MCA Sem. I Core

Mathematical Techniques for Computer Applications(MCAC 103) L 3

Descriptive Statistics and Data Visualization

7 Jan 2022

Vasudha Bhatnagar
Department of Computer Science,University of Delhi
Delhi, India.

Outline

1. Describing Data

2. Measures of the Location of the Data

3. Measures of the Spread and Shape of Data

4. Data Visualization

5. Paired Data

Goals of Descriptive Statistics

1. Qualitative and Quantitative analyis (visualizing data, understand the patterns, to make quick statements about the system's behavior)
2. Characterize the behavior in simple terms and quantities
3. Understand relations among variables
4. Fit suitable models and use them to make forecasts

Terminology

1. A **population** consists of all units of interest.
2. Any numerical characteristic of a population is a **parameter**.
3. A **sample** consists of observed units collected from the population to make statements about the population.
4. Any function of a sample is called **statistic**

Goals of Descriptive Statistics

1. Qualitative and Quantitative analyis (visualizing data, understand the patterns, to make quick statements about the system's behavior)
2. Characterize the behavior in simple terms and quantities
3. Understand relations among variables
4. Fit suitable models and use them to make forecasts

Terminology

1. A **population** consists of all units of interest.
2. Any numerical characteristic of a population is a **parameter**.
3. A **sample** consists of observed units collected from the population to make statements about the population.
4. Any function of a sample is called **statistic**

Collected/observed data : **Sample**

Samples are analysed to make statements about the **population**

Types (Categorization) of Variables

1. Based on Values
    1. Discrete (Number of children in a family)
    2. Continuous (Winter temperature in Leh)
2. Based on Scale
    1. Nominal (Color of eye, Nearest Metro line)
    2. Ordinal (Product rating)
    3. Interval Scale - be added, subtracted, can have values below Zero (Temperature, Calender time)
    4. Ratio Scale - can be added, subtracted, multiplied, and divided, no values below Zero (Age, Money, Weight)
    5. Binary (Gender)
3. Based on Role
    1. Independent (Years of experience, highest qualification ⟶ Salary)
    2. Response (Years of experience, highest qualification ⟶ Salary)

Measures of the Location

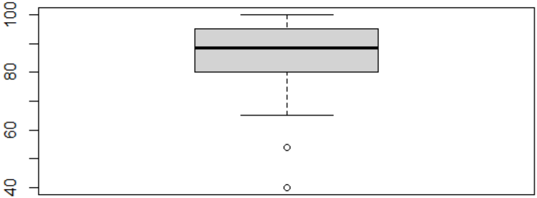
Common data location descriptors

1. Five number summary and Quartiles
2. Inter-Quartile Range
3. Outliers: All observations above Q3+1.5*IQR or below Q1 - 1.5*IQR are outliers
4. Percentiles: useful for comparing values,
   may or may not be part of the data,
   indicate the relative standing of a data value when data are sorted into numerical order from smallest to largest,
   interpretation of whether a certain percentile is "good" or "bad" depends on the context of the situation to which the data applies
   Quartiles are special percentiles

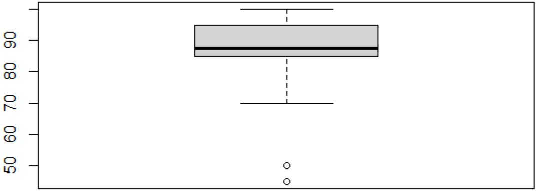
Given marks of 38 **students in Programming course 54.0 87.0 76.0 90.0 100.0 95.0 95.0 90.0 80.0 90.0 100.0 85.0 76.0 85.5 90.0 100.0 90.0 85.5 100.0 70.0 95.0 95.0 79.0 95.0 85.0 90.0 100.0 86.0 95.0 100.0 65.0 85.5 76.0 100.0 66.5 85.0 40.0 85.5**

Sorted marks of 38 students in Programming course
40.0 54.0 65.0 66.5 70.0 76.0 76.0 76.0 79.0 80.0 85.0 85.0 85.0 85.5 85.5 85.5 85.5 86.0 87.0 90.0 90.0 90.0 90.0 90.0 90.0 95.0 95.0 95.0 95.0 95.0 95.0 100.0 100.0 100.0 100.0 100.0 100.0 100.0
Five number summary
Min: 40.0 Q1: 80.0 Q2: 88.5 Q3: 95.0 Max: 100.0
IQR: 95 -80 = 15
Outliers: 40, 54
Percentiles: 90^{th} =100, 80^{th} = 95, 60^{th} = 90

