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:
   Based on the boxplots you provided for the categorical variables, here are the inferences about their effect on the dependent variable cnt:

   1. Season: Summer and Fall seasons appear to have higher median cnt values compared to Spring and Winter.

   Winter has the lowest median cnt and also shows some outliers. This suggests that the season has a significant effect on the dependent variable, with more rentals occurring in warmer months.

   2. Holiday: There doesn't seem to be a significant difference in the cnt distribution between holidays and non-holidays, although the median is slightly higher on holidays. This suggests that holidays may not have a major effect on bike rentals.
   3. Working Day: The cnt distribution appears similar between working days and non-working days. This implies that whether it's a working day or not may not have a large impact on the rentals.
   4. Weather Situation (weathersit): Weather Situation A (likely representing clear or good weather) has the highest median cnt, while Weather Situation C (possibly indicating bad weather like rain or snow) has the lowest. This is expected, as poor weather conditions typically reduce bike rentals.
   5. Month: There is a clear seasonal pattern, with the highest median cnt in the months of July to September. Winter months (December, January, February) have lower median cnt values. This reinforces the earlier inference that warmer months are associated with higher bike rentals.
   6. Weekday: The day of the week doesn't show much variation in the cnt distribution. This suggests that rentals are fairly consistent across different weekdays, with perhaps slight increases or decreases depending on the day.
   7. Year: The year 2019 has a significantly higher median cnt compared to 2018. This could indicate growth in bike rentals over time, possibly due to increased popularity or changes in the population or infrastructure.
      In summary,
      - Season, weather situation, and month have a noticeable impact on the dependent variable cnt, with warmer months and clear weather being associated with higher bike rentals.
      - Year also shows an upward trend, suggesting increasing rentals over time.
      - Holiday, working day, and weekday do not seem to have a strong impact on bike rentals.

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

Ans: Using `drop_first=True` when creating dummy variables is important because it helps prevent multicollinearity in linear regression models. By dropping the first category, we avoid perfect collinearity among the dummy variables, which keeps the model stable and the coefficients interpretable relative to the baseline category.

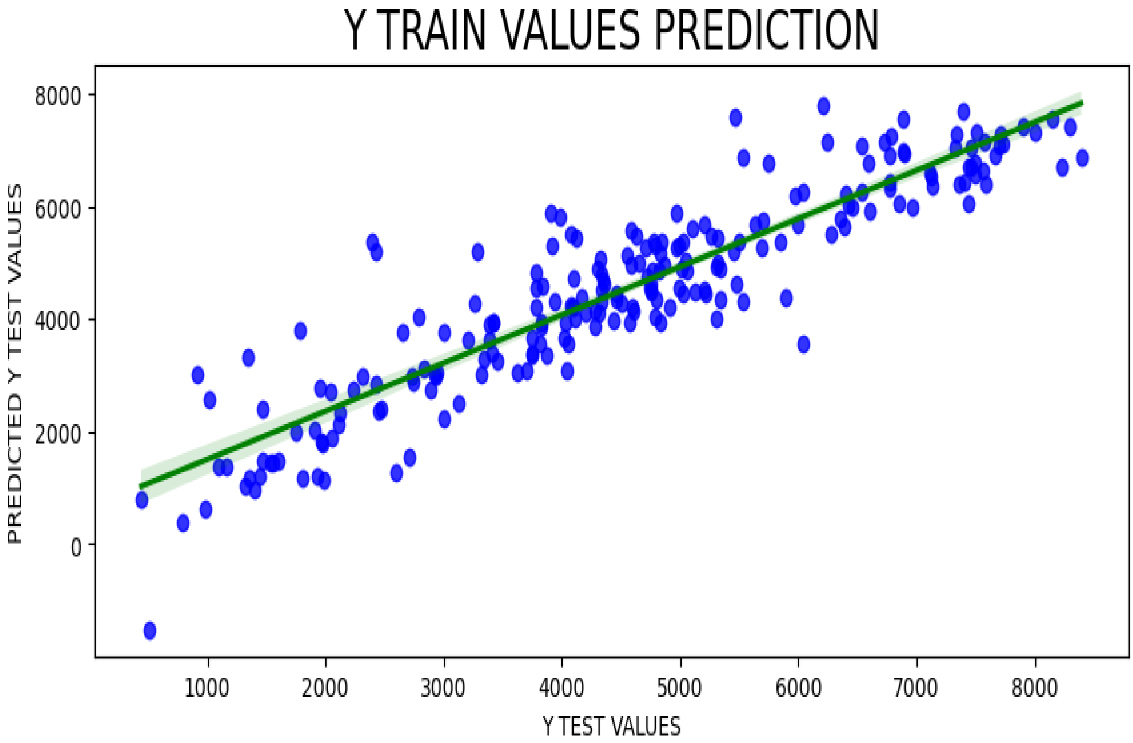
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

