Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

   From the univariate analysis:

   1) Demand for bike is less during start of the year and it increases and reaches peak at the mid of the year and reduces at the end of the year
   2) Demand for bike is less in winter compared to other seasons
   3) No variation of demand for weekday or weekend

   From Bivariate Analysis:

   4) Looks like temp and atemp are having similar relation with count. So dropping atemp from analysis

   From Multivariate Analysis:

   5) Cnt as positive correlation with temp, But negative correlation with windspeed. Other variables can be ignored due to very less correlation value

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**

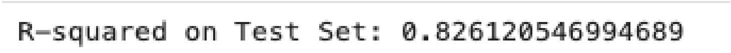
   Using drop_first=True during dummy variable creation (one-hot encoding) helps to avoid the "dummy variable trap," a scenario where dummy variables are highly correlated (multicollinearity). By dropping one dummy variable from each categorical feature, we effectively remove this perfect multicollinearity, ensuring that our model's coefficients remain interpretable and stable. This approach reduces the number of features without losing information, as the dropped category can be inferred from the others.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

   ‘temp’ variable has highest correlation with the target variable

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

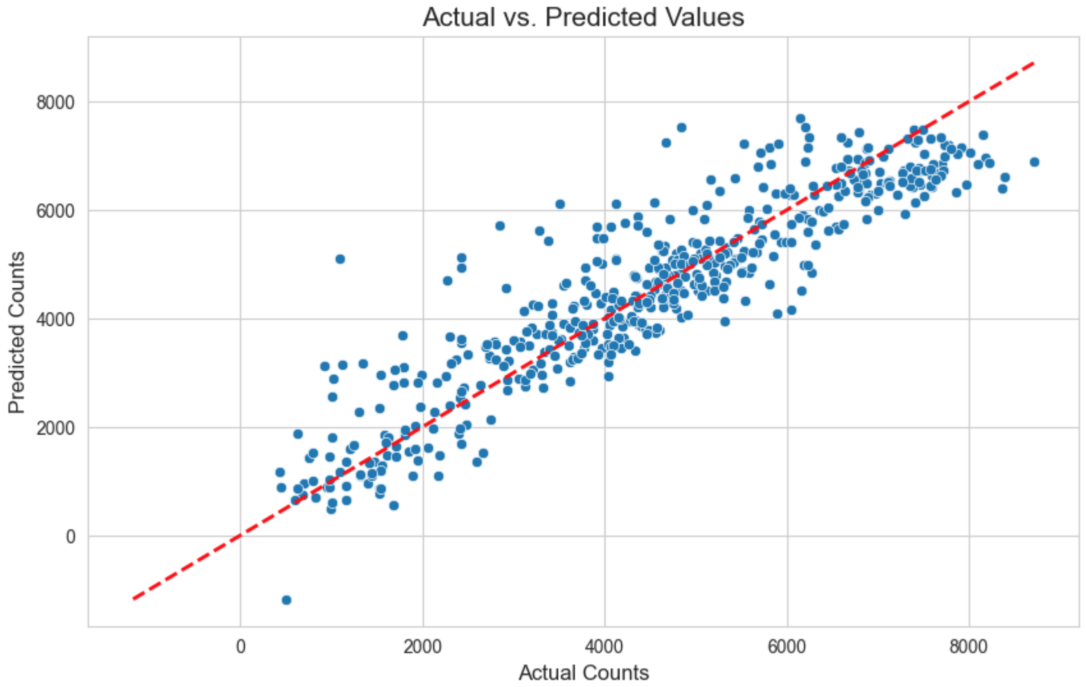
   1) R2_score_test is 0.826. This indicates that model is capable of explaining of 82.6% of the variance in bike rental demand with the given independent variables.



```
R-squared on Test Set: 0.826120546994689
```

