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)

Answer: As per my analysis using seaborn and matplotlib libraries box and pair plots. One can infer from visualization.

- Weekdays, Thursday, Friday & Saturday have more number of bookings on boom bike.
- When the weather is sunny it has attracted more riders than when it was raining.
- The count of riders is the lowest in January and the highest in June. May, June, July, August, September and October record the highest number of riders.

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

    - drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
    - When categorical variables with n-levels are present. It is recommended to use n-1 columns, so we need to drop 1 column. In general it could be any column but as general rule of thumb we drop the first.

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

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
    - No Multicolinearity: The independent variables shouldn't be correlated. If multicollinearity exists between the independent variables, it is challenging to predict the outcome of the model.
    - Distribution of error terms: The errors terms should be normally distributed.
    - Residual Analysis: Check the residuals (the differences between predicted and actual values) against the fitted values.
    - Homoscedasicity: Check for consistent spread of residuals across all levels of the independent variable. A cone-shaped pattern indicates heteroscedasticity, which might violate the assumption.
    - Recursive Feature elimination: RFE works by recursively removing features, fitting the model, and evaluating the model performance until the desired number of features is reached.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
    - Holiday
    - Temperature
    - Season

General Subjective Questions

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

   - A statistical model that examines the linear connection between a dependent variable and a given set of independent variables is known as linear regression. When there is a linear relationship between variables, the dependent variable's value likewise changes in response to changes in the values of one or more independent variables (increase or decrease).
   - The relationship can be mathematically stated using the equation − Y = mX + c.
   - In this case, Y is the dependent variable that needs to be predicted.
   - The independent variable that we are predicting with is called X.
   - The regression line's slope, or m, shows how much of an impact X has on Y, whereas the Y-intercept, or c, is a constant. Y would be equal to c if X = 0.
   - Additionally, as will be shown later, the linear relationship might have a positive or negative character.
   - Positive Linear Relationship: When both the independent and dependent variables rise, a linear relationship is said to be positive.
   - Negative Linear relationship: A linear relationship will be called positive if independent increases and dependent variable decreases.

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

   - **Anscombe's quartet** comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph. The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.
   - The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.
   - **Anscombe's quartet** is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

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

- The Pearson correlation coefficient (r) is the most widely used correlation coefficient and is known by many names:

- Pearson's r
- Bivariate correlation
- Pearson product-moment correlation coefficient (PPMCC)
- The correlation coefficient

- The Pearson correlation coefficient is a statistics, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.
- Although interpretations of the relationship strength vary between disciplines, the table below gives general rules of thumb:

Pearson correlation coefficient (*r*) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
    - In machine learning, scaling is a preprocessing step that involves converting numerical information to a standard scale. Because many machine learning algorithms are sensitive to the size of input features, it is essential. By ensuring that each feature contributes equally to the training process, scaling keeps certain features from predominating over others based only on scale.
    - **Normalized scaling**: also known as Min-Max scaling, transforms the values of numerical features to a specific range, typically between 0 and 1. This scaling method is achieved by adjusting the values based on the minimum and maximum values within the dataset. Normalized scaling is useful when the distribution of features is approximately uniform. It ensures that all features have the same scale, preventing some features from dominating others based on their magnitudes.
    - Adjusts values to a specific range (e.g., 0 to 1).
    - Formula = Xnormalized = X-min(X)/max(X)-min(X)

- **Standardized scaling**: also known as Z-score normalization, transforms numerical features to have a mean of 0 and a standard deviation of 1. This scaling method involves subtracting the mean of the feature from each value and then dividing by the standard deviation. Standardized scaling is suitable when features have different units and follow a normal distribution. It helps algorithms that are sensitive to the scale of features converge faster and perform better. Additionally, it makes the interpretation of coefficients in linear models more straightforward.
- Adjusts values to have a mean of 0 and a standard deviation of 1.
- Formula: Xstandardized = X-mean(X)/std(X)

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
   - If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
   - A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence