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)

All the categorical variables have –ve coefficients except for month September and Winter season.

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

If there is a categorical column with only n values i.e season in the dataset, without drop_first = True, get_dummies function would create 4 columns for 4 seasons and each season will be having either 0 or 1 based on the season corresponding to the row. Ideally we can ignore one column, as we could get the final column value from the other columns . Drop_first= true will ignore one column and only create n-1 columns. Also having more columns increases r-square and if the column is not necessary it's better to remove it.

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

Temp has the highest correlation with the target.

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

- Residuals should be Normally distributed i.e. Mean should be cantered around zero
- No Multi collinearity i.e VIF should be < 5
- The relationship between target and independent variable should be linear . This can be observed using scatter plot.

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

Temp , yr and winter season contribute significantly towards demand of the model.

General Subjective Questions

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

It's a supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent variables. The aim of algorithm is to find the best fit linear equation that predicts the dependent variable from one or more independent variables. It has the following stages

- Data pre-processing i.e. convert the categorical columns to numerical by using One Hot Encoding technique
- Apply scaling on all the numerical columns so that all of them are on the same scale.
- Check whether the assumptions of linear regression are satisfied
- Build a model using Trained Data Set
- Test the model by predicting the Test Data
- Evaluate the model by comparing r-square value from train and test data sets

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

It consists of 4 datasets that have same statistical values i.e mean, variance, standard deviation, correlation but still will be showed very differently when plotted on a graph. This was intended to counter the impression that calculations are exact but graphs are rough.

So its always a good practice to plot the graph initially before proceeding with model building.

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

It's a numerical summary of the strength of linear association between the variables. The value of Pearson coefficient is always between -1 and 1. Using this value , we can determine

- Correlation between two variables
- How strong the correlation is
- Is the correlation positive or negative

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

Scaling is a way to normalize the range of independent numerical variables in data so that they are all in the same range. If the scaling is not done, then algorithm takes magnitude into account and not the accounts.

Standardized:

It brings all the variables such that the mean is zero and variance is one. Should be applied for variables which follows a Gaussian distribution when plotted

Normalized:

It brings the variables between 0 and 1. Should be applied for variables which doesn't follow a Gaussian distribution when plotted

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

This happens when one variable is able to create perfect multiple regression on other variables.

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

This is used to plot the quantiles of the one dataset against the second data set.It is used to identify the following

- Does 2 data sets come with a common distribution
- Does 2 data sets have a common location and scale
- Does 2 data sets have similar distribution/shape