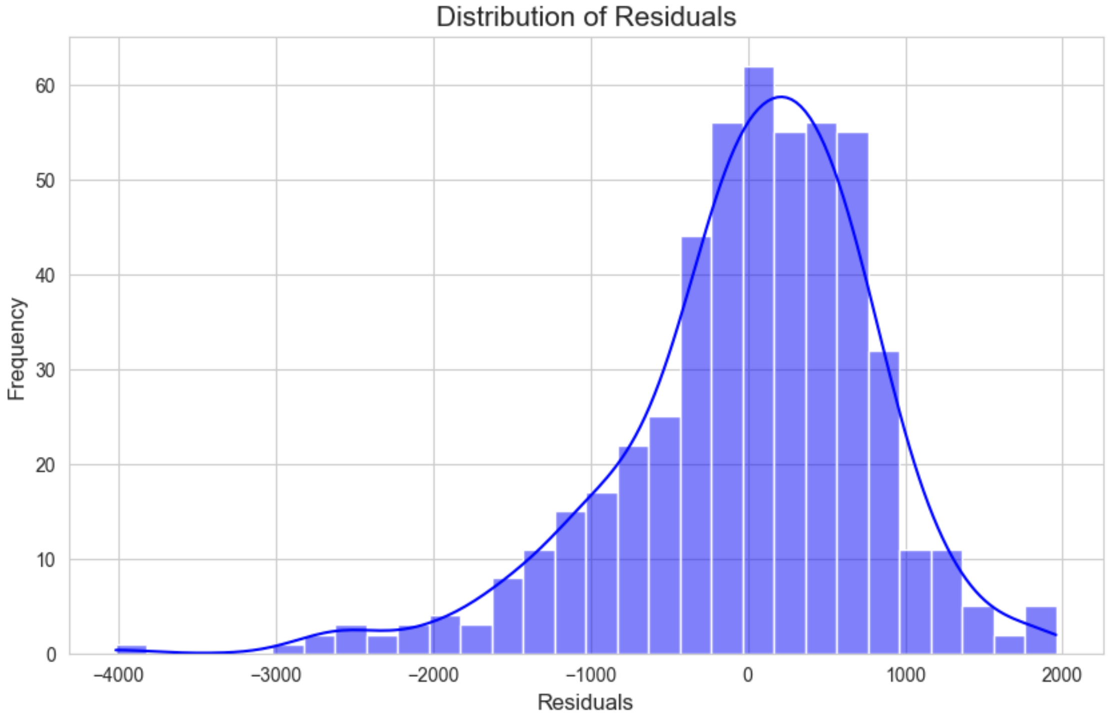
   2) Linearity

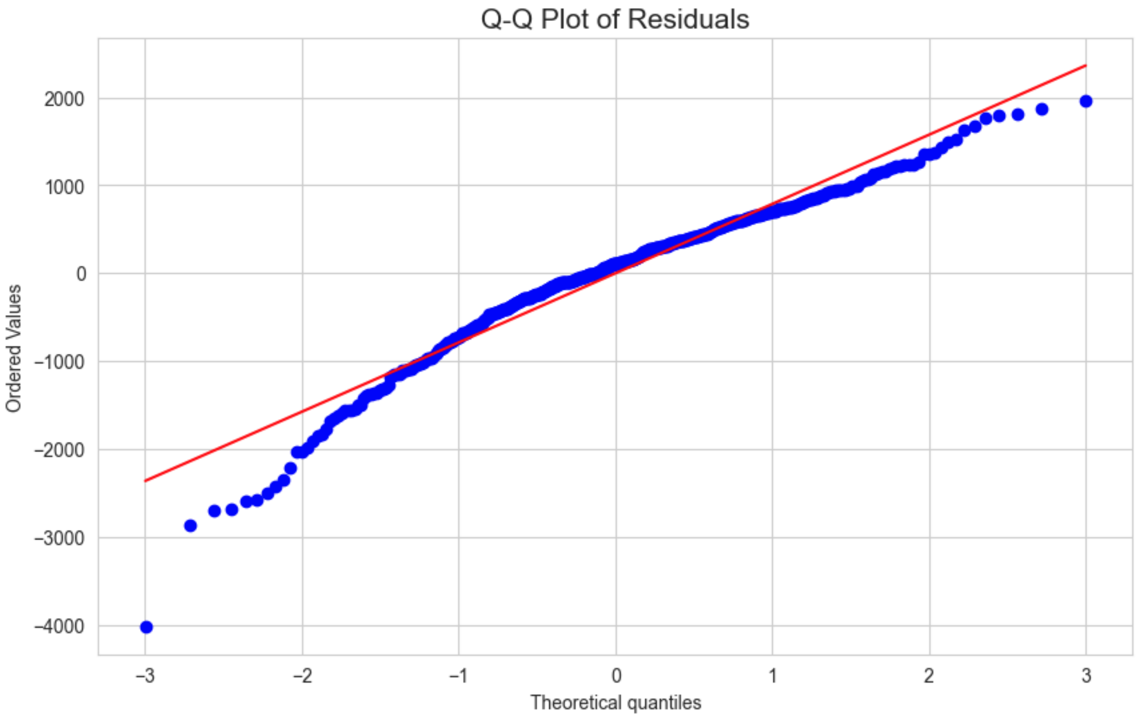
   Assumption: The relationship between the independent variables and the dependent variable is linear.



Validation: I visually inspected scatter plots of the predicted values against actual values and looked for linear patterns. A linear pattern suggests that the linearity assumption holds. Non-linear patterns might indicate that a linear model is not appropriate.

3) Normality of Residuals
Assumption: The residuals (differences between observed and predicted values) of the model are normally distributed.





Validation:
I plotted a histogram of the residuals. A bell-shaped curve centered around zero indicates that the residuals are approximately normally distributed.
I also used a Q-Q plot (quantile-quantile plot) to compare the distribution of residuals against a normal distribution. Points that closely follow the diagonal line in a Q-Q plot suggest that the residuals are normally distributed.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

A) temp (Temperature):

Coefficient: 3892.2865
P-value: Significantly low (p < 0.001)
Interpretation: Temperature is the most significant predictor of bike demand. The positive coefficient suggests that higher temperatures are associated with increased bike rentals, making it the top contributing feature.

B) yr_1 (Year: 2019):

Coefficient: 1917.1324
P-value: Significantly low (p < 0.001)
Interpretation: The year 2019 (encoded as yr_1) significantly impacts bike rental demand, indicating a notable increase in demand from 2018 to 2019. This reflects the growing popularity or expansion of the bike-sharing service.

C) weathersit_Light_Snow_Rain (Weather Situation: Light Snow or Rain):

Coefficient: -1434.9070

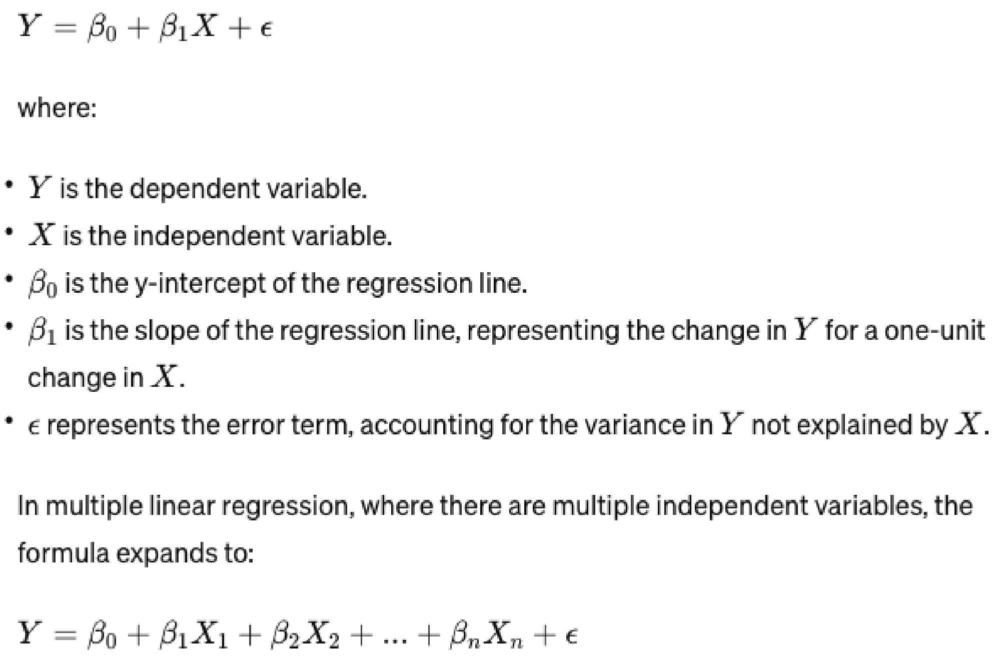
P-value: Significantly low (p < 0.001)
Interpretation: This feature has a substantial negative impact on bike demand. It suggests that light snow or rain significantly reduces the number of bike rentals, making it a critical factor affecting demand, albeit in a negative way.

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**

Linear regression is a fundamental algorithm in statistical and machine learning for modeling the relationship between a dependent variable and one or more independent variables. The goal is to find a linear equation that best predicts the dependent variable from the independent variables.

