Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

There are 7 categorical variables present in the dataset and their impact can be visualized through the box plot:

i. **Season**: During spring, the bike demand is lowest and during fall its highest.

ii. **Year(2018 and 2019):** The bike demand increase in the later year(2019) than 2018. It gives an impression about increased business year by year.

iii. **Month(JAN to DEC):** The bike demand is lowest in JAN and keeps on increasing till OCT after which the demands drops till December.

iv. **Holiday**: The plot shows that the median of bike rental is more on working days than the holidays. The body for the 'non-holiday' being shrinked around median shows same number of users using the bike rental services.

v. **Weekday:** The weekday does not have much impact on the demand is the distribution for all days are around same and median is around the same range.

vi. **Workingday:** The average demand is same on working and non-working days. Also, on the working days, almost same number of people are availing the bike service.

vii. **Weathersit:** The demand is highest in clear weather.

The impact of these categorical data can be visualized in the boxplot as below:

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

We use drop_first=True to reduce the number of columns. This reduces the multicollinearity among the dummy variables.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

The highest correlation from the numerical variables with the target variable 'cnt' is with "temp" and "atemp". There is high +ve correlation between them.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)

By checking below parameters:

i. Normal distribution of Error Terms.

ii. Residual Plot

iii. VIF for multicollinearity

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

The 3 most important features contributing towards the bike demands are:

i. **Temperature (temp):** This is positively impacting the bike rental demand.
ii. **Light snow & Rain weather:** This is impacting negatively to the bike rental demand.
iii. **Year-2019:** This has a positive correlation with the bike rental demand.

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Linear Regression algorithm is a statistical method to find out the target/predictive variable by using the independent variable.

It gives the relationship between how change in one independent variable impacts the target variable.

The equation for linear regression algorithm is:

$Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3$

Where Y = Target Variable

And X_1, X_2,X_3.. = Independent variable

B_0 = Intercept/Constant

B_1, B_2, B_3.. = Coefficient of X_1, X_2, and X_3 respectively.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Anscombe's quartet consists of 4 datasets with almost same statistical summary but different graphs when presented over scatterplot.

It is to emphasize that numerical summary could be misleading, and visualization of the data is important.

The major observation is done on the mean, standard deviation and regression line where each of these are same for each dataset but qualitatively the datasets are different.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

It is the correlation coefficient that measures the linear relationship between 2 datasets. The value lies between -1 and 1 and shows relative strength.

The Formula for the person's R is:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

The scaling is done to bring the variable which is of high unit under same range as other variables. This is to get every variable on same scale without other getting higher weightage.

Normalized Scaling:

Xnorm = (Xi – mu)/(Xmax – Xmin)

Here, all the values ranges from -1 to 1. The only drawback here is that even outliers are confined to 1 or -1 instead of showing significantly on the graph.

Standardized Scaling:

Xstd = (x - mu)/sigma

It is not bounded under any range but gets normalized around mean and standard deviation.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

VIF provides the amount of multicollinearity between 2 variables. If the VIF is infinity, it shows that are 2 variables which are having very high correlation. This will impact the model reliability. In such cases, we should drop one variable out of 2 which are with high correlation.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

A Q-Q plot or quantile-quantile plot is a scatterplot which is obtained by plotting 2 sets of quantities against each other. It helps us in assessing that the set of data came from same population or not. If the distribution follows as x=y,then the dataset follows the normal distribution otherwise data sanity needs to be checked.