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)
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
2. Explain the Anscombe's quartet in detail. (3 marks)
3. What is Pearson's R? (3 marks)
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answers to Assignment based Subjective Questions

Question 1.

Categorical variables needs to be converted to quantitative variables/dummy variables to get meaningful relationships on the dependent variables.

Question 2.

drop_first = True is required, as the absence of all other variables (zero values) can be used to encode the first column.

Question 3.

Registered, casual, atemp, temp are the highly correlated values with target variable.

Question 4.

By using R2 values and IHF values the model is refined. Then using the test dataset values are validated and verified, if the predicted values are close to the actual values.

Question 5.

Temp

Answers to General Subjective Questions

Question 1.

Linear regression algorithm

Linear regression algorithm is used to predict linear relationships between target variable and independent variable.

Simple Linear regression

$$y = mx + c,$$

Where m is the slope and c is the constant, in relationship between two variables which is plotted to be a straight line.

Multi variate Linear regression

$$y = m1x1 + m2x2 + m3x3 + c$$

A suitable value for coefficients are selected and cost function is checked, if minimal.
The coefficients m and c are selected randomly.

There are several cost functions like Root Mean Squared Error. The plotted values are compared against actual values by applying the cost function and the error is determined.

m & c are updated incrementally, so that we find the values for which cost function is as minimal as possible.

This is done first on a training data set. Using the identified linear regression, values are predicted for a test data set.

If the predictions are off, there may not be a linear relationship.

Question 2.

Anscombe quartet is 4 sets of 11 datapoints (x,y) with almost same standard deviation and mean for both x and y as well as correlation between x and y.

However, the nature of these curves when plotted in a graph are completely different.

This shows, why we must analyse and plot datapoints to gather maximum possible insights into the data.

Question 3.

Pearson's R is correlation coefficient with values between -1 and +1, that identifies the degree of linear relationship between two variables x, y.

R = -1 , Negative correlation - x decreases with increase in y
R=1, Positive correlation - x increases with increase in y
R= 0, No linear correlation - Some other non-linear correlation might exist.

This can be easily calculated in Excel, Google sheets, Python etc.

Question 4.

All independent variables might have different ranges of values. So the linear relationship might not be predicted accurately on the same scale. So we use scaling to control the deviations.

Normalization is performed to make sure all the values are in the range between 0-1.

Xi - Xmin / Xmean -Xmin

The above formula can be applied to all the values to perform normalization.

Standardisation is performed by using the z table and distribution is changed in such a way that Nth percentile of values can be easily determined.

Question 5.

If there's collinearity between two or more independent variables, i.e, they've a perfectly linear relationship, it causes potential issues with the linear regression model.

To avoid this, only one of this variable will be used in model building.

VIF (Variance Inflation Factor) is used to identify the effects of the high collinear variables on the model.

If VIF is infinity, it means variables are highly collinear, and one of the variable can be eliminated.

Another approach can be Principle component analysis (PCA) to derive more independent variables from the collinear variables to better explain the model.

Question 6.

Q-Q plot (Quantile Quantile plot) is the plot in which quantiles of two data sets are plotted against each other to ensure that the distribution is similar or to identify the distribution followed by the datasets.