Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer:

To analyse the effect of categorical variables on the dependent variable, we can refer to the regression results shown in the assignment. Here is a detailed analysis:

Regression Results Summary

1. **yr (Year)**
   - **Coefficient:** 1009.4777
   - **P-value:** 0.000
   - **Inference:** The year variable is highly significant (P-value < 0.05). The positive coefficient indicates that bike rentals have increased in the second year compared to the first.
2. **holiday**
   - **Coefficient:** -74.4549
   - **P-value:** 0.058
   - **Inference:** The holiday variable is not significant at the 0.05 level but is close (P-value slightly above 0.05). The negative coefficient suggests that bike rentals decrease on holidays.
3. **workingday**
   - **Coefficient:** 69.3053
   - **P-value:** 0.016
   - **Inference:** The workingday variable is significant (P-value < 0.05). The positive coefficient indicates that bike rentals are higher on working days compared to non-working days.
4. **season_spring**
   - **Coefficient:** -672.9164
   - **P-value:** 0.000
   - **Inference:** The season_spring variable is highly significant (P-value < 0.05). The negative coefficient indicates that bike rentals are lower in spring compared to the base season (which is not listed but usually is summer or winter).
5. **weathersit_Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds**
   - **Coefficient:** -378.6048
   - **P-value:** 0.000
   - **Inference:** This weather situation variable is highly significant (P-value < 0.05). The negative coefficient indicates that adverse weather conditions significantly reduce bike rentals.
6. **weathersit_Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist**
   - **Coefficient:** -299.7704
   - **P-value:** 0.000
   - **Inference:** Like the previous weather variable, this one is also highly significant with a negative impact on bike rentals.
7. **weekday variables (weekday_1 to weekday_5)**

- **Coefficients and P-values:**
  - **weekday_1:** -24.4769 (P-value: 0.467)
  - **weekday_2:** -39.4155 (P-value: 0.232)
  - **weekday_3:** 28.2838 (P-value: 0.498)
  - **weekday_4:** 45.8732 (P-value: 0.163)
  - **weekday_5:** 48.8756 (P-value: 0.144)
- **Inference:** None of the weekday variables are significant (P-values > 0.05), suggesting that there is no substantial difference in bike rentals among these weekdays compared to the base weekday (typically weekend or Monday).

Summary

- **Year, workingday, spring season, and adverse weather conditions (both light and misty)** have significant effects on the dependent variable (bike rentals).
- **Holidays** have a marginally significant negative impact.
- **Weekdays** do not show a significant difference in bike rentals, indicating that rentals are relatively stable across different weekdays.

Q2. Why is it important to use drop_first=True during dummy variable creation?

Answer:

Using `drop_first=True` during dummy variable creation is important for several reasons related to statistical modelling and multicollinearity:

1. Avoiding Multicollinearity

- Multicollinearity occurs when one or more predictor variables in a regression model are highly correlated. This can make the model estimates unstable and difficult to interpret.
- When creating dummy variables for a categorical variable with `n` categories, you typically get `n` binary columns (one for each category). If you include all `n` dummy variables in the model, one of them can be perfectly predicted by the others (perfect multicollinearity).
- By using `drop_first=True`, you drop one of the dummy variables, thus reducing the number of dummy variables to `n-1`. This helps to avoid the issue of perfect multicollinearity.

2. Creating a Baseline (Reference Category)

- Dropping the first category creates a baseline or reference category against which the effects of the other categories are measured.
- The coefficients of the remaining dummy variables indicate the change in the dependent variable relative to this baseline category.
- This makes the interpretation of the model coefficients more meaningful. For example, if you have a categorical variable for "season" with categories [Spring, Summer, Fall, Winter] and you drop "Spring", the coefficients for "Summer", "Fall",

and "Winter" will indicate how the dependent variable differs in these seasons compared to Spring.

3. Model Efficiency

- Reducing the number of dummy variables by dropping one can led to a more efficient model in terms of computation and storage.
- This can be particularly important when dealing with large datasets or many categorical variables.

Example

Let's consider a simple example with a categorical variable "season" with four categories: [Spring, Summer, Fall, Winter].

Without `drop_first=True`:

- Four dummy variables: `season_Spring`, `season_Summer`, `season_Fall`, `season_Winter`

With `drop_first=True`:

- Three dummy variables: `season_Summer`, `season_Fall`, `season_Winter`
- "Spring" is the reference category.

interpretation

- The coefficient for `season_Summer` will show the effect of Summer relative to Spring.
- The coefficient for `season_Fall` will show the effect of Fall relative to Spring.
- The coefficient for `season_Winter` will show the effect of Winter relative to Spring.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer:

Based on the pair-plot and the correlation heatmap among the numerical variables, the variable that has the highest correlation with the target variable (`cnt`) is `registered`.

In the provided heatmap, the correlation coefficient between `registered` and `cnt` is 0.95, indicating a very strong positive correlation. This means that as the number of registered users increases, the total count of bike rentals (`cnt`) also tends to increase.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer:

To validate the assumptions of Linear Regression after building the model on the training set, several diagnostic checks and plots can be checked. Here's how I have approached this:

1. Linearity

Assumption: The relationship between the predictors and the response variable is linear.

- **Method:** Plot residuals versus fitted values.
- **Check:** Built a scatter plot and looked for the systematic pattern in the plot.

2. Normality of Residuals

Assumption: The residuals are approximately normally distributed.

Validation Method:

- Plot a histogram of the residuals.
- Plot a Q-Q plot (Quantile-Quantile plot) of the residuals.

3. Homoscedasticity

Assumption: The residuals have constant variance (homoscedasticity).

Validation Method:

- Plot the residuals versus the fitted values again (same plot used for linearity check).

3. Homoscedasticity

Assumption: The residuals have constant variance (homoscedasticity).

Validation Method:

- Plot the residuals versus the fitted values again (same plot used for linearity check).

4. Independence of Residuals

Assumption: The residuals are independent of each other.

Validation Method:

- Plot residuals over time or in the order of data collection.
- Perform the Durbin-Watson test.

Interpretation: The Durbin-Watson statistic should be between 1.5 and 2.5. A value close to 2 indicates that there is no significant autocorrelation.

5. No Perfect Multicollinearity

Assumption: There is no perfect multicollinearity among the predictors.

Validation Method:

- Calculate the Variance Inflation Factor (VIF) for each predictor.

Interpretation: VIF values should be below 10. High VIF values indicate multicollinearity, which can be problematic.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer:

Based on the final model summary, the top three features contributing significantly towards explaining the demand for shared bikes are:

1. **Year (`yr`)**:
   - **Coefficient**: 1009.4777
   - **p-value**: 0.000
   - **Explanation**: This indicates that the demand for bikes significantly increases by approximately 1009 units for each subsequent year. This could be due to increased popularity or expansion of the bike-sharing program over time.
2. **Apparent Temperature (`atemp`)**:
   - **Coefficient**: 696.9909
   - **p-value**: 0.000
   - **Explanation**: This suggests that higher apparent temperatures are associated with a significant increase in bike rentals. For each unit increase in apparent temperature, bike demand increases by approximately 697 units.
3. **Weather Situation (`weathersit_Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds`)**:
   - **Coefficient**: -378.6048
   - **p-value**: 0.000
   - **Explanation**: Adverse weather conditions such as light snow, light rain, thunderstorms, and scattered clouds significantly reduce bike demand. Specifically, these weather conditions lead to a decrease in bike rentals by approximately 379 units.

These three features have the highest absolute coefficients with significant p-values, indicating their strong impact on bike demand in the dataset.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail?

