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:

1. Season 3 has highest demand for rental bikes.
2. Weekday does not give clear understanding of demand
3. The clear weather has highest demand
4. In holidays the demand is decreased
5. Demand is continuously growing till June. September,has highest demand.

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

Ans: It helps in reducing extra column created during dummy variable creation. If not done can create multicollinearity issues.

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

Ans: The feature temp

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

Ans:

1. Linearity Check
2. Multicollinearity check using VIF
3. Error terms do not follow pattern
4. Error terms are normally distributed with mean 0
5. Rechecking R2 and Adjusted R2 values

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: 'holiday', 'temp' and 'hum' are highly related with target model

General Subjective Questions

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

Ans: In Regression, we plot a graph between the variables which best fit the given data points. The machine learning model can deliver predictions regarding the data. In naïve words, "Regression shows a line or curve that passes through all the data points on a target-predictor graph in such a way that the vertical distance between the data points and the regression line is minimum." It is used principally for prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables.

Linear regression is a quiet and simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables.

To calculate best-fit line linear regression uses a traditional slope-intercept form.

Y= b0 +b1*X

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

Ans: Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

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

Ans : Pearson's Correlation Coefficient is a linear correlation coefficient that returns a value of between -1 and +1. A -1 means there is a strong negative correlation and +1 means that there is a strong positive correlation. A 0 means that there is no correlation (this is also called zero correlation).

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

Ans : It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm. Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

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

Ans : If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Ans : Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution.