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)

Bike Rentals is higher:

- *During the Fall season*
- *In year 2019*
- *In clear weather*
- *On Holidays*

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

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Thus, reducing the correlations among dummy variables.

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

From looking at the pair-plot among the numerical variables, 'registered' 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)

Residual analysis, linear relationship, Normal Distribution of error terms.

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

'*Light Snow*', '*windspeed*', '*temp*'

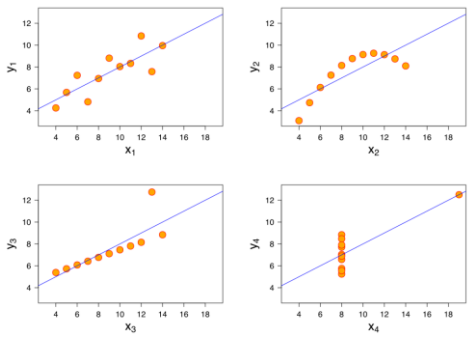
General Subjective Questions

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

It is a statistical technique to understand the relationship between one dependent variable and several independent variables. The objective of regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.

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

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.



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

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus, it is a normalised 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 means transforming the data in terms of the range of independent variables or features of data.
Scaling is performed during the data pre-processing to handle highly varying magnitudes or values or units.
Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling. Standardization is another scaling technique where the values are centred around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

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

It 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 (which show an infinite VIF as well).

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

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. 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.

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.