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? (3 marks)

The categorical variables have a significant effect on the demand .

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

If drop_first = False , it creates dummy variables one for each level and 10 levels will have 10 dummy variables. drop_first = True will remove the dummy variable for 1st categorical variable and gives number of levels minus 1 variables only.

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

The registered variable has the highest correlation with the cnt variable which is the target.

This is seen both in pair plot and heatmap as well.

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

Calculated the r2 .

Plotted the y pred and y test to see if it is following a linear relation or not.

checked the residual and seen whether it is normally distributed or not across the zero.

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

The top 3 features is defined based on the highest coefficients in the final model.

Yr_2019, windspeed, mnth_Jan,weathersit_light rain

General Subjective Questions

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

Linear regression algorithm is used to predict continuous values. It has independent and dependent variables . The dependent variable is predicted using the independent variables. The linear regression basically tries to fit the independent and dependent variables in a straight line equation.

Eg: Y = a1 * X1 + a2 * X2 + a3 * X3 + ….+K

So the values Y is dependent variable and X1,X2,X3… are independent variables. a1, a2, a3 ,… are coefficients.

The coefficient defines the extent to which an independent variable tied to is effecting the dependent variable . High coefficient value specifies this independent variable is highly impacting the dependent variable . The sign of coefficient tells in which direction.

The Y can be profit and X1 can be taken as sales in above equation.

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

Anscombe has plotted 4 completely different datasets all having equal statistical values like mean, variance, straight line equation etc… But all these when plotted will appear completely different visualizations.

So he actually asked us not to depend on mere statistical parameters and declare the datasets are similar, but instead plot the datasets on a graph to understand the dataset before actually building a

Model.

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

Pearson's corelation coefficient is using to define how strong the linear relation between 2 variables is. Its values usually lies between -1 and +1. +1 and -1 are signifies high correlation but in opposite direction.

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

Scaling is bringing all the variables/features in a dataset to same range. To bring different units for variables to same range by removing their units.It improves the performance of the algorithm

Standardized scaling = (x-xmin)/(xmax-xmin), xmin and xmax are maximum and minimum values of x in the same column of dataset. x is actual value.

Normalized scaling = (x-mean)/standard deviation

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

Vif = 1/(1-r2)

Vif= infinity means , r2 = 1 and the variables are highly correlated with each other. As there is multicollinearity its better to remove the one of the 2 variables.

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

When we take the quantiles (0.1,0.2,0.5,0.75,0.9,etc..)of 2 distributions and plot them graphically it is called a Q-Q plot.

The Q-Q plot is checked with respect to the y=x line. If the plotting is all in sync with the y=x line it means both the distributions are almost similar.