# Bike Sharing Assignment

## 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)**

- There were more reservations made in 2019 than there were in 2018, which suggests that business is positive.
- Bookings for the fall season have increased significantly. In addition, there has been a significant increase in booking counts from 2018 to 2019 in all seasons.
- The number of reservations is typically smaller on non-holiday days, which makes sense given that some people would rather spend the holidays at home with their families.
- It's obvious that having clear whether contributed significantly to the increase in Booking.
- The distribution of bookings between working and non-working days seems to be about equal.
- When it came to Bookings, Thursday, Friday, Saturday, and Sunday were busier than the first few days of the week.
- The months of May, June, July, August, September, and October saw the most bookings. From the start of the year until about midway through, the trend showed an increase, and then near the conclusion of the year, it showed a decline.

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

A method for representing categorical variables with binary values (0 or 1) in statistical modeling and machine learning is the construction of dummy variables.

For every category in the original categorical variable, additional binary (dummy) variables must be created. These dummy variables function as markers for the existence or non-existence of a particular category.

The number of categories in the original category variable determines how many dummy variables need to be made. Usually, you create $n - 1$ dummy variables for a categorical variable with $n$ categories.

Here's the rationale:

**1. $n$ – 1 Dummy Variables Rule:**

Multicollinearity is introduced by creating $n$ dummy variables since the data in one category can be anticipated from the others. Regression model coefficient estimation may become problematic as a result. You can avoid perfect multicollinearity by constructing dummy variables with a value of $n - 1$, as this allows for the implicit collection of information about the omitted category.

**2. Avoiding Redundancy:**

The constant term in the model implicitly captures the information about the omitted category. Redundancy would be introduced if all dummy variables were included.

**3. Enhancing Interpretability:**

The answer variable's change in relation to the excluded category is represented by the coefficients of the dummy variables. This facilitates a more straightforward interpretation of the model.

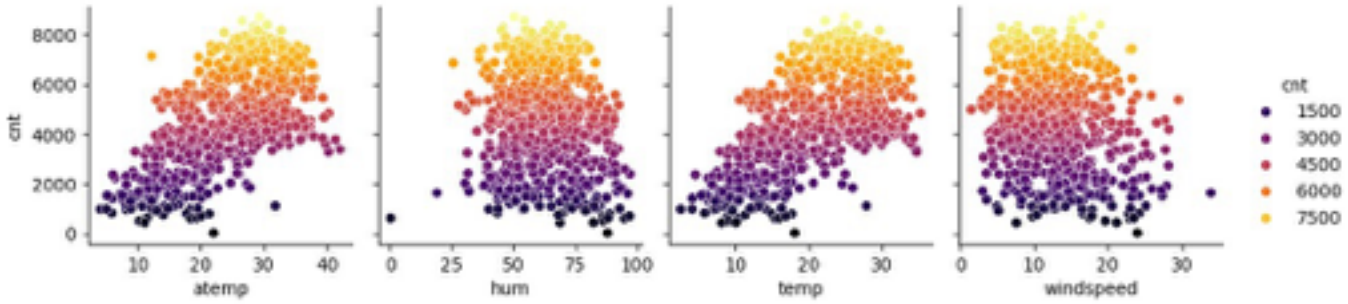
For example, if you have a variable "Colour" with categories "Red," "Blue," and "Green," you would create two dummy variables, say "Is_Blue" and "Is_Green." The absence of both dummy variables implies that the colour is "Red."

When Utilizing pandas libraries in Python, we may specify `drop_first=True} to automatically drop one of the dummy variables in order to follow the $n - 1$ rule when constructing dummy variables.

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

**Answer:**

The graph below shows that the variable "temp" has the strongest correlation with the target variable. Only one of the duplicate variables, "atemp" and "temp," is chosen for determining the best fit line.



