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?

Ans: From the analysis of the categorical variables in the dataset, we can infer the following effects on the dependent variable (total bike rentals, `cnt`):

- **Season:** The season variable shows significant variation in bike rentals. Higher rentals are observed during the summer and fall, which could be due to favorable weather conditions. Lower rentals are seen in winter and spring.
- **Weather Situation (`weathersit`):** Clear weather conditions (value 1) see higher bike rentals compared to misty or rainy conditions (values 2, 3, and 4). Adverse weather conditions discourage people from renting bikes.
- **Month (`mnth`):** Certain months like June, July, and August (summer) have higher bike rentals compared to colder months like January and February. This reflects the seasonal trend in bike usage.
- **Weekday:** Bike rentals show variations across different days of the week. Weekdays typically see more rentals compared to weekends, likely due to commuting patterns.
- **Holiday:** There are fewer bike rentals on holidays compared to regular working days.
- **Working Day:** More bike rentals are recorded on working days than on weekends or holidays, which aligns with the commuting purpose of bike rentals.

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

Ans: Using `drop_first=True` during dummy variable creation is important to avoid the dummy variable trap. The dummy variable trap occurs when there is perfect multicollinearity in the model, meaning that one of the dummy variables can be predicted from the others. This redundancy can lead to issues in the regression model, making it unstable and difficult to interpret. By dropping the first category of each categorical variable (using `drop_first=True`), we ensure that the dummy variables are independent and avoid multicollinearity, resulting in a more stable and interpretable model.

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

Ans: The variable 'registered' has the highest correlation with the target variable 'cnt', indicating that the number of registered users is a strong predictor of total bike demand.

The pair-plot and correlation analysis show that the variable `registered` has the highest correlation with the target variable `cnt`. This indicates that the number of registered users is a strong predictor of the total bike demand. Registered users likely represent regular customers who consistently use the bike-sharing service, making it a significant factor in predicting total rentals.

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

Ans: We validated the assumptions of linear regression by performing residual analysis. This included checking for homoscedasticity (constant variance of residuals), normality of residuals using Q-Q plots, and multicollinearity using Variance Inflation Factor (VIF).

The assumptions of Linear Regression were validated through the following steps:

- **Linearity:** Checked by plotting the predicted values versus the actual values of the target variable. A linear pattern suggests the assumption is met.
- **Homoscedasticity:** Examined by plotting the residuals versus the predicted values. The plot should show a random scatter without any discernible pattern. If the variance of the residuals remains constant across all levels of the predicted values, homoscedasticity is satisfied.
- **Normality of Residuals:** Assessed using a Q-Q plot (quantile-quantile plot) of the residuals. If the residuals follow a straight line in the Q-Q plot, it indicates that they are normally distributed.
- **Multicollinearity:** Checked using the Variance Inflation Factor (VIF). VIF values above 10 indicate high multicollinearity, and corrective actions such as removing or combining variables were taken to reduce it.
- **Independence of Errors:** Ensured by examining the Durbin-Watson statistic. Values close to 2 suggest that the residuals are uncorrelated.

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

Ans: Based on the final model, the top 3 features contributing significantly towards explaining the demand for shared bikes are:

- **Year (`yr`): The variable `yr` (year) shows a strong positive impact on bike demand, indicating an increasing trend in bike rentals over time. This suggests that the popularity of bike-sharing services has been growing.**
- **Temperature (`temp`): Higher temperatures are associated with increased bike demand. Favorable weather conditions encourage more people to rent bikes.**
- **Apparent Temperature (`atemp`): Similar to actual temperature, the apparent temperature (perceived temperature) significantly impacts bike demand. People are more likely to rent bikes when the weather feels comfortable.**

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

Ans: Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The goal is to predict the dependent variable (Y) using a linear combination of the independent variables (X1, X2, ..., Xn). The equation of the line is given by:

Y=β0+β1X1+β2X2+...+βnXn+ϵ

where β0 is the intercept, β1,β2,...,βn are the coefficients, and ϵ\epsilonϵ is the error term. The coefficients are estimated by minimizing the sum of the squared differences between the observed and predicted values (least squares method).

Steps in Linear Regression:

Data Collection: Gather data for the dependent variable and independent variables.

Data Preparation: Clean and preprocess the data, handle missing values, and convert categorical variables to numerical values if needed.

