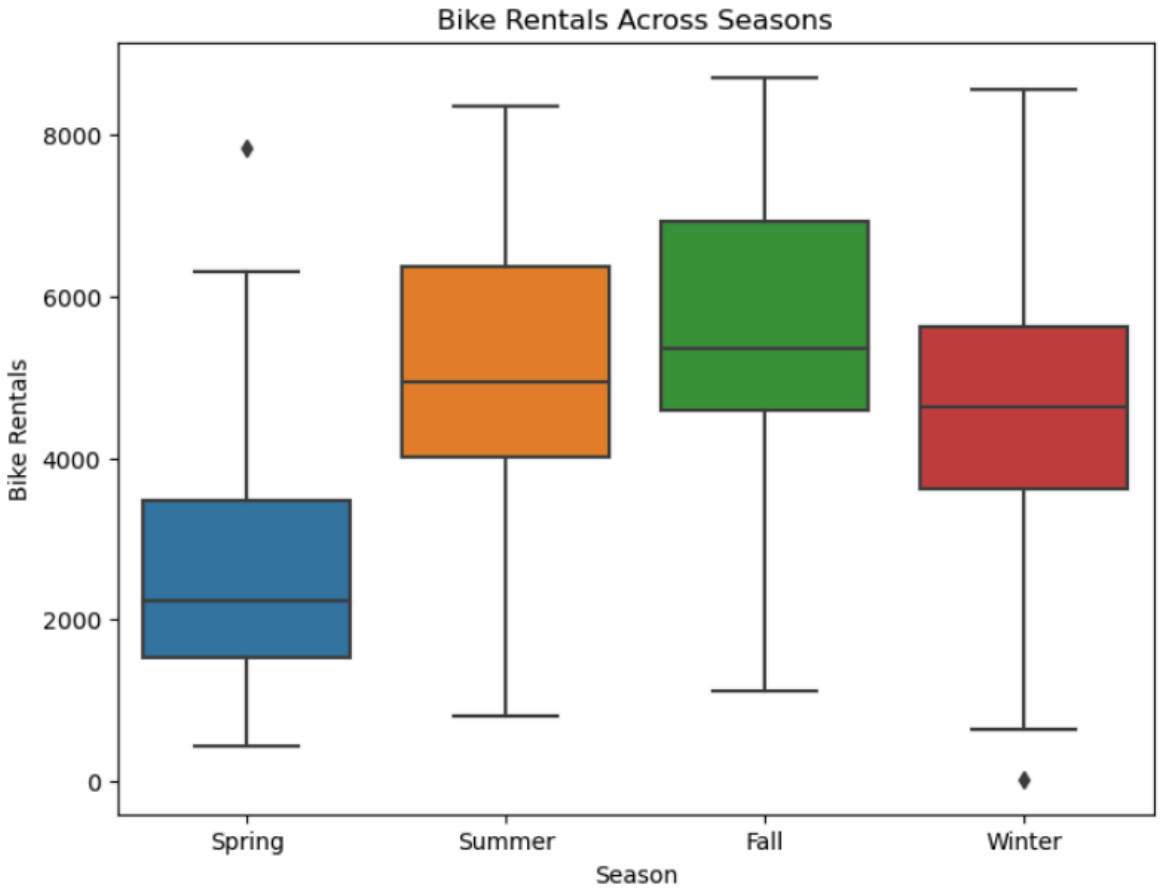
Assignment-based Subjective Questions

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

Ans: From the analysis of the categorical variables, here's what can be inferred generally:

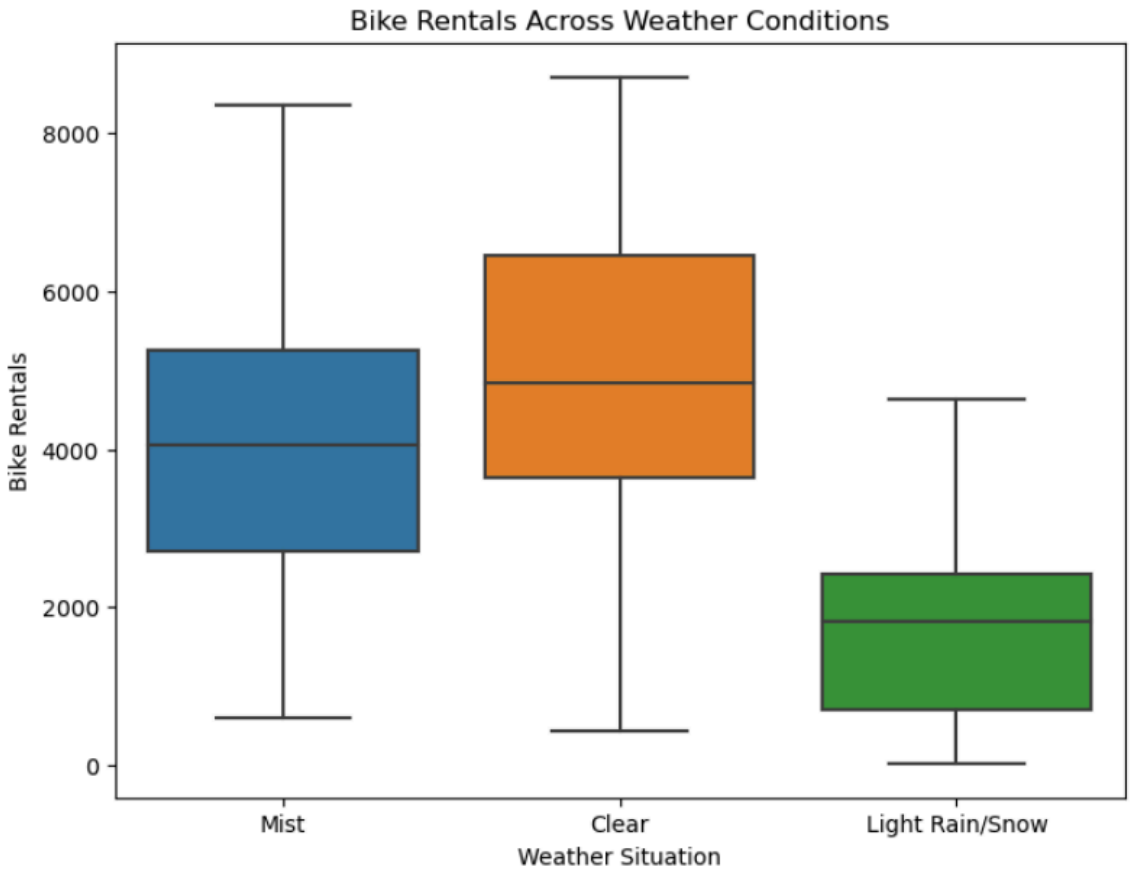
1. Season (season):

- **Inference**: Bike rentals tend to vary significantly across seasons. **For example:** Rentals are highest in Summer and Fall, likely due to favorable weather, and drop in Winter and Spring.
- **Impact**: Seasonality has a clear cyclical influence on bike rentals, indicating a strong relationship.



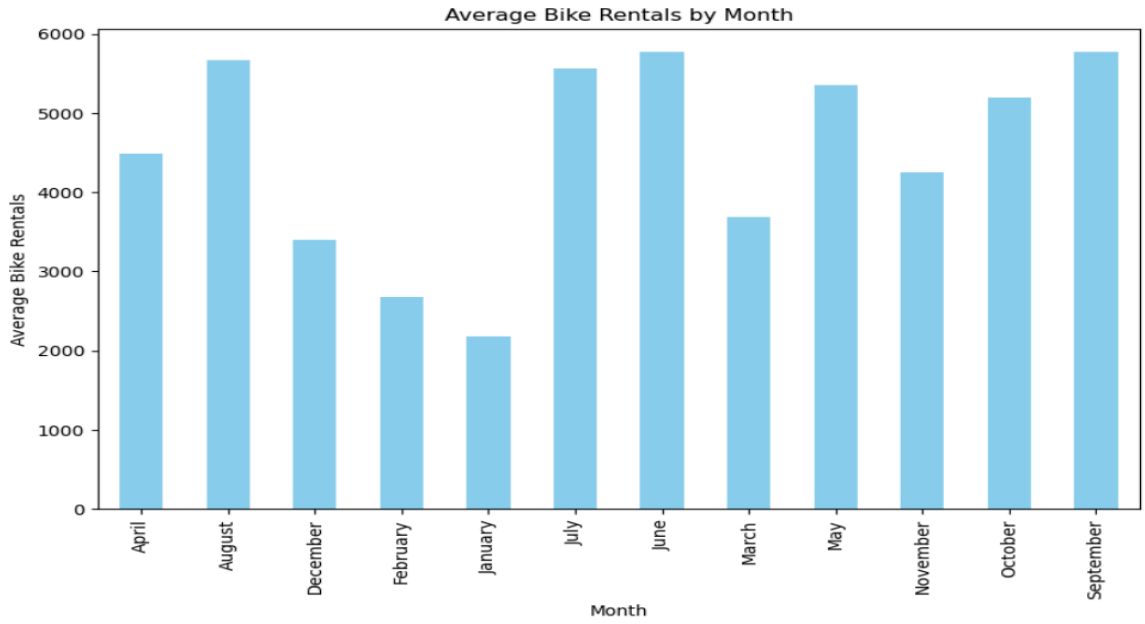
2. Weather Situation (weathersit):

- **Inference**: The weather has a noticeable effect: Clear weather conditions increase rentals, while adverse weather (like rain or snow) significantly reduces them.
- **Impact**: Weather conditions directly influence user behavior and rental demand.



