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 our analysis of categorical variables , we saw that bike rental demands are highly impacted by categorical

features like season (fall has higher bike rental demands compared to other seasons like summer, winter, least is spring),

also categorical features like month, weekday,weathersit are highly impacting the bike rental demands.

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

Ans: We should use drop_first = True to avoid dummy variable trap

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

Ans: Temp. Casual and registered are not considered since target (cnt) is the sum of both.

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

Ans: a. Using pairplot to check if any numerical features are linearly distributed

b. Used distribution plot to check if residuals are normally distributed

c. Used scatter plot to check residual variance and residual patterns.

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

Ans: Temp, holiday, summer.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear Regression algorithm tries to find the best line using equation y=mx+c. It is an unconstrained minimisation problem in which it tries to reduce the cost function ∑(i=0,n){y-m0-m1x1). It tries to find the optimal weights and intercepts by updating coefficient values in every iteration. There are two types of linear regression:

a) Simple Linear Regression : In simple linear regression we have one independent variable and one target variable for e.g. given experience predict the salary of employees
b) Multiple Linear Regression : In Multiple Linear Regression we have multiple independent variable like experience, qualification, skillset and other features to predict salary of employees.

There are four assumptions of Linear Regression:

1) Features must be linearly related with the target variable.
2) Residuals must follow normal distribution with mean 0 and standard deviation 1.
3) Error terms must not follow any pattern and are independent.
4) There must be little or no variance in error terms.

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet resembles that descriptive statistics can be same for few similar datasets but graphical/plots of those datasets can be different. Anscombe came up with four similar datasets having nearly identical descriptive statistics like mean, variance & standard deviation but different plots.

3. What is Pearson's R?

Ans: Pearson's R is a method used to find correlation between features. The value of correlation in Pearson's R lies between -1 and 1. If p<0 it resembles positive correlations means if x increase y will also increase. If p >0 it resembles negative correlation means if x increase y will decrease.

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

Ans: Scaling refers to bringing data on same scale to achieve optimal weights and intercepts in linear regression equation quickly. For eg: units of weights, units of height are two different scale, when we scale them we will bring them to same scale either between 0 and 1 (minmax scaler) or with mean 0 and standard deviation 1 (standard scaler). In Normalized scaling we scale data between 0 and 1 whereas in standardized scaling we scale data in such a way that means of that features will be 0 and standard deviation will be 1.

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

Ans: When there is a perfect correlation between variables then VIF will reach infinity since R2 will be 1 and as per formula VIF = [1/(1-R2)].

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

Ans: Q-Q plot resembles the plot of two quantiles against each other. We use Q-Q plot to check if data is coming from same distribution by checking if all points lie on the straight line at an angle of 45^0. if all points lie away from straight line at an angle of 45^0 it is coming from different distributions. We use this to check in linear regression if after train test split both the sets are coming from same distributions.