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.

- Among all four season in the season of spring the demand of bike is low as compared to other three.
- The demand of bike has increased in year 2019 as compared with 2018.
- The demand of bike is lowest in month jan while Jun to Sep is the period when bike demand is high.
- The demand of bike is almost similar in all weekdays.
- There is no change in demand of bike from working day and non working day.
- The demand of bike is higher in 1: Clear, Few clouds, Partly cloudy, Partly cloudy while the demand of bike is lowest in 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

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

Ans.It important to use **drop_first=True** during dummy variable creation as

- It reduces the extra column which is unneeded .
- If do not drop that dummy column it will not give proper correlation results.

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 The numerical variable **'registered'** has the highest correlation with the target variable 'cnt'
(1 mark)

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

Ans.

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

- Temperature ,
- Weathersit,
- Year

General Subjective Questions

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

Ans. Linear Regression is used for predictive analysis . This algorithm comes under supervised learning . Linear Regression shows a linear relationship between dependent and independent variables which means it finds out if the value of dependent variable (y) is changing according to the value of independent variable(x).

Example -

- Salary and years of experience
- Price of potato depend on no of kg we buy

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

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

Ans.

- Pearson's Correlation coefficient measures the linear correlation between two sets of data .
- It is the ratio between the covariance of two variables and the product of their standard deviation.
- The value of Pearson's R lies between -1 to +1 where -1 shows strong negative relationship between two variables 0 means there is no correlation and +1 shows the strong positive relationship between variables.

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

Ans.

Scaling is one of the step of Data preprocessing .

- Which is performed to scale the data in particular range.
- For Normalized scaling MinMaxScaler is used from sklearn library(sklearn.preprocessing.MinMaxScaler) which brings all the data in 0 to 1 range .

- standardized scaling is used to bring data intp standard normal distribution which has mean as meu and standard deviation as sigma by replacesing values by z-scores .

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
(3 marks)
Ans. If VIF is infinite means there is perfect correlation between two independent variables.
When R2=1 it makes 1/(1-R2) infinity.

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 plot is used to find out if two datasets are come from same distribution
- Q-Q plot shows graphical representation of statistical properties such as scale, skewness are same of two distributions .