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:

- The rides are high during the Mid of the year and low during the beginning and the end of the year.
- The rides are low during the Spring season compared to others.
- The rides are comparatively higher during holidays but is not having much impact.
- The rides are almost equally distributed during weekdays.

----

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

Ans:

This is done to reduce the number of columns created for dummy variables for the model. It is enough because the value will not be anything other than the missed variable if not for the existing ones. This also reduces the correlation between the variables.

----

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

Ans:

temp, wind speed, and yr columns are in high correlation. Registered and Casual columns are also in correlation as the sum gives the cnt.

----

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

Ans:

- A Residual Analysis was performed on the training data to evaluate the assumption in Linear Regression that the errors are normally distributed.
- The R2 value is above 0.80, indicating that the independent variables explain most of the dependent variable proportion.
- The Prob(F-Statistic) value is almost Zero which also tells that the model is reliable.

----

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

Ans:

Year, temp, and windspeed seemingly contribute more towards the demand for shared bikes.

General Subjective Questions

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

Ans:

A Linear Regression algorithm is a way to derive a linear relationship between a target variable and a single variable (Single Linear Regression) or Multiple variables (Multiple Linear Regression).

The relationship is assumed to be linear under a few conditions and there is a tolerance above which the model cannot be assumed Linear.

It is based on the formula

$$y = mx + c$$

There are 4 assumptions in Linear Regression

1. The relationship between the dependent variable and the independent variable is linear.
2. Homoscedasity: The variance is considered to be similar across the ranges of values.
3. In multiple linear regression the variables are independent of each other.
4. Normality: For any value of X, Y is normally distributed.

----

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

Ans:

Anscombe’s quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

This quartet consists of 4 datasets which when statistical values are but will be very different when plotted in a scatter plot.

Each one also fits the model very differently when a model is built for the same. It emphasizes the need for visualizing the data before building the model. This also tells that Regression when we should consider regression.

----

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

Ans:

Pearson's R is the correlation coefficient that gives the correlation between two variables. It can range from -1 to 1. If the value is above 0.5 then the correlation is considered to be strong. The correlation is very useful in identifying the relationship between the independent variable and the target variable and also the correlation among the independent variables which can be avoided.

----

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

Ans:

Scaling is a process where all the data are compressed into a particular range. This is done to improve the performance of prediction and to have values in close range. There are two types of scaling Normalized and Standardized scaling.

Normalized Scaling:

Brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization.

Standardized Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

sklearn.preprocessing.scale helps to implement standardization.

----

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

Ans:

VIF score shows the correlation between independent variables. When VIF is infinite that means the values are highly correlated and it is better to drop one among those. 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.

(3 marks)

Ans:

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential, or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have the training and test data set received separately and then we can confirm using a Q-Q plot that both the data sets are from populations with the same distributions. This will help in the reliability of our model as the output is related to the input and if the input is different the predicted values will be different leading to an error.