Answer:

Linear Regression Algorithm

Linear regression is a statistical method used for modeling the relationship between a dependent variable (target) and one or more independent variables (predictors). The primary goal is to find the best-fitting linear relationship that can be used to predict the value of the dependent variable based on the values of the independent variables. Here's a detailed explanation of the linear regression algorithm:

1. Model Representation

The linear regression model assumes a linear relationship between the dependent variable yyy and the independent variables XXX. The equation of the line is given by:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

Where:

- y is the dependent variable.
- $X_1, X_2, \ldots, X_n$ are the independent variables.
- β_0 is the intercept.
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients (slopes) of the independent variables.
- ϵ is the error term (residual) which accounts for the variability in y that cannot be explained by the linear relationship.

2. Assumptions

Linear regression relies on several assumptions:

- **Linearity**: The relationship between the independent and dependent variables is linear.
- **Independence**: The residuals (errors) are independent.
- **Homoscedasticity**: The residuals have constant variance at every level of X.
- **Normality**: The residuals of the model are normally distributed.

3. Objective Function

The objective of linear regression is to find the coefficients $(\beta_0, \beta_1, \ldots, \beta_n)$ that minimize the sum of the squared residuals (errors) between the observed and predicted values of the dependent variable. This is known as the **Ordinary Least Squares (OLS)** method. The objective function is:

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

Where:

- y_i is the actual value of the dependent variable.
- $\hat{y}_i$ is the predicted value of the dependent variable.
- m is the number of observations.

4. Solving the Linear Regression

To find the best-fitting line, we solve for the coefficients that minimize the objective function. This can be done using calculus (derivatives) to derive the following normal equations:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Where:

- $\hat{\beta}$ is the vector of estimated coefficients.
- X is the matrix of independent variables (including a column of ones for the intercept).
- y is the vector of the dependent variable.
- X^T is the transpose of the matrix X.
- $(X^T X)^{-1}$ is the inverse of the matrix $X^T X$.

5. Prediction

Once the coefficients are estimated, predictions can be made using the linear regression equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_n X_n$$

6. Evaluation

- **R-squared (R^2)**: Measures the proportion of variance in the dependent variable that is predictable from the independent variables.
  $R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2}$
  Where $\bar{y}$ is the mean of the observed values.

- **Mean Squared Error (MSE)**: The average of the squared differences between the observed and predicted values.
  $MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$

- **Root Mean Squared Error (RMSE)**: The square root of the mean squared error.
  $RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}$

To evaluate the performance of the linear regression model, we use several metrics:

7. Validation of Assumptions

After building the model, it is crucial to validate the assumptions of linear regression by analyzing residuals:

- **Linearity**: Check residuals vs. fitted values plot for any patterns.
- **Normality**: Use Q-Q plot and histogram of residuals.
- **Homoscedasticity**: Check residuals vs. fitted values plot for constant variance.
- **Independence**: Use Durbin-Watson test for autocorrelation of residuals.

Conclusion

Linear regression is a fundamental and widely used statistical method for understanding and predicting relationships between variables. By carefully validating assumptions and interpreting coefficients, it provides valuable insights into the data.

Q2. Explain the Anscombe's quartet in detail?

Answer:

Anscombe's Quartet

Anscombe's Quartet is a set of four datasets that have nearly identical simple descriptive statistics but reveal very different distributions and relationships when graphed. The quartet was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analysing it and to illustrate how different datasets can exhibit similar statistical properties but have distinct characteristics.

The Four Datasets

Each of the four datasets in Anscombe's Quartet consists of 11 (x, y) points and has the following nearly identical statistical properties:

- **Mean of x values**: 9
- **Mean of y values**: 7.50
- **Variance of x values**: 11
- **Variance of y values**: 4.125
- **Correlation between x and y**: 0.816
- **Linear regression line (y on x)**: $y=3.00+0.500x$

Despite these similarities, the datasets differ significantly when plotted graphically:

1. **Dataset I**: A typical linear relationship with some random noise.
2. **Dataset II**: A perfect quadratic relationship, demonstrating that non-linear relationships can have the same correlation as linear ones.
3. **Dataset III**: A linear relationship, but with one outlier that can significantly affect the results.
4. **Dataset IV**: Vertical line of points with one horizontal outlier, showing how a single point can greatly influence correlation and regression results.