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

**Answer:**

One of the most important steps in ensuring the model's reliability is to validate the assumptions of linear regression. Following the model's construction on the training set, I took the following actions to verify my assumptions:

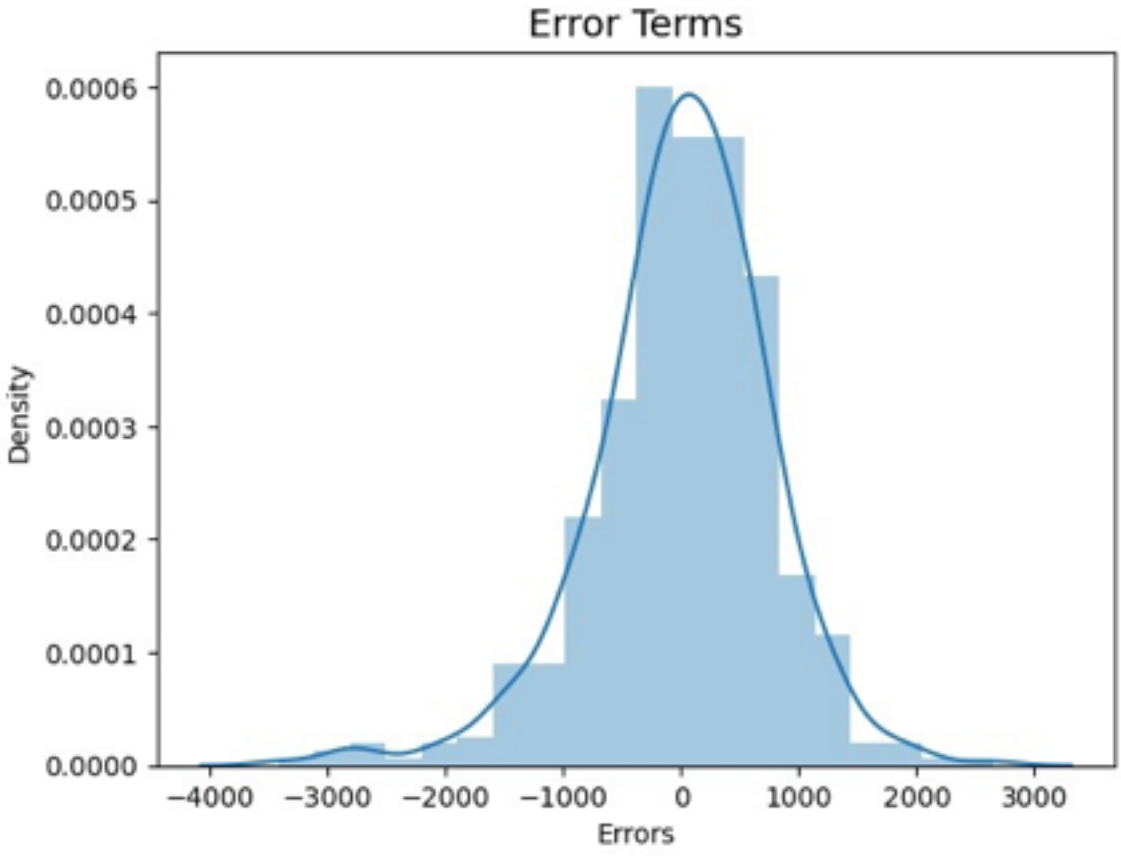
**1. Residual Analysis:**

- Process: Examine the residuals (the differences between observed and predicted values).

- Check: Residuals should be approximately normally distributed, and there should be no discernible paKerns in the residual plot.

