Assignment-based Subjective Questions

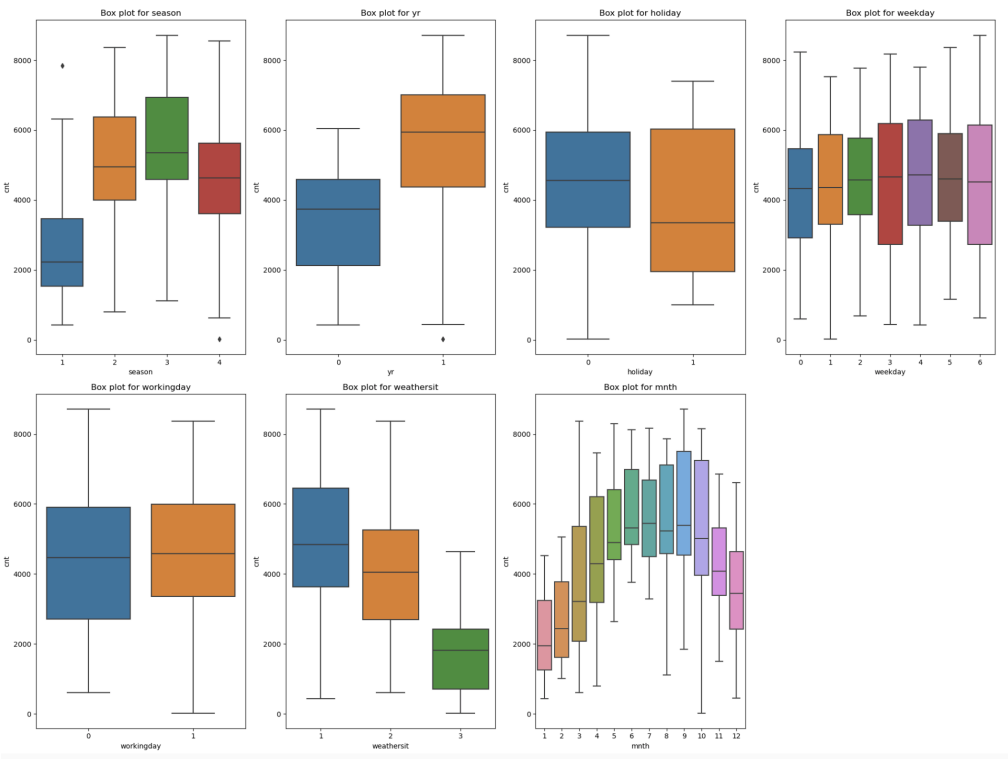
Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

In the dataset, the following categorical variables were used: **season**, **yr** (year), **holiday**, **weekday**, **workingday**, **weathersit** (weather situation), and **mnth** (month). These variables were visualized using boxplots to analyze their impact on the dependent variable, bike demand. Here's a breakdown of the effects:

1. **Season**: For the **season** variable, we observe that **category 3 (Fall)** has the highest median, indicating that bike demand was highest during this season. In contrast, **category 1 (Spring)** has the lowest median, showing the least demand during this time.
2. **Yr (Year)**: The year **2019** shows a higher count of users compared to **2018**, suggesting an increase in bike demand in 2019.
3. **Holiday**: There is a noticeable drop in bike rentals during holidays, as indicated by the boxplot.
4. **Weekday**: The bike demand remains fairly consistent throughout the week, with no significant fluctuation observed across weekdays.
5. **Workingday**: The **Workingday** boxplot reveals that the majority of bookings are concentrated between **4,000 and 6,000** rentals. This suggests that the median number of users is stable throughout the week, and there isn't much variation in bookings between working days and non-working days.
6. **Weathersit (Weather Situation)**: The boxplot shows no rentals during **heavy rain/snow** conditions, indicating that such weather situations are quite adverse for bike rentals. The highest demand is seen when the weather is **Clear** or **Partly Cloudy**.
7. **Mnth (Month)**: Bike rentals peaked in **September**, while **December** saw a decline. This aligns with the weather patterns—**September** is typically a favorable month for bike rentals, whereas **December**, with its significant snowfall, likely caused a drop in demand.
8. Below are box plot for the same:

