

# Exploring and Processing Data - Part 2

---



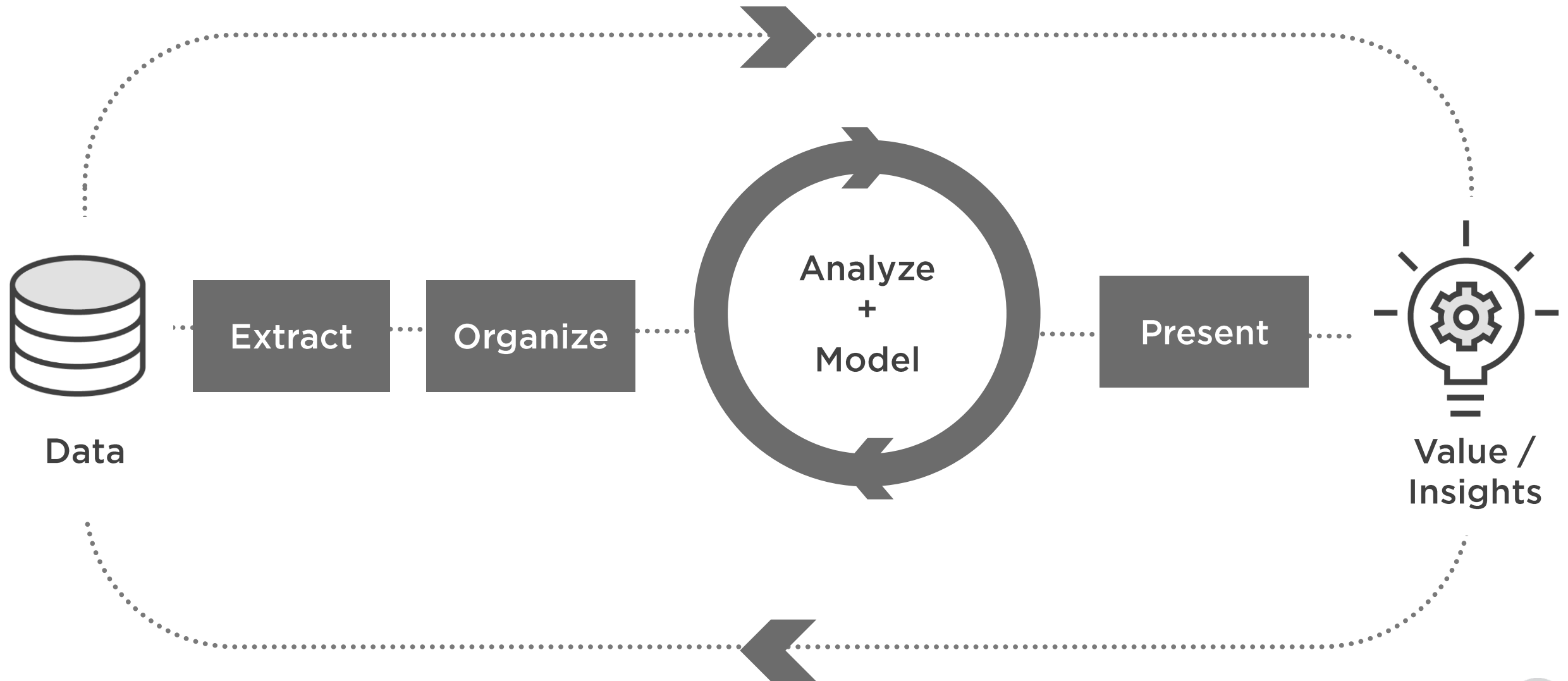
**Abhishek Kumar**

AUTHOR

@meabhishekkumar



# Data Science Project Cycle



```
graph TD; DM[Data Munging] --> FE[Feature Engineering]; FE --> AV[Advanced Visualization]; AV --> EDA[Exploratory Data Analysis]; EDA --> DM; Organize((Organize))
```

The diagram illustrates a cyclical data science process. It consists of four rectangular boxes arranged in a circle, connected by curved arrows indicating a clockwise flow. The boxes are labeled: "Data Munging" (top), "Feature Engineering" (right), "Advanced Visualization" (bottom), and "Exploratory Data Analysis" (left). The "Exploratory Data Analysis" box is highlighted in orange, while the others are dark gray. In the center of the cycle is the word "Organize".



# Overview (Concepts)

## Exploratory data analysis

- Distributions
- Grouping
- Crosstabs
- Pivots



# Overview (Tools)

## Python

- NumPy
- Pandas



# Exploratory Data Analysis

**Basic structure**

**Summary  
statistics**

**Distributions**

**Grouping**

**Crosstabs, Pivots**



# Distributions

## Univariate

- Histogram
- Kernel Density Estimation (KDE) plot

## Bivariate

- Scatter plot



# Histogram



Age

10

10

11

9

10

8

12

12

8

10

Bin

3

3

3

2

3

2

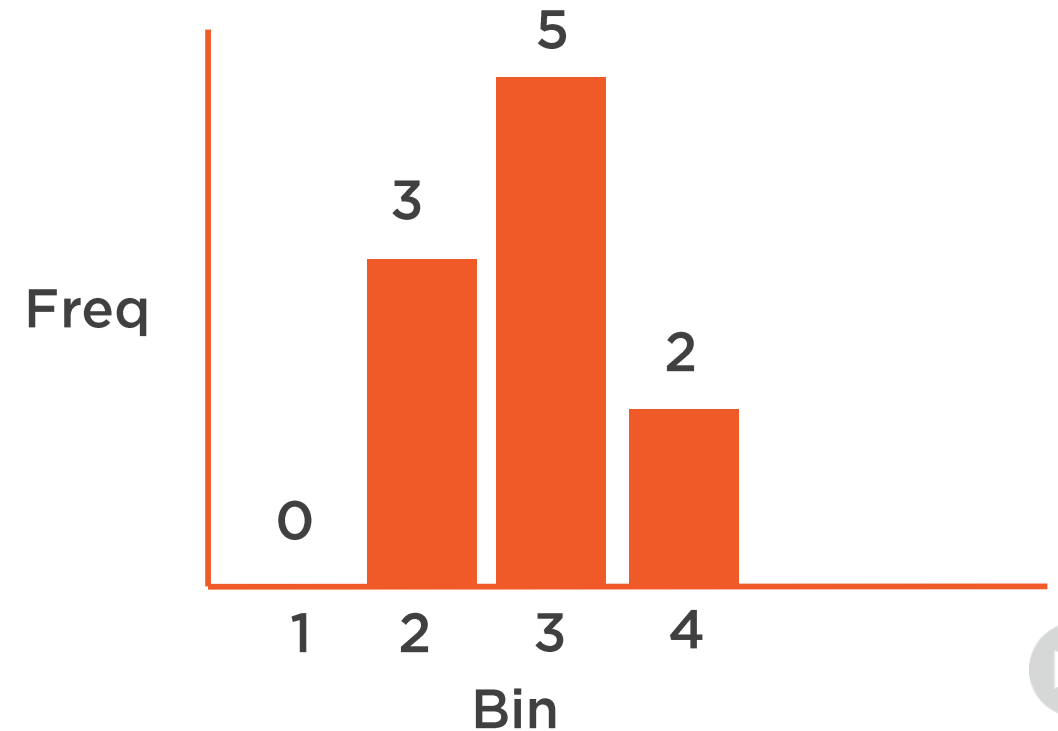
4

4

2

3

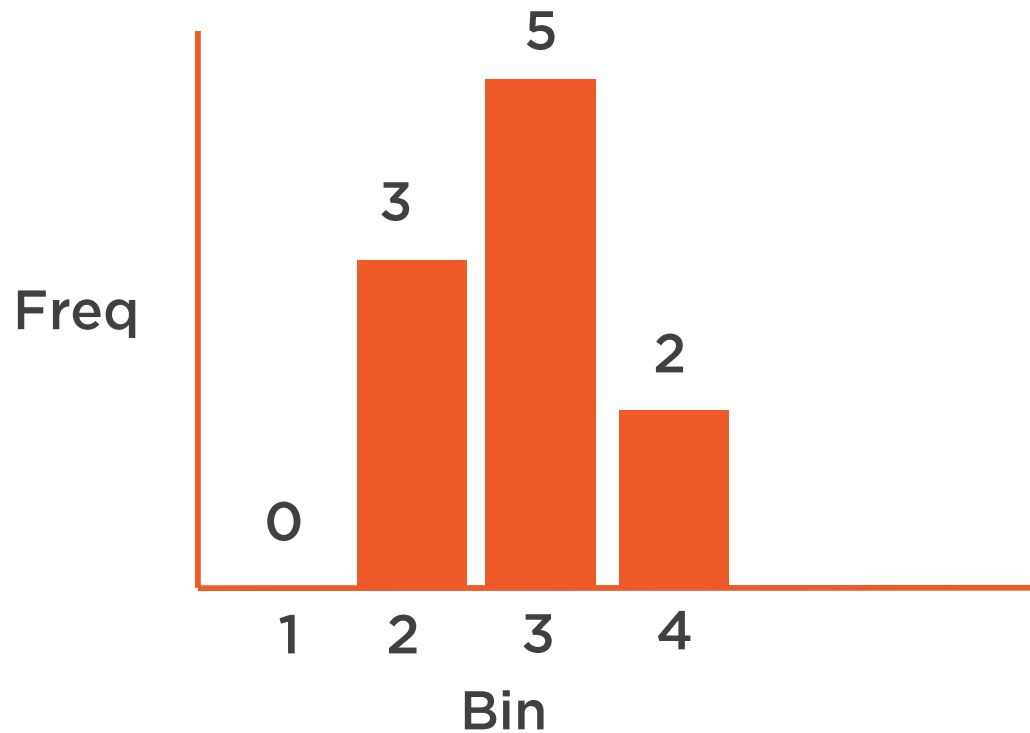
Bucket (bin) number	Bucket (bins)	Frequency
1	6-8	0
2	8-10	3
3	10-12	5
4	12-14	2



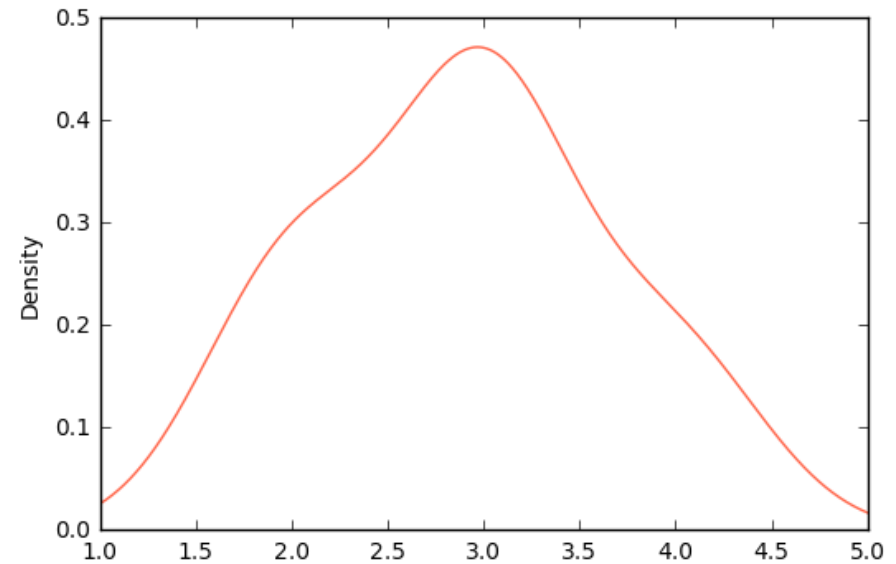


# Kernel Density Estimation (KDE) Plot

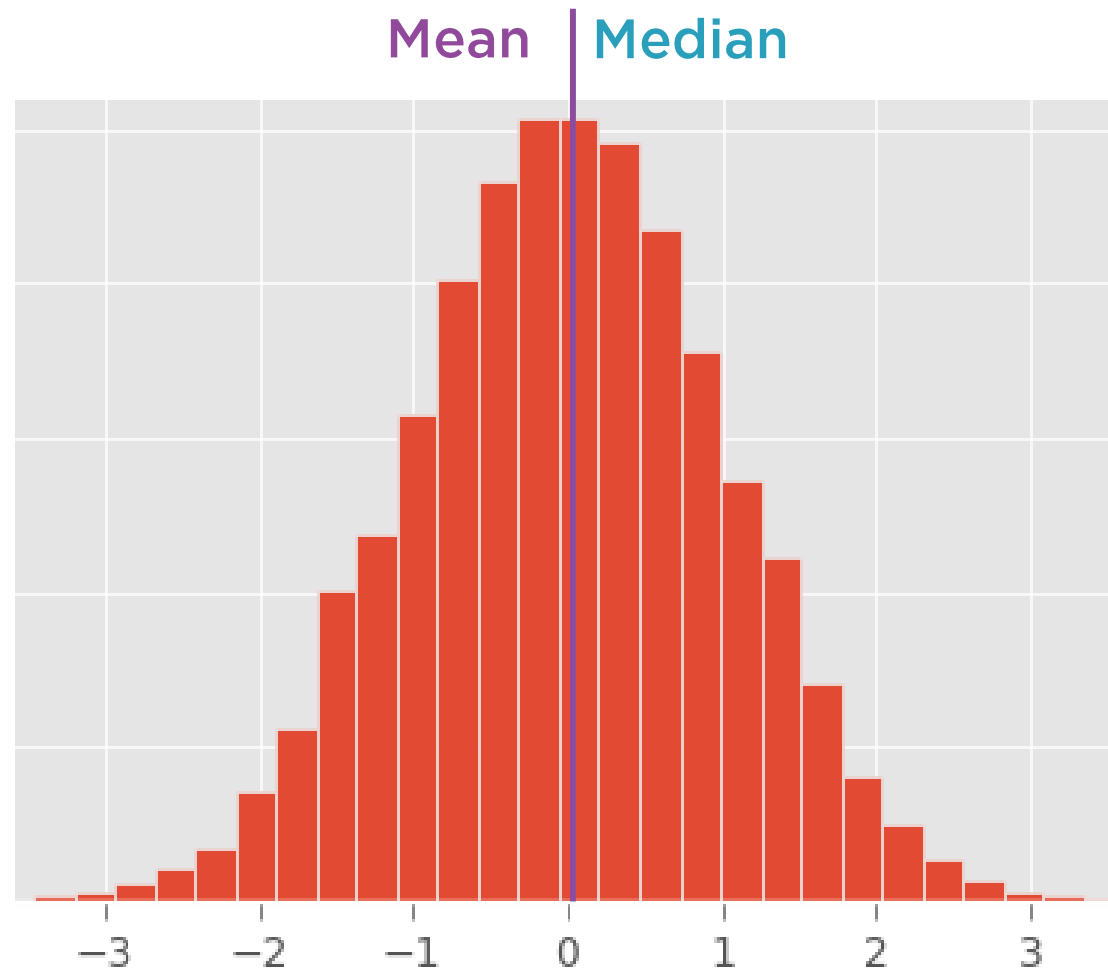
Histogram



KDE Plot



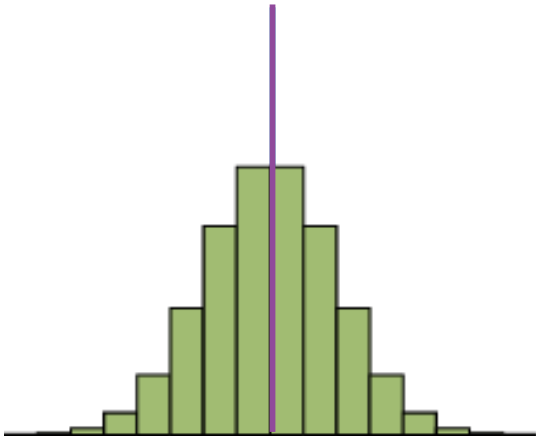
