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)

Season have a very significant impact, the dependent variable (cnt) i.e. total no of bike rentals varied across seasons with fall being high and spring being the least. Relatively spring is significantly impacted compared fall, summer and winter

Similarly, weather situation has impact on bike rentals with clear weather being the most preferable and light snow being the least.

Year 2018 have less bike rentals than 2019, probably in one year bike rentals got more publicity and its usage increased

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

We don't need first column because that values can be derived with the help of other columns. This will also help to avoid perfect multicollinearity and improve the stability of model

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

Temp have more correlation with target variable

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

Residual analysis – plotting the error terms distribution and checking if they are normally distributed. Anyway, while building the model, it is also ensured to drop the independent variables with high VIF and reduce multicollinearity.

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

Yr, temp and light snow are the variables with high magnitude coefficients, and they are significantly contributing to the demand of shared bikes. These are followed by wind speed and spring

General Subjective Questions

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

Linear regression algorithm is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data . Here linear equation will be in the form of y = m1x1 + m2x2 + m3x3 + ... + C

y – dependent variable , x1,x2,x3.... are independent variables, m1,m2,m3 are slopes and C is intercept

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

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation and linear regression) but differ significantly when graphically visualized.

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

Pearson's correlation coefficient (R) is a measure of linear relationship between two variables it quantifies the strength and direction of a linear association between two continuous variables.

R = 1 -> perfect positive linear relation ship

R = -1 -> perfect negative linear relationship

R = 0 - > No linear relationship

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

Scaling is the process of transforming variables to a specific range. It is performed to ensure the units of different scales are comparable. Normalized scaling transforms the data to a specific range often between 0 and 1. standardized scaling transforms data to have a mean of 0 and standard deviation of 1.

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

As VIF = 1/(1-R^2) and as R tends to get equal to 1 VIF will become infinite. This is because of variables which have perfect multicollinearity. That means one or more variables can be exactly predicted by a linear combination of others.

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 used to assess whether a given dataset follows a specific theoretical distribution. In linear regression they are used to check the normality of residuals. This helps in checking the linear regression model assumptions