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?
    I. cnt vs season : Rental bike usage experiences an increase during fall, while it reaches its lowest point in spring
    II. cnt vs yr : The number of rental bikes showed an increase in 2019 (represented as 1) compared to 2018
    III. cnt vs holiday : The bike demand is increased during working days as compared to holidays
    IV. cnt vs weekday : There is no notable distinction between weekends and weekdays in terms of rental bike demand
    V. cnt vs weathersit : Rental bike quantities are higher on clear weather days and decrease to their lowest point during light snow and rain

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

    Using drop_first=True when creating dummy variables is crucial for avoiding multicollinearity issues in regression models. By dropping the first category, we prevent high correlation among variables and also to reduce the number of variables during dummy variable creation

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

    The variable “temp” has the highest correlation with the target variable looking at the pair-plot among the numerical variables

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
    To validate the assumptions of Linear Regression after building the model on the training set, Below mentioned criteria can be checked;
    I. Checking the linearity
    II. Checking the R squared value
    III. checking distribution of error terms using histogram
    IV. DW (Durbin-Watson) for the model to check if there is any presence of autocorrelation
    V. Checking the P- values for the variables whether they are significant
    VI. Checking the VIF (<5)

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
    “temp”, “yr” and “Light_Snow + Rain” are the top 3 features contributing significantly towards explaining the demand of the shared bikes

General Subjective Question:

1. Explain the linear regression algorithm in detail.

   Linear Regression is a widely used machine learning technique for predicting continuous outcomes. Its objective is to find the best-fitting line that represents the relationship between input variables and the target variable. The equation of a simple linear regression model is given by:

   $y = b0 + b1 * x$

   where:

   y is the target variable (dependent variable).

   x is the predictor variable (independent variable).

   b0 is the y-intercept (the value of y when x is 0).

   b1 is the slope (the change in y for a unit change in x).

   The goal is to estimate the values of b0 and b1 that best fit the data.

   We start with a dataset containing input variables (X) and their corresponding target variable (Y). The goal is to minimize the difference between our predicted Y values and the actual Y values in the dataset by finding an optimal line.

   To achieve this, we employ Ordinary Least Squares (OLS), a method that calculates the sum of squared differences between predicted and actual Y values. By minimizing this sum, we can identify the line that best fits the data.

   The coefficients of the regression line, including slope and intercept, are determined through calculus. We optimize these coefficients by taking partial derivatives of the sum of squared differences with respect to them, setting the derivatives to zero, and solving for the optimal values.

   Once the coefficients are obtained, we can make predictions by plugging new input values into the regression line equation. This yields the predicted values of the target variable.

   To evaluate our model's performance, we utilize statistical metrics such as the coefficient of determination (R-squared). R-squared helps us understand the goodness of fit of the line to the data. Additionally, hypothesis testing can be employed to assess the significance and strength of the relationship between input variables and the target variable.

2. Explain the Anscombe’s quartet in detail.

   Anscombe's quartet is a collection of four datasets that have identical summary statistics but exhibit distinct patterns when visualized. It was introduced to emphasize the importance of data visualization and the limitations of relying solely on summary statistics.

Each dataset in the quartet consists of 11 data points with two variables: X and Y. Despite having the same means, variances, and correlation coefficients, the datasets differ significantly when plotted.

Dataset I forms a linear relationship, where Y increases steadily with X. Dataset II shows a non-linear relationship with an apparent quadratic curve. Dataset III has a strong outlier that heavily influences the regression line and correlation. Finally, Dataset IV appears to have no clear relationship between X and Y, with most points clustering around a single Y value.

The quartet aims to challenge the notion that summary statistics alone provide a comprehensive understanding of the data. While the summary statistics may appear similar across the four datasets, the visual patterns reveal substantial differences in the underlying relationships.

Anscombe's quartet highlights the importance of data visualization in exploratory data analysis. Visualizing the data allows us to observe patterns, trends, and anomalies that summary statistics might overlook. It emphasizes the need to complement statistical analysis with visual examination to gain a more complete understanding of the data.

3. What is Pearson’s R?

Pearson's R, also known as Pearson's correlation coefficient, is a statistical measure used to determine the strength and direction of the linear relationship between two variables. It quantifies how closely the variables move together and provides a standardized value between -1 and +1.