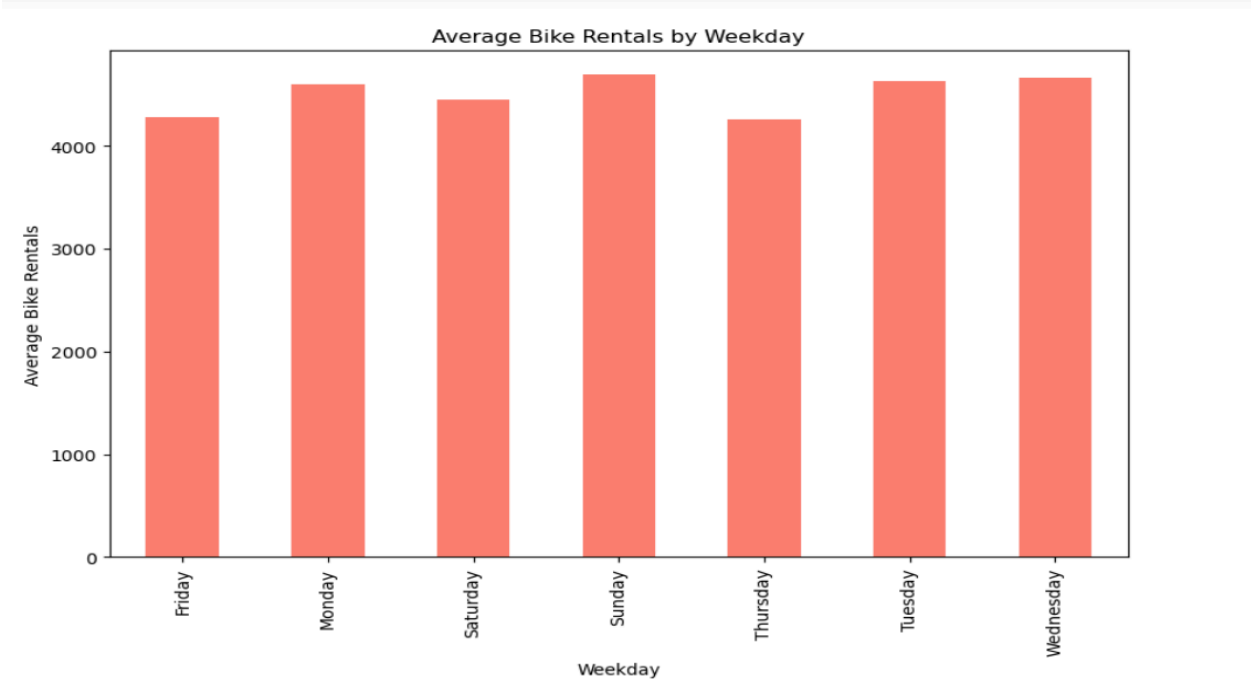
3. Month (mnth):

- **Inference**: Certain months (e.g., June, July, August, September) might show higher rentals, especially in summer, while colder months (e.g., January, February, December) typically show lower rentals.
- **Impact**: The month is a proxy for seasonality and specific trends like holiday seasons or school breaks.



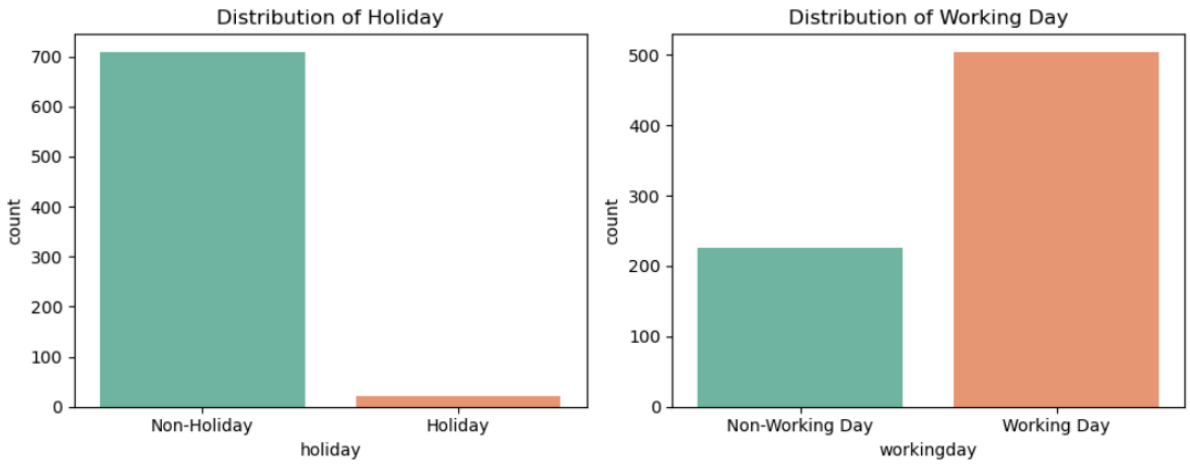
4. Weekday (weekday):