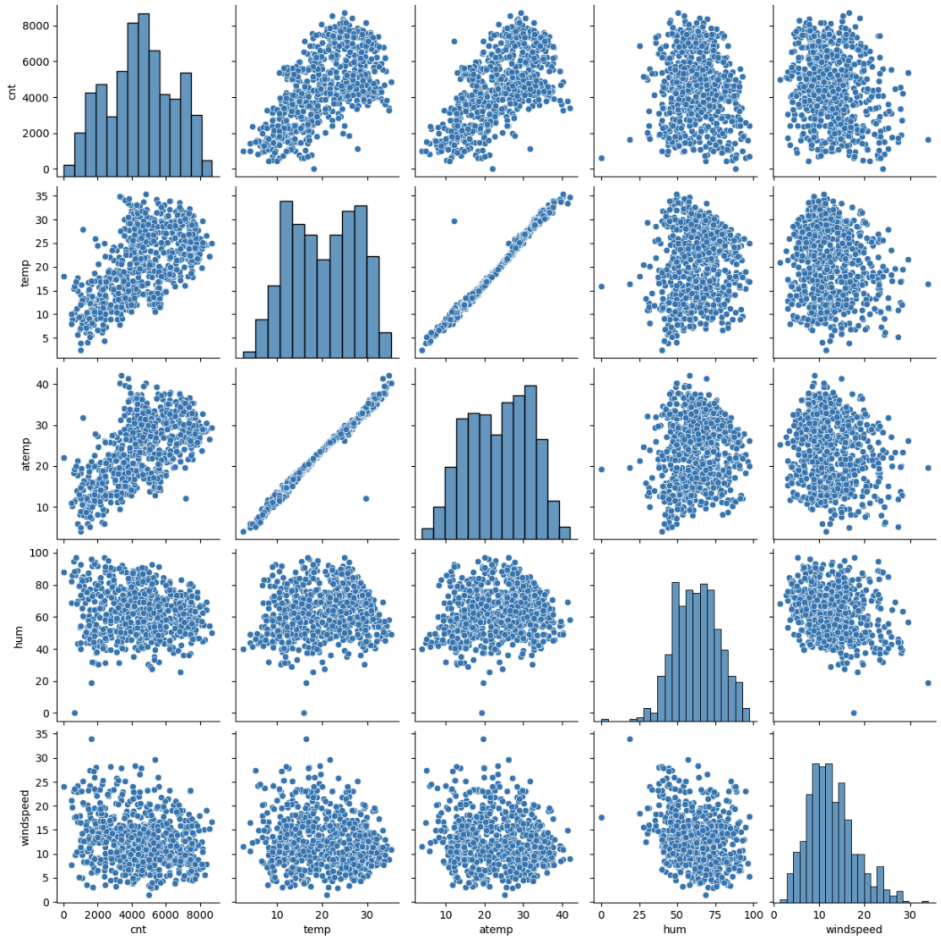


Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)
Using drop_first=True is crucial when creating dummy variables because it helps avoid the creation of an extra column, which in turn reduces multicollinearity among the dummy variables. When we have a categorical variable with **n** levels, we only need **n-1** dummy variables to represent it.

For example, let's consider a categorical column with three possible values. If we create dummy variables for all three values, we would end up with three columns. However, if one of the variables is neither "furnished" nor "semi_furnished," it is implicitly "unfurnished." In this case, we don't need a separate column to represent "unfurnished" because it can be inferred from the absence of the other two variables.