By calculating Pearson's R, we can assess the degree to which the variables are related. A positive value indicates a positive correlation, meaning that as one variable increases, the other tends to increase as well. Conversely, a negative value indicates a negative correlation, where an increase in one variable corresponds to a decrease in the other. A value close to zero suggests little to no linear relationship between the variables.

To compute Pearson's R, we divide the covariance of the variables by the product of their standard deviations. This normalization allows for easy comparison of correlation values across different datasets.

It's important to remember that Pearson's R measures only linear relationships and assumes the relationship is constant. Non-linear relationships may not be accurately captured by this coefficient.

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

Scaling is a preprocessing technique in data science used to transform numerical features into a common scale. It involves adjusting the values of the variables to a specific range or distribution. Scaling is performed to ensure that all features have equal importance and to prevent certain variables from dominating others in the analysis.

Scaling is important for several reasons. First, it helps in comparing variables with different units and scales. Scaling ensures that all variables contribute equally to the analysis and avoids bias towards variables with larger values.

There are two commonly used scaling techniques: normalized scaling and standardized scaling. Normalized scaling, also known as min-max scaling, rescales the values of the variable to a range between 0 and 1. It calculates the ratio between the difference of each value and the minimum value to the difference between the maximum and minimum values. This technique preserves the relative relationships between the values but compresses the range.

Standardized scaling, also referred to as z-score scaling, transforms the values of the variable to have a mean of 0 and a standard deviation of 1. It subtracts the mean from each value and divides it by the standard deviation. Standardized scaling maintains the relative relationships between the values and preserves the overall distribution shape.

The main difference between normalized scaling and standardized scaling lies in the scale and distribution of the transformed values. Normalized scaling brings the values to a specific range, whereas standardized scaling centers the values around zero with a standard deviation of 1.

The choice between normalized scaling and standardized scaling depends on the requirements of the analysis and the characteristics of the data. Normalized scaling is suitable when preserving the original range is important, while standardized scaling is useful when maintaining the distribution shape and comparing variables with different means and variances is desired

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

   The occurrence of an infinite value for VIF typically happens when there is perfect multicollinearity in the dataset. Multicollinearity refers to a high correlation between predictor variables in a regression model. Perfect multicollinearity occurs when one or more variables can be perfectly predicted using a linear combination of other variables.

   In such cases, when a variable can be precisely predicted using other variables, the VIF calculation for those variable results in an infinite value. This happens because the VIF formula involves dividing the variance of a predictor variable by its residual variance, and if there is no residual variance due to perfect prediction, the division by zero leads to an infinite VIF value.

   Perfect multicollinearity can be problematic for regression models because it hinders the interpretation of individual variable effects. It also causes instability in estimating the regression coefficients and inflates their standard errors.

   To address perfect multicollinearity, it's essential to identify and handle the correlated variables appropriately. This can involve removing one of the correlated variables, transforming the variables.

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

A Q-Q plot, short for Quantile-Quantile plot, is a graphical tool used to assess whether a dataset follows a particular theoretical distribution. It helps us understand how well our data matches the expected distribution.

The Q-Q plot works by comparing the quantiles of our dataset against the quantiles of a theoretical distribution. It plots the observed quantiles on the x-axis and the expected quantiles on the y-axis. If the data closely adheres to the theoretical distribution, the points on the plot will fall along a straight line.

In linear regression, a Q-Q plot is employed to check the assumption of normality. Linear regression assumes that the residuals, which are the differences between the observed and predicted values, are normally distributed. By examining the Q-Q plot of the residuals, we can evaluate whether this assumption holds true.

If the points on the Q-Q plot align closely to the straight line, it suggests that the residuals follow a normal distribution. This indicates that the linear regression model is appropriate and the assumption of normality is met. On the other hand, if the points deviate significantly from the straight line, it indicates a departure from normality, suggesting that the linear regression model may not be the best fit for the data.

The Q-Q plot is important in linear regression as it helps us validate the assumption of normality for the residuals. If the assumption is violated, it could impact the accuracy and reliability of the regression model. By identifying departures from normality, we can explore potential remedies such as data transformation or considering alternative modelling techniques.