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: From analysis we can infer:

- March to Oct month seems to have more cnt of ride
- Spring have less count
- Year 2019 have more entries than 2018, So it seems there is upward trend yearwise.
- There are very few rides where there is Rainy situation

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

Ans: Extra column gets created during dummy variable creation, So drop_first=True helps in reducing that extra column.

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

Ans: “temp” has the highest correlation with the target variable

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

Ans: Linear Regression Model validation was done using its basic assumptions as below:

a. Errors are normally distributed
b. Multicollinearity checks among the variables.
c. Linear relationship was shown in plot
d. Homoscedasticity, there was no visible pattern in residual values
e. Independence of residuals, there was no auto-correlation

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: These 3 features showed highest significance: holiday, temp and sept month.

General Subjective Questions

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

Ans:

Linear regression can be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables.

Linear relationship between variables means that when the value of one or more independent variables will change, the value of dependent variable will also change accordingly.

Mathematically the relationship can be represented with the help of following equation −

$$Y = mX + c$$

Here,

Y is the dependent variable we are trying to predict.
X is the independent variable we are using to make predictions.
m is the slope of the regression line which represents the effect X has on Y
c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

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

Ans:

Anscombe’s Quartet comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

3. What is Pearson’s R? (3 marks)

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.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

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

Ans:

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 feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Example**:** If an algorithm is not using scaling method, then it can consider the value 300 meter to be greater than 1 km but that’s actually not true and, in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

Normalized scaling	**Standardized scaling**
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
Scales values between [0, 1] or [-1, 1].	Not bounded to a certain range
It is greatly affected by outliers	It is much less affected by outliers.
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

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

Ans:

If there is perfect correlation, then VIF is infinite. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 3, this means that the variance of the model coefficient is inflated by a factor of 3 due to the presence of multicollinearity.

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
Ans:
The Q-Q(quantile-quantile) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Usage:
A Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction of points below the given value. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The more the departure from this reference line, the greater the evidence
for the conclusion that the two data sets have come from populations with different distributions.

Importance:
When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences.