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 – Based on the analysis of categorical variable count of bike rent increases mainly

- Summer and fall session
- More bike renting in the year 2019, which mean renting bike business is growing as compared with 2018 and post covid, the business may back to normal.
- Month of Sep have more bike renting compared to any other month
- Saturdays, Wednesday & Thursday have more bikes on rent

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

Ans – Dummy variable creates separate variable for each category, for example if we have n categories, then it will create n features (columns). It means while creating model we need to take care of an additional variable. The same effect can be obtained by dropping one of the columns from dummy variable using drop_first, in that way we need to analyse n-1 variable instead of n, which will save effort, time & avoid redundancy.

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

Ans. 1. Temp / atemp

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

Ans.

- Validate the linear regression by validating the VIF and P values,
- by analysing the Residual error analysis
- and linear relationship between dependent and target variable.

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

Ans – Top 3 variables Actual temperature (atemp), year , spring

General Subjective Questions

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

In statistical modelling, linear regression is a process of estimating the relationship among

variables. The focus here is to establish the relationship between a dependent variable and one or more independent variable(s). Independent variables are also called ‘predictors’.

Regression helps you understand how the values of dependent variable changes as you change the values of 1 predictor, holding the other predictors static (or same). This means that simple linear regression in its most basic form doesn’t allow you to change all the predictors at a time and measure the impact on the dependent variable.

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

- Anscombe’s Quarter can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, yet appear very different when graphed. Each dataset consists of eleven points. It majorly demonstrates the importance of graphing data before analysing it and effect of outlier on statistical properties.

3. What is Pearson’s R? (3 marks)

- It is the most common way of measuring a linear correlation. It’s a number between -1 to 1 that measures the strength and direction of relationship between two variables.

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

Scaling is a technique by which independent feature variables in dataset standardise to a fixed range.
In a dataset there are chance that the independent variable varies in different unit, in that scenario if we create model with this values, chances are the model will predict incorrect values.
Normalization is a process where every independent variable are fixed in between 0-1. This is also called min-max scaling
minMaxScalling : x = (x-min(x))/max(x)-min(x)
Standardization is a technique where the values are centred around the mean with a unit standard deviation.

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

VIF = 1/(1-R^2), if R^2 is 1 Then VIF is inifinity, this means a perfect correlation, an infinite VIF value indicates that the corresponding variable may be expressed exactly by linear combination of other variable.

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to assess if a set of data plausibly came from same theoretical distribution such a Normal, Exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.