Thus, by using drop_first=True, we drop one of the columns, reducing redundancy and making the model more efficient.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)
From the pair plot, 'temp' and 'atemp' have highest correlation with the target variable.
Below is the plot snapshot:



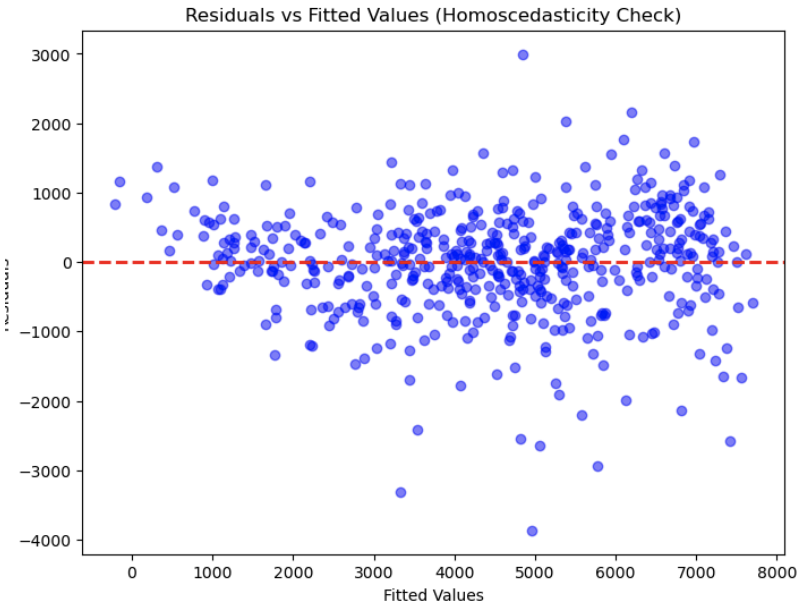
Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

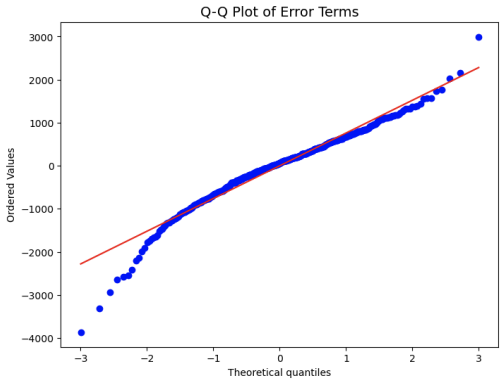
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

After building the linear regression model on the training set, the assumptions of linear regression were validated through the following steps:

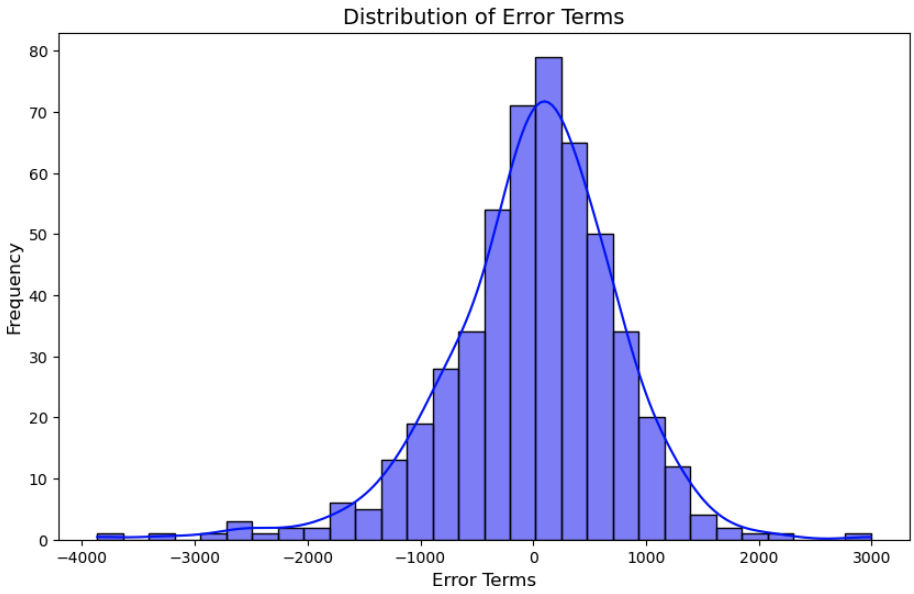
1. **Linearity**: To check the linearity assumption, I plotted the residuals against the predicted values. Ideally, the residuals should be randomly scattered around the horizontal axis with no discernible pattern. A linear pattern in the residual plot would indicate a violation of the linearity assumption.
2. **Homoscedasticity**: Homoscedasticity refers to constant variance of the residuals. I used a residual vs. fitted values plot to visually inspect if the spread of residuals is constant across all levels of fitted values. If the residuals fan out or contract in a systematic way, it indicates heteroscedasticity, which violates this assumption.