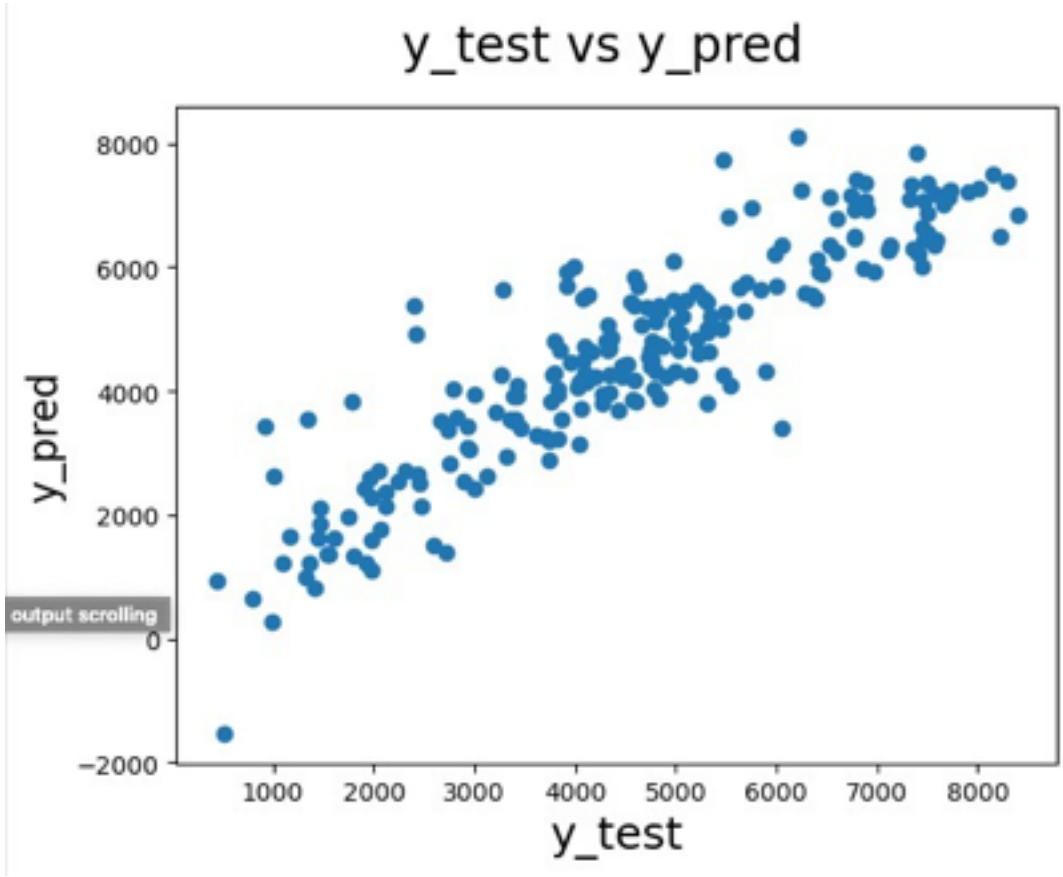
- The training set's error terms are regularly distributed .

- Mean is near to 0, the model has a constant variance i.e. Homoscedasticity.



**2. Linearity:**

- Method: Produce a scatterplot showing the actual values compared to the predictions.




**3. Independence of Residuals:**

Procedure: Look for autocorrelation in the residuals.

Verify: When the residuals are plotted against time or other pertinent factors, there shouldn't be any observable pattern.

**4. Cross-Validation:**

- Process: Validate the model on a test set or through cross-validation.

- Check: Assess the model's performance on new data to ensure generalizability and consistency.

**5. Check for Overfitting:**

- Process: Evaluate model performance on a test set.

- Check: Ensure that the model generalizes well to new, unseen data without overfifng the training set.

**6. Multicollinearity: -**

Procedure: Determine Variance Inflation Factors (VIF) for the variables that are predictors.

Verify: VIF readings must fall below a predetermined level (usually 5 or 10) to guarantee that multicollinearity is not an issue.

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

**Answer:**

From the equation of the best fit line:

cnt = 1763.3081 -521.0958 x holiday -1427.8111 x hum + 4934.2029 x temp -1679.0976 x windspeed + 381.3737 x workingday + 1993.5720 x yr + 648.1735 x season_Summer + 1100.2682 x season_Winter -340.3520 x mnth_January -387.1024 x mnth_July + 804.8554 x mnth_September + 469.5340 x weekday_Sunday -2126.2583 x weathersit_Light Rain/Snow -483.5670 x weathersit_Mist/Cloudy

The following three features significantly contribute to explaining the demand for shared bikes:

- Temperature (temp)
- Winter season (winter)
- Calendar year (year)

# General Subjective Questions

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

A statistical technique for simulating the relationship between a dependent variable and one or more independent variables is called linear regression. Predicting the value of the dependent variable from the values of one or more independent variables is a common application for it. The fundamental concept is identifying the optimal line (or hyperplane in the event of several independent variables) that minimizes the total squared differences between the dependent variable's observed and predicted values.

