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: All of the independent, categorical variables have some statistically significant effect on the dependent variable.

I chose to:

1. First, replace all the integers with their actual meaning using the Data Dictionary
2. Then, I created dummy variables out of the updated columns

And I observed that because of choosing this approach, I was able to get a nuanced view on the categorical variables and was able to see what categories within a categorical variable had a statistically significant effect on the dependent variable.

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

Ans: It is important to write drop_first = True during creating dummy variables because we can explain the first dummy variable using other dummy variables. Example - if a categorical variable had three levels, we will have 3 dummy variable columns. However, we can explain the first 1st column using the other two, meaning that if the other two are 0 then it would definitely mean that the first column is 1.

This would increase multicollinearity and in order to avoid that, we should drop the first column while creating dummy variables.

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

Ans: The highest correlation is with the Registered variable. However, that would obviously be the case as ‘cnt’ is the sum of Registered and Casual. Apart from these 2 variables, atemp has the highest correlation among the numerical variables.

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

Ans: By performing residual analysis and creating a histogram to see if the error terms are normally distributed or not and the mean is towards 0.

Apart from that, I used VIF to eliminate the chances of multicollinearity.

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

1. Year
2. Spring
3. Light Snow

General Subjective Questions

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

Linear Regression is a statistical way to predict the value of a dependent variable using one or more independent variables.

First, you choose all the independent variables that you want to work with and then run an OLS analysis, which minimizes the sum of squares of the difference between predicted values of the dependent variable and the actual values.

The goal is to create a line of best fit. It would be presented in terms of $Y = a + bX1 + cX2 + dX3$…., where Y is a dependent variable and X1, X2, X3…are an individual independent variable.

After that, we can go through the summary to see what all independent variables are able to explain a portion of the dependent variable in a statistically significant manner.

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

It is a way of visualizing your data and then analysing it instead of only relying on the summary statistics to make conclusions.

Because of Anscombe's quartet, we are making sure that we are accounting for curvature, anomalies and outliers through data visualization.

If two datasets had the exact same summary statistics, it doesn't mean that they would look the same when plotted on a graph and it is important that we account for all the details that we can only get once we have conducted data visualization.

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

It is the coefficient of the linear correlation between two columns. It tells us whether the two columns are inversely or directly related and tells us in what intensity they are related, meaning higher the absolute value of the coefficient, stronger the correlation.

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

It is a way to process the data before running analysis on it to make sure that the entire data is within a certain range. It helps in generalizing a certain range for all the variables and it also helps in expediting the running of code.

Normalized scaling brings everything between 0 and 1. However, Standardized scaling replaces the value with how many standard deviations far the value is from its mean, i.e. the Z score.

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

If there is a perfect correlation, i.e. 1, among variables, then the VIF value would be infinite.

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

Q-Q Plots are plots of two quantiles against each other. It assesses if a set of data came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.