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- Season, Weather situation, holiday, month, working day and weekday are the categorical variables in the dataset.

Season- The boxplot shows that spring season has the lowest value of 'cnt', while the fall season has the highest value of cnt.

Weather situation- When there is heavy rain/snow, there are no users indicating that the weather is extremely unfavourable. The highest count was observed when the weather forecast was clear, partly, cloudy.

Holiday- Rentals were found to be lower during the holidays.

Month- September had the most rentals, while December had the fewest. The observation is comparable to the one made in weather situation.

Weekday- Weekends saw a significant increase in book hiring compared to weekdays.

Working day- It had little effect on dependent variable.

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

Ans- **drop_first=True** is important to use, as it helps in reducing the extra column created during dummy variable creation. It reduces the correlation among dummy variables.

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

Ans- 'Temp' and 'atemp' are the numerical variables which has highest correlation with target variable 'cnt'.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans-By checking- Normality or errors through Residual analysis. The distribution of residuals should be normal and centred around 0. We test this residuals assumption by producing a distplot of residuals to see if they follow a normal distribution or not.

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

Ans-Temp, weather situation and yr are the features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions

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

Ans- Linear regression is defined as an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events.

It is divided into two parts:-

Simple Linear Regression: SLR is used when the dependent variable is predicted using only one dependent variable.

Multiple Linear Regression: MLR is used when the dependent variable is predicted using multiple independent variables.

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

Ans- Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different. It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.

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

Ans- The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

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

Ans- Scaling a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

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-squared, etc.

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1.

sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

***Standardization* Scaling:**

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

sklearn.preprocessing.scale helps to implement standardization in python.

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

Ans- 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.

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

Ans- Q-Q plots are graphical tools that help you assess the validity of some assumptions in regression models, such as normality, linearity, and homoscedasticity.
The advantages of the q-q plot are: The sample sizes do not need to be equal. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.