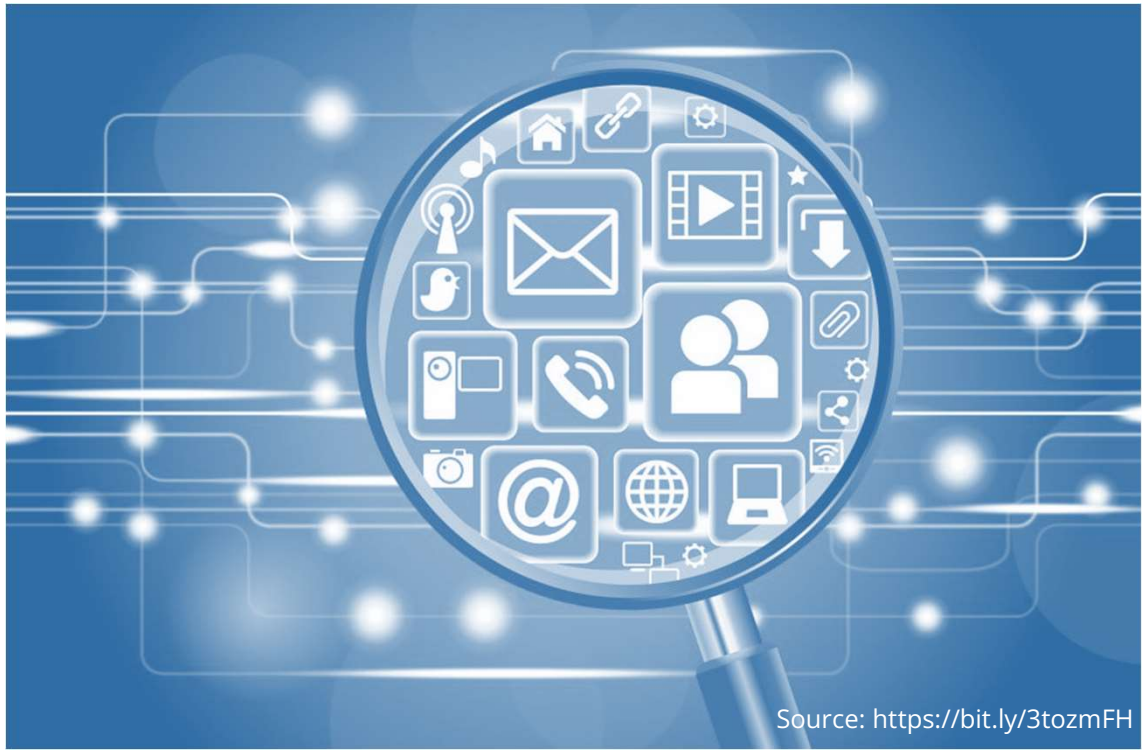


Chapter 4 Data exploration

CSS 341 Introduction to Data Science

Chukiat Worasucheep

Important Notice

การเรียนการสอนหัวข้อนี้ ผ่านทางสื่อออนไลน์ (Online meeting)

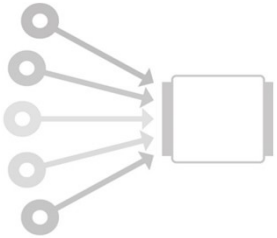
และมีการบันทึกภาพและเสียงเพื่อประโยชน์ทางการศึกษาต่อไปในอนาคต.

หากท่านไม่ยินยอมให้มีการเผยแพร่การบันทึกดังกล่าว ขอให้แจ้งให้ผู้สอนทราบภายใน 36 ชั่วโมง.