- **Inference**: There may not be significant variation in rentals across weekdays. However: **Weekends** (e.g., Saturday, Sunday) might have slightly higher rentals due to leisure activities. **Working** days might have consistent but slightly lower rentals.
- **Impact**: Weekday effects are subtle but can influence trends.



5. Holiday (holiday):

- **Inference**: Rentals tend to decrease on holidays compared to non-holidays, as fewer people commute to work.
- **Impact**: Being a holiday negatively affects rental demand, but leisure use may offset this.



6. Working Day (workingday):

- **Inference**: Rentals on working days are generally higher due to regular commuting patterns. However, non-working days may show increased leisure-based rentals.
- **Impact**: Working day status is a significant determinant of rental trends.

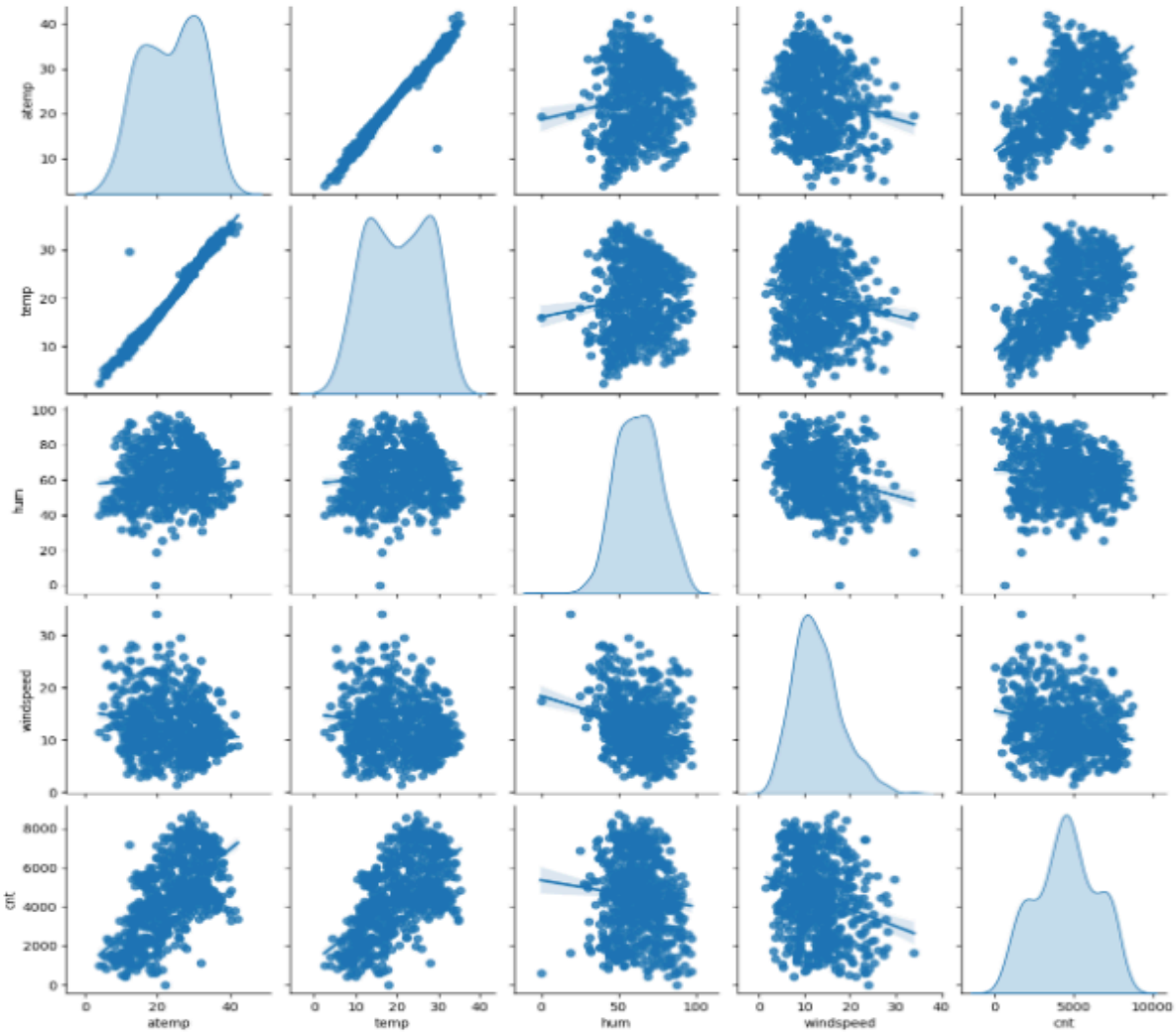
2. Why is it important to use drop_first=True during dummy variable creation? (2 marks)

