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)

Count = 0.4648 * atemp + 0.2357 * yr + 0.0258 * workingday + 0.0092 * weekday + 0.0474 * winter -0.1252 * spring - 0.0752 * (Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist) -0.2915 * (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)

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

Ans: There are few reasons, why "drop_first=True" is required to be used during dummy variable creation, which have been mentioned below;

Multicollinearity and Interpretability:

While dummy variables are being created for a full data set, the resultant dataset consists of a redundant dummy variable, which introduces multicollinearity in the data set. This situation dilutes the accuracy of coefficient estimations of statistical models.

Multicollinearity leads to straight forwards interpretability of coefficients, which is nothing but dilution of coefficient estimation.

This situation can be avoided by the said function.

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

Ans: Among the numerical variables, "atemp" is having the highest correlation with target variable. More rentals are happening at moderate to high temperature, i.e. around 15 to 35 degrees. This can be interpreted that, people prefer more biking in warmer weather.

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

Ans: Post taring the data set and making the model, did the test run on it. Which gave almost same r2 value. Which concludes model is valid. Also r2 is 81%, that means 81% , which is a decent score for a model.

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: 1. Temperature, 2. Weather- (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds), 3. Season - winter

General Subjective Questions

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

Ans: Algorithm of linear regression is to find out relationship between dependent and independent variable. Which in tern helps to get the more accurate predictions.

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

Ans: It's a set of four different datasets that have almost identical simple descriptive statistics but differ significantly in graphical representation. This has been created by a statistician Francis Anscombe in 1973, the quartet is used to demonstrate the importance of graphing data before analyzing it and to show that different datasets can have the same statistical properties yet be qualitatively different.

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

Ans: It's a correlation coefficient, which measures the strength and direction of linear relationship between two variables, and ranges between -1 to +1.

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

Ans: Scaling is making the data values in a common and comparable range. It is performed for accurate model building and model works on numerical form of ant information. Normalized scaling makes the values between 0-1, which helps to take care of outliers too. Standardized scaling is changing the values around the mean of available data.

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

VIF = 1/(1-r2), infinite VIF means r2 is 1, than means 100 of the variables can be explained by other variables, which is exactly opposite of the basic requirement in any linear regression model building, which in multicollinearity.

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

It is a graphical tool used to compare the distribution of a dataset with a the normal distribution. It helps to assess whether a dataset follows a given distribution, highlighting deviations from the expected distribution.