Assignment-based Subjective Questions

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

Answer: From the analysis of the categorical variables from the dataset it could be inferred the bike rental rates are likely to be higher in summer and the fall season, are more prominent in the months of September and October, more so in the days of Sat, Wed and Thurs and in the year of 2019. Additionally we could discern that bike rental are higher on holidays

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

Answer: While creating dummy variable, all the levels are created with zeros and ones. It is by default understood that for two variables only 1 indicater is enough but we are returned with all indicators. To remove the first indicator drop_first=True is very important.

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

Answer: The temp variable has the highest correlation with the target variable

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

Answer: I have validated the assumptions of linear regression by checking the VIF against each variable, plotting the error distribution of residuals and linear relationship between the dependent variable and a feature variable.

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

Answer: Temperature, year and holiday variables are the top 3 features contributing significantly towards explaining the demand of the shared bike.

General Subjective Questions

Question 1: Explain the linear regression algorithm in detail ?

Answer: Linear Regression is a machine learning algorithm which is based on **supervised learning** category. It finds a best linear-fit relationship on any given data, between independent (Target) and dependent (Predictor) variables. In other words, it creates the best straight-line fitting to the provided data to find the best linear relationship between the independent and dependent variables. Mostly it uses **Sum of Squared Residuals** Method.

Linear regression is of the 2 types:

i. **Simple Linear Regression:** It explains the relationship between a dependent variable and only one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.

Formula for the Simple Linear Regression:

$Y=\beta 0+\beta 1X1 +\epsilon$

ii. **Multiple Linear Regression:** It shows the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X. It fits a 'hyperplane' instead of a straight line.

Formula for the Multiple Linear Regression:

$Y=\beta 0+\beta 1X1+\beta 2X2+\ldots+\beta pXp+\epsilon$

The equation of the best fit regression line $Y = \beta_o + \beta_1 X$ can be found by the following two methods:

· Differentiation

· Gradient descent

We can use statsmodels or SKLearn libraries in python for the linear regression

Question 2: Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet consists of four data sets that have nearly identical simple descriptive statistics but have very different distributions and appear very different when presented graphically. Each dataset consists of eleven points. The primary purpose of Anscombe's quartet is to illustrate the importance of looking at a set of data graphically before beginning the analysis process as the statistics merely does not give the an accurate representation of two datasets being compared.

Question 3. What is Pearson's R?

Answer: Pearson's R was developed by Karl Pearson and it is a correlation coefficient which is a measure of the strength of a linear association between two variables and it is denoted by 'r'. it has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

. Positive Correlation: If one value changes, the other also changes in the same direction (increase/decrease).

· **Negative Correlation:** If one value changes, the other changes in the opposite direction.

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

Answer: Scaling is the process to normalize the data within a particular range. Many times, in our dataset we see that multiple variables are in different ranges. So, scaling is required to bring them all in a single range.

There are two types of scaling methods: **Normalization** and **Standardization**. Normalization typically scales the values into a range of [0,1]. Standardization typically scales data to have a mean of 0 and a standard deviation of 1 (unit variance).

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

Answer: The value of VIF is infinite when there is a perfect correlation between the two independent variables. The Rsquared value is 1 in this case. This leads to VIF infinity as VIF equals to 1/(1-R2). This concept suggests that is there is a problem of multi-collinearity and one of these variables need to be dropped in order to define a working model for regression.

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

Answer: The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. It also helps to find out if the errors in dataset are normal in nature or not.