# Normal Distribution



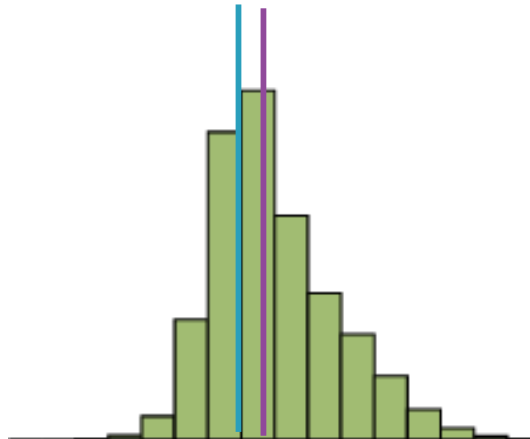
Skewness : zero



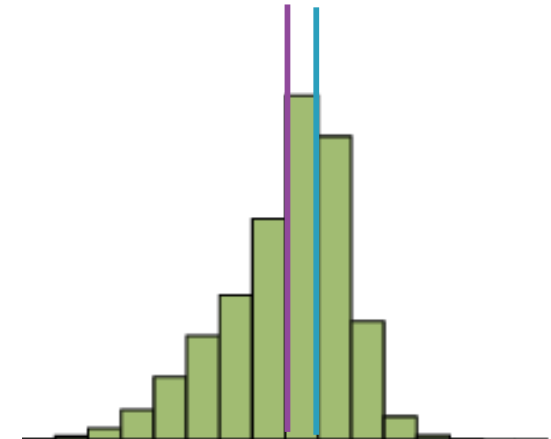
# Univariate Distribution : Skewness



Normal distribution