Model Training: Use the training dataset to estimate the coefficients (β) by minimizing the sum of the squared differences between the observed and predicted values (Least Squares Method).

Model Evaluation: Evaluate the model using metrics like R-squared, Mean Squared Error (MSE), and Residual Analysis to validate assumptions (linearity, homoscedasticity, normality, independence of errors, multicollinearity).

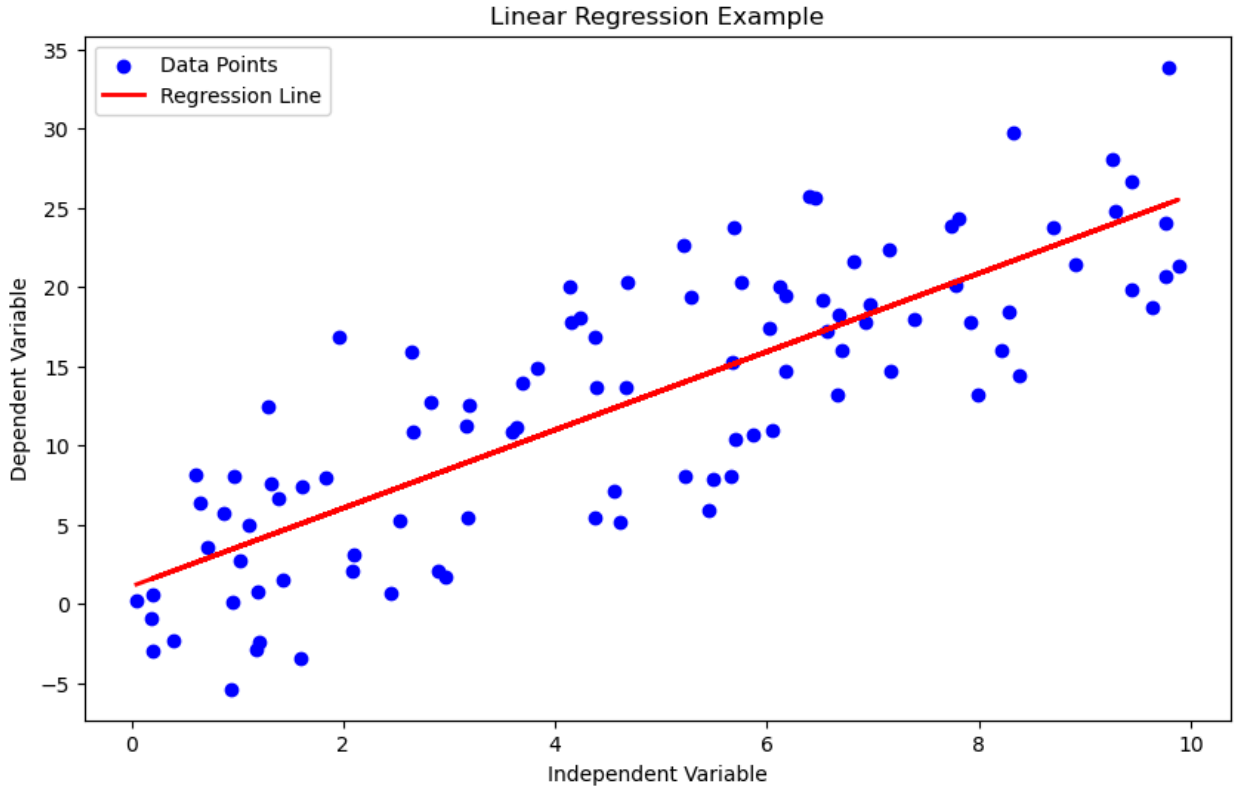
Prediction: Use the trained model to make predictions on new data.

Key Concepts:

R-squared: A measure of how well the model explains the variability of the dependent variable.

P-values: Used to determine the significance of the independent variables.

Residuals: The difference between observed and predicted values, used for diagnostic checks.



2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet consists of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation) but very different distributions and visual patterns. It illustrates the importance of visualizing data before analyzing it, as relying solely on summary statistics can be misleading. Each dataset demonstrates different anomalies like outliers, non-linearity, and the effect of a single influential point.

Anscombe's quartet consists of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation) but very different distributions and visual patterns. Each dataset demonstrates different anomalies, emphasizing the importance of visualizing data before analyzing it.