The steps of a linear regression algorithm are as follows:

## 1. Model Representation:

**Simple Linear Regression:** For a single independent variable, the model is written as follows:

represented as:

$y = b_0 + b_1 \cdot x + \varepsilon$

where:
- $y$ is the dependent variable,
- $x$ is the independent variable,
- $b_0$ is the y-intercept (constant term), - $b_1$ is the slope of the line, and
- $\varepsilon$ represents the error term.

**Multiple Linear Regression:** A statistical method called multiple linear regression (MLR), or just multiple regression, makes use of many explanatory variables to forecast the value of a response variable.

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots\ldots + \beta_p x_{ip}$

where, for i=n observations:

Yi=dependent variable

xi=explanatory variables

β0 = y-intercept (constant term)

βp = slope coefficients for each explanatory variable

ϵ = the model's error term (also known as the residuals)

**2. Objective Function:**

The aim is to determine the values of ($b_1$, $b_2$, $b_3$,..., $b_n$) that reduce the total squared deviations between the values that are seen and those that are anticipated. The mean squared error (MSE) or sum of squared errors (SSE) are common ways to describe this:

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

where ( $m$ ) is the number of data points,

($y_i$) is the observed value,

and ($\hat{y}_i$) is the predicted value.

**3. Minimization:**

The procedure use gradient descent and other optimization techniques to determine the ideal values for the coefficients. Iteratively updating the coefficients in the direction of cost function minimization is the aim.

**4. Training the Model:**

The values of the coefficients that best suit the data are discovered by the algorithm during the model's training on a dataset. Input-output feeding of the algorithm is required here, where:

Updating and modifying the coefficients until the model generates forecasts that are somewhat close to the real results.

**5. Prediction:**

After training, the model can be used to new, unobserved data sets to generate predictions. By entering the new input values into the regression equation that was learned, the predicted values are obtained.

**6. Evaluation:**

The model's performance is assessed using metrics such as ($R^2$) (coefficient of determination), MSE, or other relevant metrics, depending on the context.

**7. AssumpJons:**

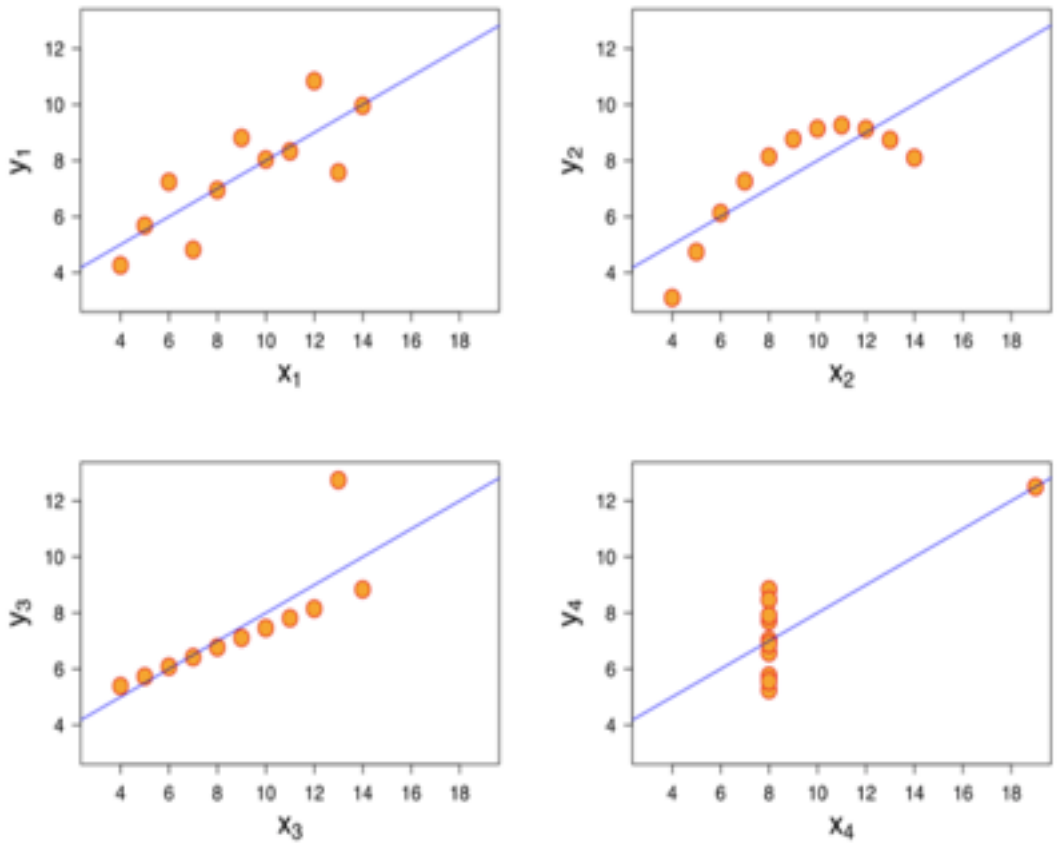
Linear regression relies on the assumption of a linear relationship between independent and dependent variables, normally distributed errors, constant error variance (homoscedasticity), and the absence of perfect multicollinearity, ensuring that there is no perfect linear relationship among the predictors.

Linear regression is a versatile and widely used algorithm, but it's important to check whether its assumptions hold in each dataset and consider more advanced techniques when those assumptions are violated.

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

Four datasets make up Anscombe's Quartet; they have nearly identical basic descriptive statistics, but their graphs show substantial differences. This demonstrates the value of data visualization and the drawbacks of depending only on summary statistics. This quartet emphasizes the idea that, when graphed, datasets with comparable statistical features can show a variety of patterns. The statistician Francis Anscombe created them in 1973 to illustrate the significance of charting data during analysis as well as the impact of outliers and other noteworthy findings on statistical features.

---




*Graphical representation of Anscombe's quartet*

All four sets are identical when examined using simple summary statistics but vary considerably when graphed.

From the above diagram:

- The initial scatter plot (top left) suggests a straighforward linear relationship, depicting two correlated variables, where y could be characterized as Gaussian with a mean linearly dependent on x.

- In the second graph (top right), although a relationship between the variables is evident, it is not linear, rendering the Pearson correlation coefficient irrelevant. A more general regression and the corresponding coefficient of determination would be more suitable.

- Moving to the third graph (bottom left), the modelled relationship is linear, but a different regression line is warranted (considering a robust regression). The calculated regression is skewed by a single outlier, significantly reducing the correlation coefficient from 1 to 0.816.

- Lastly, the fourth graph (bottom right) exemplifies a scenario where a lone high- leverage point can yield a high correlation coefficient, even when the other data points fail to indicate any relationship between the variables.

The datasets are as follows. The *x* values are the same for the first three datasets.

| I | | II | | III | | IV | |
|---|---|---|---|---|---|---|---|
| *x* | *y* | *x* | *y* | *x* | *y* | *x* | *y* |
| 10.0 | 8.04 | 10.0 | 9.14 | 10.0 | 7.46 | 8.0 | 6.58 |
| 8.0 | 6.95 | 8.0 | 8.14 | 8.0 | 6.77 | 8.0 | 5.76 |
| 13.0 | 7.58 | 13.0 | 8.74 | 13.0 | 12.74 | 8.0 | 7.71 |
| 9.0 | 8.81 | 9.0 | 8.77 | 9.0 | 7.11 | 8.0 | 8.84 |
| 11.0 | 8.33 | 11.0 | 9.26 | 11.0 | 7.81 | 8.0 | 8.47 |
| 14.0 | 9.96 | 14.0 | 8.10 | 14.0 | 8.84 | 8.0 | 7.04 |
| 6.0 | 7.24 | 6.0 | 6.13 | 6.0 | 6.08 | 8.0 | 5.25 |
| 4.0 | 4.26 | 4.0 | 3.10 | 4.0 | 5.39 | 19.0 | 12.5 |
| 12.0 | 10.84 | 12.0 | 9.13 | 12.0 | 8.15 | 8.0 | 5.56 |
| 7.0 | 4.82 | 7.0 | 7.26 | 7.0 | 6.42 | 8.0 | 7.91 |
| 5.0 | 5.68 | 5.0 | 4.74 | 5.0 | 5.73 | 8.0 | 6.89 |