3. **Normality of Errors through Q-Q plot**: The residuals should be normally distributed. To check this assumption, I used a **Q-Q plot** (quantile-quantile plot) y. A normal distribution of residuals would imply that the points lie along the reference line in the Q-Q plot.



4. **Normality of Errors through distplot:** To further assure our assumption of normal distribution of Error term, the distplot of residual was plotted and confirmed for normal distribution.



5. **Multicollinearity**: I checked for multicollinearity between predictor variables using the **Variance Inflation Factor (VIF)**. A VIF value greater than 10 suggests high multicollinearity, indicating that some predictor variables are highly correlated with each other. This could lead to instability in the coefficient estimates and affect model performance.

   By performing these diagnostic checks, I ensured that the linear regression assumptions were not violated, thereby improving the reliability and validity of the model.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

Below are the final model coefficients:
const 1482.047663
season_Spring -621.733748
season_Summer 401.178082
season_Winter 600.363076
mnth_Jul -409.339063
mnth_Oct 298.866962
mnth_Sep 690.780061
weekday_Saturday 511.711302
weathersit_Light Snow & Rain -2569.472989
weathersit_Mist & Cloudy -720.918959
yr 2036.473928
holiday -473.836667
workingday 410.497305
temp 4137.213976

windspeed -1310.882841

Since we want the features contributing towards the demand, we will consider top 3 features with with highest positive coefficients:
1. **Temperature (temp)**: The coefficient for temp is **4137.21**, which is the highest among all variables. This indicates that temperature has a strong positive effect on bike demand — as the temperature increases, bike demand is expected to increase significantly.

2. **Year (yr)**: The coefficient for yr is **2036.47**, also a large positive value. This shows that as the year progresses, bike demand increases, suggesting a growing trend in demand for shared bikes over time.

3. **Month of September (mnth_Sep)**: The coefficient for mnth_Sep is **690.78**, indicating a strong positive correlation with bike demand during the month of September. This suggests that bike demand peaks in September, making it a key month for the company to focus on.

These features are the most significant predictors of bike demand according to the model, with **temperature** and **year** showing the largest effects on increasing demand, and **September** standing out as a key month for higher bike rentals.

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Linear regression is a machine learning algorithm based on supervised learning. It performs a regression task, which means it predicts a continuous output variable (y) based on one or more input variables (x). It is mostly used for finding out the linear relationship between variables and forecasting.

The basic idea of linear regression is to find a line that best fits the data points, such that the distance between the line and the data points is minimized. The line can be represented by an equation of the form:
$y = \theta 0 + \theta 1x$
where θ0 is the intercept (the value of y when x is zero) and θ1 is the slope (the change in y for a unit change in x).
These are called the parameters or coefficients of the linear model.

