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]:
   1. Demand has increased significantly with year
   2. Working day has positive corelation
   3. Higher demand in September
   4. Higher demand on Saturdays
   5. Windspeed has negative impact on demand
   6. Clear weather has drives high demand
   7. Higher demands in fall season

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
   [Ans]:
   Removes unnecessary columns from the dataset and this reduces multicollinearity among dummy variables. For n categorical variable values, n-1 or less dummy variables should be created.

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

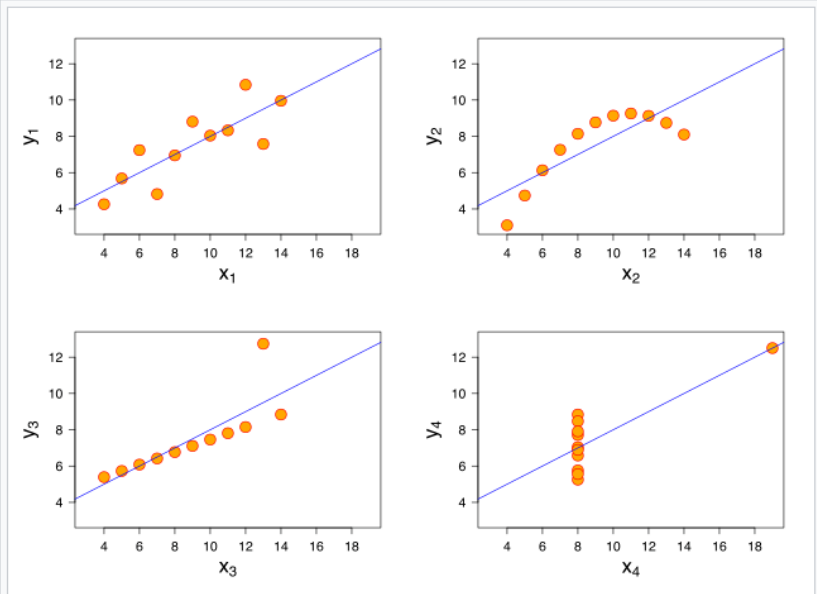
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   [Ans]:
   1. Residuals have a normal distribution
   2. Mean of residuals is zero
   3. R^2 value is significant (~75%)

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]:
   1. yr
   2. weathersit
   3. season
   4. windspeed

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   [Ans]: Linear relationship between independent variable(s) to dependent variable.

2. Explain the Anscombe's quartet in detail. (3 marks)
   [Ans]: Four datasets having similar summary statistics like mean, median, mode, variance etc. but different visual behaviours when plotted.



(image source: Wikipedia)

3. What is Pearson's R? (3 marks)
   [Ans]: Corelation between variables. Values are between -1 and 1. Value near -1 implies high negative corelation, +1 implies high positive corelation, zero implies independence.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
   [Ans]: Scaling is done by standardization (mean=0 and std dev=1) or normalization (highest value and lowest values spread between 1 and 0). It is done to remove undue favour towards high value variables.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
   [Ans]: Perfect corelation

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
   [Ans]: Probability plot: Quantile-to-quantile plot