Basic Concept:
The linear regression model assumes a linear relationship between the dependent variable using the formula:



$$Y = \beta_0 + \beta_1 X + \epsilon$$

where:

- Y is the dependent variable.
- X is the independent variable.
- β_0 is the y-intercept of the regression line.
- β_1 is the slope of the regression line, representing the change in Y for a one-unit change in X.
- ϵ represents the error term, accounting for the variance in Y not explained by X.

In multiple linear regression, where there are multiple independent variables, the formula expands to:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon$$

Estimating the Coefficients:
Linear regression aims to estimate the coefficients (β) that minimize the difference between the observed values and the values predicted by the linear model. This is commonly achieved through the method of least squares, which minimizes the sum of the squared residuals (differences between observed and predicted values).

In our case, we applied linear regression to model the demand for shared bikes as a function of various features like temperature, year, weather conditions, etc. The analysis involved:

Feature Selection: Identifying relevant features using Recursive Feature Elimination (RFE), focusing on those significantly influencing the dependent variable.

Model Fitting: Estimating the coefficients of the linear equation by fitting the model to the training data, using OLS (Ordinary Least Squares) regression.

Validation of Assumptions: Checking the assumptions of linear regression through residual analysis, including examining the distribution of residuals, checking for homoscedasticity, and assessing the linearity of the model.

Interpretation: Interpreting the coefficients to understand the influence of each feature on bike demand. Positive coefficients indicate features that increase demand, while negative coefficients suggest features that decrease demand.

Linear regression offers a clear interpretation of the relationship between variables, making it a valuable tool for understanding and predicting outcomes in various domains, including demand forecasting, as illustrated in this shared bikes dataset analysis.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet comprises four distinct datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x, y) points. This quartet was constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties. It serves as a powerful reminder that statistical summaries can sometimes be misleading, and visualizing data is a crucial step in data analysis.

Key Properties:
When analyzed statistically, the four datasets in Anscombe's quartet have the same or very similar values for the following properties:

- ➔ Mean of both x and y variables.
- ➔ Variance of x and y variables.
- ➔ Correlation between x and y variables.
- ➔ Linear regression line (y = mx + c) that best fits the data, including the slope (m) and y-intercept (c).
- ➔ Coefficient of determination (R^2), which measures the proportion of the variance in the dependent variable that is predictable from the independent variable.

Despite these similarities in statistical summaries, the datasets have very different distributions and appear distinct when graphed. Each set illustrates a different case or problem in regression analysis.

The Four Datasets:

Dataset I: Follows a pattern that appears to be a simple linear relationship, corresponding closely to the assumptions of linear regression.

Dataset II: Demonstrates a curve (quadratic relationship) rather than a linear relationship. Linear regression is not appropriate here, but the statistical properties mirror those of the first dataset.

Dataset III: Contains an outlier that influences the slope of the regression line. Without this outlier, the dataset would have a very different linear relationship.

Dataset IV: Shows a case where one outlier is driving the entire correlation. Without the outlier, there would be little to no correlation between x and y variables.

Importance:
Anscombe's quartet is a foundational lesson in data analysis, emphasizing the following points:

Visualize Data: Always graph your data before starting the analysis. Visual inspection can reveal data properties and structures that summary statistics cannot.
Beware of Outliers: Outliers can significantly affect the results of your analysis, leading to misleading conclusions.
Understand the Data: Knowing the underlying assumptions of statistical methods is crucial. For instance, linear regression assumes a linear relationship between variables, which might not always hold true.

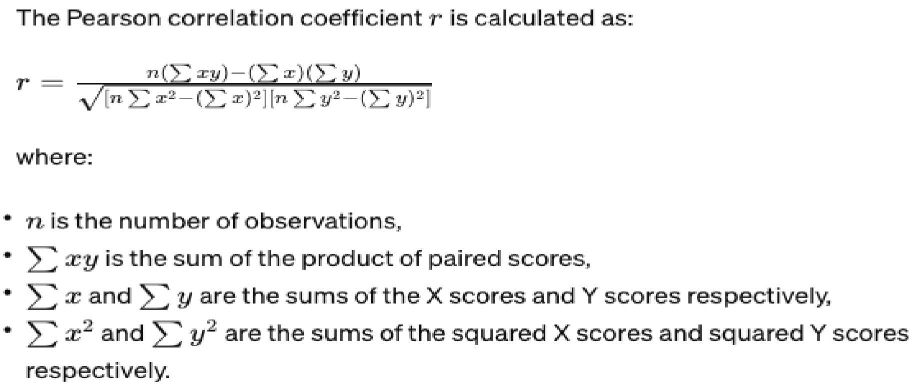
The quartet serves as a cautionary tale, reminding analysts and statisticians to look beyond numerical summaries and to use visualization tools as an integral part of their data analysis workflow.

