Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer: As per the analysis of categorical variables, I can say that

Bike rent is higher in 2019 than 2018

Bike rent is high in normal week days than holiday or week ends

Bike rent is high in month of September

Bike rent is high on Saturday compared to other days.

Bike rent is high in clear weather.

Q2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Answer: It is important to use drop_first=True during dummy variable creation because it will reduce an extra column creation while dummy columns creation.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer: temp variable has the highest correlation with the target variable.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer: By looking into p-value to make sure that selected features will have minimum value and must be < 0.05.
By looking into VIF value to make sure that selected features will have minimum values and should not be > 5
By looking into R-Squared and Adjusted R-Squared where the difference between them should be less.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer: Based on the final model, below 3 are the top features contributing significantly towards explaining the demand of the shared bikes:

Temp (positive corelation)
Year (Positive Corelation)
weather_moderate (Negative Corelation)

General Subjective Questions

Q1. Explain the linear regression algorithm in detail. (4 marks)

Answer: Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous or numeric variables such as sales, salary, age, product price**,** etc.

Linear regression can be further divided into two types of the algorithm:

Simple Linear Regression:
If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

Multiple Linear regression:
If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression

Q2. Explain the Anscombe's quartet in detail. (3 marks)

Answer: 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 analysing it and the effect of outliers on statistical properties.

Q3. What is Pearson's R? (3 marks)

Answer: In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus, it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer: 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.

Normalization/Min-Max Scaling:
It brings all of the data in the range of 0 and 1. **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.

Standardization Scaling:
Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer: The value of VIF is calculated by the below formula:

VIFi = 1 / 1-Ri^2

Where, 'i' refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer: 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. A 45-degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.