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?

- Bike demand does not get effected whether it's working day or not
- Bike demand in spring is highest
- Bike demand is high when the weather is clear and mist cloudy and it is low for light snow
- Bike demand is almost same for all weekdays.

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

- It is used delete the extra columns after creating dummy for particular column
- It helps to reduce co linearity

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

Ans: temp and atemp both have the same correlation with target variable and also highest

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

Ans: The distribution of residuals should be normal

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

- Temperature (temp),
- Weathersit
- Year(yr)

General Subjective Questions

1. Explain linear regression algorithm in detail.

Ans: Linear Regression is a type of supervised Machine Learning Algorithm that is used for prediction of numeric values. It is most basic form of regression analysis and is most commonly used predictive analysis model. The equation of linear regression is

"Y=mx+c"

It has an assumption that there is a linear relationship between dependent variable(y) and independent variable(x). Here in linear regression we calculate the best fit line that describes the relationship between independent and dependent variables. Linear regression is performed on dependent variable is of continuous type and the independent may of continuous, categorical etc.

Regression is of two types:

1) Simple Linear Regression: It is used when dependent variable is predicted using only one independent variable.
2) Multiple Linear Regression: It is used when dependent variable is predicted using multiple independent variables.

The equation of SLR:

Y=B0+B1x

The equation of MLR:

Yi=B0+B1x1+B2x2+…..+Bpxp

B0=intercept

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's Quartet was developed by statistician Francis Anscombe. It contains 4 data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasize both the importance of graphing data before analyzing it and the effect of outliers and other influential observations

3. What is Pearson's R?

Ans: Pearson's r is a numerical representation of the strength of the linear relationship between the variables. Its value ranges from -1 to +1. It explains the linear relationship of two sets of data.

If r=1 means the data is perfectly linear with a positive slope

If r=-1 means the data is perfectly linear with a negative slope

If r=0 means there is no linear relationship

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

Ans: Scaling is a method used to normalize the range of independent variables of features of data. It is performed during preprocessing stage to deal with varying values in the dataset.

If scaling not done, then machine learning algorithm tends to weigh greater values.

Normalization is used when you know that the distribution of your data does not follow of Gaussian distribution.

Standardization is helpful in cases where the data follows a Gaussian distribution. Standardization does not have any range.

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

Ans: VIF-variance inflation factor-It indicates how much co-linearity has increased the variance

$$VIF=1/ (1-RI^2)$$

VIF=infinity if there is perfect correlation.

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

Ans: The quantiles of the first data set are plotted against the quantiles of the second data set in a q-q graphic. It's a tool for comparing the shapes of different distributions. A scatter plot generated by plotting two sets of quantiles against each other is known as a Q-Q plot.