   Ans: Among numerical variables, registered has the highest correlation with the target variable which is cnt (corr between cnt and registered is 0.945411)

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   Ans: Key Assumptions of Linear Regression are

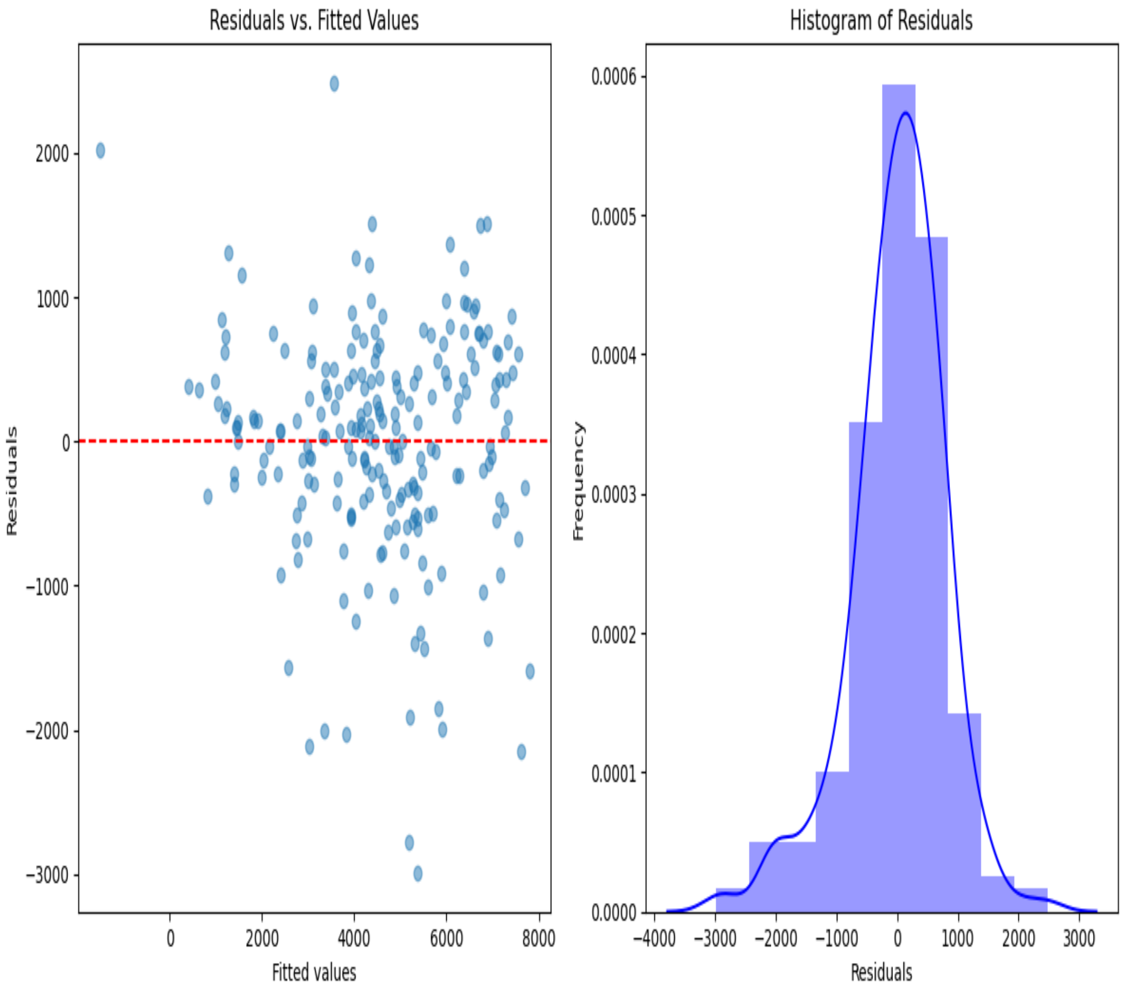
   - Linearity: The relationship between independent and dependent variables should be linear.
   - Independence: Errors are independent.
   - Homoscedasticity: Constant variance of errors.
   - Normality: Errors should be normally distributed.
   - No Multicollinearity: Independent variables should not be highly correlated.

