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?

Answer –

- More rental in 2019 than in 2018
- Rentals are more when there is Mist+Cloudy or Clear Weathersit
- Rentals are more in summer and fall season
- More rentals in some of the months (Aug & September predominantly)

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

Answer – It is important as it reduces one extra variable created for the analysis, as the value of the third category can be inferred using the two categories available (considering we have three categories available).

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer - Temperature

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer – There are 4 assumptions which we take to perform linear regression:

1. **There should be linear relationship between X & Y** – we performed scatter plot and saw that there has been a linear relationship with some variables like temp, humidity etc.
2. **Residuals should be normally distributed** – Performed the residual analysis at the end and from the plot we can conclude that the errors are normally distributed
3. **Error should have constant variance** – We can perform a scatter plot between the fitted variable and the error(residue) if it is not scattered along the different values of predicted variable then it holds Homoscedasticity
4. **Residuals should be independent of each other –** An autocorrelation can be performed for this

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer - Using RFE we saw that the 3 most important variables are Temperature, humidity and windspeed

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Answer - It performs regression to predict the target variable. It falls under supervised learning. The coefficients are calculated based on the reducing the cost function which is Residual sum of squares here

2. Explain the Anscombe's quartet in detail. (3 marks)

Answer - It is the combination of 4 datasets with 11 data points which have similar descriptive statistics but have different distribution, it can fool the linear regression, thus data visualization is needed to understand the relationship between X & Y variables

3. What is Pearson's R? (3 marks)

Answer - Person's R also referred as 'r' is the correlation coefficient, it tells the relation between 2 numeric variables, it varies between the range -1 to 1. If the value of r =1, it means there is very high correlation. For every units increase of the variable X there will be increase for variable Y in a fixed proportion. Similarly, if the value is -1 it means that one value increases the other decreases in a fixed proportion. If the value is 0 then there is no relationship between X & y.

However, r doesn't help us understand the cause of one over the other. Like whether it is X increases Y or Y increases X is not clear.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer - Scaling is a normalization technique which brings the independent variables in the same range. It is performed to make the coefficients of independent variables in same range and makes the calculation faster.

Normalization Scaling also known as Min-Max scaling it brings independent variable in the range 0 to 1

Standardized Scaling brings the independent variable in a normal distribution, its mean would be 0 and standard deviation = 1

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer - When the value of VIF is infinite, it means that there is a perfect correlation between 2 independent value. Thus, the value of r-square becomes 1. As ViF = 1/(1-R-square), as r-square = 1 (1-R-square) = 0, so VIF = infinite

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer - Q-Q plot helps in understanding if there is any theoretical distribution such as normal, exponential, uniform etc. in the data. It is done by plotting the quantiles of the dataset along 45 degree line