Graphical Representation

To understand the differences, let's describe the plots of each dataset:

1. **Dataset I**:
    - Appears to be a simple linear relationship between x and y, with some random variation around the line.
    - The linear regression model is appropriate here.
2. **Dataset II**:
    - Shows a clear non-linear (curved) relationship between x and y.
    - A quadratic model would be more appropriate than a linear regression model.
3. **Dataset III**:
    - Most points lie on a straight line, but there is one significant outlier.
    - The outlier can greatly affect the slope and intercept of the linear regression line.
4. **Dataset IV**:
    - Nearly all the points have the same x value except for one outlier.
    - The outlier dramatically affects the correlation and regression line, which does not represent the main data cluster.

Importance of Anscombe's Quartet

Anscombe's Quartet highlights several key lessons in data analysis:

1. **Graphical Analysis**: Always visualize your data before analysing it. Graphs can reveal patterns, trends, and outliers that descriptive statistics might not capture.
2. **Outliers and Leverage Points**: Outliers and leverage points can have a significant impact on statistical measures like the mean, variance, correlation, and regression coefficients. Identifying and understanding these points is crucial.
3. **Model Appropriateness**: Statistical models must be chosen based on the data's characteristics. A linear regression model is not always appropriate, especially if the relationship between variables is non-linear.

4. **Data Integrity**: Similar statistical summaries can come from very different data distributions. Relying solely on summary statistics without understanding the underlying data can lead to incorrect conclusions.

Conclusion

Anscombe's Quartet serves as a powerful reminder of the importance of exploratory data analysis. It emphasizes the need for visualization and thorough examination of data to avoid misleading interpretations based on summary statistics alone. By understanding and applying the lessons from Anscombe's Quartet, analysts can make more accurate and meaningful interpretations of their data.

Q3. What is Pearson's R?

Answer:

Pearson's R (Pearson Correlation Coefficient)

Pearson's R, also known as the **Pearson correlation coefficient** or **Pearson's product-moment correlation coefficient**, is a measure of the linear correlation between two variables. It quantifies the strength and direction of the linear relationship between two continuous variables.

Formula

The formula for Pearson's R is:

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

Where:

- r is the Pearson correlation coefficient.
- n is the number of data points.
- x_i and y_i are the individual data points for variables x and y.
- $\bar{x}$ and $\bar{y}$ are the means of the variables x and y.

Interpretation

Pearson's R ranges from -1 to +1:

- **r=+**: Perfect positive linear correlation. As x increases, y increases perfectly.
- **r=−**: Perfect negative linear correlation. As x increases, y decreases perfectly.
- **r=0**: No linear correlation. The variables x and y do not have a linear relationship.

The closer the value of r to +1 or -1, the stronger the linear relationship between the two variables.

Significance Testing

To determine if the observed correlation is statistically significant, a hypothesis test can be conducted. The null hypothesis (H0) is that there is no linear correlation between the variables (r=0). The significance of Pearson's R can be tested using a t-distribution:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Where:

- t is the t-statistic.
- r is the Pearson correlation coefficient.
- n is the number of data points.

The t-statistic can be compared against a critical value from the t-distribution table with n−2 degrees of freedom to determine the p-value.

Assumptions

For Pearson's R to be a valid measure of correlation, the following assumptions must be met:

- **Linearity**: The relationship between the variables should be linear.
- **Homoscedasticity**: The variability of the residuals (errors) should be constant across all levels of the independent variable.
- **Normality**: The variables should be approximately normally distributed (especially for small sample sizes).