Learning objectives

- เข้าใจความสำคัญและ data exploration
- สามารถใช้ libraries ต่าง ๆ ของ python (เช่น pandas, numpy, scipy.stats) ในการทำ Data acquisition and Exploratory Data Analysis
  - คำสั่ง read_csv()
  - Basic row/col filtering
  - คำสั่งทางสถิติพื้นฐาน เช่น min, max, mean, standard deviation (s.d.)
  - Skewness, kurtosis
  - Correlation, histogram
  - Groupby()
  - Numpy
  - Scipy.stats

Data science process



Data acquisition 02

Data Preparation 04

06 Data modeling

07 Deployment and maintenance

05 Data visualization

03 Data Exploration

01 Understand business and problems

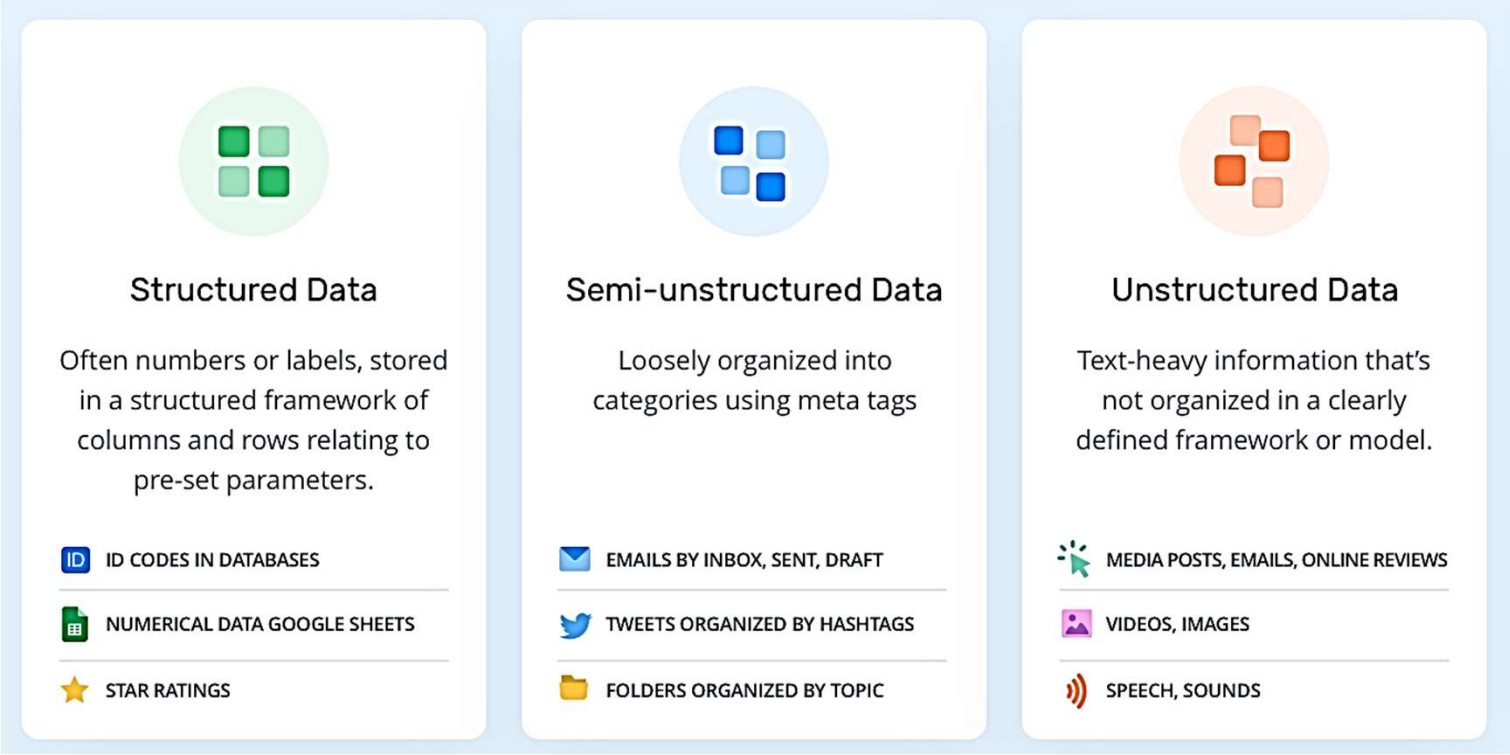
Exploration Data Analysis (EDA)

(การวิเคราะห์ข้อมูลเชิงสำรวจ)

EDA is to analyze data to summarize their main characteristics, sometimes with visual methods, to have a better understanding.