Ans: During creating a dummy variable using **drop_first=True** is very important because it helps to prevent Dummy Variable Trap. Which occurs when there is multicollinearity among dummy variables. It drops the first dummy category and ensures the dummies are independent which makes the model mathematically stable, prevents the redundancy of the feature set.
For example, in the *season* column with categories Spring, Summer, Winter, and Fall, creating dummies without **drop_first** would generate 4 variables. If Spring, Summer, and Winter are all 0,

the model can infer it's Fall, leading to redundancy. Using drop_first=True removes one category, creating only 3 dummies and avoiding this problem.

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

Ans: The variable “temp” has the highest correlation with the target variable. This indicates that as the temperature increases, the number of bike rentals tends to increase, Which suggests that warmer weather positively impacts bike usage.



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

Ans: After building the model on the training set, I validated the key assumptions of Linear Regression through various diagnostic tests and visualizations:

1. **Linearity:**

I confirmed that the relationship between the predictors and the target variable was linear by plotting the residuals vs fitted values (predicted values). A random scatter of residuals around zero, without any distinct pattern, indicated that the linearity assumption was valid. This ensures that the model has correctly captured the relationship between predictors and the target variable.

2. **Homoscedasticity (Constant Variance of Errors):**

I verified homoscedasticity by plotting the residuals vs fitted values (predicted values). In this plot, the residuals should spread evenly across all levels of the fitted values. If the spread of residuals was consistent (i.e., no funnel or cone shape), it would confirm that the assumption of constant variance was met. Any deviations from this could signal heteroscedasticity, which would require further adjustments like using weighted least squares.

3. **Independence of Errors:**

I assessed the independence of residuals using the Durbin-Watson test, which tests for autocorrelation in the residuals. A value close to 2 indicates that the errors are independent, which is an important assumption for reliable model performance.

4. **Multicollinearity:**

To address multicollinearity, I calculated the Variance Inflation Factor (VIF) for each predictor variable. VIF values above 5 (or 10) would indicate high multicollinearity, suggesting that the variable is highly correlated with other predictors. I ensured that the VIF values for all variables were within an acceptable range to prevent issues with inflated standard errors and unreliable coefficient estimates. Additionally, I checked the correlation matrix to identify any predictors with high correlations and considered removing or transforming them if necessary.

I also used **drop_first=True** when creating dummy variables for categorical features, preventing the problem of the dummy variable trap, which occurs when multicollinearity arises from the inclusion of all levels of a categorical variable.

By validating these assumptions, I ensured that the linear regression model was robust, reliable, and interpretable, with stable coefficient estimates that accurately reflect the relationship between the features and the target variable.

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

Ans: Based on the final model, the top 3 features contributing significantly towards explaining the demand for shared bikes are:
1. **Temperature (temp):**

- Coefficient: 0.4405
- p-value: 0.000 (which is statistically significant, as it is less than the significance level of 0.05)
- VIF: 43.90 (The high VIF suggests some multicollinearity, but temp remains a significant feature with a strong positive correlation with bike demand).

The temperature variable has a strong positive impact on bike demand. As the temperature rises, bike rentals increase, which is reflected in its positive coefficient of 0.4405. This feature is highly significant as its p-value is much smaller than 0.05, confirming its importance in predicting bike demand.

2. **Year (yr):**

- Coefficient: 0.2311
- p-value: 0.000 (significant)
- VIF: 2.13 (indicating low multicollinearity)

The year variable shows a strong positive relationship with bike demand, with a coefficient of 0.2311. This suggests that as the year progresses, the demand for shared bikes increases. Its p-value is also statistically significant, confirming that it has a meaningful effect on bike rental demand. The VIF value is low, suggesting minimal multicollinearity with other features.

