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:

- Weekend Days like Saturday and Sunday. Working Days like Friday see high number of bookings
- 2019 has highest number of bookings compared to other years.
- Weather like Clear has high number of bookings compared to other weather days.
- Months like July-September has high number of bookings compared to other.

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

Ans: It reduces the extra columns created during dummy variable creation. Useful during proper correlation calculations.

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

Ans: Temp and atemp has highest correlation with target variable.

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

Ans:

- Normality of error terms- Error terms are normally distributed.
- Linear relation – Linearity exists among the factors/variables
- Plotting of variables – The Tested and Predicted are plotted against each other to validate the model.

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:

- Temp
- Light Rain
- Sep

General Subjective Questions

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

Ans:

A model to analyse and validate the relationship of a target variable with given set of data having independent variables or factors.

A Linear equation mathematically defined as

Y=mX+C

Y – Target Variable

m- Slope of the Linear line

X- Independent Variable

C- Y-Intercept Constant.

Linearity: It expresses either a Positive or Negative Linear relationship between the Target and Independent variable.

If both increase or both decrease then – Positive Linear relationship.

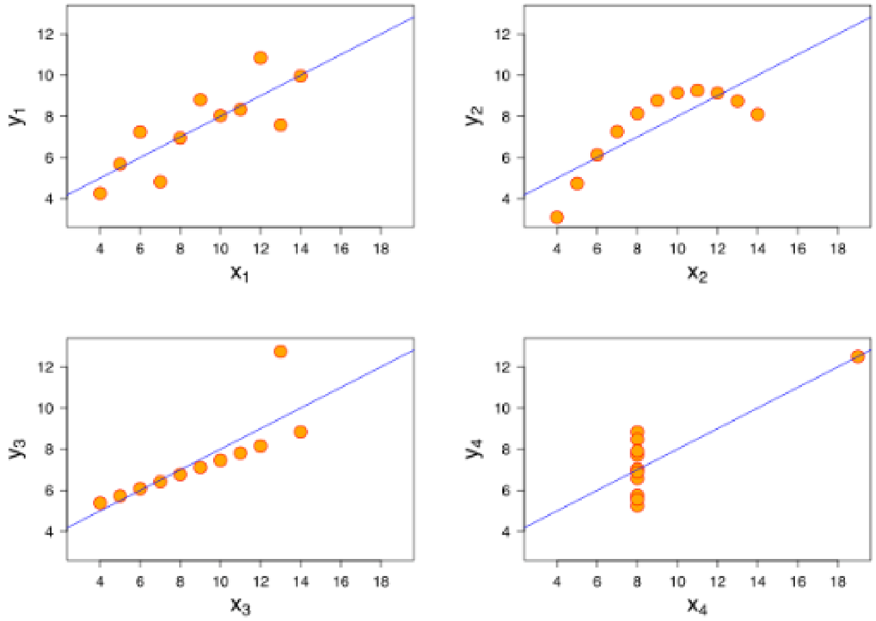
If one increase and the other decrease then – Negative Linear relationship.

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

Ans:

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics. But changes completely when graphed.

Example Graph:



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

Ans:

- 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.

- Scatter Plots are mainly used to analyse.
- The Pearson's correlation coefficient varies between -1 and +1

  r = 1 means the data is perfectly linear with a positive slope

  r = -1 means the data is perfectly linear with a negative slope

  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)

Ans:

1. Normalized Scaling:
    - Min and Max are used.
    - Scale Values are defined in a range [-1,1] or [0,1]
    - Scikit-Learn provides MinMaxScaler for Normalization

2. Standardized Scaling:

    - Mean is used.
    - Scale values are not fixed in a range or region.
    - Outliers has no effect.

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

Ans: If VIF is infinite then it means there is a perfect correlation between independent variables.

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

Ans: Referred to "quantile-quantile (q-q) plot".

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

The advantages of the q-q plot are:

a. The sample sizes do not need to be equal.
b. Many distributional aspects can be simultaneously tested.

q-q Plot determines:

a. Whether two data sets come from populations with a common distribution.
b. Whether two data sets have common location and scale
c. Whether two data sets have similar distributional shapes
d. Whether two data sets have similar tail behavior