To find the best values of θ0 and θ1, we need to define a cost function that measures how well the line fits the data. A common choice is the mean squared error (MSE), which is the average of the squared differences between the actual y values and the predicted y values:
$MSE = (1/n) * \sum(y - y')^2$
where n is the number of data points, y is the actual value, and y' is the predicted value.
The goal is to minimize the MSE by adjusting θ0 and θ1. There are different methods to do this, such as gradient descent, normal equation, or using libraries like scikit-learn.

Linear regression can also be extended to multiple input variables (x1, x2, …, xn), in which case the equation becomes:
$y = \theta 0 + \theta 1x1 + \theta 2x2 + \ldots + \theta nxn$

Limitations are: it assumes a linear relationship between the input variables and the output variable, which may not always be the case. Another limitation is that it may be sensitive to outliers or multicollinearity.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)
Anscombe's Quartet

Anscombe's Quartet was developed by the statistician **Francis Anscombe** to demonstrate the importance of graphing data before analyzing it. It consists of four datasets that have nearly identical statistical properties, such as the same mean, variance, and correlation, but their distributions and visual appearances are vastly different when plotted. This highlights how

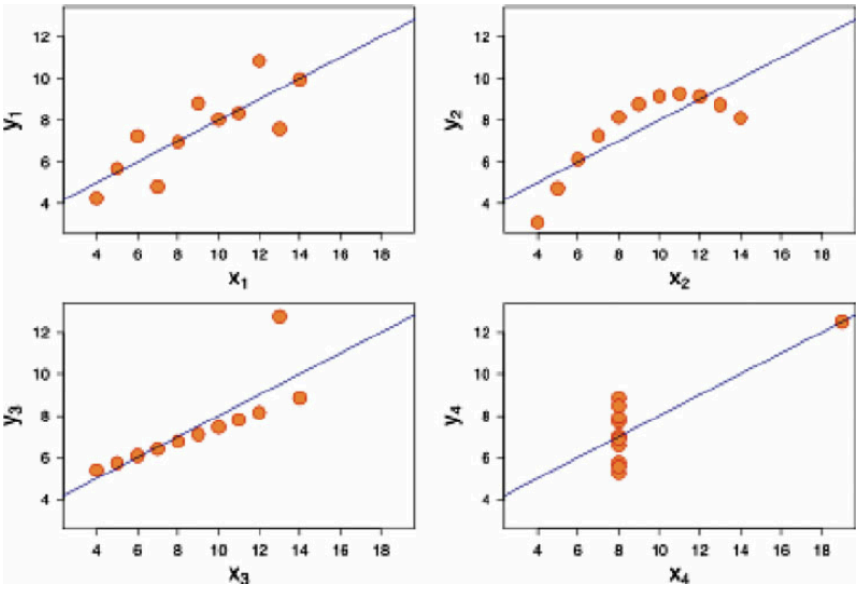
important it is to visualize data in addition to relying on summary statistics, especially when outliers and influential observations are present.

The four datasets in Anscombe's Quartet all have the following statistical properties:

- Mean of x: 9
- Mean of y: 7.5
- Variance of x: 11
- Variance of y: 4.12
- Correlation between x and y: 0.816
- Regression line equation: y=3+0.5x

However, when plotted, the datasets reveal very different patterns and relationships:

1. **First Scatter Plot (Top Left)**:
    - This graph shows a **simple linear relationship** between x and y. The points follow a clear upward trend that is consistent with the linear regression model. It visually confirms the positive correlation and appears to fit well with the regression line.
1. **Second Scatter Plot (Top Right)**:
    - In this plot, the relationship between x and y is **non-linear**. While there is some relationship between the two variables, it does not follow a simple linear trend. Instead, the data might show a curved or more complex form of association that linear regression cannot capture effectively.
1. **Third Scatter Plot (Bottom Left)**:
    - This plot shows a **linear relationship** between x and y, but with a noticeable outlier. The regression line is **influenced by the outlier**, which causes the line to be offset from where it would have been without the outlier. Despite a visually apparent linear trend, the presence of the outlier reduces the correlation coefficient from 1.0 (perfect correlation) to 0.816. This illustrates how a single influential point can affect the overall correlation and regression model.
1. **Fourth Scatter Plot (Bottom Right)**:
    - This plot presents a scenario where a **single high-leverage point** produces a **high correlation coefficient**, even though the remaining data points show no clear relationship between x and y. The regression line in this case is heavily influenced by the high-leverage point, leading to an artificially high correlation, despite the fact that the other data points do not suggest any real relationship. This demonstrates how high-leverage points can distort statistical analysis and lead to misleading conclusions.



