Python for Data Science Comprehensive Workshop

Part 02 – Data Cleaning & Manipulation Using Pandas

H.M. Samadhi Chathuranga Rathnayake

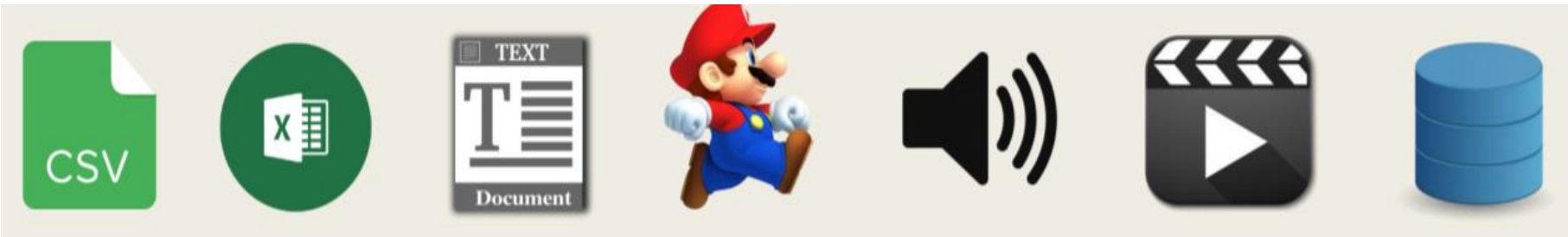
B.Sc(Hons).Special in Industrial Statistics (1st Class) (UOC),
B.Eng (Hons) in Software Engineering (LMU),
CLSSWB, Dip SE, Dip IT, Dip IT & E-Com, Dip B.Mgt, Dip HRM, Dip Eng

Data Wrangling

Where does data come from?

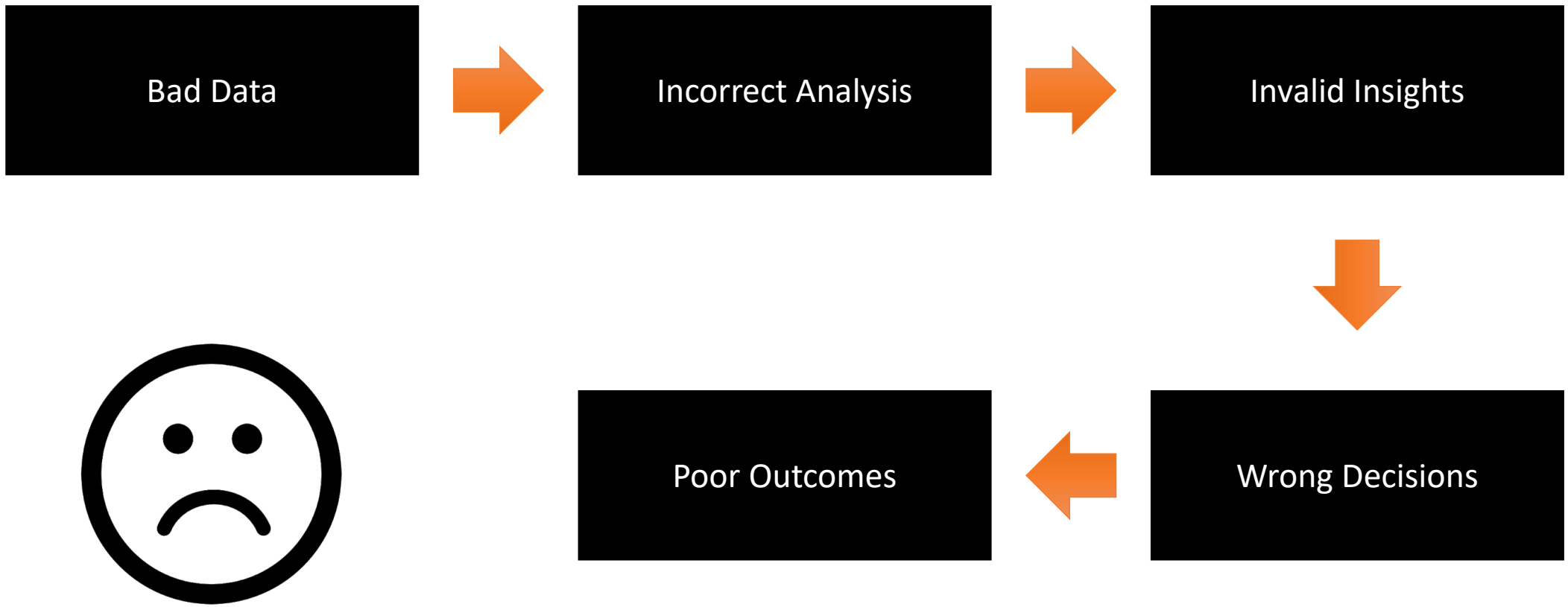
- Proprietary data sources
- Government data sets
- Academic data sets
- Web search
- Sensor data
- Crowdsourcing
- By researcher (Creating own datasets)

What are the formats?



Data Wrangling

Why this is important?



Data Wrangling

Data Wrangling (Data Munging) is the process of converting “raw” data into data that can be explored and analyzed to generate valid actionable insights.

Data Cleaning

Data Manipulation

Data Wrangling

Common problems with data

- Missing values
- Outliers
- Duplicates
- Untidy data

Data Wrangling

Missing values

Name	Age	Height (cm)	Weight (kg)
Jane	23	167	50
David	24	168	70
Scott	21	170	
Harry		182	50
Anne	20	153	38

Data Wrangling

Dealing with missing values

- Removing missing value columns & rows
- Filling missing values

Data Wrangling

Duplicate values

Name	Age	Height (cm)	Weight (kg)
Jane	23	167	50
David	24	168	70
Scott	21	170	68
Harry	22	182	50
Scott	21	170	68

Data Wrangling

Dealing with duplicate values

- Removing duplicate rows

Data Wrangling

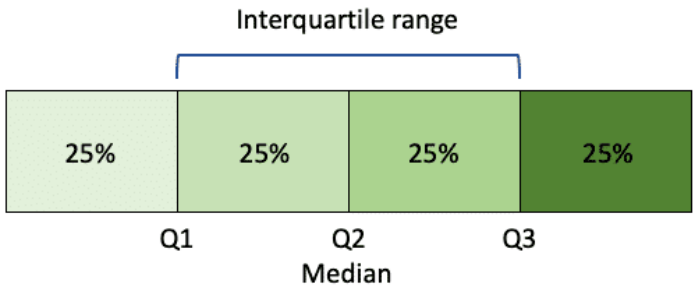
Outliers values

Name	Age	Height (cm)	Weight (kg)
Jane	23	167	50
David	24	168	150
Scott	21	170	68
Harry	22	182	50
Anne	20	153	38

Data Wrangling

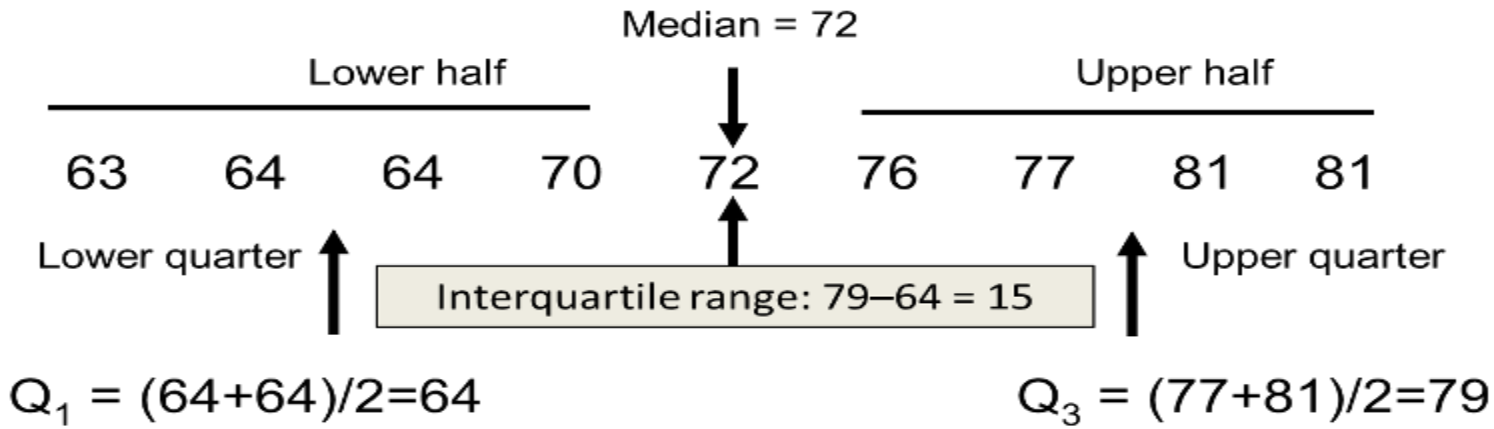
Dealing with outliers

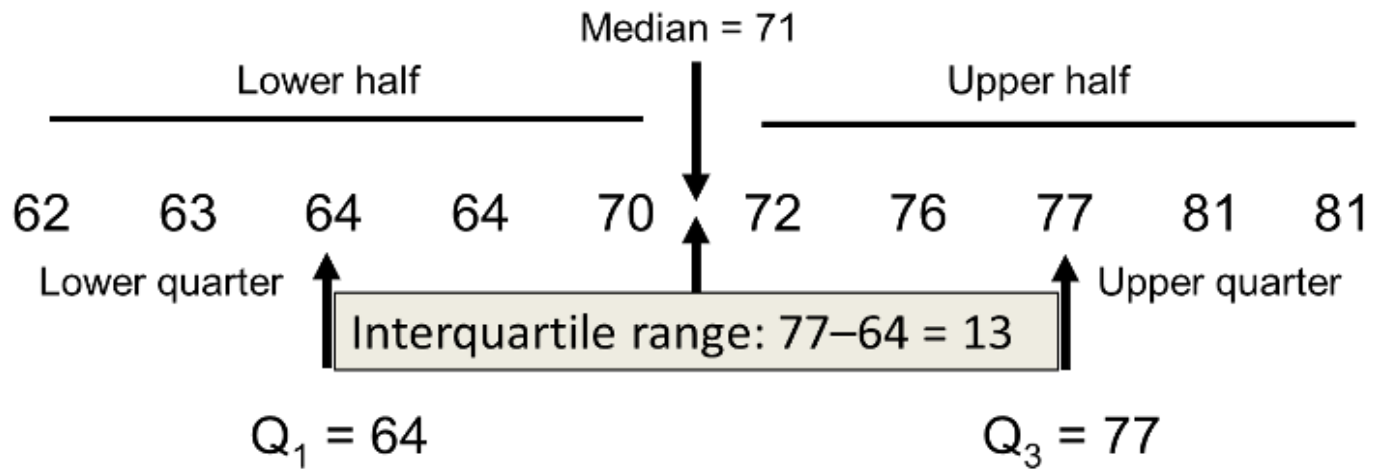
- Detect the outliers
- Remove the outliers
  - Calculate quartiles
  - Calculate the Inter Quartile Range (IQR)
  - Remove the values outside 1.5 times IQR



Data Wrangling

IQR





Data Wrangling

Untidy data

Uni	2015	2016	2017
UOC	8902	9221	9021
UOK	6789	7834	7634
UOM	5600	5467	6234

Uni	Year	Intake
UOC	2015	8902
UOC	2016	9221
UOC	2017	9021
UOK	2015	6789
UOK	2016	7834
UOK	2017	7634
UOM	2015	5600
UOM	2016	5467
UOM	2017	6234

Data Wrangling

Dealing with untidy data

- Data manipulation

Pandas library

pandas is a fast, powerful, flexible and easy to use open-source data analysis and manipulation tool, built on top of the Python programming language.