Bike sharing Assignment

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?**

   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.

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

   It helps to reduce the extra column created during dummy variable creation. It removes the first column which is created for the first unique value of a column.

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

   Temperature variable 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?**

   Validated the assumptions of linear regression by checking the VIF, error distribution of residuals and linear relationship between the dependent variable and a feature variable.

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

   The top 3 features contributing significantly towards the demand of the shared bikes are the temperature, the year and the holiday variables.

General Subjective Questions.

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

   Linear Regression is an ML algorithm used for supervised learning, which investigates the relationship between a dependent and independent variable. Linear regression analysis involves in graphing over set of data points that most closely fits the overall shape of the data. There are two types of regressions – linear regression and logistic regression. Linear regression will be used with continuous variable, where as logistic regression with categorical variables.
   There are two types of linear regression – simple linear regression and multiple linear regression. Multiple linear regression is when multiple independent variables are used to predict the numerical value of the dependent variable.

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

   Anscombe's quartet is a group of four data sets that provide a useful caution against applying individual statistical methods to data without first graphing them. They have identical statistical properties, but look total different when graphed.

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

   The relation coefficient that doesn't just tell us whether two variables move in the same or opposite direction like the covariance, it also indicates how strong the relationship is and its value range from -1 to 1.

   R = covariance / (std.deviation of X * std.deviation of Y)

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

   Scaling is a step of data pre-processing which is applied to independent variables to normalize the data within a particular range. It helps in speeding up the calculations in an algorithm.

   The difference between normalization and standardization is that while normalization helps you to scale down the feature between 0 to 1, where as standarized scaling helps to scale down the feature based on the standard normal distribution.

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

   Variance inflation factor(VIF), $(1 / (1 – R^2))$ is a measure of multicollinearity in the set of multiple regression variable. Multicollinearity occurs when the x variables are themselves related.The value of VIF is infinite when there is a perfect correleation between two independent variables. We need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

   Q-Q plot is a probability plot, to visualize how close a sample distribution is to a normal distribution. This helps us to determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential. It also helps to find out if the error in the dataset are normal in nature or not.