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?

Ans. Conclusions drawn for the categorical variables could be –

1. Bikes are rented the most in fall season
2. Bikes are rented the most in the month of September
3. Bikes are rented the most on Saturday
4. Bikes are rented the most on days with clear weather
5. Bikes are rented the most in 2019

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

Ans. When dummy variables are created using python code, for n categories, n dummy variables are created but we require only n-1 variables. This makes a requirement to drop the one dummy variable created. And that is achieved by drop_first function.

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

Ans. Temp variable

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

Ans. Had the following steps –

a. Check the p-values of the independent variables
b. Check the VIF
c. Do the residual analysis to ensure that the errors follow a normal distribution
d. Predicting on the test set

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

Ans. Temperature, Year and season variables

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans. Linear regression is a supervised machine learning algorithm that defines a linear relationship between dependent variable and independent variable(s). When we have only one independent variable, we call it as simple linear regression while in case of multiple variables, we call it as multiple linear regression.

For any linear regression, the equation of the model should be of the form,

y = c + m1x1 + m2x2 + ...

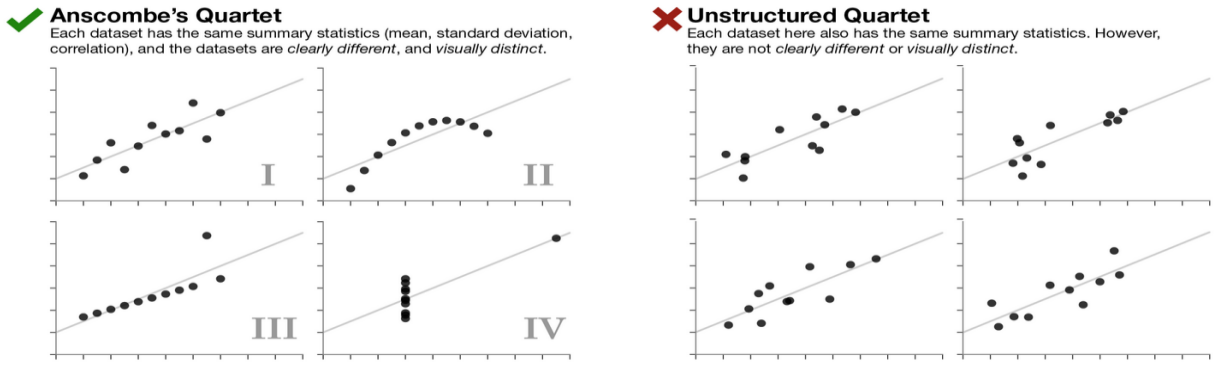
where y is the dependent variable or the target variable and the x1, x2 ... are independent variables.

c is the constant or the intercept and m1, m2 ... are the intercepts.

2. Explain the Anscombe's quartet in detail.

Ans. Anscombe's quartet is a group of four data sets that all produce the same statistical properties (mean, standard deviation, and correlation); This leads to believe that the data sets are similar. However, once the data is displayed (graphed), it is clear that the data sets have different trends.

In contrast, the "unstructured quartet" on the right side of Figure 1 (below) also has the same statistical power as the Anscombe quartet. But since there is no obvious pattern underlying every data set, these four are proof of that. will be difficult to see Not as good as original data



3. What is Pearson's R?

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

Pearson correlation coefficient (r)	Correlation type	Interpretation
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the same direction.
0	No correlation	There is no relationship between the variables.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the opposite direction.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans. Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Normalized Scaling – Bring all values of data within the range of 0 and 1

Standardized scaling - It brings all of the data into a standard normal distribution which has mean zero and standard deviation one

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

Ans. In VIF, each feature is regression against all other features. If R2 is more which means this feature is correlated with other features.

- VIF = 1 / (1 – R2)
- When R2 reaches 1, VIF reaches infinity

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

Ans. A quantile-quantile (Q-Q) chart is a graphical tool that helps us evaluate whether data fits a distribution such as a normal, exponential, or random distribution. It also helps determine whether two data sets taken from the population have a distribution.

This is useful in linear regression cases where we train and test the data separately, then we can use the Q-Q plot to verify that both datasets have the same distribution from the population.