Box pot for Programming Marks



Box pot for Data Structure marks



Measures of centrality

Given n data elements $x_1, x_2, \ldots, x_n$

1. Mean: Average, Computed as $\frac{\sum_{i=1}^{n} x_i}{n}$
2. Median: Middle most data value
   How to find: Arrange in ascending order and pick the middle most data value (what if n is even)?
3. Mode: Most frequent data value
   How to find: Make a frequency table
   Pick the value with highest frequency

Given sorted marks of 38 students in Programming course
40.0 54.0 65.0 66.5 70.0 76.0 76.0 76.0 79.0 80.0 85.0 85.0 85.0 85.5 85.5 85.5 85.5 86.0 87.0 90.0 90.0 90.0 90.0 90.0 90.0 95.0 95.0 95.0 95.0 95.0 95.0 100.0 100.0 100.0 100.0 100.0 100.0 100.0

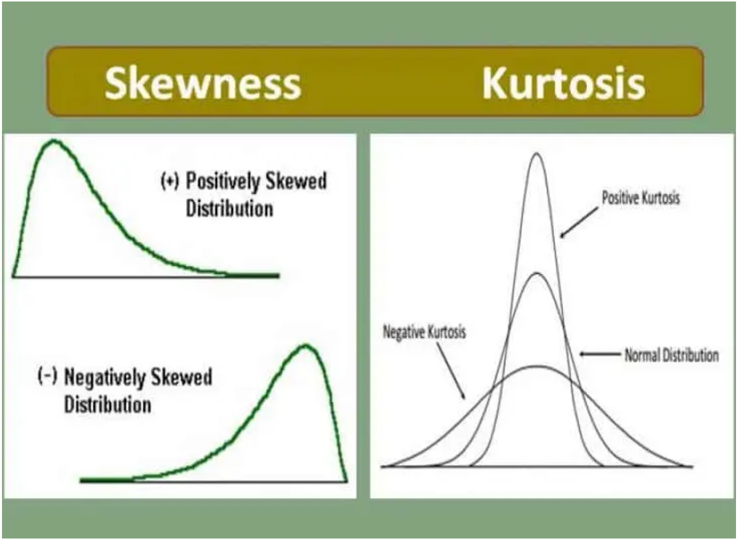
1. Mean: 85.85526
2. Median: 88.50
3. Mode: 100

Measures of Spread and Shape

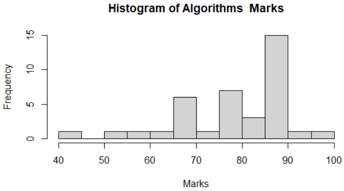
1. **Variance (σ^2): $\frac{1}{n}\Sigma_i(x_i - \bar{x})^2$**
2. **Standard Deviation (σ): $\sqrt{\sigma^2}$**
3. **Skewness (β): $\frac{1}{n}\Sigma_i(x_i - \bar{x})^3/\sigma^3$ (measure of symmetry)**
4. **Kurtosis : $\frac{1}{n}\Sigma_i(x_i - \bar{x})^4/\sigma^4 - 3$ (measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution)**

For Programming marks

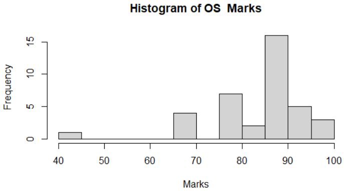
1. σ^2 = 177.3096 , σ = 13.31576
2. Skewness = -1.422479 (*left tail is long relative to the right tail*)
3. Kurtosis = 2.189252


Skewness
Kurtosis
(+) Positively Skewed Distribution
(-) Negatively Skewed Distribution
Positive Kurtosis
Negative Kurtosis
Normal Distribution

Algorithms and Operating Sysytems Marks

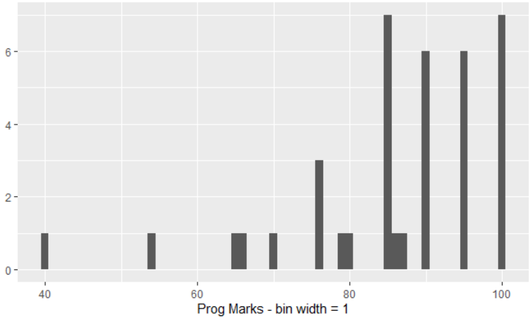


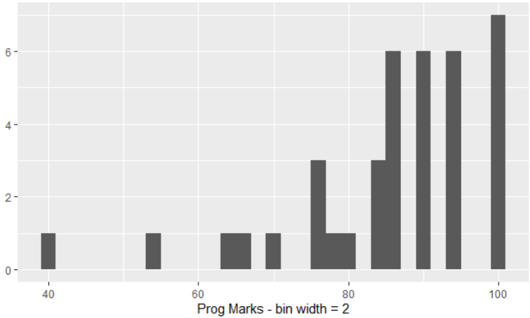
Mean = 79.23684, Var = 162.0235, SD = 12.72884, Skewness = -0.9541266, Kurtosis = 0.6076563