Right (positive)  
skewed



Left (negative)  
skewed

— Median

— Mean



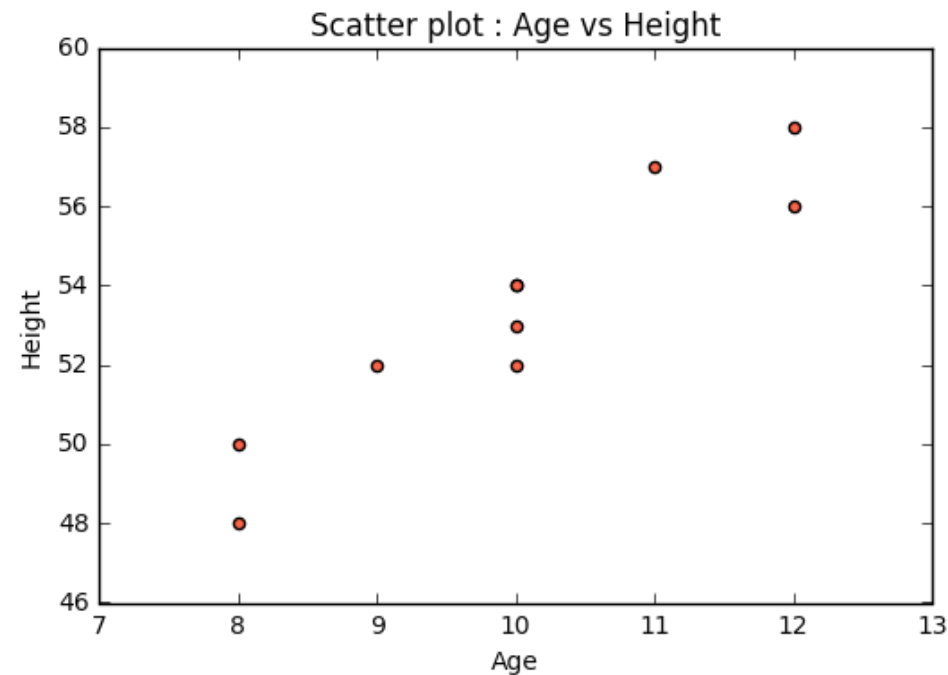
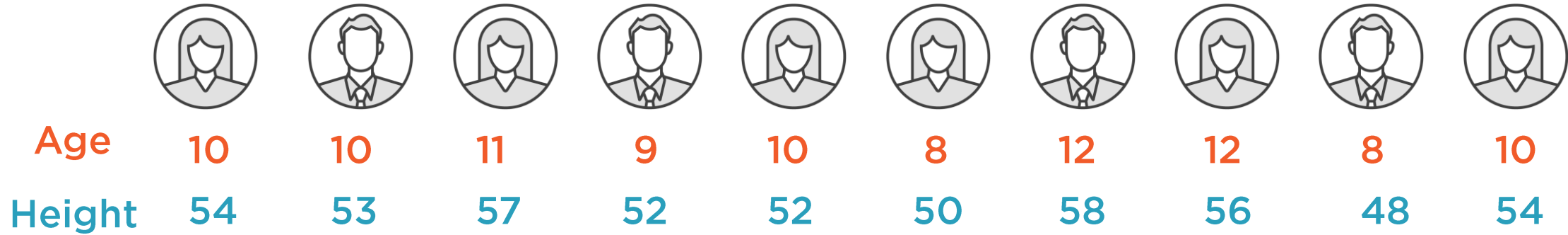
# Demo



