Assignment-based Subjective Questions

Que1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans: From the analysis, we got to know that count of the total rental bikes are higher:

- When there is fall season.
- When there is no holiday.
- When weather is clear.
- When the day is working.
- When month is June that is summers.
- There is no significantly higher affect of the weekday. For all the weekday the count is higher.

Que2: Why is it important to use drop_first=True during dummy variable creation?

Ans: It is important to use drop_first = true for dummy variable to drop first dummy variable because we don't really need it. The logic behind it is when other dummy variables will have zero in their values, there will surely be the first variable to be true. In other words we can say that by using it, we can reduce the correlation that is created between dummy variables.

Que3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: "temp" and "atemp" is having almost similar correlation with the target variable by looking at the pair plot.

Que4: How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: we just do the residual analysis by plotting a displot of errors and saw if the errors are normally distributed or not.

Secondly, we checked the p-value that should be close to 0 and checked VIF that should be less than 5.

Que5: . Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans: 'temp' and 'scatteredClouds' are the variables that are highly correlated with the target variable in our final model.

General Subjective Questions

Que1. Explain the linear regression algorithm in detail.

Ans: Linear Regression is a supervised machine learning algorithm that finds a linear relationship between one dependent and one or more independent variable by finding a line between them. Mathematically, we can represent linear regression as:

$$y = mx + C$$

Where:

C = Intercept

m= slope,

x= independent variable

y= dependent variable

Linear regression is of two types:

1. Simple Linear Regression: In simple linear regression, there is only one independent variable and relationship between *x and y* is a linear function. So, changes in *y* are assumed to be caused by changes in *x.* Assumptions for this regression are:
    - Error terms are normally distributed that have constant variance.
    - Error terms are having zero mean and are independent.

2. Multiple Linear Regression: In multiple linear regression, there are more than one independent variables. We can represent it as:
    $$y = c + m1x1 + m2x2 + m3x3 + \cdots \ldots \ldots \ldots \ldots \ldots mnxn$$
    Instead of line, this model fits a hyperplane. For inferences, the assumptions from simple linear regression holds. Other than that, New considerations are also added:
    - Adding more is not always helpful.
        - Model may overfit by becoming too complex. It will have high train accuracy and low test accuracy.
        - Association between predictor variables i.e. multicollinearity can be increased.
    - Feature Selection plays a important role in multiple linear regression.

Que2: Explain the Anscombe's quartet in detail.

Ans: Anscombe's Quartet is used to mark the value of plotting data before analysing it on the basis of its statistical properties. So basically, tells us about the data visualization importance. It is a combination of four data-set, each data-set contains eleven dependent (y) and independent (x) points. Those data-sets are having same mean, standard deviation, variance and etc. In other words, we can say that they all are same in their descriptive statistics, but for each of them graphical representation is different.

For example, if we have four datasets, each of them is having same mean, standard deviation and mean. But when we plot them on a graph, they show totally different behaviour from other. One of them can show linear relationship between x and y. second dataset can have a linear relationship but with the outlier. Third one can show a curve shape between x and y. third one can have stable x for all y except at one outlier. So, all are not same. This shows us the importance of data visualization.

Que3: What is Pearson's R?

Ans: Correlation is the relationship between two variables that lies between -1.0 to +1.0. Pearson's R can be called as the measurement of the strength of that correlation and how the variables are associated with each other. So, when a variable is changed, we can calculate the effect of its change in other variable by using Pearson's R or Pearson Correlation. It is very significant in data statistics. It basically draws a line through the data point of two variables so that it can show a linear relationship between them. The strength of that relationship is measured by Pearson correlation Coefficient calculator. That relationship that we found can be positive or negative. Pearson correlation between two variables x and y can be found by using the formula given below:

$$Correlation\ (x, y) = \frac{Sx}{\sqrt{Sx * Sy}}$$

Where s_{xy} is the covariance of X and Y and s_x and s_y are the standard deviation of X and Y

- **Positive linear relationship:**
  When the relationship between two variables is such that one variable is increased with the increment of another variable.
- **Negative linear relationship:**
  When the relationship between two variables is such that one variable is decreased with the increment of another variable.

The correlation coefficient is interpreted as follows:

If r = 0.85, it will be positive correlation.

If r = -0.85, it will be negative correlation.

If r = 1.00, it will be perfect positive correlation.

If r = -1.00, it will be perfect negative correlation.

If r = 0.00, there will be no correlation.

Que4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: Scaling is referred as the standardization of independent variables (that are present in the data) in the fixed range. Scaling is recommended but not necessary. Scaling of features is performed for several reasons given below:

- **Ease of Interpretation:** suppose we have two independent variables in a dataset. One variable range is 1000 – 5000 and the range for another variable is 1 to 5. So, coefficient of first variable will be higher and for second variable, it will be less. That does not mean first variable is explaining the target variable well. Therefore, we need to scale both variable in fixed range.
- **Faster Convergence for gradient descent method.**

There are two scaling methods:

1. **Standardized Scaling:** It is a scaling technique in which the values are arranged in such a way that they are centred at zero with the standard deviation of 1. The formula for standardization is given below:

$$X' = \frac{X - mean}{S.D.}$$

2. **Normalized Scaling:** It is scaling technique in which the values are arranged in such a way that all of them lie between zero and one. The formula for normalized scaling is given below:

$$X' = \frac{X - Xmin}{Xmax - Xmin}$$

Que5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: If VIF = infinity, it means that there is perfect correlation between two variables. In that case, we will get R2 = 1, then 1/(1-R2) will be infinity. We have to drop a variable that is causing multicollinearity from the dataset to overcome that problem.

Que6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression

Ans: When two different quantiles are plotted against each other, that plot is called as Quantile-Quantile plot or Q-Q plot. The first quantile is of the variable for which hypothesis testing is done and the second quantile is the actual distribution against which testing is done.

For example:

If we are performing testing to see whether the employees age is normally distributed or not. So, we will take the first quantiles as employees age and we will take second quantile from a

normally distributed curve to see that if both the quantiles are having the same distribution or not. If they are having same distribution, they will be on straight line roughly. The steps to generate a Q-Q plot for members age to test for normality is given below:

1. First of all, variable of interest is taken i.e. employees age. Suppose there are 13 employees in this scenario.

2. Normal curve is plotted and this plot is divided into 14 equal segments (n+1; where n=#data points)

3. Z score is calculated for every points.

4. The z-score obtained is plotted against the sorted variables. Usually, the z-scores are in the x-axis and the ordered values are variable quantiles are in the y-axis. The variables on x-axis is called as theoretical quantiles and the values on y-axis are called as variable quantile.

5. Now see if data points are lined up in a straight 45-degree line or not.

6. If it does, the age is normally distributed. If it is not, check it against other possible distributions.