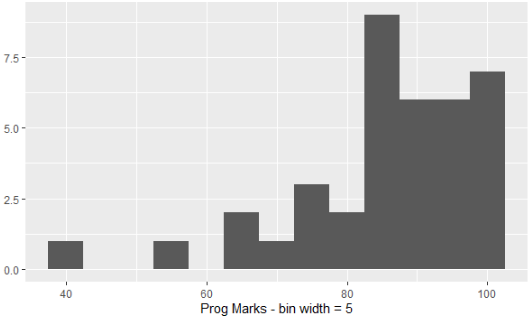
Mean = 84.42763, Variance = 134.4896, S D = 11.59696, Skewness = -1.560437, Kurtosis = 3.554787

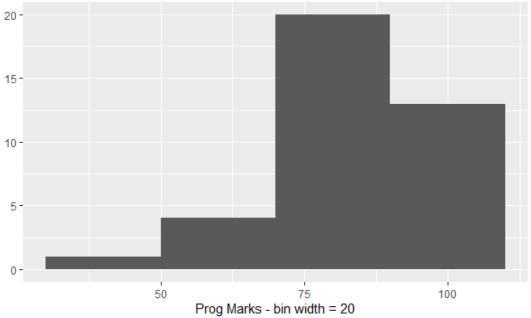
Making Sense out of Data by Grouping





Programming Marks - Class interval 5 Vs. 20





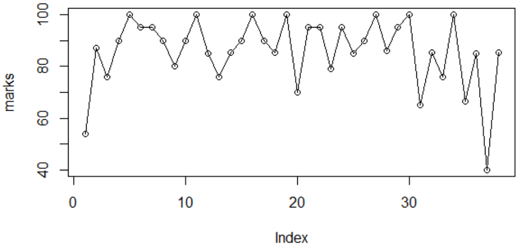
If the dataset is large and the number of distinct values is too large, it is useful to divide the values into grouping (class intervals)
Then plot the number of data values falling in each class interval
The number of class intervals chosen depends on goal of analysis
Choosing too few classes leads to lossing of information about the actual data values in a class
Choosing too many classes will not distinguish between the classes

Mean and Variance of grouped data are weighted by class frequencies

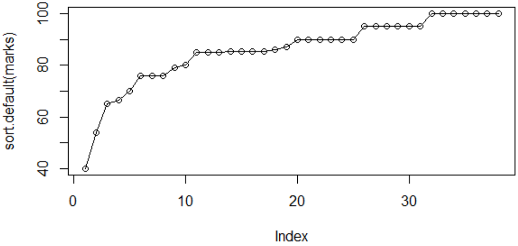
Stem and Leaf Plot
Data: Marks in Programming

```
The decimal point is 1 digit(s) to the right of the |
4 | 0
5 | 4
6 | 57
7 | 06669
8 | 0555666667
9 | 000000555555
10 | 0000000
```

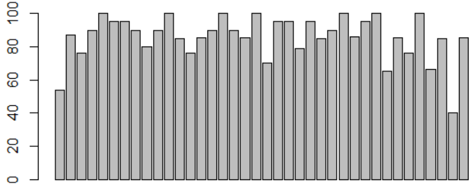
Programming Marks - Line Graph



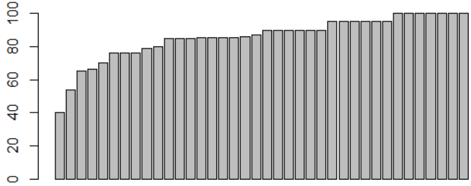
Programming Marks Sorted



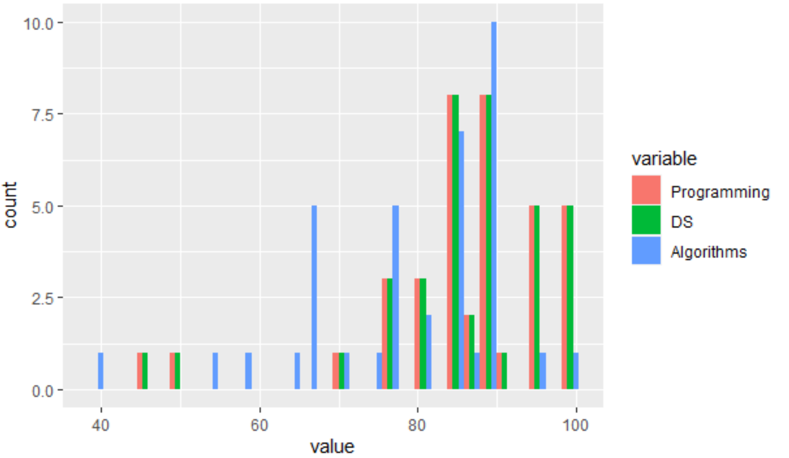
Programming Marks -Bar Plot

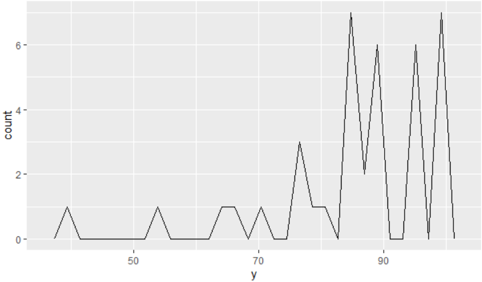


Programming Marks Sorted

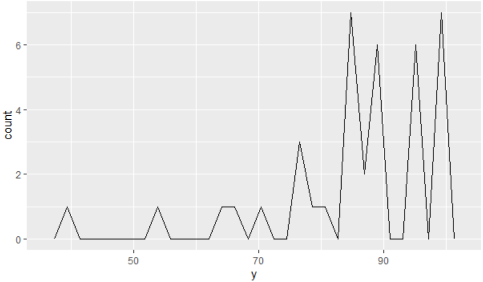


Comparison of Prog, Data Structure and Algorithms Marks

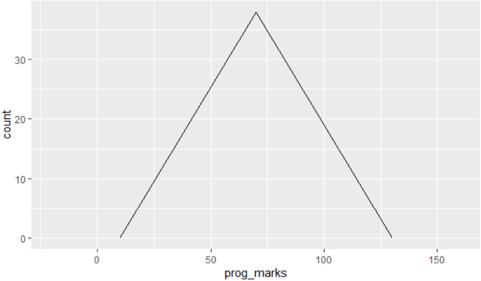




Frequency Polygon



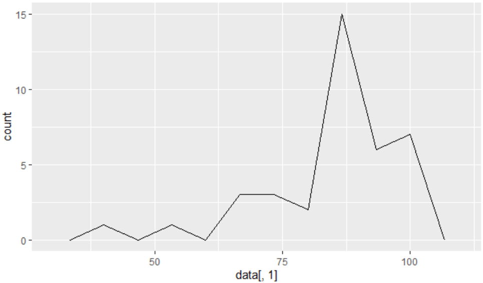
Frequency Polygon



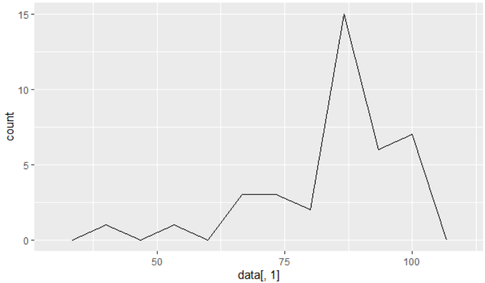
Frequency Polygon -

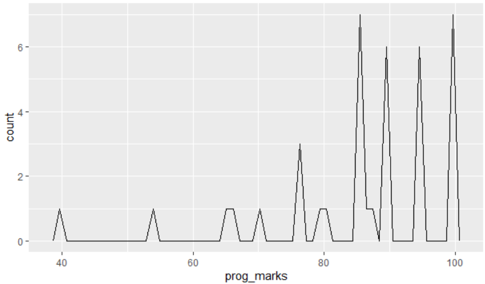
Bin = 1

Frequency Polygon - Bin = 10



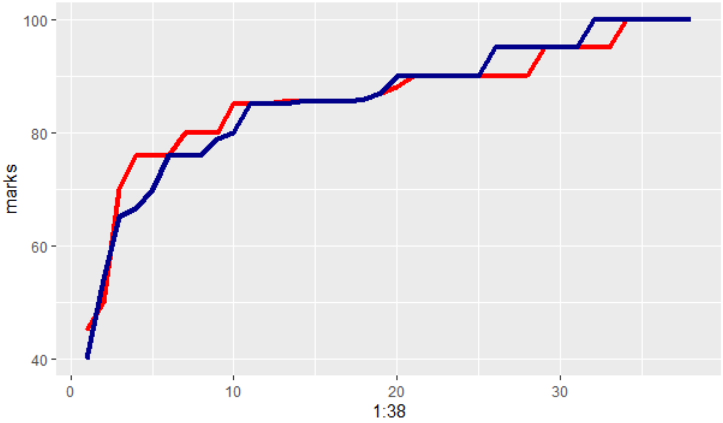
Frequency Polygon - Bin = 10



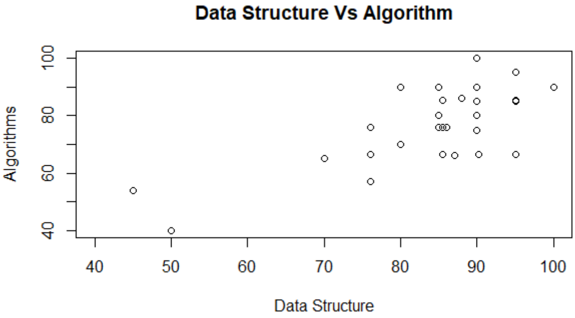


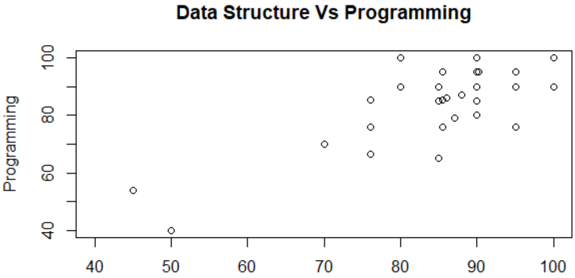
