COM 506 P ANALYTICS & SYSTEMS OF BIGT DATA PRACTICE – PROBLEM SET II

1. On New Year's Eve, Tina walked into a random shop and surprised to see a huge crowd there. She is interested to find what kind of products they sell the most, for which she needs the age distribution of customers. Help her to find out the same using **histogram**. The age details of the customers are given below

7, 9, 27, 28, 55, 45, 34, 65, 54, 67, 34, 23, 24, 66, 53, 45, 44, 88, 22, 33, 55, 35, 33, 37, 47, 41,31, 30, 29, 12.

Identify the type of histogram (eg. Bimodal, Multimodal, Skewed..etc). Use different bin sizes.

2. A Coach tracked the number of points that each of his 30 players on the team had in one game. The points scored by each player is given below. Visualize the data using ordered **stem-leaf plot** and also detect the outliers and shape of the distribution.

22, 21, 24, 19, 27, 28, 24, 25, 29, 28, 26, 31, 28, 27, 22, 39, 20, 10, 26, 24, 27, 28, 26, 28, 18, 32, 29, 25, 31, 27.

3. For a sample space of 15 people, a statistician wanted to know the consumption of water and other beverages. He collected their average consumption of water and beverages for 30 days (in litres). Help him to visualize the data using **density plot**, **rug plot** and identify the mean, median, mode and skewness of the data from the plot.

WATER	3.2, 3.5, 3.6, 2.5, 2.8, 5.9, 2.9, 3.9, 4.9, 6.9, 7.9, 8.0, 3.3, 6.6, 4.4
BEVERAGES	2.2, 2.5, 2.6, 1.5, 3.8, 1.9, 0.9, 3.9, 4.9, 6.9, 0.1, 8.0, 0.3, 2.6, 1.4

4. A car company wants to predict how much fuel different cars will use based on their masses. They took a sample of cars, drove each car 100km, and measured how much fuel was used in each case (in litres). Visualize the data using **scatterplot** and also find co-relation between the 2 variables (eg. Positive//Negative, Linear/ Non-linear co-relation) The data is summarized in the table below.

(*Use a reasonable scale on both axes and put the explanatory variable on the x-axis.)*

Fuel used (L)	3.6	6.7	9.8	11.2	14.7
Mass (metric tons)	0.45	0.91	1.36	1.81	2.27

5. The data below represents the number of chairs in each class of a government high school. Create a **box plot** and **swarm plot** (add jitter) and find the number of data points that are outliers.

35, 54, 60, 65, 66, 67, 69, 70, 72, 73, 75, 76, 54, 25, 15, 60, 65, 66, 67, 69, 70, 72, 130, 73, 75, 76

6. Generate random numbers from the following distribution and visualize the data using **violin plot.**

(i) Standard-Normal distribution.
(ii) Log-Normal distribution.

7. An Advertisement agency develops new ads for various clients (like Jewellery shops, Textile shops). The Agency wants to assess their performance, for which they want to know the number of ads they developed in each quarter for different shop category. Help them to visualize data using **radar/spider charts.**

Shop Category	Quarter 1	Quarter 2	Quarter 3	Quarter 4
Textile	10	6	8	13
Jewellery	5	5	2	4
Cleaning Essentials	15	20	16	15
Cosmetics	14	10	21	11

8. An organization wants to calculate the % of time they spent on each process for their product development. Visualize the data using **funnel chart** with the data given below.

Product Development steps	Time spent (in hours)
Requirement Elicitation	50
Requirement Analysis	110
Software Development	250
Debugging & Testing	180
Others	70

9. Let's say you are the new owner of a small ice-cream shop in a little village near the beach. You noticed that there was more business in the warmer months than the cooler months. Before you alter your purchasing pattern to match this trend, you want to be sure that the relationship is real. Help him to find the correlation between the data given.

Temperature	Number of Customers
98	15
87	12
90	10
85	10
95	16
75	7