Conclusion:

Anscombe's Quartet emphasizes the importance of visually inspecting data before drawing conclusions based solely on summary statistics. It also highlights how **outliers** and **high-leverage points** can significantly affect the results of statistical analyses like correlation and linear regression. The dataset is a reminder that **correlation does not imply causation**, and the relationship between variables can be much more complex than it appears based on simple statistics.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson's r (also known as the **Pearson correlation coefficient**) is a numerical measure that quantifies the strength and direction of the linear relationship between two variables. It ranges from **-1 to +1**, with the following interpretations:

- **r = +1**: Perfect positive linear correlation — as one variable increases, the other also increases in a perfectly linear fashion.
- **r = -1**: Perfect negative linear correlation — as one variable increases, the other decreases in a perfectly linear fashion.
- **r = 0**: No linear correlation — no linear relationship exists between the two variables.

Key Points:

- **Strength**: The closer the value of r iis to **+1** or **-1**, the stronger the linear relationship between the variables.
    - **r = +1** indicates a **perfect positive** linear relationship.
    - **r = -1** indicates a **perfect negative** linear relationship.
    - **r = 0** indicates **no linear relationship** between the variables.
- **Direction**: The sign of r (positive or negative) indicates the **direction** of the relationship.
    - **Positive r** means that as one variable increases, the other also increases.
    - **Negative r** means that as one variable increases, the other decreases.

Interpretation:

Pearson's r is commonly used to answer questions like **"Can we draw a straight line to represent the data?"** It helps determine if there is a **linear association** between two variables and, if so, how strong and in which direction that relationship is.

Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

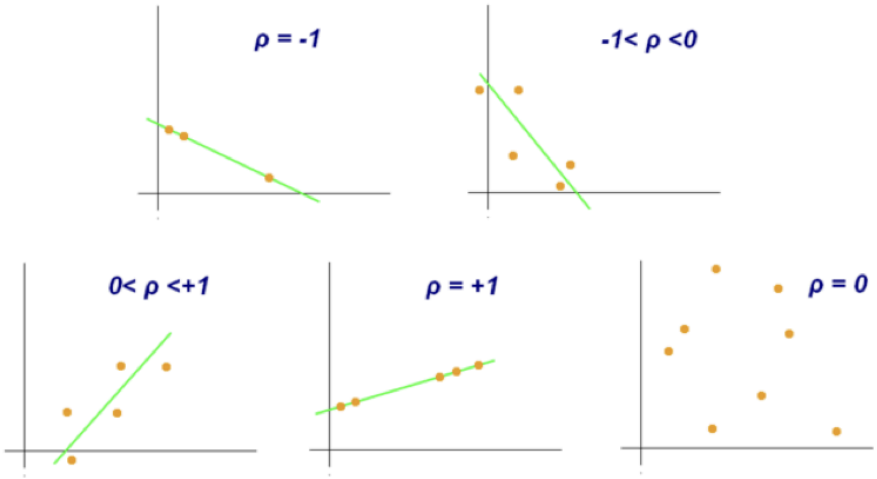
x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

As can be seen from the graph below, r = 1 means the data is perfectly linear with a positive slope r = -1 means the data is perfectly linear with a negative slope r = 0 means there is no linear association



Conclusion:

In essence, **Pearson's r** measures the degree of **linear relationship** between two variables. It tells us how well one variable can be predicted based on another in a linear context, and it is crucial for understanding correlations in data analysis.