3. **Weather Condition - Light Rain/Snow (weathersit_Light Rain/Snow):**

- Coefficient: -0.2499
- p-value: 0.000 (significant)
- VIF: 1.33 (low multicollinearity)

The weather condition Light Rain/Snow has a negative impact on bike rentals. As shown by the negative coefficient of -0.2499, bike demand decreases when the weather is misty or rainy. This

is statistically significant with a p-value of 0.000, indicating its strong effect on bike demand. The VIF value is low, indicating minimal multicollinearity

General Subjective Questions

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

Ans: Linear regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (predictors). The algorithm assumes that there is a linear relationship between the target and predictors, which is represented by a straight line in the case of one predictor.

1. Model Equation:

The equation of a simple linear regression model with one predictor is:

$Y = \beta 0 + \beta 1X + \epsilon$

Where:

- Y is the dependent variable (target).
- X is the independent variable (predictor).
- β0 is the intercept (the value of Y when X=0).
- β1 is the coefficient of the independent variable, showing the change in Y for a one-unit change in X.
- ϵ represents the error term, accounting for variability not explained by the model.

For multiple predictors, the equation becomes:

$Y=\beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta nXn + \epsilon$

2. Objective:

The objective of linear regression is to find the best-fitting line (in the case of simple linear regression) or hyperplane (in the case of multiple regression) that minimizes the difference between the predicted values and the actual values. This difference is known as the **residual**.

The model is trained by adjusting the coefficients (β0,β1,…,βn) such that the sum of squared residuals is minimized.

3. Fitting the Model:

The model is fit using a method called **Ordinary Least Squares (OLS)**. The process involves:

- Calculating the best values for the coefficients $\beta 0, \beta 1, \ldots, \beta n$ that minimize the sum of squared errors (the difference between predicted values and actual values).

4. Assumptions of Linear Regression:

For the model to be reliable, certain assumptions must be met:

- **Linearity:** The relationship between predictors and target variables is linear.
- **Independence:** Observations are independent of each other.
- **Homoscedasticity:** Constant variance of errors (residuals) across all levels of independent variables.
- **Normality of Errors:** The residuals should be normally distributed.
- **No Multicollinearity (for multiple regression):** Predictors should not be highly correlated with each other.

5. Evaluation:

After training, the model's performance is evaluated using various metrics:

- **R-squared (R^2):** This represents the proportion of the variance in the target variable that is explained by the predictors. It ranges from 0 to 1, with 1 meaning perfect fit.
- **Mean Squared Error (MSE)** and **Root Mean Squared Error (RMSE):** These metrics measure the average of the squared differences between predicted and actual values.

6. Use Cases:

Linear regression is widely used for:

- Predicting continuous numerical values.
- Understanding relationships between variables, such as how advertising expenditure impacts sales or how temperature affects bike rentals.

7. Limitations:

- Linear regression assumes a linear relationship, so it may not perform well if the true relationship is nonlinear.
- It is sensitive to outliers, which can distort the model's accuracy.
- It requires the assumptions to be met for the results to be reliable.

In summary, linear regression is a simple yet powerful tool for predicting continuous variables, understanding relationships, and providing insights into the strength and nature of relationships between variables.

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

Ans: Anscombe's Quartet is a set of four datasets that have nearly identical simple descriptive statistics, such as mean, variance, correlation, and linear regression results, yet they are very different when visually examined. The purpose of Anscombe's Quartet is to demonstrate the importance of visualizing data before making assumptions and drawing conclusions.

1. Components of Anscombe's Quartet:

Anscombe’s Quartet consists of four distinct datasets, each containing 11 data points with two variables X and Y. While the datasets share the same summary statistics, their distributions and relationships are very different. The four datasets are:

- **Dataset 1:**
    - **Linear relationship:** Y has a clear linear relationship with X, making it a good candidate for linear regression.
- **Dataset 2:**
    - **Nonlinear relationship:** Despite the same summary statistics as Dataset 1, this dataset has a curved or quadratic relationship between X and Y, which cannot be accurately captured by linear regression.
- **Dataset 3:**
    - **Outlier effect:** Here, most of the data points are linear, but there is one extreme outlier. This outlier strongly affects the correlation and regression results, even though the other data points follow a clear linear trend.
- **Dataset 4:**
    - **Vertical line pattern:** In this dataset, the Y-values are the same for each value of X, except for one point. This scenario demonstrates that even when the relationship between the variables is not linear, the statistical metrics can still suggest a strong correlation.

