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?

Ans – From our EDA, we can say that the rentals are more in the year 2019. Users are more likely to rent in Clear weather during Summer and fall season. During May to Oct month, bike rentals are more.

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

Ans – We use drop_first= True during creating dummy variables to reduce redundancy. For a categorical column with n distinct values, n-1 dummy variables are enough for interpretation.

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

Ans – Temperature column has the highest correlation with the target variable.

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

Ans – We have done residual analysis which showed the error terms are normally distributed around 0. We found linear relationship between the features and the target. We have made sure all VIFs are less than 5. Our model has a linear equation.

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

Ans – 1. Temperature

2. Year

3. Light Snow(weathersit)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans – Linear Regression is a Machine Learning algorithm used for supervised learning. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. There are 2 types of linear regressions – simple linear regression and multiple linear regression. Simple linear regression has only one independent variable whereas multiple linear regression has more than 1 independent variables. We find a best-fit line to show the relationship between the independent and dependent variable using linear regression model.

2. Explain the Anscombe's quartet in detail.

Ans –Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed(scatterplots). 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.

3. What is Pearson's R?

Ans – The Pearson correlation method is the most common method to use for numerical variables; it assigns a value between − 1 and 1, where 0 is no correlation, 1 is total positive correlation, and − 1 is total negative correlation.

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

Ans –Scaling of the data comes under the set of steps of data pre-processing when we are performing machine learning algorithms in the data set.

The datasets can be having different units and based on that the numerical values can be in high range. So, to bring all datasets to a comparable range.(0 to 1 or -1 to 1), This helps in better performing models.

Normalization brings the datasets in the range 0 to 1 or -1 to 1. Standardization does not have any range of values , we replace the value with the Z scores.

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

Ans – This shows high multicollinearity between two independent variables and one needs to be dropped to proceed with Linear Regression modelling.

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

Ans –The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.