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?

- The categorical variables in the dataset are weathersit and season which plays important role in demand of bikes. As we can see from dataset, the demand of bikes increases with Clear, Few clouds, Partly cloudy, Partly cloudy and decrease with Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog. The same goes with season, from the dataset its evident that the highest demand of bikes are in fall and lowest in spring.

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

- **drop_first=True is important to use, as** it helps in reducing the extra column created during dummy variable creation**. Hence it reduces the correlations created among dummy variables.**

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

- Temp/atemp and registered numerical variable has one of the highest correlations with target variable.

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

- Assumption 1: There is linear relationship between predicted and test values.
- Assumption 2: Error terms are normally distributed.
- Assumption 3: Error terms are independent to each other. We checked it via VIF.
- Assumption 4: The variance between the error terms did not follow ant pattern.

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

- **atemp:** A coefficient value of 0.387364 indicates that a unit increase in atemp variable, increases the bikes demand by 0.387364 units.

- **yr:** Basically there is a positive correlation between yr and bikes hire number. This is evident from EDA as well. A coefficient value of 0.236879 indicates that a unit increase in yr variable, increases the bikes demand by 0.236879 units.

- **light_rain:** A coefficient value of '-0.268408' indicates that a unit increase in light_rain variable, decreases the bikes demand by 0.268408 units.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

- Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x).

2. Explain the Anscombe's quartet in detail.

- Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. 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?

- Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

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

- Scaling is step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm. Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
- **Normalization/Min-Max Scaling:**

- It brings all of the data in the range of 0 and 1.
- sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.
- **Standardization Scaling:**
- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (?) zero and standard deviation one (?).
- sklearn.preprocessing.scale helps to implement standardization in python.

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

- If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
- An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

- Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.
- 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.