2. Identical Descriptive Statistics:

Despite the obvious differences in the datasets' distributions, all four datasets have the same basic 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 = 0.82
- Linear regression line: Y = 3 + 0.5X

These identical summary statistics make it appear as if the datasets are very similar, but this is misleading without visual inspection.

3. Visualizing Anscombe's Quartet:

The key takeaway from Anscombe's Quartet is that, while summary statistics such as mean, variance, and correlation are useful, they can be misleading when the data does not follow the assumed model. Visualizing the data using scatter plots reveals the differences between the datasets:

- **Dataset 1** shows a linear trend.
- **Dataset 2** shows a curved pattern.
- **Dataset 3** shows a linear trend with an outlier.
- **Dataset 4** shows a vertical line with a single outlier.

4. Key Lesson:

Anscombe's Quartet emphasizes the importance of visualizing data before applying statistical models, particularly linear regression. It shows that:

- Descriptive statistics alone do not fully capture the data's underlying structure.
- Visualizations like scatter plots are crucial for identifying patterns, outliers, and relationships that cannot be inferred from summary statistics alone.

Conclusion:

Anscombe's Quartet teaches the valuable lesson that, while statistical summaries can provide insights, they should always be complemented with visualizations to gain a more accurate understanding of the data. This helps avoid misleading conclusions when applying models, particularly in cases where the assumptions of the model do not hold.

3. What is Pearson's R? (3 marks)

Ans: **Pearson's R**, also known as the **Pearson correlation coefficient**, is a statistical measure used to assess the strength and direction of the linear relationship between two continuous variables. It quantifies how well the two variables move in relation to each other.

1. Definition:

Pearson's R is calculated as the covariance of the two variables divided by the product of their standard deviations. The formula for Pearson's R is:

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

Where:

- Xi and Yi are individual data points of the two variables.
- X^- and Y^- are the means of the variables X and Y, respectively.

2. Interpretation of Values:

The value of Pearson's R ranges from -1 to +1, with different values representing different types of correlation:

- **r = +1:** Perfect positive correlation. As one variable increases, the other variable increases proportionally.
- **r = -1:** Perfect negative correlation. As one variable increases, the other variable decreases proportionally.
- **r =0:** No linear correlation. There is no predictable relationship between the two variables.
- **0 < r < +1:** Positive correlation. As one variable increases, the other variable tends to also increase, but not perfectly.
- **−1 < r < 0:** Negative correlation. As one variable increases, the other variable tends to decrease, but not perfectly.

3. Significance of Pearson's R:

- **Strength of Relationship:** The closer the value of Pearson's R is to +1 or -1, the stronger the linear relationship between the variables.

- **The direction of Relationship:** The sign of Pearson's R indicates the direction of the relationship. Positive values indicate a direct relationship, while negative values indicate an inverse relationship.
- **Linear Assumption:** Pearson's R only measures linear relationships. It will not detect non-linear relationships, even if they are strong.

4. Use Cases:

- Pearson's R is widely used in fields like economics, biology, social sciences, and engineering to assess how two variables are related.
- For example, it can be used to measure the relationship between the amount of rainfall and crop yield or between hours of study and exam scores.

5. Limitations:

- **Sensitive to Outliers:** Pearson's R is highly sensitive to outliers, which can distort the relationship between the variables.
- **Linear Only:** It only captures linear relationships. If the relationship between variables is non-linear, Pearson's R might be close to 0 even if a relationship exists.

Conclusion:

Pearson's R is a valuable tool for measuring the linear relationship between two continuous variables. It provides insights into both the strength and direction of the correlation but should be used with caution, particularly when the relationship might not be linear or when outliers are present.

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

Ans: **1. What is Scaling?**
Scaling refers to the process of transforming the features (variables) in a dataset so that they have certain properties, usually to make them comparable in terms of magnitude. It adjusts the range of values of the data so that features with larger values or different units do not dominate the learning process.

2. Why is Scaling Performed?

- **Improves Model Performance:** Many machine learning algorithms (like gradient descent-based methods, support vector machines, and k-nearest neighbors) perform better or converge faster when the data is scaled. This is because scaling ensures that each feature contributes equally to the model.
- **Prevents Dominance of Larger Features:** Without scaling, features with larger numeric ranges (e.g., income in thousands and age in single digits) could disproportionately influence the results.
- **Stabilizes Training Process:** Some algorithms assume that features are on the same scale for proper convergence (e.g., linear regression, logistic regression, etc.).