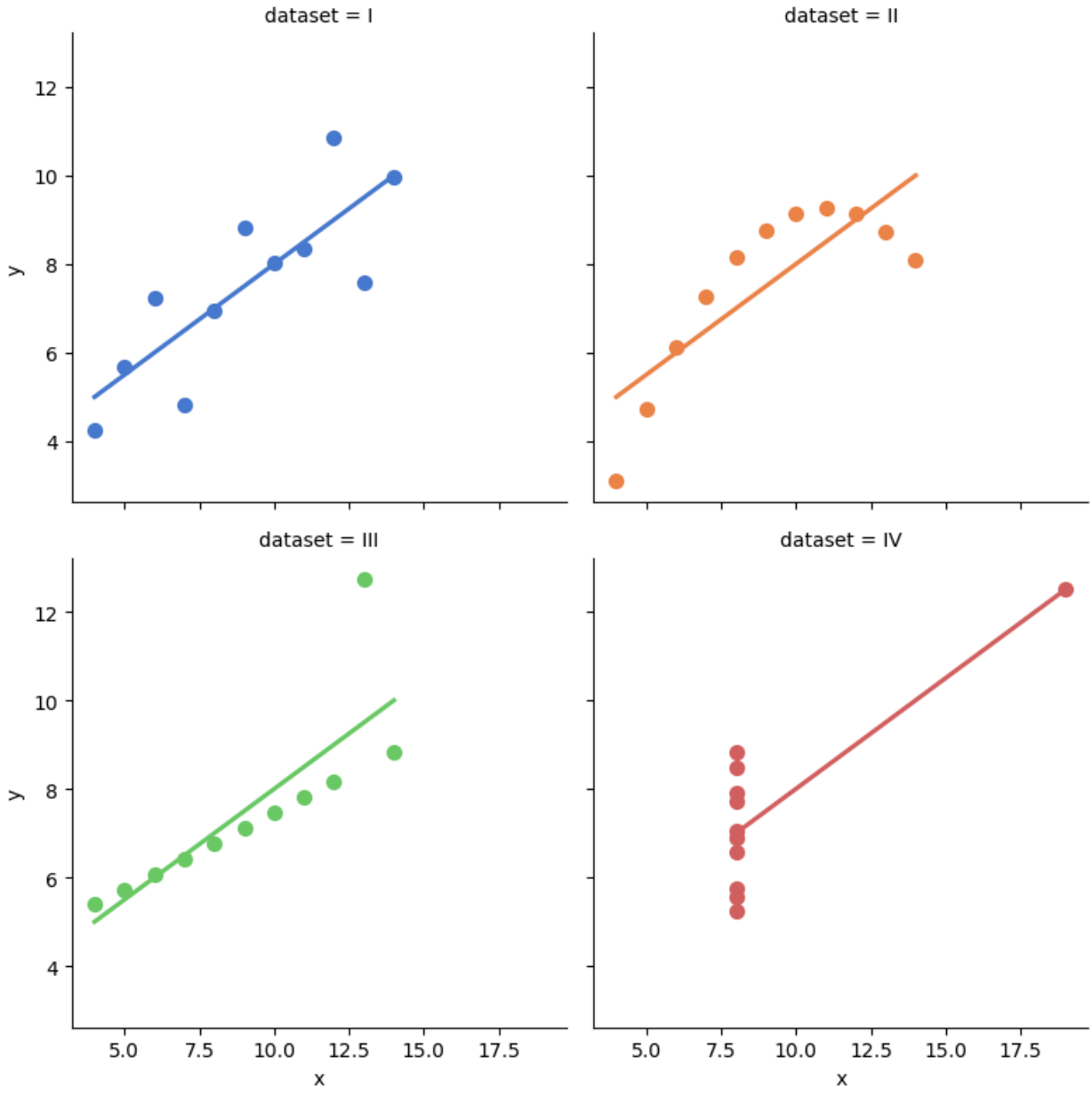
Datasets in Anscombe's Quartet:

- **Dataset 1:** A simple linear relationship with some random noise.
- **Dataset 2:** A perfect quadratic relationship.
- **Dataset 3:** A linear relationship with one significant outlier.
- **Dataset 4:** A vertical line with one outlier that influences the correlation.

