1) From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans - Based on the findings of the categorical data analysis of these variables; season, weather condition, holidays, and working days, we can conclude that these factors affect the demand for bikes. For instance, weather plays a role as people will rent bikes more often during favorable weather such as spring and summer but will rent few bikes during winter or during a rainy season. In the same manner, there are relatively less bike rents over the weekends and specific holidays, thus suggesting that casual or daily commutes contribute a significant proportion of bike rents.

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

Ans - In most cases, it is advisable to set drop_first=True when creating dummy variables to overcome the multicollinearity problem. This is done by deleting one category in each of the sets of dummy variables, which helps in avoiding the problem of perfect multicollinearity. If you omitted this step, you would be left with too many correlated endogenous variables which would only complicate the model and may not allow the correct determination of the coefficients.

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

Ans - From the pair-plot generated, it can be noticed that the variable ‘registered’, the user type has a higher correlation coefficient with the total count of bike i.e., ‘cnt’. This makes sense since majority of them provide registered user data implying that they are frequent commuters who use the bikes frequently.

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

Ans - As a form of model validation, I conducted the following tests to ensure that the assumptions of linear regression were met. First, I tested for homoscedasticity by checking the residuals and overlaying it on a graph, which should be randomly located about the center of zero. To ensure that the residuals also followed a normal distribution, I also conducted a Q-Q plot. Last of all, I conducted a preliminary analysis for multicollinearity by computing VIF for independent variables.

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

Ans - By observing that the coefficients of the probability weight are greater than 0, we suggest that the features that impact demand for bikes most notably are temperature, the number of registered users, and whether it is a working day. From the moderate temperatures have a positive effect on the number of bikes rented while users using the bike rental service require more bikes during the week due to the daily commute.

General Subjective Questions

1) Explain the linear regression algorithm in detail.

Ans - Linear regression is one of the basic algorithms which provides continuous value outcomes. The goal is to get the line that lays best through the points that have least squares between observed and predicted values of the attribute. This is done using least squares method where a line is fit in a way to minimize the sum of squares of distances between the actual data points and the projected line. The result is a linear equation that fits the input data and can be used to generate further values.

2) Explain Anscombe’s quartet in detail.

Ans - Anscombe’s quartet is a compilation of four datasets that have the same driving numerical characteristics mean and variance but appear highly dissimilar when plotted. This reveals how useful it is to plot information before computing it, as employing measures of central tendency can be deceptive. When comparing the distributions of the two datasets, it is clear to see that two different datasets, with the exact same mathematical measure of dispersion, tell two completely different stories when represented in a graphical form.

3) What is Pearson’s R?

Ans - Pearson’s R also known as Pearson correlation coefficient aims at determining relationship between two variables which is linear. It varies from -1 to 1 where, 1 means perfect positive linear association, -1 means perfect negative linear association, and 0 mean no linear association at all. It is one of the most fundamental coefficients to measure the extent of the association between two variables.

4) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans - Scaling corrects the range of the features in the dataset so that they are within the same range of magnitude. This is especially important in the case of algorithms whose efficiency depends on the range of the data, as is the case of the gradient descent based algorithms. Normalized scaling scales the data values between 0 and 1 while standardized scaling shifts the data to a mean of 0 and a standard deviation of 1, which makes it favorable for algorithms that respond swiftly to the information distribution.

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

Ans - VIF can reach a value of infinity when there is perfect Multicollinearity, that is, if one predictor variable can be re-estimated from the others in terms of a linear equation. This condition reveals the fact that the regression model has the problem of unable to distinguish variables and that results of coefficients are also instable and unreliable.

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

Ans - A Q-Q (quantile-quantile) plot compares the quantiles of the residuals of your model to the quantiles of a normal distribution. It’s used to check if the residuals follow a normal distribution, which is an assumption of linear regression. If the points on the Q-Q plot fall along a straight line, it indicates that the residuals are normally distributed, validating one of the key assumptions of linear regression.