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?

→ Inferences – Weather situation do affect the no. of sharing of bikes.

- The no. of bikes shared i.e., count is least for spring

- The number of bikes shares increased in 2019

- The count has zero values for weather situation - category-4 = 'Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog'

- The count values increased in from 3rd month and the demand remains almost same till the month 10, but the highest count in month 9

- The count values are drops during holidays

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

→ drop_first=True is important to use, it is helpful in removing/dropping the unnecessary columns which is created during the dummy variable's creation. It reduces the correlations created among the dummy variables. Another reason is, if we have all dummy variables it leads to Multicollinearity between the dummy variables. To keep this under control, we lose one column

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

→ temp and atemp has highest correlation with target variable (cnt).

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

→ Distribution should follow normal distribution and centred around 0.(mean = 0). We validate this assumption about residuals by plotting a plot of residuals and see if residuals are following normal distribution or not. The above diagram shows that the residuals are distributed about mean = 0.

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

→ 1. Temp = 0.570037

2. Yr = 0.232127

3. Season Winter = 0.126393

General Subjective Questions

1. Explain the linear regression algorithm in detail.

→ Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values. Linear Regression is the most basic form of regression analysis. Regression is the most commonly used predictive analysis model.

Linear regression is based on the popular equation **"y = mx + c".**

we calculate the best fit line which describes the relationship between the independent and dependent variable. Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc. Regression method tries to find the best fit line which shows the relationship between the

dependent variable and predictors with least error.

In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term.

Regression is broadly divided into simple linear regression and multiple linear regression.

1. Simple Linear Regression : SLR is used when the dependent variable is predicted using only **one** independent variable.

2. Multiple Linear Regression :MLR i s used when the dependent variable is predicted using multiple independent variables.

The equation for MLR will be:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$$

β1 = coefficient for X1 variable

β2 = coefficient for X2 variable and so on…

β0 is the intercept (constant term).

2. Explain Anscombe's quartet in detail.

Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four 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

3. What is Pearson's R?

Pearson's r is a numerical summary of the strength of the linear association between the variables. It value ranges between -1 to +1. It shows the linear relationship between two sets of data.

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

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

r = 0 means there is no linear association

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

→ Feature **scaling** is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data pre-processing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

● Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbours and Neural Networks.

● Standardization can be helpful in cases where the data follows a Gaussian distribution. Standardization does not have a bounding range. So, if you have outliers in your data, they will not be affected by standardization.

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

→ **VIF - the variance inflation factor -** The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. (VIF) =1/(1-R_1^2). If there is perfect correlation, then VIF = infinity. Where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables- If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1. So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity"

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

→ A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. It is used to compare the shapes of distributions. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

The q-q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behaviour?