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type

of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

**3. What is Pearson's R? (3 Marks)**

The linear link between two variables is measured by Pearson's correlation coefficient, sometimes known as $r$. It measures how strongly and in which direction two continuous variables are associated linearly. Values between -1 and 1 are accepted by the coefficient,

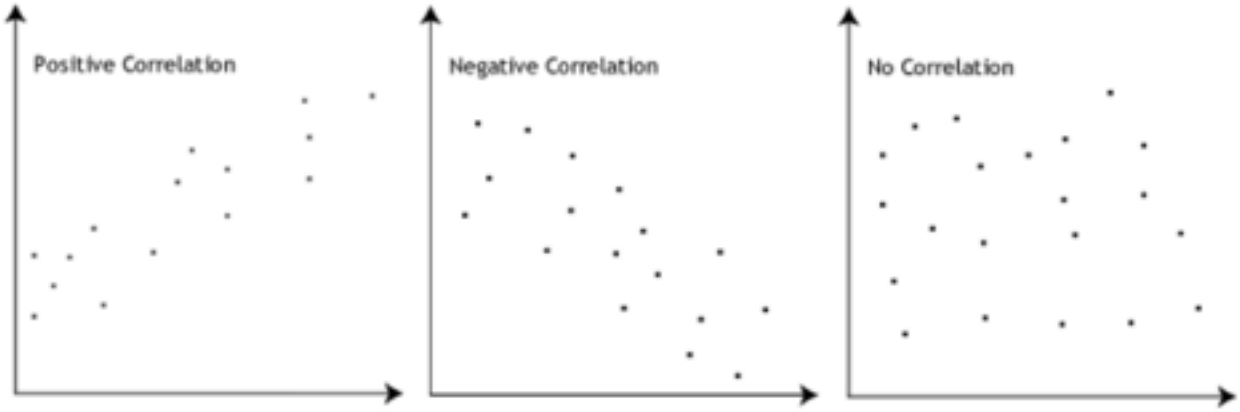
where:

- $r = 1$: Perfect positive linear correlation.
- - $r = -1$: Perfect negative linear correlation.
- - $r = 0$: No linear correlation.

A positive correlation is indicated by a value larger than 0, meaning that as one variable's value rises, the other variable's value also rises.

A negative relationship is shown by a number smaller than 0, meaning that when one variable's value rises, the other variable's value falls.

The following figure illustrates this:




*Correlation Graphs*

The formula for Pearson's correlation coefficient between two variables, $X$ and $Y$, with $n$ data points, is given by:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

- $X_i$ and $Y_i$ are the individual data points.
- $\bar{X}$ and $\bar{Y}$ are the means of $X$ and $Y$.
- The numerator represents the covariance between $X$ and $Y$.
- The denominator is the product of the standard deviations of $X$ and $Y$.

Pearson's correlation coefficient is widely used in statistics to assess the strength and direction of the linear relationship between two variables. It's important to note that correlation does not imply causation, and a correlation coefficient close to zero does not necessarily mean the absence of a relationship; it only indicates the absence of a linear relationship.

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

Scaling is the process of converting variable values to a given range or distribution in the context of data pre-processing. In order to make all the variables comparable and to stop one from dominating the others, the goal is to scale them all to a similar value.

**Advantages of Scaling:**

**Equitable Weightage**: By ensuring that each variable contributes equally to the study, scaling helps to avoid the results of larger-magnitude variables being unduly influenced by them.

**Convergence**: When features are on a comparable scale, many machine learning methods perform better, especially those that rely on gradients or distances (such as k-nearest neighbors, support vector machines, and gradient descent-based algorithms). Scaling helps the Optimization process to converge more quickly.

**Interpretability**: Because the coefficients in linear models show how much the dependent variable changes with a one-unit change in the predictor variable, they are easier to understand.

