Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer abouttheir effect on the dependent variable? (3 marks)

Ans: In the provided data set the categorical variables are: 'season', 'yr', 'mnth', 'holiday', 'weekday', 'workingday' and 'weathersit'. And The explains the following.

- Season Spring(fall) has the strongest demand for bike rentals.
- I've seen that demand for next year has increased.
- Demand keeps rising month by month till June. September month has the highest demand for rental bikes. After September, demand is falling.
- During holidays, there is a decline in demand.
- The weekday does not provide a good picture of demand.
- Clear weathershit is the most in-demand.
- During September, bike sharing is greater. It is lower towards the end and beginning of the year, perhaps owing to extreme weather conditions.

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

Ans: drop first= True is essential since it reduces the extended column formed during the construction of dummy variables. Consequently, it lowers the correlations between dummy variables.

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

Ans: There are two variables which are evenly correlated with the target variable they are temp (temperature in Celsius) and atemp(feeling temperature in Celsius)

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

Ans: The following measures are taken to be consideration whilw validating a Linear Regression model:

a. There should a Linearity of the data
b. Predictors are independent and observed with negligible error
c. Residual errors have a mean value of zero
d. The VIF must be less than 0.05.
e. Residual errors are independent of each other and predictors

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

Ans: The top three features are yr(Year), temp(temperature in Celsius), and weathersit_bad (Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog). We can say that the demand for shared bikes is not fully but mostly depends upon temperature, weather conditions and the year.

General Subjective Questions

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

Ans: Linear Regression is a supervised learning-based machine learning technique. It carries out a regression operation. Regression models a predicted value based on independent features. Primarily, it is used for determining the correlation between variables and predicting values. The kind of relation considered between dependent and independent variables and the number of independent variables differs in the various regression models.

The steps involved in the linear regression model are as follows:

a. Data loading and understanding of the Data
b. Preprocessing
   I. Dropping Un-necessary or Non-Relevant columns for our analysis
   II. Mapping values of categorical columns with Actual Values for better understanding
   III. Dummy Variable creations for categorical columns
c. EDA and Visualizing the Data: To analyze the categorical and numerical feature and their co-relation
d. Splitting the Data into Training and Testing Sets
e. Missing Value Imputation
f. Rescaling the Features
g. Feature Selection
h. Model Building
i. Model Evaluation

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

Ans. Anscombe's Quartet may be described as a collection of four data sets that are almost equal in terms of basic descriptive statistics, but have idiosyncrasies that trick the regression model if it is constructed. They have vastly different distributions and scatter plots depict them differently. In our data set.

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

Ans. Pearson's R (Pearson correlation coefficient or PCC) measures the degree of the linear relationship between two variables. for example, we have two variables X & Y which are linearly correlated. The more linear X and Y are, the closer Pearson's correlation coefficient will -1 if the correlation is negative or +1 if it is positive. The PCC of perfectly linearly uncorrelated will 0.

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

Ans.

1. Feature scaling is a technique used to standardize the range of independent variables or Features. It is also known as data normalization and is often conducted during the pre-processing phase of data processing. Oftentimes, we see that the range of data values changes considerably. In certain machine learning methods, objective functions cannot operate effectively unless they are normalized. Another purpose for using feature scaling is that gradient descent converges significantly more quickly with feature scaling than without it.
2. The objective of applying feature Scaling is to ensure that features have nearly same scales so that each feature is of comparable importance and is simpler to process by the majority of ML algorithms.
3. In both instances, you are changing the values of numeric variables such that the modified data points possess certain advantageous qualities. In scaling, you alter the data's range, but in normalizing, you alter the shape of the data's distribution. Scaling only modifies the data's range. Normalization is a more fundamental change. The objective of normalization is to transform your data into a normal distribution.

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

Ans. VIF = infinite if a perfect correlation exists in the selected features. This demonstrates that there is a perfect correlation between two independent variables. R2 = 1 in the event of perfect correlation, leading to 1/(1-R2) infinity. To resolve this issue, we must exclude one of the feature that is creating this perfect multicollinearity.

A VIF score of infinity implies that the variable in query may be described precisely by a linear combination of other variables (which show an infinite VIF as well).

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 Graphs (Q-Q Plots) are plots of two quantiles against one another. A quantile is a fractional number below which certain values fall. For instance, the median is a quantile at which 50% of the data are below and 50% lay above. Q Q plots are used to determine whether two sets of data are from the same distribution.