Frequency Polygon - Bin = 60

Comparison using line graph

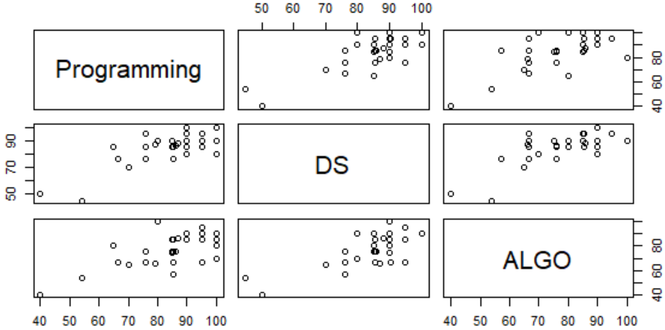


Scatter Plot



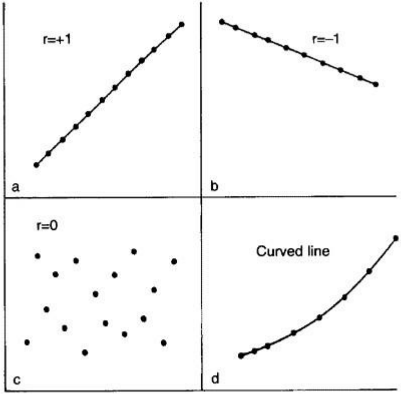


Scatterplot Matrix



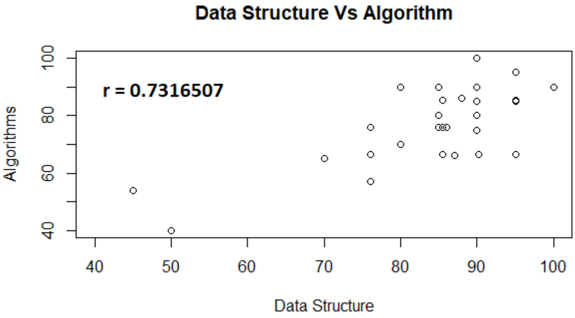
Pearsons's Correlation Coefficient

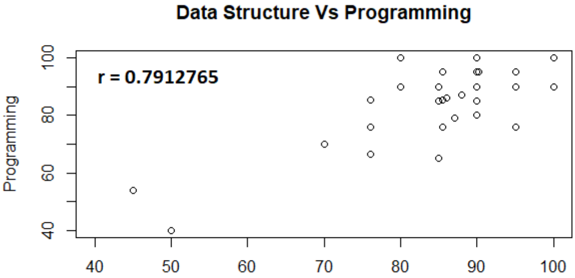
1. Measures degree of association between two variables, and denoted by *r*
2. Measure of linear association between two variables
3. Scales from + 1 through 0 to – 1
4. Computed as covariance of the two variables divided by the product of their standard deviations



$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{[\Sigma(x-\bar{x})^2(y-\bar{y})^2]}}$$

Scatter Plot


Data Structure Vs Algorithm
r = 0.7316507
Algorithms
Data Structure


Data Structure Vs Programming
r = 0.7912765
Programming