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

- In 2019 the bike rentals demand is increased
- From April to September the bike rentals are high and from November it gradually decreases.
- Working day and weekdays show no significance
- When the weather is clear and partial cloudy the demand is high.
- The demand is low in spring season.

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

Ans The use of drop_first = True is it helps in reducing the extra column created by the dummy variable.

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

Ans In the pair-plot we saw that the highest correlations are temp, fall & year with cnt.

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

Ans

1. No multicollinearity – VIF < 5
2. The residuals are normally distributed
3. Linear relationship between Dependent and independent variable
4. The variance is constant

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

Ans

1. Year
2. Fall
3. Summer
4. May (month)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans Linear regression is one of the regression technique in which a dependent variable has a linear relationship with an independent variable. The main goal of Linear regression is to consider the given data points and plot the trend line that fit the data in the best way possible.

Mathematically, we can write a linear regression equation as:

$y = a + bx$

Where a and b given by the formulas:

$$b(slobe) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

2. Explain the Anscombe's quartet in detail.

Ans Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

- **Data Set 1:** fits the linear regression model pretty well.
- **Data Set 2:** cannot fit the linear regression model because the data is non-linear.
- **Data Set 3:** shows the outliers involved in the data set, which cannot be handled by the linear regression model.

- **Data Set 4:** shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

3. What is Pearson's R?

Ans In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

• Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

• Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

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

Ans Scaling:

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling.

Why to perform scaling:

To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Difference between Normalized and Standard scaling:

Normalization typically means rescales the values into a range of [0,1].

Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance)

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

Ans If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables

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

Ans Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile.
For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q- Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q– Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two
distributions.