Example

Suppose you have two variables: the number of hours studied and the scores on a test. By calculating Pearson's R, you can determine whether there is a linear relationship between studying time and test performance, and how strong this relationship is.

Conclusion

Pearson's R is a powerful and widely used statistic for measuring the linear relationship between two continuous variables. By understanding its calculation, interpretation, and assumptions, analysts can effectively use it to draw insights about the strength and direction of relationships in their data.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer:

Scaling

Scaling is the process of transforming the features of a dataset so that they are on a similar scale. This is important in machine learning, particularly for algorithms that compute distances between data points or assume normally distributed data. Scaling ensures that features contribute equally to the model, improving performance and convergence speed.

Why Scaling is Performed

1. **Improves Model Performance**: Many machine learning algorithms, like gradient descent-based algorithms, SVMs, and k-NN, perform better and converge faster when the data is scaled.
2. **Equal Contribution**: It ensures that each feature contributes equally to the result, preventing features with larger ranges from dominating the model.
3. **Reduces Computational Complexity**: Scaling can simplify calculations and reduce computational load, especially in distance-based algorithms.
4. **Normalization of Data Distribution**: Some algorithms assume normality in the data distribution; scaling helps in achieving this assumption.

Types of Scaling

There are two main types of scaling: **normalized scaling** and **standardized scaling**.

Normalized Scaling

Normalization, also known as **Min-Max Scaling**, transforms the data into a range of [0, 1] or [-1, 1]. Each feature is scaled according to the minimum and maximum values of that feature.

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

- **Purpose**: To bound the values within a specific range, making the data easier to work with and interpret.
- **Use Case**: Useful when the data does not have outliers, and you want to preserve the relationships between values.
- **Example**: Suppose you have a feature with values ranging from 10 to 100. After normalization, the smallest value (10) will be transformed to 0, and the largest value (100) will be transformed to 1.

Standardized Scaling

Standardization, also known as **Z-score normalization**, transforms the data so that it has a mean of 0 and a standard deviation of 1.

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

Where:

- μ is the mean of the feature.
- σ is the standard deviation of the feature.
- **Purpose**: To center the data and ensure that it has unit variance, making it suitable for algorithms that assume normality.
- **Use Case**: Useful when the data has outliers or when features have different units and ranges.
- **Example**: Suppose you have a feature with a mean value of 50 and a standard deviation of 10. After standardization, a value of 60 will be transformed to (60−50)/10=1.

Differences Between Normalized and Standardized Scaling

Differences Between Normalized and Standardized Scaling

Aspect	Normalized Scaling (Min-Max Scaling)	Standardized Scaling (Z-score Normalization)
Range	Transforms data to [0, 1] or [-1, 1]	Transforms data to have mean = 0 and std = 1
Formula	$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$	$x' = \frac{x - \mu}{\sigma}$
Use Case	When you need a specific range and no outliers are present	When the data has outliers or different units
Effect on Outliers	Can be heavily influenced by outliers	Less affected by outliers, but may not bound data
Application	Suitable for neural networks and image processing	Suitable for regression, clustering, and PCA

Conclusion

Scaling is a crucial step in preprocessing data for machine learning models. It helps in ensuring that all features contribute equally, improves model performance, and speeds up convergence. The choice between normalization and standardization depends on the nature of the data and the specific requirements of the machine learning algorithm being used.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer:

Infinite Variance Inflation Factor (VIF)

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

Why VIF Can Be Infinite

An infinite VIF value indicates perfect multicollinearity, meaning that one predictor variable is a perfect linear combination of one or more other predictor variables. This perfect collinearity leads to the inability to estimate regression coefficients uniquely, causing the VIF to be mathematically undefined or infinite.

Causes of Infinite VIF

1. **Exact Linear Relationship**: When one predictor variable can be expressed as an exact linear combination of other predictor variables, the matrix used to compute VIF becomes singular, leading to an infinite VIF.

