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)

Basis the initial analysis, following categorical variables seem to be affecting dependent variable:

a. Season 3 has the highest median of bike sharing users

b. Month 9 shows similar trend, of highest median

c. Weather situation 1 and year 2019 show the increased usage

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

Because the extra column that gets added during dummy variable creation gets dropped using this option

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

Temperature has the highest correlation with count

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

a. Checking the residue plot spread

b. Mean of the residuals

c. Correlation of the residuals

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

a. Month August

b. Month January

c. Weekend

General Subjective Questions

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

a. Basis the input features, it helps predict the value of target variable.

b. It assumes linear relationship between the input features and the target variable.

c. Basis the linear regression equation, $y = b0 + b1x1 + b2x2 + b3x3 + ...$, from the given dataset, it fits a model and computes the coefficients for dependent variable and constant, if present.

d. In this process of computing the coefficients, an OLS based approach will seek to minimise the sum of square of the error or residuals.

e. Basis the approach chosen, say Gradient method, OLS, etc., the algorithm will converge basis the defined minimisation criteria.

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

a. It consists of 4 datasets that have nearly similar descriptive statistics, i.e., mean, variance, standard deviation, but have a different distribution when they are plotted.

b. Basis data visualisation, the difference in their behaviour could be noted.

3. What is Pearson's R? (3 marks)
a. It is a correlation coefficient used to measure linear correlation.
b. The value lies from -1 to 1.
c. Value 0 indicates no correlation.
d. (0,1] indicates that both the variables change in same direction.
e. [-1,0) indicates that both the variables have opposite behaviour.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
a. Scaling indicates pre processing of data to bring the entire dataset in a similar range.
b. This helps in speeding up the algorithm and reducing computational power.
c. Normalized scaling: $x = (x - \min(x))/(\max(x) - \min(x))$
d. The above helps bring the data in the range [0,1].
e. Standardization replaces the values by Z-score. It is given by: $x = (x - \text{mean}(x))/\text{sd}(x)$

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
a. Larger the VIF, higher the correlation between the variables.
b. If VIF is infinite, it indicates that there is a perfect correlation between the 2 variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
a. A Q-Q plot is used to compare a data set to a theoretical model.
b. Goodness of a fit can be graphically viewed using this.
c. These could also be used to compare 2 theoretical distributions.