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]: Not all the categorical variables had an effect on target variable but some of the categorical variables like

Seasons-“spring”,”winter”,

months - “aug”,”jun”,”sep”,”mar”,may”,”oct”

Weekdays-”weekday_6” have contributed positively to the count while

weathersit_2(Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist) and weathersit_3(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) have contributed negatively to the total count.

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

[Ans]:N categorical variables can be explained through N-1 dummy variables,so drop_first=True can be used for the same of limiting the number of variables.

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

[Ans]:” atemp” variable has the highest correlation

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

[Ans]The assumptions are validated by doing residual analysis on the data and plotting a distribution plot to find out whether the mean is at 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)

[Ans] The top 3 features are

1.Year – Yr – Positive impact

2.WeatherSit_3 – (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) — Negative impact

3.Season – Spring – Negative impact

General Subjective Questions

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

Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis. The independent variable is also the predictor or explanatory variable that remains unchanged due to the change in other variables. However, the dependent variable changes with fluctuations in the independent variable. The regression model predicts the value of the dependent variable, which is the response or outcome variable being analyzed or studied.

Thus, linear regression is a supervised learning algorithm that simulates a mathematical relationship between variables and makes predictions for continuous or numeric variables

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

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.
Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyze it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.

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

In statistics, the **Pearson correlation coefficient** (**PCC**) is a correlation coefficient that measures linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations ; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

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

Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If scaling is not done, then a machine algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Scaling guarantees that all features are on a comparable scale and have comparable ranges.This process is known as feature normalisation. This is significant because the magnitude of the features has an impact on many machine learning techniques. Larger scale features may dominate the learning process and have an excessive impact on the outcomes. You can avoid this problem and make sure that each feature contributes equally to the learning process by scaling the features.

Normalization or Min-Max Scaling is used to transform features to be on a similar scale. The new point is calculated as:

$$X\_new = (X - X\_min)/(X\_max - X\_min)$$

This scales the range to [0, 1] or sometimes [-1, 1]. Geometrically speaking, transformation squishes the n-dimensional data into an n-dimensional unit hypercube. Normalization is useful when there are no outliers as it cannot cope up with them. Usually, we would scale age and not incomes because only a few people have high incomes but the age is close to uniform.

Standardization is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

$$X\_new = (X - mean)/Std$$

Standardization can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Geometrically speaking, it translates the data to the mean vector of original data to the origin and squishes or expands the points if std is 1 respectively.

We can see that we are just changing mean and standard deviation to a standard normal distribution which is still normal thus the shape of the distribution is not affected.

Standardization does not get affected by outliers because there is no predefined range of transformed features.

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

A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity.

$$VIF=1/(1-Ri2)$$

If Ri2 value tends to 1,VIF tends to infinity,which means there is a perfect correlation

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

The quantile-quantile(q-q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution. They are commonly used in statistics, data analysis, and quality control to check assumptions and identify departures from expected distributions.

Q-Q plot can also be used to test distribution amongst 2 different datasets. For example, if dataset 1, the age variable has 200 records and dataset 2, the age variable has 20 records, it is possible to compare the distributions of these datasets to see if they are indeed the same. This can be particularly helpful in machine learning, where we split data into train-validation-test to see if the distribution is indeed the same.