    - For example, if $X_1 = 2X_2 + 3X_3$, then X_1 is perfectly cc

2. **Duplicated Variables**: If a variable is duplicated in the dataset, it will have perfect collinearity with itself, causing VIF to be infinite.
3. **Dummy Variable Trap**: In the context of dummy variables for categorical data, including all dummy variables for a categorical feature can lead to perfect multicollinearity.
    - For instance, if you have a categorical variable with 3 levels (A, B, C) and you create dummy variables for all three (A, B, C), one of these dummy variables can be perfectly predicted by the others (C = 1 - A - B), leading to infinite VIF.

Preventing Infinite VIF

1. **Remove Perfectly Collinear Variables**: Check for and remove or combine perfectly collinear variables.
2. **Avoid Dummy Variable Trap**: When creating dummy variables, use \text{drop_first=True} to avoid perfect multicollinearity by dropping one of the dummy variables.
3. **Regularization**: Use regularization techniques like Ridge Regression (L2

regularization), which can handle multicollinearity by adding a penalty term to the regression equation.

Conclusion

An infinite VIF value is a clear indication of perfect multicollinearity in the dataset. It is essential to identify and address this issue by examining the relationships between predictor variables, removing redundant variables, and using appropriate techniques to ensure that the regression model can be estimated correctly.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression?

Answer:

Q-Q Plot (Quantile-Quantile Plot)

A **Q-Q plot** (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, often the normal distribution. It plots the quantiles of the dataset against the quantiles of the theoretical distribution. If the points on the Q-Q plot fall approximately along a straight line, it suggests that the data follows the specified theoretical distribution.

Use of Q-Q Plot in Linear Regression

In the context of linear regression, a Q-Q plot is primarily used to check the assumption of normality of the residuals. This assumption states that the residuals (errors) should be normally distributed for the statistical tests and confidence intervals to be valid.

Steps to Create a Q-Q Plot

1. **Calculate Quantiles**: Determine the quantiles of the dataset (residuals) and the quantiles of the theoretical normal distribution.
2. **Plot Quantiles**: Plot the quantiles of the residuals on the y-axis and the quantiles of the normal distribution on the x-axis.
3. **Assess Linearity**: Analyse how closely the points follow a straight line.

Importance of Q-Q Plot in Linear Regression

1. **Assess Normality of Residuals**: The normality assumption is crucial for hypothesis testing and constructing confidence intervals. A Q-Q plot helps visually assess whether the residuals are approximately normally distributed.
2. **Diagnose Model Fit**: Deviations from the straight line in a Q-Q plot can indicate issues with the model, such as skewness or kurtosis in the residuals, suggesting that the model might not be adequately capturing the data structure.
3. **Identify Outliers**: Outliers can be detected as points that deviate significantly from the straight line in a Q-Q plot. These outliers can adversely affect the linear

regression model and its assumptions.

Interpretation of Q-Q Plot

- **Straight Line**: If the points on the Q-Q plot follow a straight line, it indicates that the residuals are normally distributed.
- **Systematic Deviations**: If the points deviate systematically from the straight line (e.g., a curve), it suggests that the residuals are not normally distributed, indicating potential skewness or kurtosis.
- **Outliers**: Points that fall far from the straight line at either end of the plot indicate outliers.

Example of Q-Q Plot Interpretation

1. **Normal Residuals**: Points lie on or near the diagonal line.
2. **Right-Skewed Residuals**: Points curve upwards away from the diagonal line on the right side.
3. **Left-Skewed Residuals**: Points curve downwards away from the diagonal line on the left side.
4. **Heavy Tails**: Points deviate from the diagonal line at both ends, suggesting kurtosis (heavy tails).

Conclusion

A Q-Q plot is a valuable diagnostic tool in linear regression analysis. It provides a visual method to assess the normality of residuals, diagnose potential problems with the model, and identify outliers. Ensuring that residuals are normally distributed is crucial for the validity of statistical inferences made from the regression model.