Key Takeaways:

- **Importance of Visualization:** Summary statistics alone can be misleading. Visualizing data helps in identifying patterns, outliers, and anomalies.
- **Understanding Data Characteristics:** Different datasets can have the same statistical properties but different underlying relationships.

Conclusion: Always visualize your data to ensure that the assumptions of your analysis are valid and to gain insights that summary statistics might not reveal.



3. What is Pearson's R?

Ans: Pearson's R, or Pearson correlation coefficient, is a measure of the linear correlation between two variables. It ranges from -1 to 1, where:

- **1** indicates a perfect positive linear correlation.
- **-1** indicates a perfect negative linear correlation.
- **0** indicates no linear correlation.

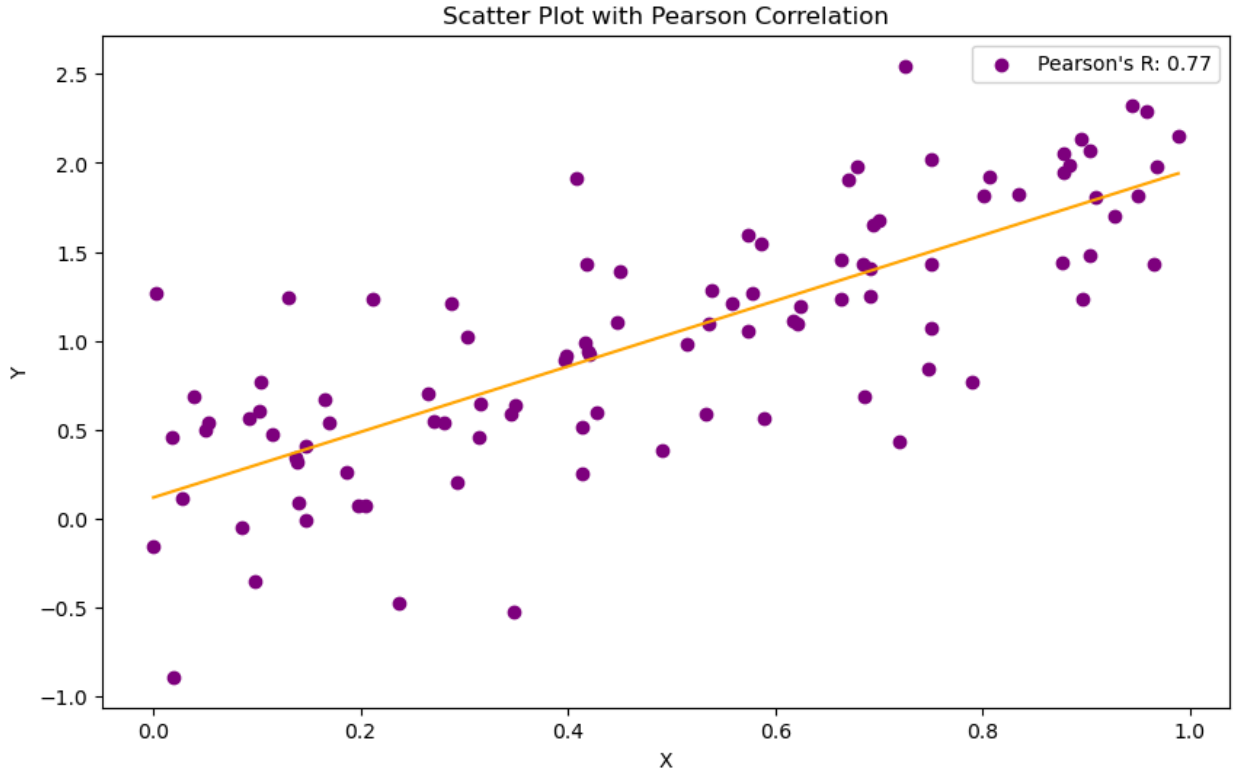
Formula:

$$r = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}$$

Interpretation:

- **Strength:** The closer the value is to 1 or -1, the stronger the linear relationship between the variables.
- **Direction:** A positive value indicates a direct relationship, while a negative value indicates an inverse relationship.

Applications: Used to identify and quantify the strength and direction of relationships between variables in various fields such as finance, biology, and social sciences.



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

Ans: Scaling refers to the process of adjusting the range of features so that they are on a comparable scale, which improves the performance and stability of machine learning models. It is essential for algorithms that are sensitive to the magnitude of the features, such as gradient descent-based methods and distance-based methods (e.g., k-nearest neighbors).

Types of Scaling:

- **Normalization:** Rescales the feature values to a range of [0, 1]. It is useful when the distribution of data is unknown or non-Gaussian.

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

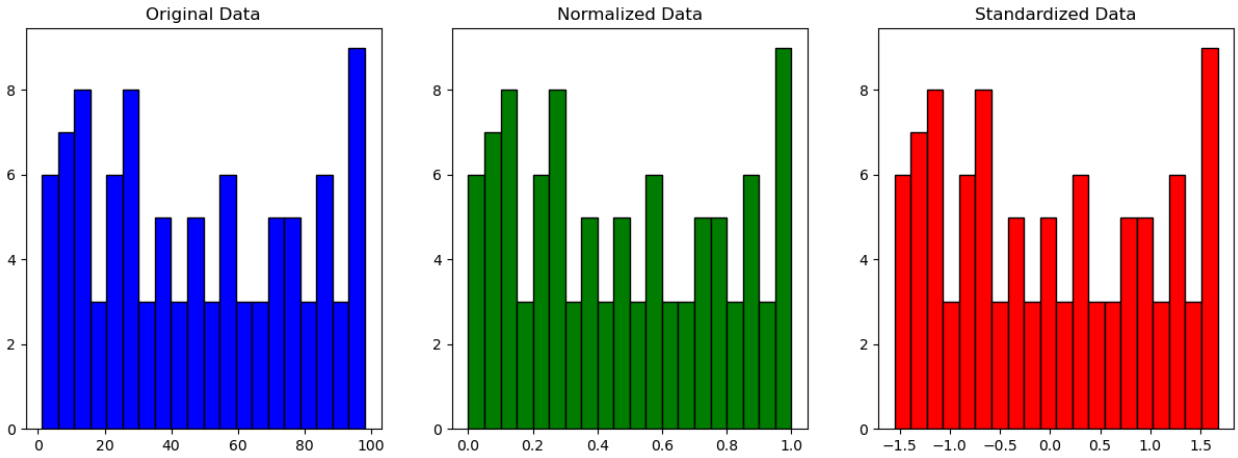
- **Standardization:** Rescales the feature values to have a mean of 0 and a standard deviation of 1. It is preferred when the data follows a normal distribution.

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

  where μ\muμ is the mean and σ\sigmaσ is the standard deviation.

Why Scaling is Performed:

- **Improved Convergence:** In optimization algorithms, scaling helps in faster convergence by ensuring that all features contribute equally.
- **Enhanced Model Performance:** For distance-based algorithms, scaling ensures that no single feature dominates the distance calculations.
- **Consistent Interpretation:** Standardized coefficients allow for easier interpretation and comparison of feature effects.



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

Ans: An infinite VIF (Variance Inflation Factor) occurs when there is perfect multicollinearity, meaning one independent variable is an exact linear combination of other independent variables. This indicates redundancy in the predictors, leading to instability in the regression coefficients.

Causes:

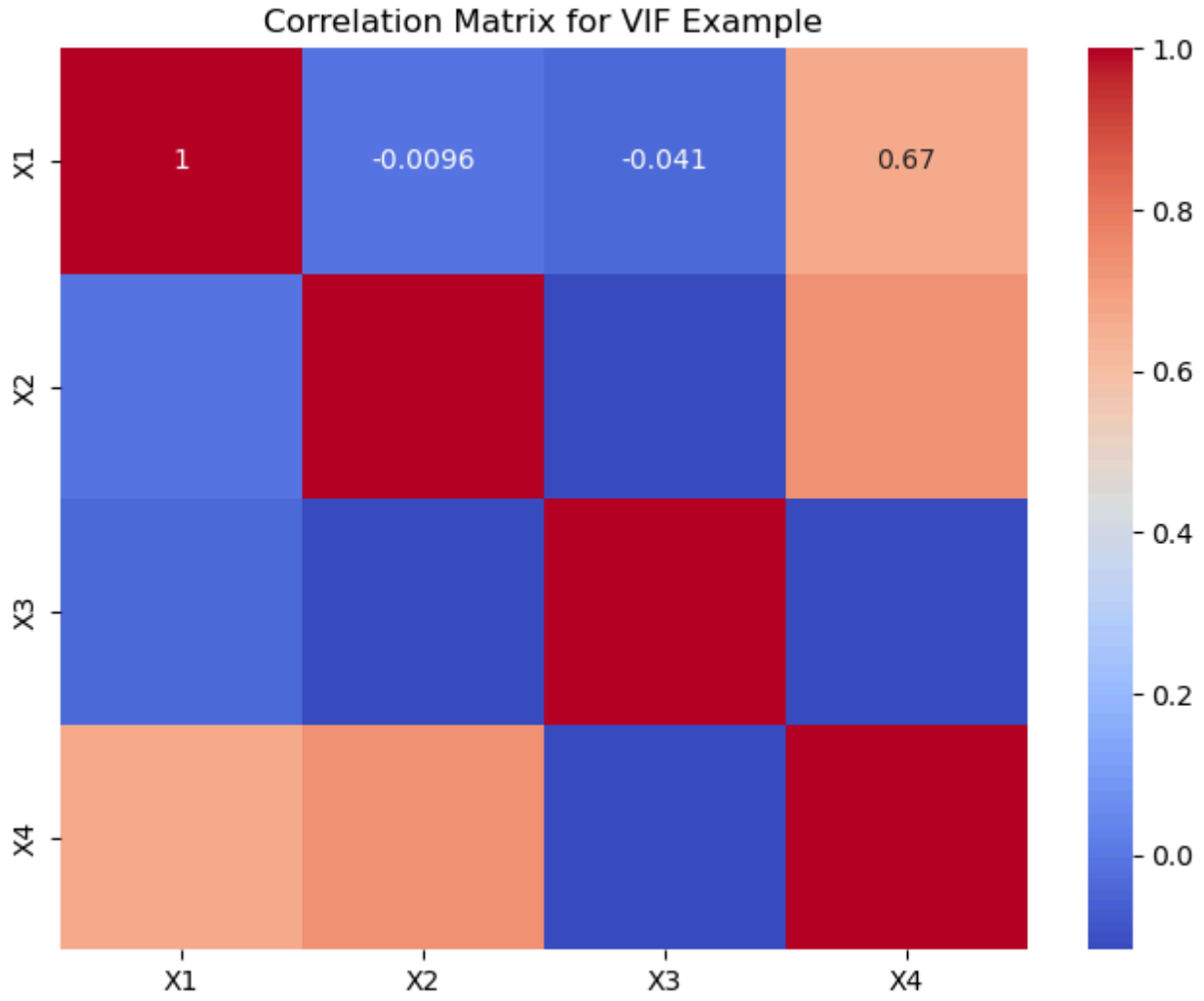
- **Exact Linear Relationship:** If one predictor can be perfectly predicted from others.
- **Duplicate Variables:** Including the same variable more than once.

Consequences:

- **Unstable Coefficients:** High standard errors for the coefficients.
- **Overfitting:** Reduced generalizability of the model.

Solution:

- **Remove One of the Correlated Variables:** Simplifies the model and reduces multicollinearity.
- **Combine Variables:** Use techniques like Principal Component Analysis (PCA) to combine correlated variables into a single feature.



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

Ans: A Q-Q (quantile-quantile) plot is a graphical tool to assess if a dataset follows a particular theoretical distribution, typically the normal distribution. It plots the quantiles of the sample data against the quantiles of the theoretical distribution.

Steps to Create a Q-Q Plot:

- **Sort the Data:** Arrange the sample data in ascending order.
- **Calculate Theoretical Quantiles:** Compute the quantiles of the theoretical distribution.
- **Plot:** Plot the sample quantiles against the theoretical quantiles.

Interpretation:

- **Straight Line:** If the points lie on a straight line, the data follows the theoretical distribution.
- **Deviations:** Systematic deviations from the line indicate departures from the theoretical distribution.

Importance in Linear Regression:

- **Normality of Residuals:** In linear regression, one of the assumptions is that the residuals are normally distributed. A Q-Q plot of the residuals helps in assessing this assumption.
- **Model Diagnostics:** Identifies potential outliers and deviations from normality, guiding further model refinement.

Conclusion: A Q-Q plot is a valuable diagnostic tool in linear regression to ensure that the residuals meet the assumption of normality, which is crucial for valid inference and reliable model predictions.