## Creating univariate distribution plots using Pandas



# Scatter Plot



# Demo



## Creating scatter plots using Pandas



# Exploratory Data Analysis

**Basic structure**

**Summary  
statistics**

**Distributions**

**Grouping**

**Crosstabs, Pivots**



# Grouping



Group : F




Group : M

Mean age	10.17	9.75
Median age	10	9.5
Count	6	4





# Grouping

										
Age	10	10	11	9	10	8	12	12	8	10
Class	1	1	2	2	3	3	2	1	1	3

Group	Summary
M - 1	...
M - 2	...
M - 3	...
F - 1	...
F - 2	...
F - 3	...



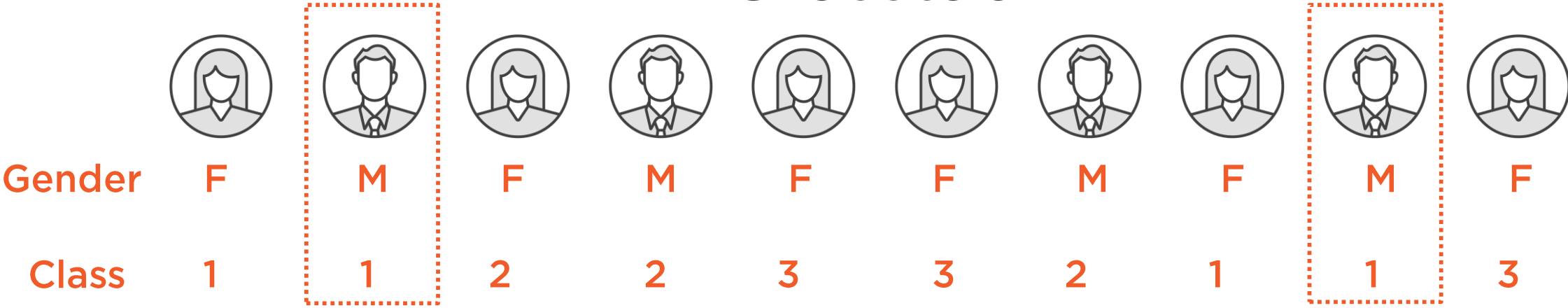
# Demo



## Grouping and aggregation using Pandas



# Crosstab



Class	1	2	3
Gender			
M	2	2	0
F	2	1	3













# Demo



## Crosstab using Pandas



# Pivot Table

										
Gender	F	M	F	M	F	F	M	F	M	F
Class	1	1	2	2	3	3	2	1	1	3
Age	10	10	11	9	10	8	12	12	8	10

3

Age

4

Mean

2

1

Class	1	2	3
Gender			
M	9.0	10.5	NaN
F	11.0	11.0	9.33



# Demo



## Pivot table using Pandas



# Summary



**Distributions**

**Grouping**

**Crosstabs**

**Pivots**

