Exploring Data With R

Abhishek Kumar
ItsAbhishekKumar.com
@MeAbhishekKumar

pluralsight
hardcore dev and IT training

Outline

Overall structure

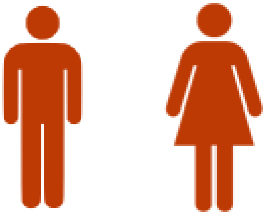
Continuous data

Categorical data

Types of Data

Categorical data

Colors

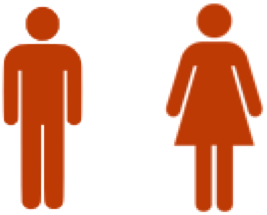


Gender

Use factor

Continuous data

Mileage



Height, Weight, Age

Use numeric / integer

Overall Structure

Number of observations

Number of features

Data types

Sample data

Dataset

Iris dataset

50 samples

Iris-setosa

50 samples

Iris-virginca

50 samples

Iris-versicolor

Features : sepal length, sepal width, petal length, petal width

Available in datasets package

Images: http://en.wikipedia.org/wiki/Iris_flower_data_set

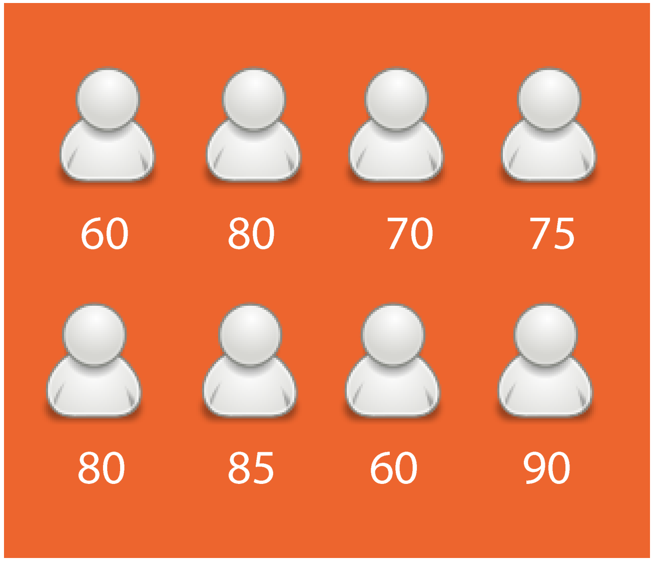
Analysis of Continuous Data

Central tendency

Spread or dispersion

Central Tendency

Mean (Average)



Set A

$$\text{Mean} = \frac{\textbf{Sum of all values}}{\textbf{Number of values}}$$

Mean = **75**

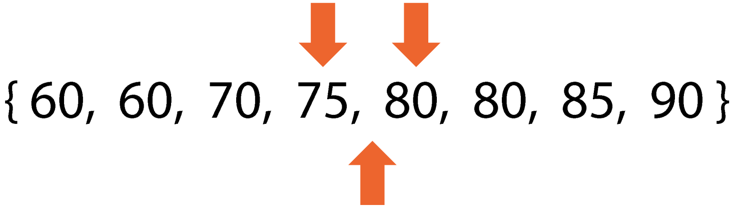
Central Tendency

Median

60 80 70 75

80 85 60 90

Set A

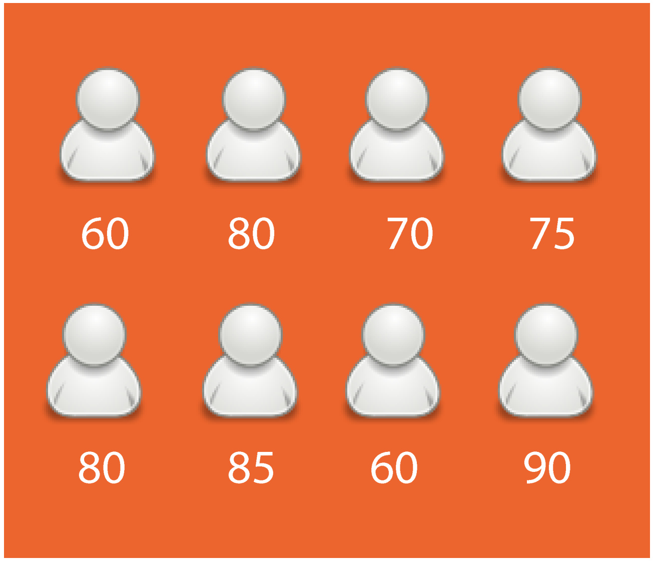


{ 60, 60, 70, 75, 80, 80, 85, 90 }

Median = **77.5**

Central Tendency

Why Not Sufficient?



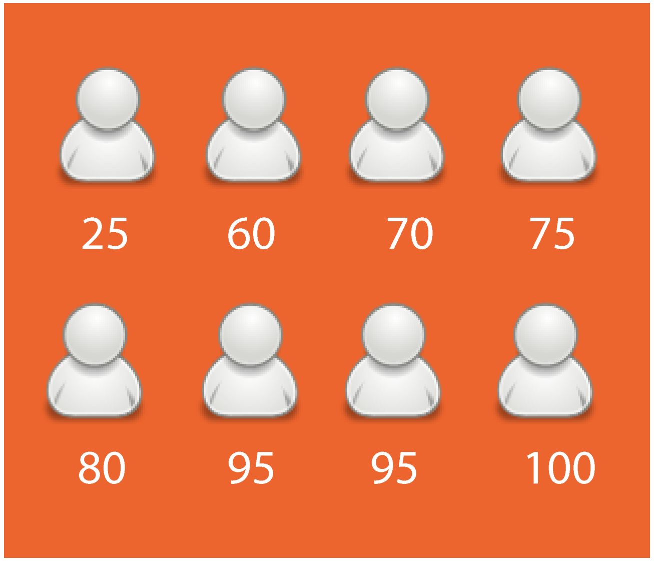
60 80 70 75

80 85 60 90

Set A

Mean = **75**

Median = **77.5**



25 60 70 75

80 95 95 100

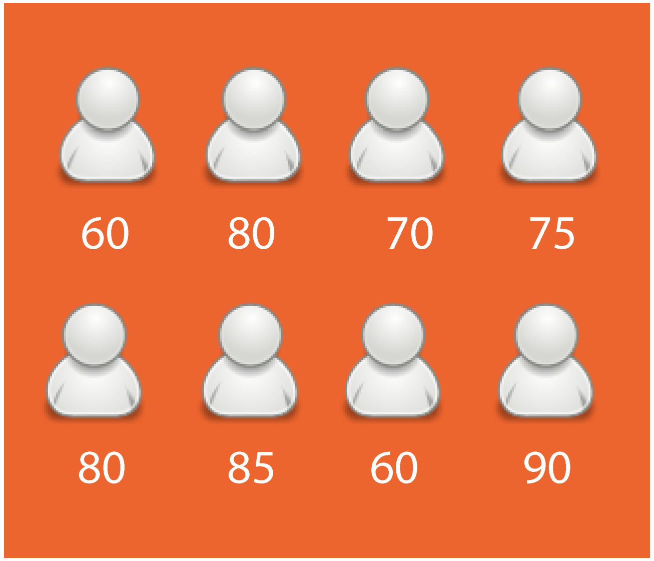
Set B

Mean = **75**

Median = **77.5**

Spread

Range



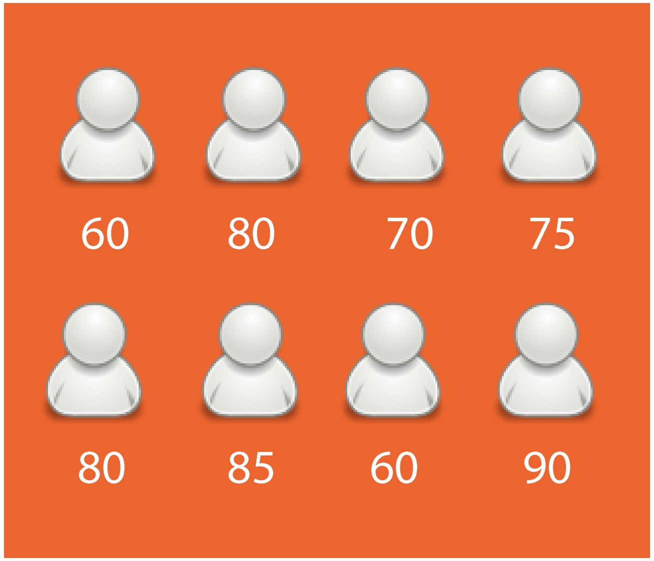
Set A

Range = **maximum - minimum**

Range = **90 – 60 = 30**

Spread

Range



60 80 70 75

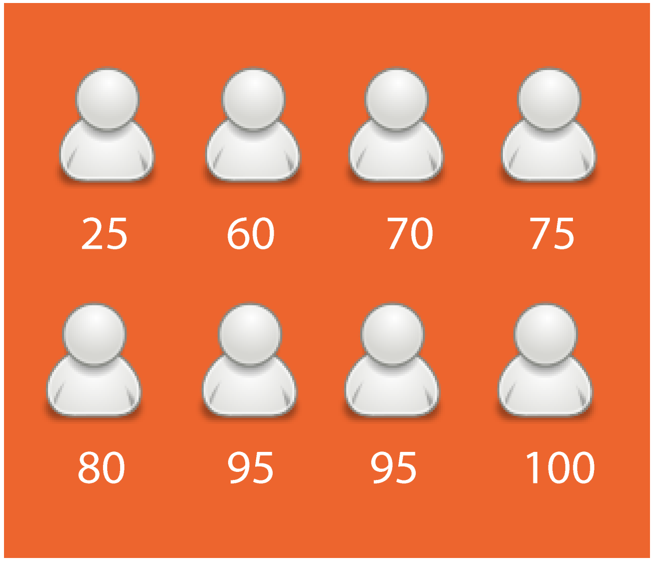
80 85 60 90

Set A

Mean = **75**

Median = **77.5**

Range = **90 – 60 = 30**



25 60 70 75

80 95 95 100

Set B

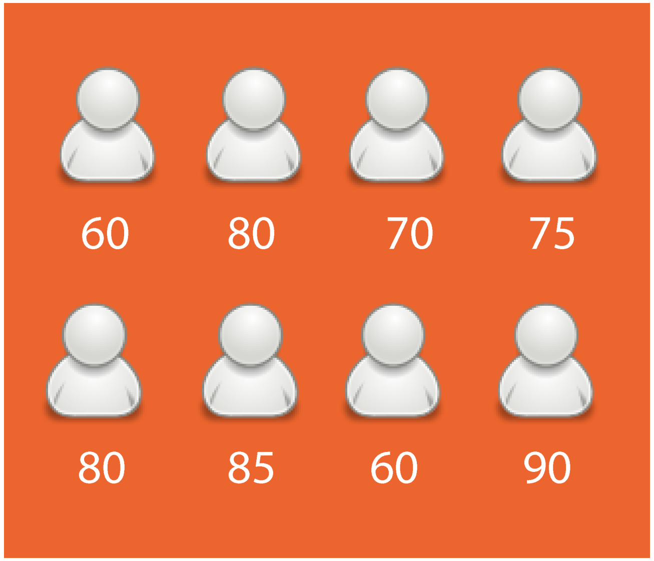
Mean = **75**

Median = **77.5**

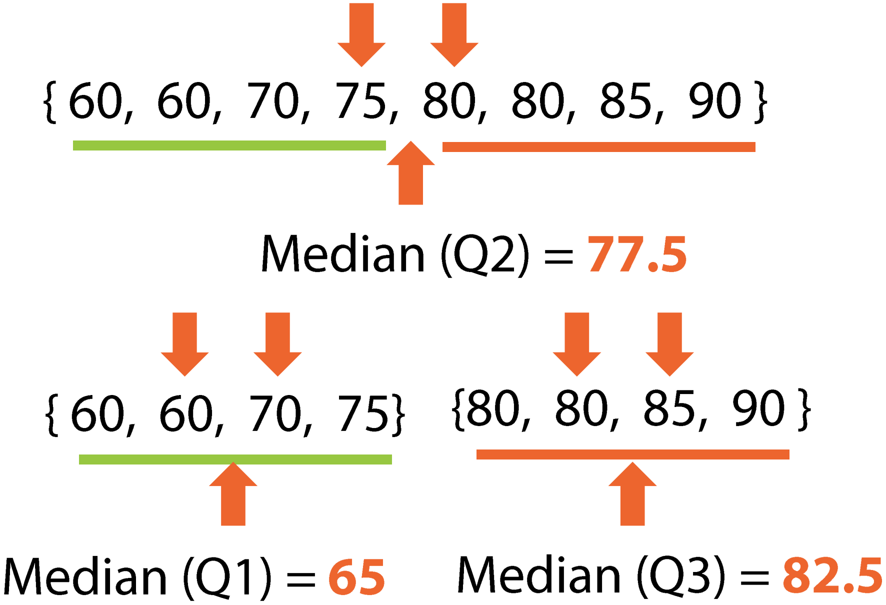
Range = **100 – 25 = 75**

Spread

Quartiles



Set A



{ 60, 60, 70, 75, 80, 80, 85, 90 }

Median (Q2) = **77.5**

{ 60, 60, 70, 75} {80, 80, 85, 90 }

Median (Q1) = **65**

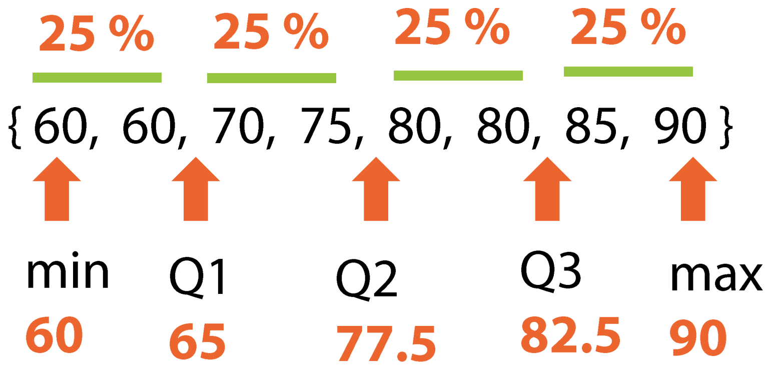
Median (Q3) = **82.5**

Spread

Quartiles

60
80
70
75
80
85
60
90

Set A



25 % 25 % 25 % 25 %

{ 60, 60, 70, 75, 80, 80, 85, 90 }

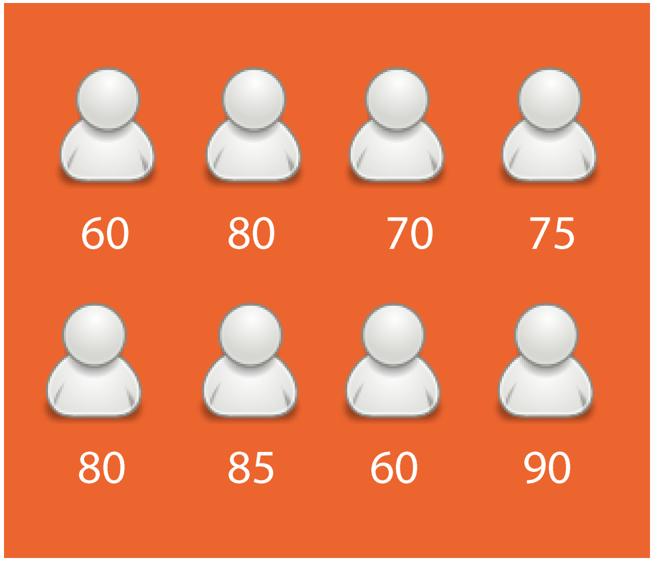
min Q1 Q2 Q3 max

60 65 77.5 82.5 90

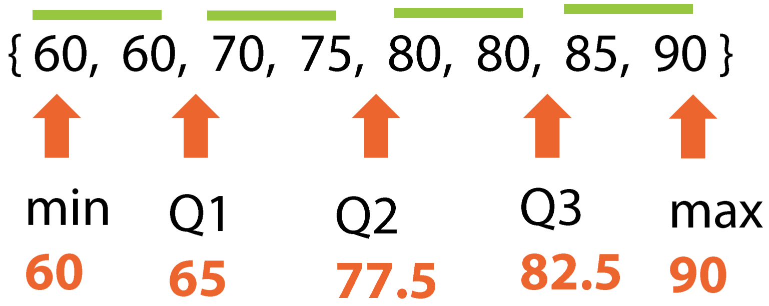
Five point summary **(min, Q1, Q2, Q3, max)**

Spread

Box Plot (Box – Whisker Plot)

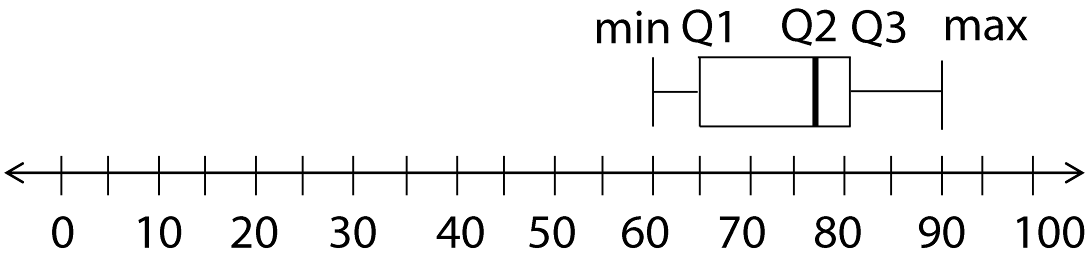


Set A



{ 60, 60, 70, 75, 80, 80, 85, 90 }

min	Q1	Q2	Q3	max
60	65	77.5	82.5	90



Spread

Box Plot

60 80 70 75

80 85 60 90

Set A

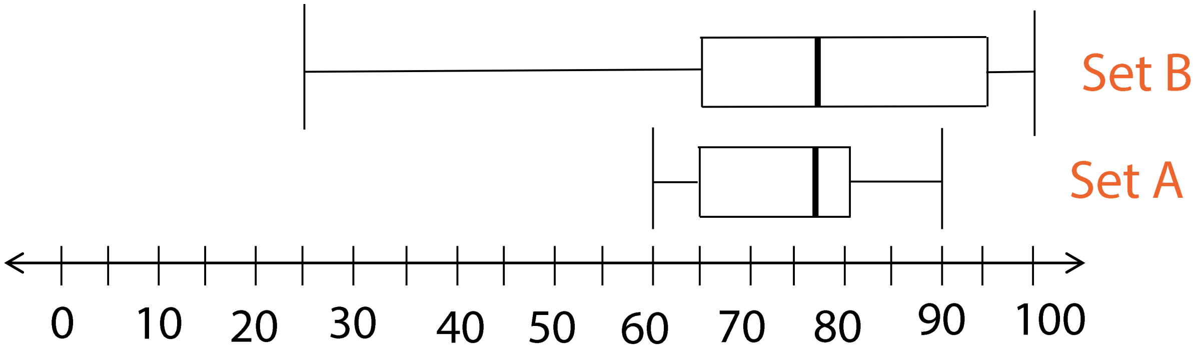
min	Q1	Q2	Q3	max
60	**65**	**77.5**	**82.5**	**90**

25 60 70 75

80 95 95 100

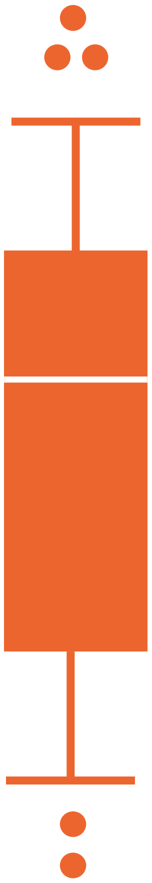
Set B

min	Q1	Q2	Q3	max
25	**65**	**77.5**	**95**	**100**



Spread

Box Plot



→ Outliers (higher than 1.5 times Q3)

→ Maximum : Highest value (excluding outliers)

→ Q3 : Quartile 3 (upper quartile)

→ Q2 : Quartile 2 (median)

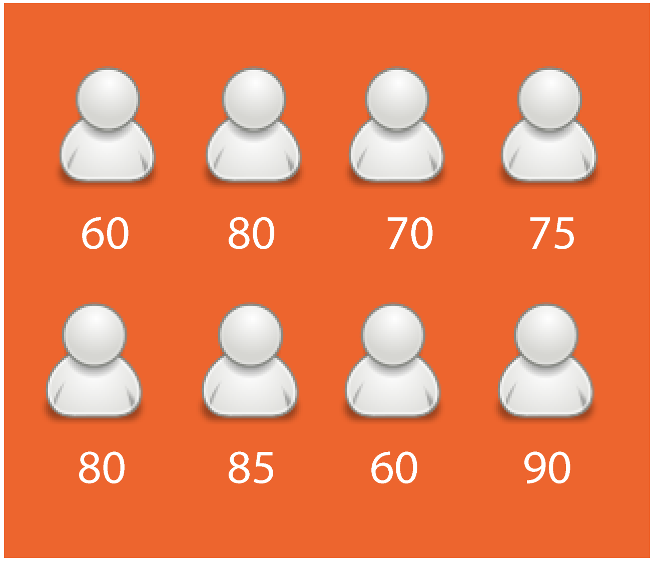
→ Q1 : Quartile 1 (lower quartile)

→ Minimum : least value (excluding outliers)

→ Outliers (less than 1.5 times Q1)

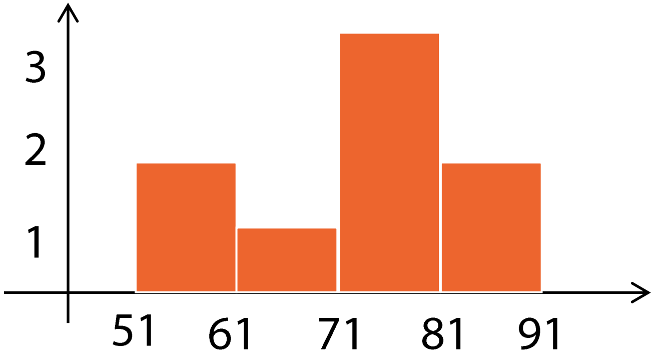
Spread

Histogram



Set A

Range	Count
51-61	2
61-71	1
71-81	3
81-91	2



Histogram

Spread

Variance & Standard deviation

60
80
70
75
80
85
60
90

Set A

{ 60, 80, 70, 75, 80, 85, 60, 90} Mean = **75**

{ -15, 5 , -5 , 0 , 5 , 10 , -15, 15}

{ 225 , 25 , 25 , 0 , 25 , 100 , 225, 225}

850

850 / 8 = 106.25

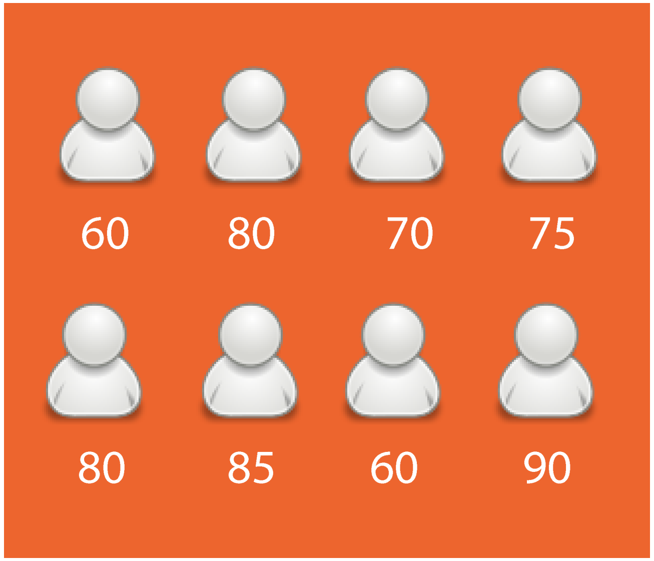
Variance = **106.25**

Sqrt(106.25) = ~10.30

Std. dev = **~ 10.30**

Spread

Variance & Standard Deviation



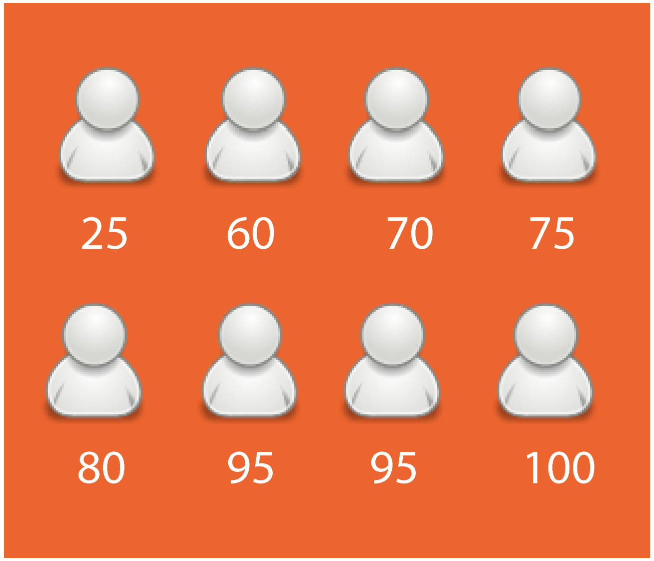
Set A

Mean = **75**

Median = **77.5**

Std. deviation = **~10.3**

Variance = **106.25**



Set B

Mean = **75**

Median = **77.5**

Std. deviation = **~ 22.9**

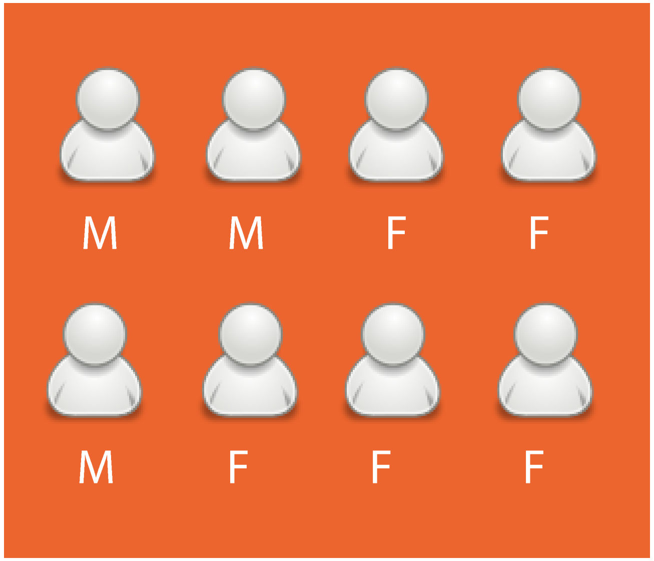
Variance = **525**

Analysis of Categorical Data

Frequency distribution

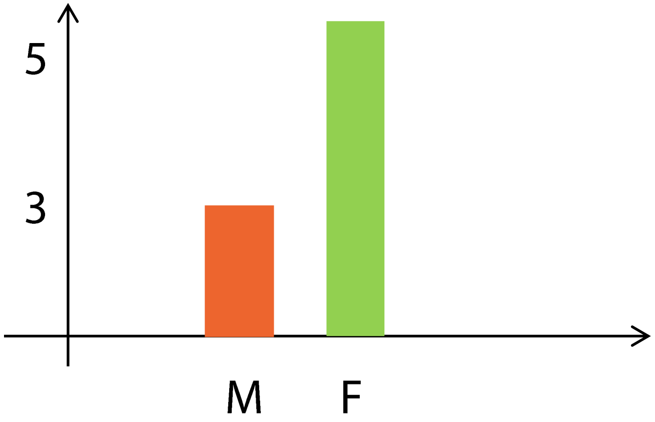
Category statistics

Frequency Distribution



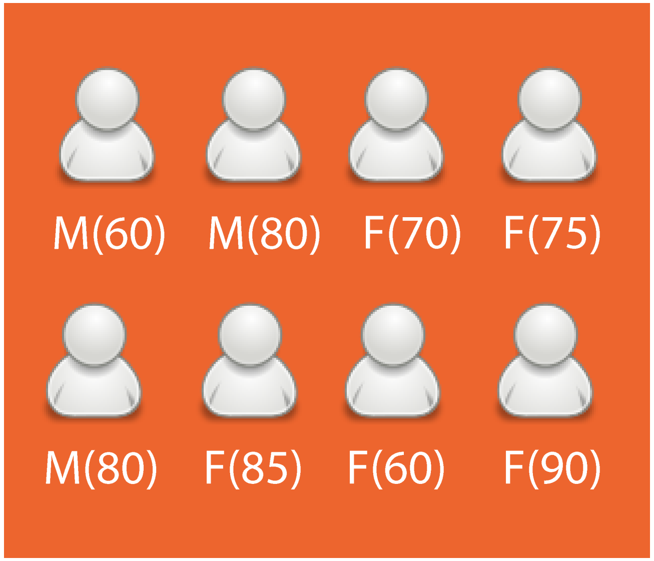
Set A

Category	Count	Proportion
Male	3	3/8 = 0.375
