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)

- on clear or partial cloudy weather situation bike demand is high.
- on fall season people tend to use the bike sharing more and second place goes to summer season.
- from 2018 to 2019 bike sharing demand increased to 65%.

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

- For n level of category variable, creating n-1 dummy variable are enough to describe the categorical variable completely.
- More over we are trying to keep our model light weight, so we are preferring to user drop_first, which helps to improve Adjusted R squared value.

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

- Temperature has higher correlation with target variable.
- Obviously Registered user may have high correlation with target variable, because more than half of the target values comes from Registered user. But it is not right choice.

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

- Verify the R squared value close to the training set.
- Using scatter plot validate the training and test model predictions have similar pattern. It matches with linear regression pattern.

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

- Temperature
- Summer season
- Light_Snow weather situation

General Subjective Questions

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

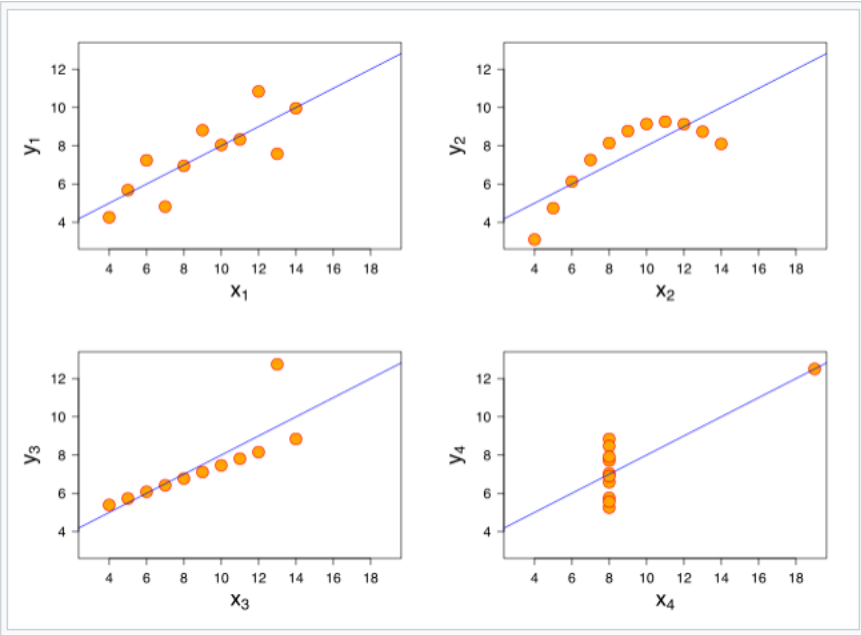
It is a supervised Machine learning algorithm helps to predict the relationship between dependent and independent variables. It assumes that the dependent and independent variables are linearly connected and explains the correlation between of them.

Y = mX + b explains the linear relationship between X and Y variables. Here b is intercept and m is slope or gradient.

Example: - The weight of the person is linearly related to their height.
So, this shows a linear relationship between the height and weight of the person. According to this, as we increase the height, the weight of the person will also increase.

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

Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph. The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.



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

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation between numerical variables. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
r = 0 means there is no linear association
r > 0 < 5 means there is a weak association

r > 5 < 8 means there is a moderate association
r > 8 means there is a strong association

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

Scaling is the process of bringing all (dependent and independent) variables values within certain range. So that they are easy to compare and make the predictions.

Independent and dependent variables value might be in different units and different range. To calculate accurate statistical results all the variable values should be in same units and within similar range.
So that machine learning calculations will be much faster in the background.

Normalized scaling brings all the data within 0 to 1 using below formula.
(x - xmin) / (xmax - xmin)

Standardized scaling brings all data under certain range using mean as 0 with standard deviation.
It uses below formula
(x - mean) / std

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

When two independent variables are perfectly correlated VIF value will be infinite.
Because we get R squared value as 1 and if we apply the formula of VIF $=1/(1-R^2)$ will result in infinity.

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

(3 marks)

Q–Q plot (quantile–quantile plot) is a probability plot, a graphical method for comparing two probability distributions by plotting their quantiles against each other.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.