3. **What is Pearson’s R? (3 marks)**

Pearson's R, also known as the Pearson correlation coefficient or Pearson's product-moment correlation coefficient (PPMCC), is a measure of the linear correlation between two variables X and Y. It gives a value between +1 and -1 inclusive, where:

+1 indicates a perfect positive linear relationship,
-1 indicates a perfect negative linear relationship, and
0 indicates no linear correlation between the variables.

Formula:



The Pearson correlation coefficient r is calculated as:

$$r = \frac{n(\sum xy)-(\sum x)(\sum y)}{\sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2]}}$$

where:

- n is the number of observations,
- $\sum xy$ is the sum of the product of paired scores,
- $\sum x$ and $\sum y$ are the sums of the X scores and Y scores respectively,
- $\sum x^2$ and $\sum y^2$ are the sums of the squared X scores and squared Y scores respectively.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**

Scaling is a data preprocessing technique used to standardize the range of independent variables or features of data. In the context of machine learning and data analysis, scaling can be crucial because the range of data values can vary widely. If one feature has a range from 0 to 1, while another has a range from 0 to 1000, the model might unduly weigh the latter feature more heavily than the former, potentially leading to inaccurate predictions or analyses.

Why is Scaling Performed?
Improves Algorithm Performance: Many machine learning algorithms that use distance calculations (e.g., k-nearest neighbors, k-means clustering) or optimization (e.g., gradient descent in neural networks, support vector machines) perform better when the data is scaled.

Increases Computational Efficiency: Algorithms converge faster when features are on similar scales, particularly in optimization algorithms.

Prevents Skewed Influence: Prevents features with larger scales from overshadowing those with smaller scales in models where feature weighting is important.

Required by Some Models: Some algorithms, like Support Vector Machines (SVM) and Principal Component Analysis (PCA), explicitly require scaling for correct execution.

Normalized Scaling vs. Standardized Scaling:
1. **Normalized Scaling (Min-Max Scaling):**
Normalization adjusts the data values to a specific scale, typically 0 to 1, without distorting differences in the ranges of values or losing information. It is performed using the formula:

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

Normalization is useful when you need to bound your values between two numbers, e.g., 0 and 1.

2. **Standardized Scaling (Z-score Normalization):**
Standardization transforms the data to have a mean of zero and a standard deviation of one. The formula for standardization is:

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

σ is the standard deviation of the feature values. Standardization does not bound values to a specific range, which may be a problem for certain algorithms (e.g., neural networks often expect an input value bounded between 0 and 1).

Choosing between normalization and standardization depends on the specific requirements of the model, the data distribution, and the presence of outliers.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

The Variance Inflation Factor (VIF) is a measure used to detect the presence of multicollinearity among independent variables in a regression model. Multicollinearity occurs when two or more independent variables are highly correlated, meaning they contain similar information about the variance. VIF quantifies how much the variance of an estimated regression coefficient increases if your predictors are correlated.

A VIF value of 1 indicates no correlation between a given independent variable and any other independent variables. VIF values between 1 and 5 suggest moderate correlation, but they are often considered acceptable. Values greater than 5 or 10 indicate potentially problematic levels of multicollinearity, depending on the context and the source.

Reasons for an Infinite VIF:
An infinite VIF (or a VIF value that is exceedingly large) can occur for several reasons, primarily related to perfect or near-perfect multicollinearity:

Perfect Multicollinearity: This happens when an independent variable is an exact linear combination of one or more independent variables. For example, if one variable is the sum or difference of two others in your dataset, it will lead to a situation where the VIF for at least one of those variables could be infinite.

Redundant Variables: Including variables in the model that are redundant (duplicate information in another format) can cause infinite VIFs. A common scenario is when dummy variables are created from a categorical variable but all categories are included without dropping one as a reference. This creates a perfect linear relationship among the dummy variables due to the "dummy variable trap."

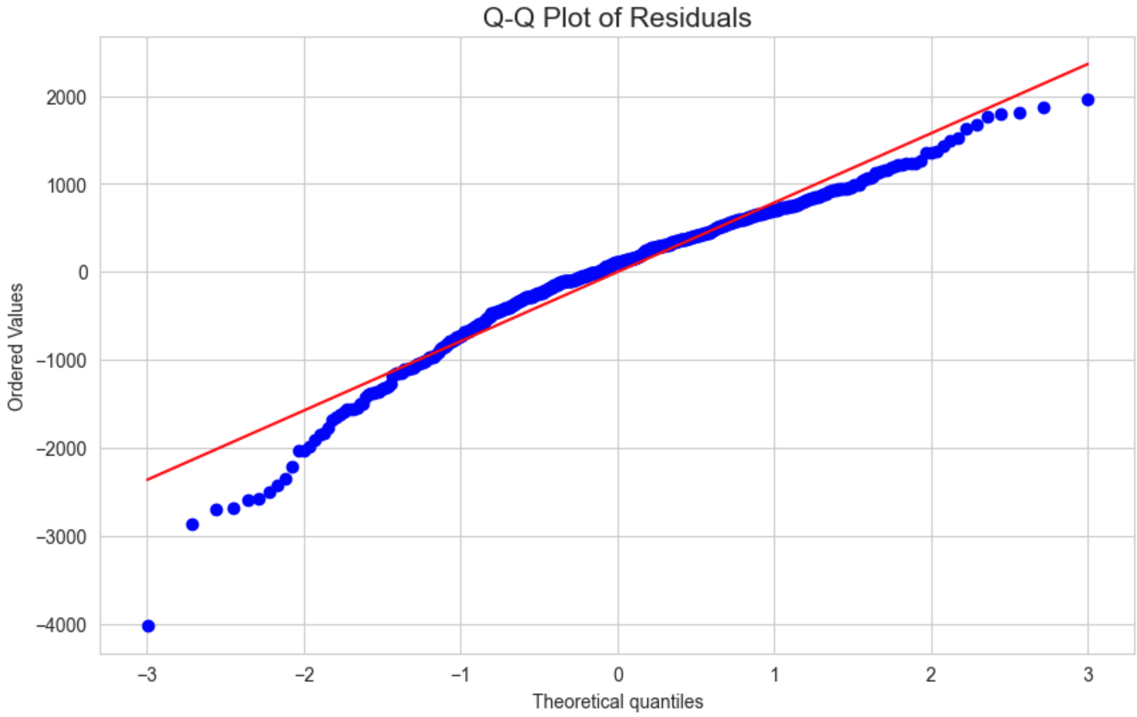
Highly Correlated Variables: Even if not perfectly linear, variables that are highly correlated can produce very large VIF values, approaching infinity as the correlation approaches perfect linearity.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

A Q-Q (quantile-quantile) plot is a graphical tool used to compare two probability distributions by plotting their quantiles against each other. If the two distributions being compared are similar, the points on the Q-Q plot will approximately lie on the line y = x. When using a Q-Q plot in the context of linear regression, it's most commonly used to compare the distribution of residuals (errors) from the regression model to a theoretical normal distribution. This comparison is crucial for checking the assumption of normality in linear regression models.

Use of Q-Q Plot in Linear Regression:

I have plotted Q-Q plot in the current assignment :



Assessing Normality of Residuals:

In linear regression, one key assumption is that the residuals (the differences between observed and model-predicted values) are normally distributed.
A Q-Q plot helps visually assess how closely the residuals of your model match the expected distribution (typically a normal distribution).
On the plot, the actual quantiles of the residuals are plotted on the y-axis, and the theoretical quantiles of the normal distribution are plotted on the x-axis.
Identifying Departures from Normality:

The Q-Q plot makes it easy to see deviations from normality. If the residuals are normally distributed, the points should form a straight line.
Curvature in the plot suggests that the residuals have skewness or kurtosis that deviates from normality.
Outliers appear as points that deviate significantly from the straight line.

Importance of Q-Q Plot in Linear Regression:

➔ Validity of Statistical Tests: Many statistical tests, including those used to calculate the significance of regression coefficients, assume normality of residuals. Deviations from this assumption can lead to incorrect inferences.

➔ Model Diagnostics: A Q-Q plot is a non-parametric method for checking the normality assumption and can indicate if transformations of variables may be necessary or if a different modeling approach should be considered.

➔ Identifying Outliers: It helps in identifying outliers in the data which can disproportionately influence the regression model.

➔ Improving Model Accuracy and Interpretability: Ensuring that the residuals follow a normal distribution can improve the accuracy of the regression model and the reliability of confidence intervals and predictions made from the model.