Assignment-based Subjective Questions

(By - Arshi Maheshwari)

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

Ans: 1) Demand of bikes is less in spring season, highest in the fall followed by summer. 2) Year 2019 had higher demand as compared to 2018. 3) August, September, October months have higher demands. 4) Demand for bikes is more on days when weather is clear or less cloudy.

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

Ans: 1) drop_first=True helps in reducing the extra column created during dummy variable creation. Hence it reduces correlations created among dummy variables. 2) For example, if we have 3 categorical variables named furnished, unfurnished, and semi furnished; it can be either furnished or non-furnished, so here our work can be easily done by taking semi-furnished and unfurnished variables.

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

Ans: atemp and temp variables show the highest correlation with the target variable cnt.

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

Ans: 1) Plotted the distribution for the residuals, the distribution is near to normal distribution with mean sum of 0. 2) Plotted graph to look for any pattern between the residuals and independent variables, no pattern exists.

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

Ans: Year, atemp and weather conditions are the top 3 features that contribute significantly.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Ans: 1) Linear Regression comes under Machine Learning Algorithm. 2) It is based on supervised learning. 3) This model is used to predict the target value based on independent variables. 4) This model is used for forecasting. 5) The linear regression model finds out a linear relationship between target variable (y) depending upon independent variable (x). 6) It is of the form $y = \beta 0 + \beta 1x$ where $\beta 0$ is the intercept and $\beta 1$ is the slope.

2. **Explain the Anscombe's quartet in detail.**

Ans: 1) Anscombe's quartet illustrates the importance of plotting the graphs before analysing or building any mode. 2) Anscombe's quartet can be defined as a collection of 4 data sets which are almost identical in simple descriptive statistics, but, there are peculiarities in the data sets, which fools the regression model when built. 3) These data sets have different distributions when plotted on scatter plots. 4) Anscombe's quartet thus tells us how important it is to visualize the data before putting it into any ML model.

3. **What is Pearson's R?**

Ans: 1) Pearson's R also known as Pearson correlation coefficient (PCC) or bivariate correlation, is a measure of linear correlation between two variables. 2) It is given by covariance of two variables divided by the product of their Standard deviations. 3) It lies between -1 to 1.

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

Ans: 1) Scaling is a step performed before processing the data into the ML model, which is applied on independent as well as dependent variables to normalize the data in a particular range. It helps in speeding up the calculations in an algorithm. 2) When data is collected, the different variables are in different units, magnitudes, and range. If scaling is not done, the ML model will only consider the magnitude of the variables which will lead to incorrect modelling. Therefore, it is important to perform scaling to bring the variables in same level of magnitude. 3) MixMax Scaling – Also known as normalization, brings all the data in the range of 0 and 1. 4) Standardized scaling – It replaces the values by their Z scores. It brings all the data into a standard normal distribution which has a mean = 0 and std deviation = 1.

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

Ans: 1) When the independent they show or have a perfect correlation between them, then the VIF comes out to be infinite. 2) In case of perfect correlation, R2=1, which leads to 1/(1-R2) = infinity. 3) Such variables should be dropped before building the model.

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

Ans: 1) Q-Q plot means Quantile-Quantile plots, are plots of two quantiles against each other. Q-Q plot is used to find out whether the two datasets come from a common distribution. A 45-degree angle (reference line) is plotted on the Q-Q plot; if the two data sets come from a common distribution then the points fall on that reference line. 2) In linear regression, Q-Q plot is used to compare the shapes of distributions, and graphical properties such as location, scale, skewness of the two distributions are similar or different.