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)

   As per the EDA done on the dataset, it was observed that
   - The average demand for shared bikes was highest during the fall season.
   - The year 2019 witnessed more demand for shared bikes compared to the year 2018.
   - Even though 2018 witnessed a lower number of shared rides, during the fall season it was still higher compared to other seasons.
   - The average demand for shared bikes was less during the holidays.
   - The average demand for shared bikes during the weekdays doesn't have a significant difference
   - The average demand for shared bikes was high when the weather was clear, Few clouds, Partly cloudy, Partly cloudy
   - The average demand for shared bikes was the least when there was Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds.

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

   - When we use the python library to create dummy variables, by default it creates the same number of dummy variables as that of the different levels present in the category. However, actually required dummy variables are always N-1 , where N is the number of levels/ classifications we have for an independent variable. It also reduces the correlation between the created dummy variables.

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

   - ‘temp’ has the highest correlation with the target variable ‘cnt’.

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

   It was validated that
   - There isn’t any multi-collinearity in the data set to have a better LR model.
   - The error points were normally distributed with a mean of zero.
   - The residuals did not follow any pattern. Hence, we could say it was the case of homoscedasticity.

5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes? (2 marks)

As per the model, it was observed that the top factors contributing to the demand of shared bikes were –

- Temp (temp)
- Weather situation (weathersit Light snow)
- Yr

Other variables that also contributed to the demand was the season of spring, the windspeed and the month of July.

General Subjective Questions

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

   Linear regression is a method of supervised learning process that assumes, there exists linear relationship between one or more independent variable and a dependent variable.

   It finds the best linear relationship and represents in the form of a straight line fitting the data points.

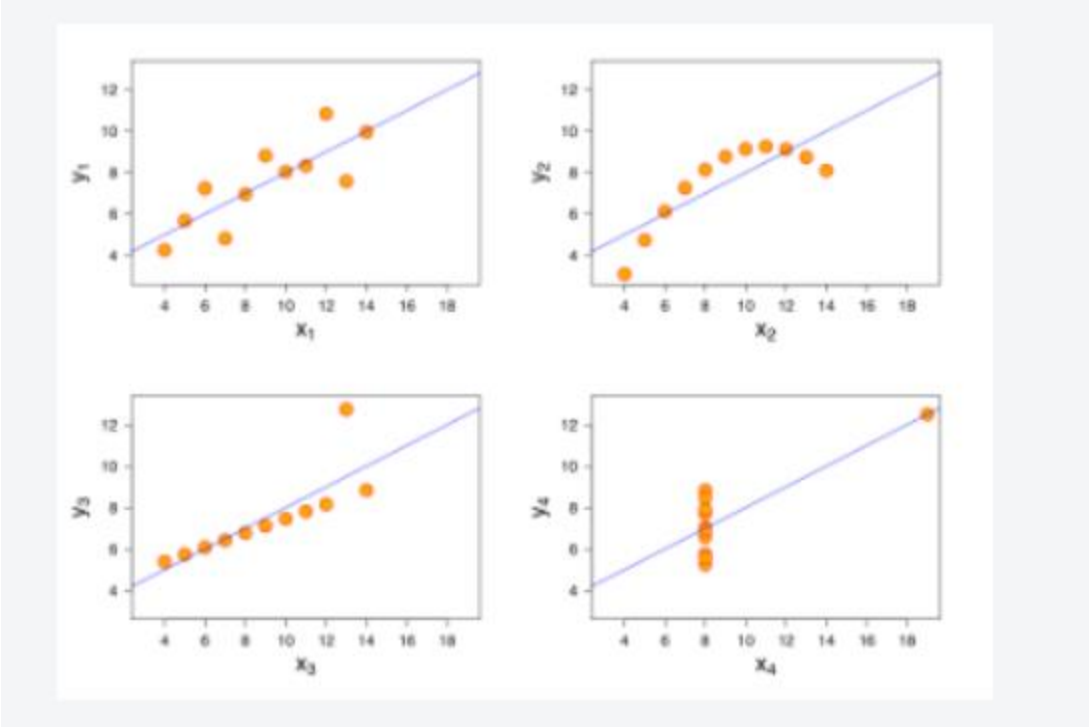
   To build a linear regression model,

   - Data is prepared ensuring we don't have any faulty records or empty records.
   - Categorical variables are handled by creating dummy variables
   - The data is then split into test and training sets and the model is trained on the training set.
   - Feature scaling is done if necessary
   - Split into X and Y sets for model building.
   - Relevant independent variables are chosen manually or automatically using (RFE).
   - P value and VIF score is checked to identify the best regression model to handle multi collinearity.
   - The final model is then used to make prediction using test dataset.
   - Residual analysis is done to ensure the assumptions for linear regression holds true.

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

Anscombe's quartet explains the shortcomings of linear regression approach. It explains

- Linear regression explains linear relationship only.
- It is sensitive to outliers
- There are assumptions that needs to be considered to model a linear regression.



As above diagram is called Anscombe's quartet. The first section explains the data points to some extent (the variance could be less since the line looks as a good fit for the data points). However, the linear relationship is not explained in the second or the last plot. The third plot shows that the model is sensitive to outlier, otherwise it would have fit the data points exactly in a straight line.

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

Pearson's R is the correlation coefficient that we use in the linear regression. It explains the linear relationship. The value ranges from -1 to 1. If it is -1, there is a negative correlation. It is 1 , then there is a positive correlation. If it is 0, then there is no relation. In some cases, where the variables are not linear, Pearson's R might not be reliable.

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

In most cases, the data set is a combination of different variables with different ranges. Scaling is a data processing done to handle high numeric values and transform it into a fixed range. This helps us to have better inferences of the dataset.

Normalized	**Standardized**
Scales values between [0, 1]	It is not bounded to a certain range

Sklearn provides a transformer called MinMaxScaler for Normalization	Sklearn provides a transformer called StandardScaler for standardization
Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1	Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation
Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution	Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution

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

It could happen when two variables show very high relationship or a exact correlation. In this case R square becomes 1, hence VIF becomes 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 Plots are plots of two quantiles against each other. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line. If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. 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.