   Linearity: Plotted residual plot with the predicted vs actual values.



. Y test values and Y predicted values have shown a strong visual semblence and hence our predicted are evaluated as a healthy fit. We do have presence of outliers, however, on a broader scale, most of the points display a healthy fit by following trend

Homoscedasticity and Normality:



Homoscedasticity:

- Mean of the residuals is exactly close to zero
- For smaller values of predictions, the residuals are found to be on higher side and focussed above the mean value of 0. However, as we move further to larger predictions.
- This validates assumptions of normal distribution of errors around mean or 0.

Normality:
Error terms seem to be approximately normally distributed, so the assumption on the linear modelling fulfilled.

Multicollinearity:
Done Variance Inflation Factor (VIF) and VIF < 5 for the

	feature	VIF
0	Temp	1.272208
1	hum	1.857075
2	Windspeed	1.179401
3	holiday	1.015213
4	season_Summer	1.192863
5	season_Winter	1.253670
6	yr_2019	1.027545
7	mnth_Sep	1.114846
8	weathersit_Light Rain or Snow	1.232582
9	weathersit_Misty and Cloudy	1.546132

Hence every assumption of linear regression is validated.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
Ans: the top three features contributing significantly are
1. weathersit_Light Rain or Snow
2. yr_2019
3. season_Winter

- Most important factor is light rain or snow with a efficient of 2518.148049 hence if a particular day has plight rain it is expected to reduce the demand by percentage from our EDA section that frequency of light rain has been lowest and heavy rains have largely been not existed October month shows the monsoon maximum number of Instances of light rains

- Second most important factor is year with a coefficient value of 2080.009542 based on the historical data given all internal and external factors remains unseen the company is expected to be a normal growth over last year at around 2080.009542 units. This helps us in factoring revenue and cost projection over a period of time

- Third most important factor is winter season which has a coefficient of 1268.277950 this signifies that every winter demand is expected to increase by the factor of 1268.277950 based on other months. Recall from our EDA technique winter season constitutes of September October November and December month company needs to work on capacity planning of these months and accordingly plan for promotional campaigns in case of a competitor exists with smaller value preposition

General Subjective Questions

1. ***Explain the linear regression algorithm in detail.***

DEFINITION:
Linear regression is a fundamental statistical method used to model the relationship between a dependent variable and one or more independent variables. It is widely used for prediction and forecasting, and it is the foundation for more complex statistical methods.

1. Purpose of Linear Regression: The primary goal of linear regression is to predict the value of a dependent variable (also called the target variable) based on the values of one or more independent variables (also called predictor variables).
   It assumes a linear relationship between the dependent and independent variables.
2. Types of Linear Regression:
   a. Simple Linear Regression: Involves one dependent variable and one independent variable.
   b. Multiple Linear Regression**:** Involves one dependent variable and two or more independent variables.

The Linear Regression Model:

The general form of a linear regression equatio is: [y= beta_0 + beta_1x_1 + beta_2x2 + …….+ beta_nx_n + epsilon]

y: Dependent variable (what you are trying to predict)

x_1, x_2, ..., x_n: Independent variables (the predictors)

beta_0: The intercept (value of y when all x's are 0)

beta_1, beta_2, ..., beta_n: The coefficients of the independent variables (these represent the change in y for a one-unit change in the corresponding x)

epsilon: The error term (captures the deviation of actual data points from the predicted line)

Assumptions of Linear Regression:

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

Independence: Observations are independent of each other.

Homoscedasticity: The variance of errors is the same across all levels of the independent variables.

Normality: The residuals (errors) of the model are normally distributed.

No Multicollinearity: In multiple regression, the independent variables should not be highly correlated with each other.

Fitting the Model:

Ordinary Least Squares (OLS): The most common method used to estimate the coefficients ((\beta_0, \beta_1, ..., \beta_n)). OLS minimizes the sum of the squared differences between the observed values and the values predicted by the model: [\text{Minimize } \sum_{i=1}^{n} (y_i - \hat{y}_i)^2]

Here, (y_i) is the observed value, and (\hat{y}_i) is the predicted value from the model.