Female	5	5/8 = 0.625



Bar plot

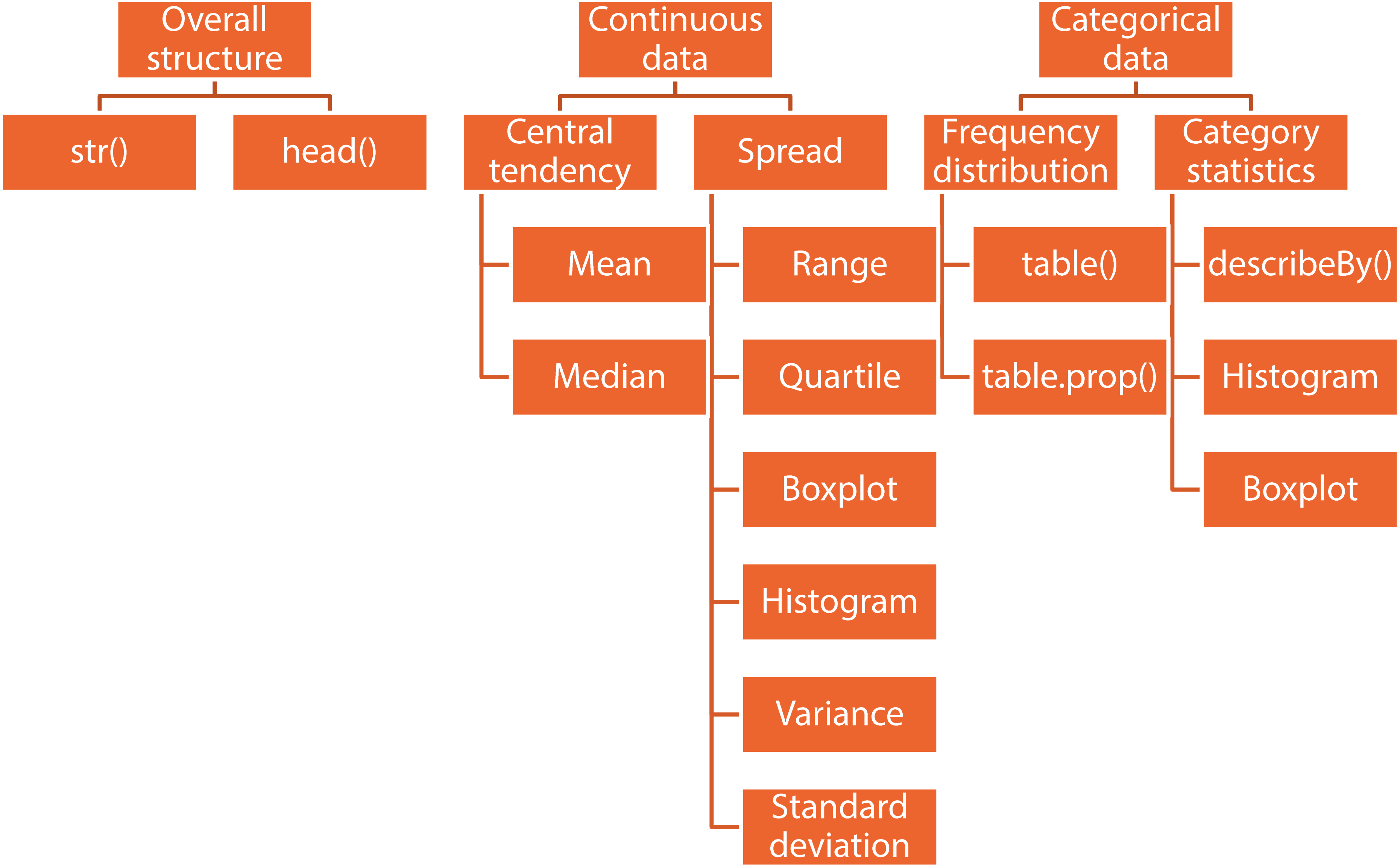
Category Statistics



Set A

Category	Values	Mean
Male	{ 60, 80, 80)	~ 73.3
Female	{ 70, 75, 85, 60, 90 }	76

Summary



- Overall structure
  - str()
  - head()
- Continuous data
  - Central tendency
    - Mean
    - Median
  - Spread
    - Range
    - Quartile
    - Boxplot
    - Histogram
    - Variance
    - Standard deviation
- Categorical data
  - Frequency distribution
    - table()
    - table.prop()
  - Category statistics
    - describeBy()
    - Histogram
    - Boxplot