A. 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?

Ans: Observations from categorical variables are:

- The year box plots indicates that more bikes are rent during 2019.
- The season box plots indicates that more bikes are rent during fall season.
- The working day and holiday box plots indicate that more bikes are rent during normal working days than on weekends or holidays.
- The month box plots indicates that more bikes are rent during September month.
- The weekday box plots indicates that more bikes are rent during Saturday.
- The weathersit box plots indicates that more bikes are rent during Clear, Few clouds, partly cloudy weather.

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

- Ans: 1.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.
- Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished.
- Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

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

Ans: By looking at the pair plot temp variable has the highest (0.63) correlation with target variable 'cnt'.

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

Ans: Assumptions of Linear Regression and their validation as per model is explained below:

- Linear Relationship: As seen from the correlation map, output variable “cnt” has a linear relationship with variables like temp, atemp.
- Multivariate Normality: It is verified in he model that the residuals are normaly distributed.
- No Multicollinearity: As seen in Correlation matrices, we find that correlation coefficients are below 0.80. In the pair plot temp variable has the highest (0.63) correlation with target variable 'cnt'.
- Variance Inflation Factor (VIF): We removed all the variables with high VIF because they were causing Multicollinearity.
- Homoscedasticity: in the plot of error terms with find that the variance of error terms (residuals) is consistent.

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

Ans: The Top 3 features contributing significantly towards the demands of share bikes are:

- weathersit_Light_Snow (negative correlation).
- yr_2019 (Positive correlation).
- Temp (Positive correlation).

B. General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear regression is a way of figuring out how one thing affects another thing. It's like when you see a pattern in numbers and use it to guess what might happen next.

Let's say you're trying to predict how much something will sell for based on how much you spend on advertising. In linear regression, you look at how these two things are related - advertising spending and sales. You notice that when you spend more on advertising, sales tend to go up. This relationship is what linear regression helps you understand.

Linear regression is useful because it helps you make predictions about things, like how much you might sell next month if you increase your advertising budget. It's kind of like drawing a line through the dots on a graph to see where they might go next.

This method works well when you have information about at least two things that might be related, like in stock market predictions or analysing scientific data.

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet is a set of four datasets that were created by the statistician Francis Anscombe in 1973 to illustrate the importance of visualizing data before analysing it and the limitations of relying solely on summary statistics. Despite having very similar statistical properties, these datasets exhibit vastly different characteristics when plotted visually. Anscombe's quartet is often used to emphasize the importance of data visualization and to caution against drawing conclusions based solely on summary statistics without considering the actual data distribution.

3. What is Pearson's R?

Ans: Pearson's correlation coefficient, often denoted as *r*, is a measure of the strength and direction of the linear relationship between two variables. It was developed by Karl Pearson, hence the name.

Pearson's *r* ranges from -1 to 1. Here's what different values of *r* represent:

- *r*= 1: Perfect positive linear relationship. When one variable increases, the other variable also increases proportionally.
- *r*= −1: Perfect negative linear relationship. When one variable increases, the other variable decreases proportionally.
- *r*= 0: No linear relationship. There is no consistent linear trend between the variables.

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

Ans: When we get a lot of independent variables in a model, many of them may be on a different scales which can lead to wrong interpretation and coefficients if processed further. So we need to scale features mainly because of two reasons:

a. Ease of interpretation
b. Faster convergence for gradient descent methods.

Two popular methods of scaling are:

i. Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.

$$x = x - \text{mean}(x) \,/\, \text{sd}(x)$$

ii. Min-Max Scaling: The variables are scaled in such a way that all the values lie between zero & one using the maximum & minimum values in the data.

$$x = x - \min(x) \,/\, \max(x) - \min(x)$$

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, p-values, R-Square etc.

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

Ans: Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in regression analysis. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, leading to issues in interpreting the model's coefficients and affecting the model's overall performance.

The formula for calculating VIF for an independent variable in a regression model is:

$$VIF = 1/(1 - R^2)$$

Where $R2$ is the coefficient of determination obtained by regressing the independent variable against all other independent variables in the model.

In some cases, the VIF value can become infinite. This happens when the coefficient of determination (R^2) is equal to 1. This indicates a perfect linear relationship between the independent variable and the other independent variables in the model. When R^2 equals 1, it implies that one or more independent variables in the model can be perfectly predicted from the other independent variables, resulting in infinite VIF values.

When VIF becomes infinite, it indicates a severe problem of Multicollinearity in the regression model. It means that one or more independent variables are redundant and can be expressed as a linear combination of other independent variables in the model. In such cases, the interpretation of the regression coefficients becomes extremely unreliable, and the model may need to be revised by removing or transforming the redundant variables to address the issue of Multicollinearity.

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

Ans: A Q-Q (quantile - quantile) plot is a graphical tool used to assess the similarity between the observed distribution of data and a theoretical distribution, typically the normal distribution. In a Q-Q plot, the quantiles (ordered values) of the observed data are plotted against the quantiles of the theoretical distribution.

In linear regression, Q-Q plots are particularly useful for:

1. Assumption Checking: They help assess the assumption of normality in the residuals (the differences between observed and predicted values). If the residuals follow a normal distribution, the points in the Q-Q plot will closely follow a straight line. Deviations suggest non-normality.

2. Detecting Outliers: Outliers in the data can be identified by observing deviations from the expected pattern in the Q-Q plot.

3. Model Evaluation: Q-Q plots aid in evaluating the adequacy of the linear regression model. Deviations from the expected pattern may prompt reassessment of the model's assumptions or structure.

In summary, Q-Q plots are essential diagnostic tools in linear regression for verifying assumptions, detecting outliers, and evaluating model performance, all of which contribute to ensuring the reliability of regression analyses.