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 the model that we built we can say the following things about the rental bike demand

1. You can see more demand for rental bikes during winter than summer
2. During the months of September and October we are likely to see more demand for rental bikes
3. It is dependent on temperature demand is less if temp is more.
4. Demand for rentals will be more on holidays.

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

Ans: While creating dummy variables we can give **drop_first=True** so that we can reduce one more extra redundant column. It is kind of optimisation.

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

Ans: In pair plot we can see variable "temp" has highest correlation with target variable.

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

Ans: We can validate the assumptions using below methods:

1. By checking the VIF
2. Linear relationship between feature variable and target variable
3. Error distribution of residuals.

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

Ans: Top 3 features are temperature, year and holidays according to model.

General Subjective Questions

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

Ans: Linear regression is a ML technique or algorithm in supervised learning. This is used in predicting a variable based on dependent variables. By using this we can establish the relationship between a independent variable and dependent variable. There are two types of linear regression algorithms in ML.

- Simple linear regression
- Multiple linear regression

In simple linear regression we derive the linear relationship between dependent variable and independent variable and by using that model we predict the output. This can be applied only of there is linear relationship among two variables.

In Multiple linear regression we derive the linear relationship between dependent variable and multiple independent variables. After building the model we can predict the target variable using it.

In positive linear relationship dependent variable or target variable will increase with increase in independent variables.
In negative linear relationship dependent variable or target variable will decrease with increase in independent variables.

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

Ans: Anscombe is a method which tells us the importance of data visualisation before applying any models. According to this model data features must be plotted to see the distribution of samples to spot the anomalies in the data. In linear regression we should only consider the data which has linear relationships.

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

Ans: It is a way of measuring a linear correlation. Its values lies between -1 and 1. If it is 1 then if one variable changes then other variable also changes in same direction. If it is -1 if one variable changes, then other variable also changes in opposite direction. If it is 0 there is no relationship between two variables.

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

Ans: Scaling is method used in linear regression to process the data so that the values of different variables are standardised so that comparison can be done. This can be implemented if the units of measurements are different for different variables.
If we don't use scaling technique, then it will result in incorrect model which cannot be used for prediction. There are mainly two scaling techniques one is normalisation, and another is standardization.
In normalisation scaling technique we bring all the values of every feature between 0 to 1 and in standardization all values of variables are brought closer to mean.

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

Ans: VIF value would be infinite among two variables if there is a perfect collinearity between them. In this case R square value would be 1. As VIF formulae is 1/(1 – R2) it will become infinite. This indicates that there is stong multicollinearity between two variables and one needs to be dropped.

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

Ans: Q-Q plot can be used to check some of the assumptions that required for valid inference. Q-Q plot compares the distribution of two sets of data. It helps us determine if two datasets follow the same kind of distribution. It also helps to find out if the errors in dataset are normal in nature or not.