However, it is important to note that Pearson's r only measures **linear relationships**, and it does not capture non-linear relationships.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)
Lets answer each section of question:

What is Scaling?

Scaling is a technique used in data preprocessing to adjust the range or distribution of feature variables in a dataset. The goal of scaling is to transform the features so they have similar units, magnitudes, or distributions, which can be critical for the performance of machine learning models.

Why is Scaling Performed?

Scaling is performed for several reasons:

1. **Model Sensitivity to Feature Magnitude:**
    - Some machine learning models are sensitive to the scale of input features. For example, distance-based algorithms like **K-Nearest Neighbors (KNN)** and **Support Vector Machines (SVM)** are heavily affected by the magnitude of input features. If one feature has a much larger range than another, it can dominate the model's performance.
2. **Improving Convergence in Optimization:**
    - Gradient-based optimization techniques, such as **Gradient Descent**, often perform better when the features are scaled. If features have very different scales, the model might take longer to converge or may not converge at all due to the uneven gradient magnitudes.
3. **Ensuring Fair Weighting of Features:**
    - Without scaling, features with larger numerical ranges could disproportionately influence the model's predictions, while features with smaller ranges might be overlooked.

Difference Between Normalized Scaling and Standardized Scaling

1. **Normalization (Min-Max Scaling):**
    - **Normalization** refers to scaling the data to a fixed range, typically [0, 1], or [-1, 1]. It is done by subtracting the minimum value of the feature and dividing by the range (max - min) of that feature.
2. **Formula:**
   ```
   Normalized value = (X - min(X)) / (max(X) - min(X))
   ```
    - **When to use:** Normalization is useful when the distribution of the data does not follow a Gaussian distribution or when you want to ensure that all features contribute equally to the model's performance. It is commonly used in algorithms like **K-Nearest Neighbors (KNN)** and **Neural Networks** where the data is expected to be on a similar scale.
3. **Advantages:**
    - Preserves the relationships between the data points in their original range.
    - Is often necessary for algorithms that rely on distance measures or when input data needs to be constrained within a fixed range.
4. **Disadvantages:**
    - Sensitive to outliers. If there are extreme values, they can skew the normalization process and distort the range of the transformed values.
5. **Standardization (Z-score Scaling):**
    - **Standardization** transforms the data to have a mean of 0 and a standard deviation of 1. This is done by subtracting the mean of the feature and dividing by the standard deviation.

6. **Formula:**
   `Standardized value = (X - μ) / σ`
   Where:
   - μμμ is the mean of the feature,
   - σσσ is the standard deviation of the feature.
   - **When to use:** Standardization is appropriate when the data follows a Gaussian (normal) distribution or when the data contains outliers. It works well for algorithms that assume the data follows a Gaussian distribution, such as **Linear Regression**, **Logistic Regression**, **Principal Component Analysis (PCA)**, and **SVM**.
7. **Advantages:**
   - Does not bound data to a fixed range, so it is less sensitive to outliers compared to normalization.
   - Useful when the features have different units or widely varying magnitudes, as it ensures all features contribute equally to the model.
8. **Disadvantages:**
   - The transformation is not bounded, meaning the transformed data could have values outside of a set range.
   - Sensitive to the presence of outliers (though less so than normalization).

Key Differences

Aspect	Normalization (Min-Max Scaling)	Standardization (Z-score Scaling)
Objective	Rescale the feature to a fixed range [0, 1] or [-1, 1]	Rescale the feature to have a mean of 0 and standard deviation of 1
Formula	`(X - min(X)) / (max(X) - min(X))`	`(X - μ) / σ`
Range	Bounded range (usually [0, 1])	Unbounded (data can have negative or large positive values)
Sensitivity to Outliers	Sensitive to outliers	Less sensitive to outliers
Use Case	Used when the distribution is not Gaussian or when a fixed range is needed	Used when the data follows a Gaussian distribution or when features have different units

Conclusion

Scaling is essential for many machine learning algorithms, and the choice between normalization and standardization depends on the nature of your data and the type of model you're using. Normalization is ideal when you need to rescale data to a specific range, while standardization is preferred when you want to adjust the data based on its statistical properties (mean and standard deviation).

