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: From the analysis of the categorical variables from the dataset it could be inferred the maximum bike rentals will be in summer and the winter season, mostly prominent in the month of September, in the year of 2019. During the weekend, especially on Sundays, there is a drop in the number of users. Also, when there is rain and mist, there is a decrease in the users.

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

Ans: drop_first=True helps in reducing the extra column created during the dummy variable creation and hence avoid redundancy of any kind. And might result in high correlation if the redundant columns are not dropped.

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

Ans: atemp and temp have the highest correlation with the target variable.

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

Ans: Validated the assumptions of linear regression by checking the VIF, error distribution of residuals and linear relationship between the dependent variable and a feature variable. No significant multicollinearity among variables.

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: The top 3 features contributing significantly towards the demand of the shared bikes are the year, the temperature and the windspeed (Negatively influencing).

General Subjective Questions

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

Ans: Linear Regression is an ML algorithm used for supervised learning. It helps in predicting a dependent variable(target) based on the given independent variables.

Mathematically the relationship can be expressed as Y = mX + c

- Y is the dependent variable we are trying to predict
- m is the slope of the regression line which represents the effect of X on Y
- c is constant, known as the Y-intercept.

Linear relation can be positive or negative

After looking into the data and cleaning it with exploratory data analysis (Univariate, Bivariate, and multivariate analysis)

Next step would be Data preparation

Load the data and understand it using dictionary provided.

Convert the columns to proper data types.
Create dummies for categorical variables.

Model Building:
Divide the data into train and test
Perform Scaling
Divide the data into X and Y
Perform Linear Regression
Use mixed approach (RFE and Manual)

Model Evaluation:
Check the various assumptions
Check the Adjusted R-square for both test and train data. (Residual analysis)
Report the model.

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

Ans: Anscombe's quartet consists of four data sets that have nearly identical simple descriptive statistics but have very different distributions and appear very different when presented graphically. Each dataset consists of eleven points. The primary purpose of Anscombe's quartet is to illustrate the importance of looking at a set of data graphically before beginning the analysis process as the statistics merely do not give accurate representation of two datasets being compared.

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

Ans: Pearson's Correlation Coefficient is used to establish a linear relationship between two quantities. It gives an indication of the measure of strength between two variables and the value of the coefficient can be between -1 and +1.

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

Ans: Scaling is a technique performed in pre-processing during building a machine learning model to standardize the independent feature variables in the dataset in a fixed range.

The dataset could have several features which are highly ranging between high magnitudes and units. If there is no scaling performed on this data, it leads to incorrect modelling as there will be some mismatch in the units of all the features involved in the model.

The difference between normalization and standardization is that while normalization brings all the data points in a range between 0 and 1, standardization replaces the values with their Z scores.

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

Ans: The value of VIF is infinite when there is a perfect correlation between the two independent variables. The R-squared value is 1 in this case. This leads to VIF infinity as VIF equals to 1/(1-R2). This concept suggests that is there is a problem of multi-collinearity and one of these variables need to be dropped in order to define a working model for regression.

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

Ans. Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.