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?

Ans: We saw from various bar plots that for different categories of variables like **Season, Holiday, yr , weathersit, mnth** the average of **cnt** variable was different but for variables like **Weekday, workingday** the demand was approximately same for **cnt**. The reason for this is high demand by casual users for these variables which compensates the decrease in demand of registered ones.

Q2: Why is it important to use drop_first=True during dummy variable creation?

Ans: It removes the extra variable resulted from category of various variables and hence reducing the correlation among dummy variables.

Q3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: From the pair plot, the variable **registered** was highly correlated with the **cnt**, then comes **casual** but these were not predictors as they were also dependent on other given variables from which we can say **temp** variable was highly correlated with **cnt** with a value of **0.63.**

Q4: How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: By checking the residuals distribution plot which should be Normal in nature centered around 0 which was in our case. Residuals are the difference of predictions of model on training set and the actual values.

Q5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans: **Temp, yr_2019, season_winter** contributed significantly in increasing the demand.

General Subjective Questions

Q1: Explain the linear regression algorithm in detail.

Ans: Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

When there is a single input variable (x), the method is referred to as simple linear regression. When there are multiple input variables, literature from statistics often refers to the method as multiple linear regression.

Different techniques can be used to prepare or train the linear regression equation from data, the most common of which is called Ordinary Least Squares. It is common to therefore refer to a model prepared this way as Ordinary Least Squares Linear Regression or just Least Squares Regression.

Equation for Linear Regression:

$Y = B_0 + B_1X_1 + \ldots\ldots + B_nX_n$

Q2: Explain the Anscombe's quartet in detail.

Ans: Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.

Q3: What is Pearson's R?

Ans: It tells about the degree of correlation among various features by its value and there are different ways of calculating correlation value among which Pearson's method is one.

Formula is:

$R_{xy} = S_{xy} / S_xS_y$

Q4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: Scaling is bringing numerical features to a similar scale so that during training the coefficient is not affected by the scale of different variables and thus making the predictions much more accurate.

Normalized Scaling – In this we scale the values between 0 and 1 giving value of 1 to the maximum value and 0 to the minimum value. It is also known as MinMax Scaling.

Standardized Scaling – In this we scale the values in terms of Standard Deviations the particular value is far from mean which is 0.

Q5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: This shows the value of r^2 = 1 which signifies perfect correlation among given variables.

Q6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans: Q-Q plots are plots two quantiles against each other.

It is used to compare the shapes of distributions, providing a graphical view how properties such as skewness and scale similar or different across quantiles.