Evaluating the Model:

R-squared (R^2): Measures the proportion of the variance in the dependent variable that is predictable from the independent variables. Values range from 0 to 1, with higher values indicating a better fit.

Adjusted R-squared: Adjusted for the number of predictors in the model, providing a more accurate measure in the context of multiple regression.

Mean Absolute Error (MAE): The average of the absolute differences between actual and predicted values.

Mean Squared Error (MSE): The average of the squared differences between actual and predicted values.

Root Mean Squared Error (RMSE): The square root of MSE, representing the average distance between the observed and predicted values.

Interpreting the Coefficients:

Each coefficient represents the change in the dependent variable for a one-unit change in the corresponding independent variable, holding all other variables constant.

The sign (+/-) of the coefficient indicates the direction of the relationship between the predictor and the outcome.

Assumption Checking:

Residual Plots: Used to check the assumptions of linear regression. For example, residuals vs. fitted values plot can help in identifying heteroscedasticity.

Q-Q Plot: Assesses whether residuals are normally distributed.

Variance Inflation Factor (VIF): Measures multicollinearity; VIF values greater than 10 suggest high multicollinearity.

Advantages and Disadvantages:

Advantages:

- Simple to implement and interpret.
- Provides insights into relationships between variables.
- Computationally efficient.

Disadvantages:

- Sensitive to outliers, which can distort results.
- Assumes linear relationships, which may not hold in all cases.
- Requires careful checking of assumptions for reliable results.

Applications of Linear Regression:

- **Predictive Modeling:** Used to predict continuous outcomes, like house prices, sales, etc.
- **Trend Analysis:** Identifying trends over time.
- **Risk Assessment:** Used in finance to assess risk factors and their relationships to outcomes.

In summary, linear regression is a powerful tool for modeling relationships between variables. However, it requires careful checking of assumptions, proper evaluation of the model fit, and consideation of potential limitations like outliers and multicollinearity.ration of potential limitations like outliers and multicollinearity.

2. ***Explain the Anscombe's quartet in detail.***

Anscombe's Quartet is a set of four distinct datasets that have nearly identical simple descriptive statistics, such as mean, variance, correlation, and linear regression lines. However, when these datasets are visualized, they reveal vastly different distributions and relationships between the variables. The quartet was constructed by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analyzing it and to show how different datasets with similar statistical properties can have different structures.

Key Points of Anscombe's Quartet:

Identical Statistical Properties:

All four datasets have the same mean for the x and y values.

They have the same variance for x and y.

Each dataset has the same correlation coefficient between x and y.

The linear regression line (y = mx + c) is nearly the same for all four datasets.

Visual Differences:

Dataset 1: Shows a typical linear relationship between x and y, which would be expected based on the regression analysis.

Dataset 2: The data is more curvilinear, indicating a non-linear relationship despite the linear regression line.

Dataset 3: Contains an outlier, which heavily influences the regression line, leading to misleading interpretations if only the statistics are considered.

Dataset 4: All x-values are the same except for one, creating a vertical line. The single differing point (an outlier) forces the regression line to fit in a misleading way.

Importance of Anscombe's Quartet:

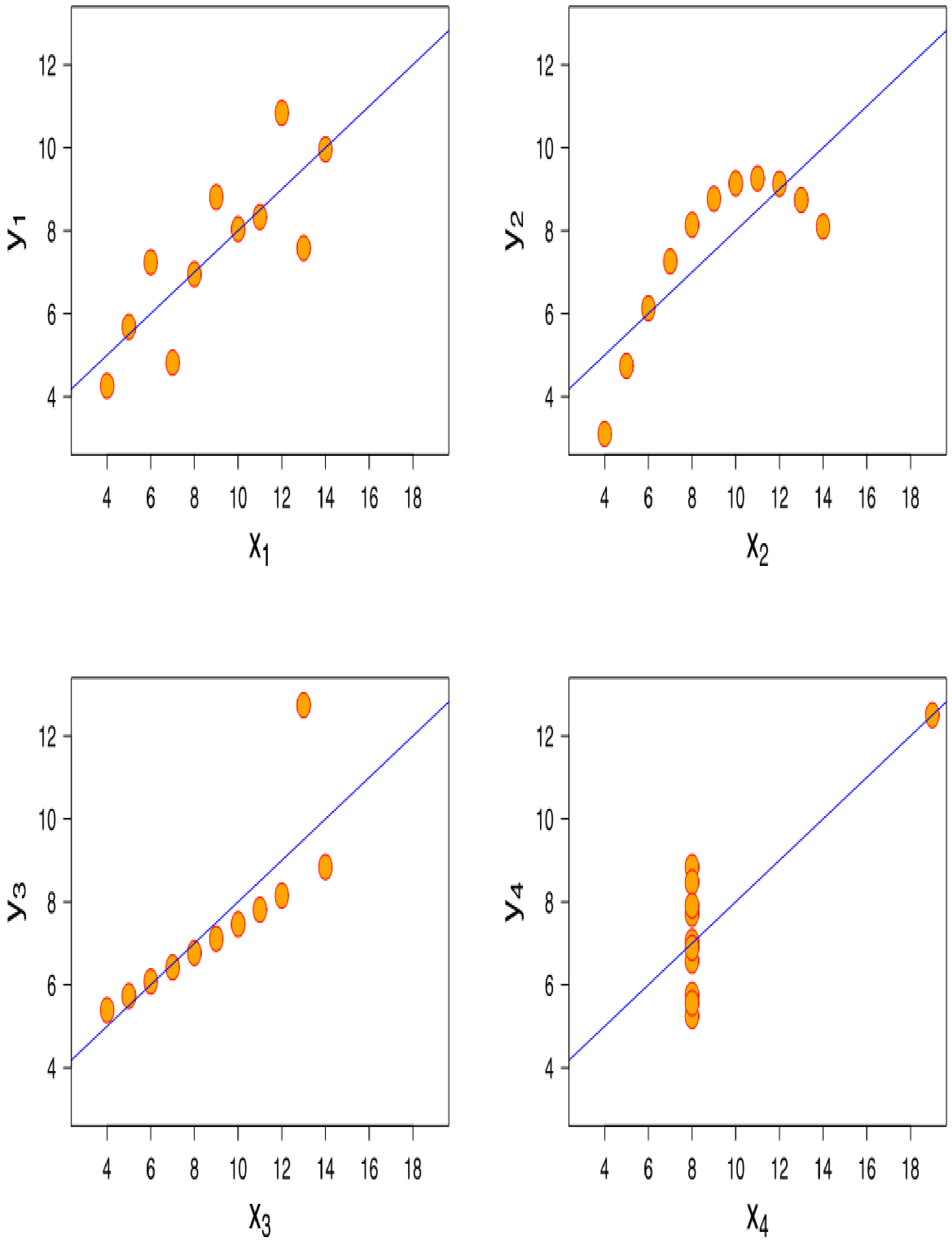
Visual Exploration: The quartet illustrates the crucial role of visualizing data before jumping to conclusions based on summary statistics. By plotting the data, one can identify patterns, outliers, or structures that simple statistics might miss.

Misleading Conclusions: If one relies solely on statistical summaries without visual inspection, one might draw incorrect or oversimplified conclusions about the data.

Teaching Tool: Anscombe's Quartet is widely used in statistics education to teach the importance of graphical analysis and to caution against the over-reliance on summary statistics.

Practical Takeaway:

Even though two datasets might have identical statistical properties, their underlying distributions and relationships can be entirely different. Always visualize your data to understand its true nature before relying solely on statistical summaries or models.



3. What is Pearson’s R?

Pearson’s R in Linear Regression is a measure of the linear correlation between two variables, typically the independent variable (predictor) and the dependent variable (outcome). It quantifies the strength and direction of the linear relationship between these two variables.

Explanation:

Definition

Pearson's R, also known as the Pearson correlation coefficient, is a statistic that ranges from -1 to 1. A value of 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship between the variables.

Calculation

The Pearson’s R is calculated as the ratio of the covariance of the two variables to the product of their standard deviations. Mathematically, it is expressed as:

[R = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}]

where:

(\text{Cov}(X, Y)) is the covariance of the variables (X) (independent) and (Y) (dependent),

(\sigma_X) and (\sigma_Y) are the standard deviations of (X) and (Y), respectively.

Importance in Linear Regression:

In the context of linear regression, Pearson's R helps to understand how well the independent variable predicts the dependent variable. A high absolute value of R indicates that the independent variable is a good predictor of the dependent variable, which supports the linear regression model. However, it’s important to remember that correlation does not imply causation; a strong correlation doesn’t necessarily mean that one variable causes change in the other.

