoretical distribution, usually the normal distribution. It plots the quantiles of the data against the quantiles of the reference distribution. If the data follows the reference distribution, the points on the plot will lie along a straight line.

Use and Importance of a Q-Q Plot in Linear Regression

In linear regression, a Q-Q plot is primarily used to check the normality of residuals:

1. Check Normality of Residuals: One of the key assumptions in linear regression is that the residuals (differences between observed and predicted values) should be normally distributed. A Q-Q plot helps visually assess whether this assumption holds. If the points in the plot lie along a straight line, it indicates that the residuals are normally distributed.

2. Identify Problems with Model: If the Q-Q plot shows significant deviations from a straight line (e.g., heavy tails or curvature), it suggests that the residuals are not normally distributed, which may violate regression assumptions and affect the reliability of statistical tests and confidence intervals.

Q-Q plot helps validate the assumption of normally distributed residuals, which is crucial for the validity of hypothesis tests and confidence intervals in linear regression.
