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)

Ans – Below are the observations from the categorical variables.

1. Year on Year, the Bike usage has increased for all seasons

2. In 2018, Bike usage increased after March and stayed over 3000 for rest of the year.

3. In 2019, Bike usage increased after Feb and stayed over 5000 till Nov. September had the highest usage.

4. Use of bikes is less at the start and end of the year.

5. Bike usage is almost same for all days of the week except on Holiday.

6. People use Bike most for weathersit 1 i.e. when its Clear, Few clouds, Partly cloudy, Partly cloudy. This is same for both years even though overall volume for 2018 is less than 2019.

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

Ans – For the model to understand the categorical variables, we encode them in the Boolean values. Boolean value of 0 and 1 can represent 2 values. **drop_first=True** drops the 1st entire column of the Boolean values when we encode categorical value in them.

Ex. Gender: [‘Male’, ‘Female’, ‘Other’]. If we represent the values using conventional Boolean columns, we will need 3 columns viz Male, Female, Other with values 0 or 1.

Same thing can be represented using only 2 columns as Male: 10, Female: 01 and Other: 00

Here for 3 values in the categorical column we needed only 2 columns i.e., n-1 columns.

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

Ans – The highest correlation we found was of temp (0.64) and atemp (0.65) which represented Temperature and Feels like Temperature.

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

Ans – We check the distribution of the errors as we know that the assumption of the LR is that the errors are normally distributed.

To check the multicollinearity, we calculate the Variance Inflation Factor. If the value of the VIF is more than 5, we remove the variable and again check the VIF.

We plot the Q-Q plot of the predicted and the actual output variables to check the variance of errors and linearity of the model.

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

Ans – Top 3 features as below

1. Temperature
2. Season – 3
3. September month

General Subjective Questions

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

Ans – Linear Regression Algorithm is a form of supervised learning where the parameters (also called as Features) of the given set (also called as independent variables) form the linear relationship with another parameter (also called as Dependent Variable).

If the dependent variable depends upon only 1 independent variable, then its called Single Linear Regression. It represented by y = mx + c where x is independent variable on which y is dependent on. The slope of the line is defined by the m and c is the Constant or Intercept of Y axis when x is zero. Using the SLR, we predict the best fit line for the given observation points.

If there are multiple variables, then its called Multiple Linear Regression. It is represented as y = c + m1.x1 + m2.x2 … mi.xi. Here xi are the multiple independent variables and mi are the coefficients. C is again same as Intercept on Y axis. Using the MLR, we predict the best fit hyperplane (since we have multiple variables, they cant be shown on 2D graph) for the given observation points.

The aim of the SLR or MLR is to observe the relationship between the input and predict the output. Being supervised learning, we have the inputs (x) and outputs (y), so the actual goal of the model becomes to predict the coefficients of x and the constant.

The goal of the model building (SLR or MLR), is to achieve the least difference between y actual and y predicted also called an error or residual. We use the methods like Least Mean Square to achieve this.

For SLR or MLR to be applicable, there are certain criterion or consideration.

There has to be a linear relationship between the IP and OP, constant variance between residuals, it should be normally distributed.

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

Ans – Consider 2 points 1 and 5. These two values have the mean of 3. If we take values 2 and 4, we would get same mean as 3. This mean that even though the points are different we get the same mean. Similarly, we can get the datapoints which not only have the same mean but also additional properties as below.

a. Same mean of X
b. Same variance
c. Same mean of Y
d. Same variance
e. Same correlation
f. Same linear regression Slope and
g. Same linear regression Intercept as well.

This means that these datasets are statistically identical but when plotted on the graph, they are not same.

Anscombe in 1973, created the 4 datasets which had the same statistic values but were actually very different representation on the graph. All these 4 datasets had same X Mean (9), X variance (11), Y Mean (7.5), Y variance (4.12), Correlation (0.8), Linear Regression slope (0.5) and Linear Regression intercept (3).

But when plotted on graph, 1st dataset had linear relationship between x and y, 2nd had non-linear relationship, 3rd dataset had perfect linear relationship for all the data points except for one point and lastly, 4th dataset with all same points except last one point as outlier.

This essentially suggests that we cannot solely depend on the statistical data and the importance of the Data visualization and in turn Exploratory Data Analysis in the model building.

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

Ans – Also called Pearson Correlation Coefficient, is a parameter which tells us the strength and the direction of the relationship between the variables. The value of Pearson's R lies within -1 and 1.

BMI for example is a function of the weight (in kg) and height (in meter).

BMI = Weight / (Height)^2

For an adult person (height is fixed), as the weight increases, BMI will also increase. This means as the Input value increases, the output also increases. This defines that weight has **Positive Relationship** between the BMI. As opposed to this, if the weight is constant and height is increased, then the BMI will decrease. This defines that height has **Negative Relationship** with BMI.

This positive or negative direction is shown in Pearson's R values as the sign.

It is obvious that BMI depends on weight and height but how much dependency is there, that is defined by the value of R. If the value is 1 (positive or negative), that means there is perfect linear relationship between the variables. Closer the value of R to 1 higher is the relationship between the variables.

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

Ans - While building the LR model we feed the observation points to the algorithm. The algorithm tries to find the optimal values of the coefficients (mi) of each input variables which are multiplied to the input values (xi). The process is doe many times to get the correct and best values of the coefficients. This multiplication and the results will vary for each variable depending upon the values of xi. These combined multiplications will affect

the output Y. If the values of these xi are very big, then it will take longer time for model to find the right coefficients (considering the vector manipulations for all features and multiple datapoints being done multiple times). So, in order to reduce this time and the efforts in the computations, scaling is required. Scaling maps your data to a range of 0-1.

Ex. Values of 0 – 100 will be fit in the scale of 0 – 1. And this is done for all the features.

In other words, scaling helps in the convergence of the variables by making their values fit same range. It also helps reduce the impact of the outliers and improve the model performance.

There are 2 ways of scaling the values.

1. **Normalized Scaling** – Also called as MinMax scaling is a technique where we use the minimum and the maximum values of the dataset to get the datapoint in 0-1 range.

$$x\_new = \frac{x - x\_min}{x\_max - x\_min}$$

Here x_new is the scaled value, x_min is the minimum value of x and x_max is the maximum value of x.

2. **Standardized Scaling** –Standardized scaling is a way to adjust data so that it has an average of zero and spreads out evenly. Meaning, we make sure the values are cantered around the mean and have unit standard variance.

$$x\_new = \frac{x - \mu}{\sigma}$$

Here x is the datapoint, μ is the mean of the value and σ is the standard deviation.

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

(3 marks)

Ans – Variance Inflation Factor is basically a degree of multicollinearity between the variables. In other words, it explains how much the predictor variable is correlated with or influenced by any other predictors in the model.

If the value of VIF is 1, that mean there is no collinearity between the variables. The higher the value of VIF, higher the dependency or collinearity.

In some extreme cases, if the predictor is perfectly correlated with other variables, then the value of VIF tends to infinity.

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

(3 marks)

Ans – A Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

It helps us identify the distribution aspects such as shift in scale or change in symmetry.

This helps in a linear regression when we have training and test data sets received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.