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

Scaling in Machine Learning:

Scaling refers to the process of adjusting the range of features in your dataset so that they can be compared on a common scale. This is particularly important in algorithms where the distance between data points or the magnitudes of features matter, such as in gradient descent optimization, support vector machines, or k-nearest neighbours.

Why Scaling is Performed:

Scaling is performed to:

Improve Model Performance: Algorithms like gradient descent converge faster when features are scaled because the optimization process is more efficient when the features are within the same range.

Ensure Equal Contribution: Scaling ensures that all features contribute equally to the model. Without scaling, features with larger ranges could disproportionately influence the model's predictions.

Enhance Interpretability: It makes it easier to interpret the results of the model, particularly in distance-based algorithms.

Difference Between Normalized Scaling and Standardized Scaling:

Normalized Scaling:

Definition: Normalization scales the data to a fixed range, typically [0, 1] or [-1, 1]. It adjusts the values to be within a specific range without affecting the relative differences between data points.

Formula:

[X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}]

Use Case: Normalization is useful when you need to bound the values within a certain range, for instance, in algorithms that need bounded inputs like neural networks.

Standardized Scaling:

Definition: Standardization scales the data based on the mean and standard deviation, resulting in features with a mean of 0 and a standard deviation of 1.

Formula:

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

where (\mu) is the mean and (\sigma) is the standard deviation of the feature.

Use Case: Standardization is preferred when features have different distributions or when the model assumes that the data is normally distributed (e.g., in linear regression or logistic regression).

In summary, normalization constrains data within a specific range, making it useful for bounded inputs, while standardization adjusts data based on statistical properties, making it more suitable when the data's distribution is important.

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

Infinite VIF Value:

The 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 the presence of multicollinearity among the independent variables. A VIF value becomes infinite when perfect multicollinearity is present, meaning that one independent variable is an exact linear combination of one or more other independent variables.

Cause of Infinite VIF:

An infinite VIF occurs when the correlation between one independent variable and a combination of the other independent variables is perfect (correlation coefficient of 1 or -1). In this case, the model cannot distinguish between the perfectly correlated variables, and the variance of the affected variable's coefficient is infinitely inflated. Mathematically, this happens when the determinant of the matrix (X'X) used to compute VIF is zero, leading to a division by zero.

Implication and Solution:

Implication: An infinite VIF indicates severe multicollinearity, which can destabilize the regression coefficients, making them highly sensitive to changes in the model. This undermines the reliability of the model and can lead to incorrect interpretations.

Solution: To address infinite VIF, you may need to:

Remove one of the perfectly correlated variables from the model.

Combine the correlated variables into a single feature through techniques like Principal Component Analysis (PCA).

Rethink the model design to avoid including redundant predictors.

In summary, an infinite VIF signifies that one variable is a perfect linear combination of others, leading to perfect multicollinearity, which must be addressed to ensure the stability and reliability of the regression model.

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

Definition:
A Q-Q plot is a scatter plot where the quantiles of a dataset are plotted against the quantiles of a theoretical distribution (e.g., normal distribution). If the data follows the theoretical distribution, the points on the Q-Q plot will approximately lie on a straight line.

Use of a Q-Q Plot in Linear Regression:

1. Assessing Normality of Residuals:

Purpose: One of the assumptions of linear regression is that the residuals (errors) of the model should be normally distributed.
Application: By plotting the residuals of a linear regression model on a Q-Q plot, you can visually check if they approximate a normal distribution. If the residuals are normally distributed, the points on the Q-Q plot should lie roughly along a straight line.

2. Detecting Deviations from Normality:
Purpose: Deviations from normality in residuals can indicate problems with the model, such as non-linearity, heteroscedasticity, or outliers.
Application: If the Q-Q plot shows significant deviations from the straight line, it suggests that the residuals do not follow a normal distribution, potentially impacting the validity of the regression model's inference.

Importance in Linear Regression:

Model Validity:
Checking the normality of residuals is crucial for valid hypothesis testing and confidence interval estimation in linear regression. Non-normal residuals may affect the reliability of the model's statistical inferences.

Model Diagnostics:
A Q-Q plot is a diagnostic tool that helps in understanding if transformations or alternative modelling approaches are needed to meet the assumptions of linear regression.
Summary

A Q-Q plot is a vital diagnostic tool in linear regression used to check if residuals follow a normal distribution, which is an important assumption of linear regression. It helps validate the model's assumptions and guides necessary adjustments to improve the model's accuracy and reliability.