3. Difference Between Normalized Scaling and Standardized Scaling:

- **Normalized Scaling (Min-Max Scaling):**
    - Normalization is the process of adjusting the data such that the values of each feature are rescaled to a fixed range, typically [0, 1].
    - Formula:
- X(normalized) = X − X(min) / X(max) − X(min)
    - **When to Use:** Normalization is best when you need to ensure all features have the same scale and when the data is not normally distributed. It is also commonly used when data needs to be represented within a fixed range, like image pixel values.
- **Standardized Scaling (Z-Score Scaling):**
    - Standardization transforms the data to have a mean of 0 and a standard deviation of 1. This method centers the data and scales it based on the standard deviation, making it less sensitive to outliers.
- Formula : X(standardized) = X −μ / σ

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

    - **When to Use:** Standardization is useful when the data has varying scales and is normally distributed. It's also better when dealing with outliers, as it does not restrict data to a certain range like normalization does.

Conclusion: Scaling is crucial for improving the performance and stability of machine learning models. **Normalization** adjusts the data within a specific range, whereas **standardization**

transforms the data to have a mean of 0 and a standard deviation of 1, with each method being suited for different types of data and models.

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

Ans: **1. VIF (Variance Inflation Factor):**

- 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 collinearity with other predictor variables.
- The formula for VIF of a predictor variable is: $\mathrm{VIF} = 1 / (1 - \mathrm{R}^2)$;

  where R^2 is the coefficient of determination obtained by regressing the predictor variable against all other predictor variables in the model.

2. Reasons for Infinite VIF:

- **Perfect Multicollinearity:** VIF becomes infinite when there is **perfect multicollinearity** between two or more predictor variables. This means that one variable can be expressed exactly as a linear combination of other variables in the model.
  - In this case, the R^2 value for the predictor will be 1, leading to the formula for VIF becoming undefined: $\mathrm{VIF} = 1 / (1 - 1) = \infty$
- **Linear Dependence:** If one variable is a direct function of others (e.g., X1 = 2 × X2), this results in infinite correlation and leads to infinite VIF.
- **Duplicate or Redundant Variables:** If you accidentally include the same variable multiple times, or if there are exact duplicates in your feature set, the correlation among those variables will be perfect, leading to infinite VIF.

3. Impact of Infinite VIF:

- **Model Instability:** Infinite VIF indicates a serious issue with the model because it suggests that the variables are redundant and carry the same information. This can lead to instability in the estimation of regression coefficients and inaccurate predictions.

- **Unreliable Interpretation:** High or infinite VIF values make it difficult to interpret the significance of the variables because the standard errors of the regression coefficients become very large, leading to misleading p-values.

Conclusion: Infinite VIF arises due to perfect multicollinearity or exact linear dependence between predictor variables. This indicates redundancy in the dataset and highlights the need to address multicollinearity, typically by removing one of the collinear variables or using dimensionality reduction techniques.

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

Ans: **1. What is a Q-Q Plot?**

- A **Q-Q plot** (Quantile-Quantile plot) is a graphical tool used to assess if a dataset follows a particular theoretical distribution, most commonly the **normal distribution**. It compares the quantiles of the dataset against the quantiles of a specified distribution.
- In a Q-Q plot:
    - The x-axis represents the theoretical quantiles (e.g., from a normal distribution).
    - The y-axis represents the actual quantiles from the sample data.
- If the data points form a straight line (typically a 45-degree diagonal line), the data is considered to follow the specified distribution, such as normality.

2. Use of Q-Q Plot in Linear Regression:

- **Checking Normality of Residuals:** In linear regression, one of the key assumptions is that the residuals (errors) of the model should be normally distributed. A Q-Q plot is used to visually check if the residuals follow a normal distribution.
    - If the residuals are normally distributed, they will appear as points on the Q-Q plot that closely follow the straight line.
    - A significant deviation from the line indicates that the residuals are not normally distributed, which may violate the assumptions of linear regression.

3. Importance of a Q-Q Plot in Linear Regression:

- **Assumption Check:** The assumption of normally distributed residuals is important for inference in linear regression. If residuals are not normally distributed, the model's significance tests (e.g., t-tests for coefficients) and confidence intervals might be unreliable.
- **Model Diagnostics:** A Q-Q plot is a quick and effective way to visually assess if the linear regression model is appropriate for the data. If the plot shows deviations from the line, it suggests that the model might not be fitting the data well or that a different model (e.g., transformation or a non-linear model) might be more appropriate.
- **Guidance for Improvements:** If the Q-Q plot indicates non-normality, it might suggest that transformations (such as log transformation or square root) could help in normalizing the residuals, thus improving the model's accuracy and validity.

Conclusion: A Q-Q plot is a diagnostic tool in linear regression to assess the normality of residuals, which is a critical assumption for valid model inference. By checking the shape of the plot, we can determine if the residuals deviate significantly from normality and take corrective actions if necessary.