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

The Variance Inflation Factor (VIF) is a metric used to measure how much the variance of a regression coefficient is increased due to collinearity with other features in the model. When perfect multicollinearity occurs, the VIF becomes infinite. Essentially, it quantifies the degree to which feature variables are correlated with one another, and is crucial for assessing the stability of a linear regression model.

$$VIF = \frac{1}{1 - R^2}$$

The formula involves R^2, which is the coefficient of determination for a given independent variable. It indicates how well this variable is explained by the other independent variables in the model. If a variable is perfectly predicted by others, its R^2 will be equal to 1, leading to a VIF of infinity because the formula for VIF becomes 1/(1-1), which results in a division by zero.

The VIF value provides insight into how much the variance (or squared standard error) of each regression coefficient is inflated compared to a scenario with no multicollinearity. For instance, a VIF of 1.9 suggests that the variance of a given coefficient is 90% larger than it would be if there were no correlation between the predictors.

Here is a general guideline for interpreting VIF:

- A VIF of 1: No correlation between the predictor and others.
- A VIF between 1 and 5: Moderate correlation.
- A VIF greater than 5: High correlation, indicating potential multicollinearity issues.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
(Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)
<Your answer for Question 11 goes here>

A **Q-Q plot** (Quantile-Quantile plot) is a graphical tool used to assess if a dataset follows a specific theoretical distribution, such as a normal distribution. It compares the quantiles of the sample data to the quantiles of a chosen reference distribution (often a normal distribution). In the plot, the x-axis represents the quantiles of the reference distribution, and the y-axis represents the quantiles of the dataset.

- **If the data follows the reference distribution**, the points will lie approximately along a straight line.
- **If the data deviates from the reference distribution**, the points will deviate from the line.

Use and Importance of a Q-Q Plot in Linear Regression

In **linear regression**, the assumptions of the model need to be checked to ensure the results are valid. One of the key assumptions is that the residuals (errors) of the model should follow a normal distribution. A **Q-Q plot** is used to visually assess this assumption.

Here's why the Q-Q plot is important in linear regression:

1. **Checking Normality of Residuals**:
    - For the linear regression model to give unbiased and reliable estimates, the residuals (the difference between the observed and predicted values) should be normally distributed.
    - A Q-Q plot of the residuals helps visually check if they follow a normal distribution. If the residuals deviate from the straight line, it suggests the data may not be normally distributed, which could violate the assumptions of linear regression.
2. **Model Diagnostics**:
    - If the Q-Q plot shows a clear deviation from the straight line, this indicates the presence of non-normality, which may lead to issues like inefficient parameter estimates, incorrect significance tests, and unreliable predictions.
    - In case the Q-Q plot shows a heavy tail or skewness, it suggests the presence of outliers or non-normal errors, which may require further investigation or transformation of the data.
3. **Detecting Outliers**:
    - Points that deviate significantly from the line on a Q-Q plot may represent outliers in the data. Identifying and addressing these outliers is crucial because they can disproportionately influence the regression model.
4. **Improving Model Performance**:
    - By using the Q-Q plot, we can assess if the assumption of normally distributed residuals holds. If not, methods like transformation of data, adding more variables, or using different regression techniques (e.g., robust regression) can be considered to improve model performance.

In conclusion, the Q-Q plot serves as an essential tool in linear regression diagnostics. It helps assess the normality of residuals, detect potential outliers, and ensure the validity of the linear regression model.