**Differences between Normalized Scaling and Standardized Scaling:**

**1. Normalized Scaling (Min Max Scaling):**

- Range: Scales the values of a variable to a specific range, usually [0, 1].

Formula:

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

- Advantages: Useful when the distribution of the variable is unknown or not Gaussian.

- Disadvantages: Sensitive to outliers.

**2. Standardized Scaling (Z-score normalizaJon):**

- Mean and Standard Deviation: Scales the values to have a mean of 0 and a standard deviation of 1.

Formula:

$$X_{\text{standardized}} = \frac{X - \text{mean}\,(X)}{\text{std}(X)}$$

- Advantages: Less sensitive to outliers; preserves the shape of the distribution. - Disadvantages: Assumes that the variable follows a Gaussian distribution.

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

A multiple regression analysis's Variance Inflation Factor (VIF) is a metric used to evaluate multicollinearity. It measures the extent to which multicollinearity increases the variance of the calculated regression coefficients. For a variable $X\%$,

the VIF formula is:

$$\mathrm{VIF}(X_i) = \frac{1}{1 - R_i^2}$$

where $R^i$ is the

$R_i^2$ value obtained by regressing $X_i$ against all other independent

variables.

When the value of VIF is infinite, it usually indicates perfect multicollinearity. Perfect multicollinearity occurs when one or more independent variables in a regression model are perfectly correlated (linearly dependent) with other variables. In such cases:

1. There is redundant information - One variable can be expressed as a perfect linear combination of others.
2. Matrix Inversion Issues - In the computation of the VIF, there's an Attempt to invert a matrix, and perfect multicollinearity leads to the matrix being singular (non- invertible).

When the matrix is singular, it means that one or more variables can be predicted exactly from the others, and as a result, the computation of the VIF becomes problematic, leading to an infinite VIF value.

To address this issue, it's crucial to identify and handle multicollinearity in the dataset. This can involve removing one of the perfectly correlated variables, combining them, or using dimensionality reduction techniques.

Addressing multicollinearity not only resolves the infinite VIF problem but also improves the stability and interpretability of the regression model.

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

**Answer:**

A graphical tool called a Q-Q (Quantile-Quantile) plot is used to determine if a dataset conforms to a specific theoretical distribution, like the normal distribution. The quantiles of the observed data and the quantiles of the predicted distribution are compared. The data may be well-represented by the selected theoretical distribution if the points in the Q-Q plot roughly follow a straight line.

**Use and Importance of Q-Q Plot in Linear Regression:**

**Normality Assessment:**

- Use: It's common knowledge in linear regression that the residuals, or the variations between observed and predicted values, follow a normal distribution. Plots with Q-Q are useful for verifying this assumption.

- Importance: The reliability of statistical conclusions drawn from the regression model may be impacted if the residuals exhibit a large departure from normality.

**Identifying Outliers:**

- Use: Points in the Q-Q plot that diverge from the anticipated straight line can be used to identify outliers in the residuals.

- Importance: Data points that the regression model is unable to adequately capture may be indicated by outliers, which can also affect the estimation of model parameters.

**Model Fit Assessment:**

- Use: Q-Q plots give an indication of how well the residuals fit into a normal distribution visually.

- Importance: Deviations from normalcy in residuals may indicate deficiencies in the regression model, which is why a strong model fit is essential for precise predictions.

**Validity of Statistical Tests:**

- Use: The assumption of residual normalcy is significant for performing hypothesis tests or creating confidence intervals.

- Importance: Breaking this assumption might result in erroneous confidence intervals and p-values, which compromise the reliability of statistical conclusions.

**Interpretation of Q-Q Plots:**

The Q-Q plot indicates that the residuals are roughly normally distributed if the points almost exactly follow a straight line.

- A divergence from normalcy is indicated by a deviation from the straight line.

When it comes to evaluating residual normality, spotting outliers, and guaranteeing the accuracy of statistical conclusions, Q-Q plots are an invaluable diagnostic tool in linear regression. They offer an easy-to-use, visual method of verifying the regression model's underlying assumptions.