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 my model, categorical variables affect a lot in model creation. Even in the final model we can see working day, weather, month, holiday contribute to model.

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

   As after creating dummy variables, number of columns increases, drop 1st reduces number of columns to be handled without affecting the model.

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

   Looking at the pairplot temp and atemp variables have highest correlation with target variable

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

   - Plotted q-q plot to check if variables show linear relationship
   - We checked if distribution of errors is normal
   - We checked if VIF is less than 5
   - We checked & removed multicollinearity

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

   Windspeed, working day & weather summer

General Subjective Questions

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

   In case of linear regression dependant and independent variables show linear relationship with each other. We predict the value of dependant variable for test data based of model we trained.

   In case of MLR, there will be many independent variables while in case of simple linear regression there will be only one independent variable. Co-efficient for each variable & constant for equation will be decided by model, only significant variables will be kept.

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

It comprises of 4 data sets which reminds us that visualizing data prior analysis is a good practice, outliers should be removed while analysing data, data stats do not entirely describe dataset.

3. **What is Pearson’s R?**
   Pearson’s R indicate the correlation between two variables. More the value of r more the correlation.

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

   Scaling is performed to bring variables on a same scale of measurement. For example, there could be 2 variables in dataset A & B. In general, if values of A are much higher than that of values of B, and in this case since values of A are already inflated, it might affect the model. That’s the reason we do scaling.

   In normalization min & max value of features is used for scaling, in standardization, mean & SD is used for scaling. In normalization values between 0 to 1 or -1 to 1. In case of standardisation, it is not bounded to a certain range.

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

   VIF infinity indicated perfect correlation between variables. More IS the correlation more is the VIF.

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

   QQ plot is quantile quantile plot. It is a graphical method to compare two probability distributions.
   In regression qq plots are very important in visualizing how numerical variables are related to each other