**PSG COLLEGE OF TECHNOLOGY**

**DEPARTMENT OF APPLIED MATHEMATICS AND COMPUTATIONAL SCIENCES**

**M.Sc (TCS / SS) – 15XTO3- Data Science**

**PROBLEM SHEET-I – Data Pre-processing**

1. Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.
2. Show a *boxplot* of the data.
3. From the box plot, find the three quartiles.
4. Add some outliers to the data and identify them from the Q-plot
5. Add another numeric attribute and identify the relationship between them using Q-Q plot and scatterplot
6. Import a dataset containing missing values from UCI / KEEL repository and perform the following:
7. Plot them using a scatter plot
8. Eliminate missing values
9. Perform smoothing using mean, median and neighbour values
10. Use chi square test to find if there are exists correlation between nominal and ordinal attributes.
11. Use covariance and correlation analysis to find if there are exists correlation between numeric attributes.
12. Create a new dataset by sampling using: Sampling without replacement, with replacement, cluster sample and stratified sample.
13. Use Principal Component Analysis to reduce dimensionality.
14. Normalize data using : min-max normalization, Z- score normalization and normalization by decimal scaling,