Types of data based on structure

- ข้อมูลแบบมีโครงสร้าง (Structured data) เช่น relational tables.
- ข้อมูลแบบไม่มีโครงสร้าง (Unstructured data) เช่น images, videos, sounds, social media posts.
- ข้อมูลแบบกึ่งโครงสร้าง (Semi-structured data) เช่น XML, JSON, NoSQL.



Structured data

- Traditional data managed in computers, e.g. spreadsheets, relational databases.
- Efficient to manage with plenty of tools, now.
- Require less storage.

customer_id	age	gender	region	income	married	children	car	response
ID12101	48	FEMALE	INNER_CITY	17546.0	NO	1	NO	NO
ID12102	40	MALE	TOWN	30085.1	YES	3	YES	NO
ID12103	51	FEMALE	INNER_CITY	16575.4	YES	0	YES	YES
ID12104	23	FEMALE	TOWN	20375.4	YES	3	NO	NO

Unstructured data

- Can be anything with various formats, e.g. images, videos, sounds.
- Take more storage space and more difficult to manage than structured data.

Text files and documents

Servers or website logs

Sensor data

Image files

Video files

Audio files

Emails

Social media data

Semi-structured data

- Data that are not relational database or tables, but still has some structure to it.
- Often organized using some *meta-tags*.
- Consists of documents held in JavaScript Object Notation (JSON) or XML formats.
- Includes key-value stores and graph databases (NoSQL).

```
{
  "image": "Cheryl-Carter.jpg",
  "firstname": "Cheryl",
  "lastname": "Carter",
  "company": "Skyble",
  "email": "ccarter@gmail.com",
  "phone": "2-(017)772-7449",
},
{ ···
```

JSON format

Cheryl Carter
Skyble
Email: ccarter@gmail.com
Phone: 2-(017)772-7449

```
<?xml version="1.0" encoding="UTF-8"?>
<employee>
    <fname>Krishna</fname>
    <lname>Rungta</lname>
    <home>London</home>
    <expertise name="SQL"/>
    <expertise name="Python"/>
    <expertise name="Testing"/>
     <expertise name="Business"/>
</employee>
```

XML data

Key	Value
Name	John
Age	34
City	Bangkok

Key-value data

Major tasks of EDA

1. Understand attribute of data
   - การเลือกข้อมูล และเตรียมให้พร้อมวิเคราะห์ เช่น การนำข้อมูลแต่ละชุด มาทำเป็น คอลัมน์ เพื่อให้เห็นถึงความแตกต่างของข้อมูล แยกออกเป็น attributes หรือ คุณลักษณะต่าง ๆ เช่น เพศ สี อายุ เป็นต้น
2. Univariate analysis การวิเคราะห์ข้อมูลตัวแปรเดียว
   - เป็นการวิเคราะห์เชิงสถิติ ที่ทำให้เห็นพฤติกรรมของ แต่ละ attribute เช่น มีค่าเฉลี่ยเท่าไหร่ มีผลรวมเท่าไหร่ มีความแปรปวนเท่าไหร่ ค่าโดยรวมคืออะไร เป็นต้น
3. Bi-/Multivariate analysis การวิเคราะห์มากกว่า 1 ตัวแปร
   - เพื่อให้เห็นถึงความสัมพันธ์ขั้นต้น เช่น การหา correlation และการเขียนกราฟ scatter plot แสดงความสัมพันธ์

Types of data based on measurement

	Scale	True Zero	Equal Intervals	Order	Category	Example
นามบัญญัติ	Nominal	No	No	No	Yes	Marital Status, Sex, Gender, Ethnicity
อันดับ	Ordinal	No	No	Yes	Yes	Student Letter Grade, NFL Team Rankings
อันตรภาค	Interval	No	Yes	Yes	Yes	Temperature in Fahrenheit, SAT Scores, IQ, Year
อัตราส่วน	Ratio	Yes	Yes	Yes	Yes	Age, Height, Weight

Source: https://thebiologynotes.com/nominal-ordinal-interval-and-ratio-data/

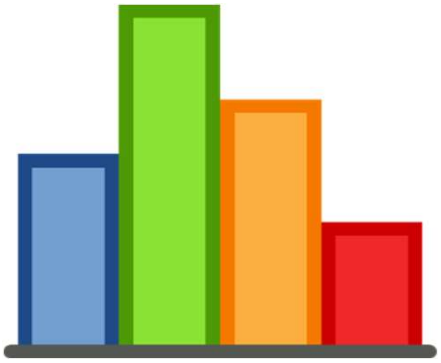
More: http://weatherwing4.6te.net/DataAnalysis%20forWeatherPatterns.pdf

Types of data to explore

- *Nominal* data (ข้อมูลแบ่งกลุ่มเท่านั้น ไม่สามารถนำมาคำนวณได้)
  - Status: single, marital, divorce
  - City: Bangkok, Lampang, Yala, Khonkan, etc.
  - Phone number: 081-1234567
- *Ordinal* data (ข้อมูลแบ่งกลุ่ม และ มีลำดับ)
  - E.g. “Excellent”, “Good”, “Average”, “Bad”, “Worst”
  - Income over expense: <= 100,000 100,001 – 500,000 500,001 – 1,000,000 1,000,001++
  - Education: Elementary, High School, Bachelor, Graduate
- *Numerical* data (ข้อมูลเชิงตัวเลข แบ่งช่วงชัดเจน)
  - Interval (ค่า 0 ไม่ได้เป็นศูนย์จริง): temperature in F, IQ
  - Ratio (ค่า 0 เป็นศูนย์จริง): age, weight

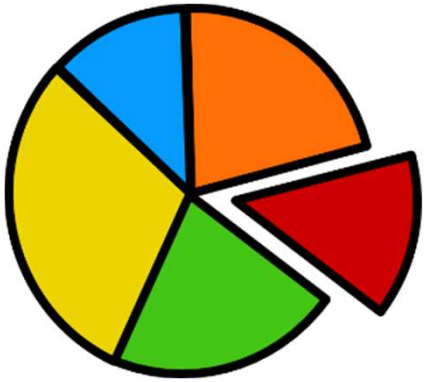
Methods to present “nominal data”

- Frequency ความถี่
  - Count the number of items of interest, fi



Bar chart

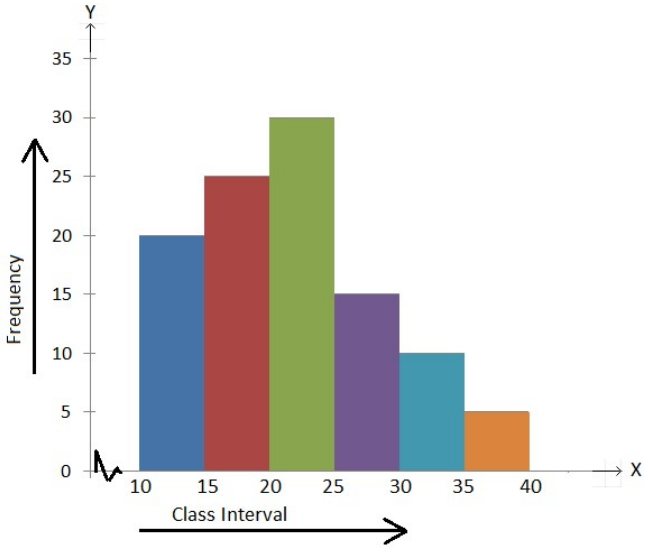
- Proportion สัดส่วน
  - Divide frequency by total number of items fi / N
- Percentage ร้อยละ
  - Proportion multiplied by 100…
  - i.e. 100 * fi / N

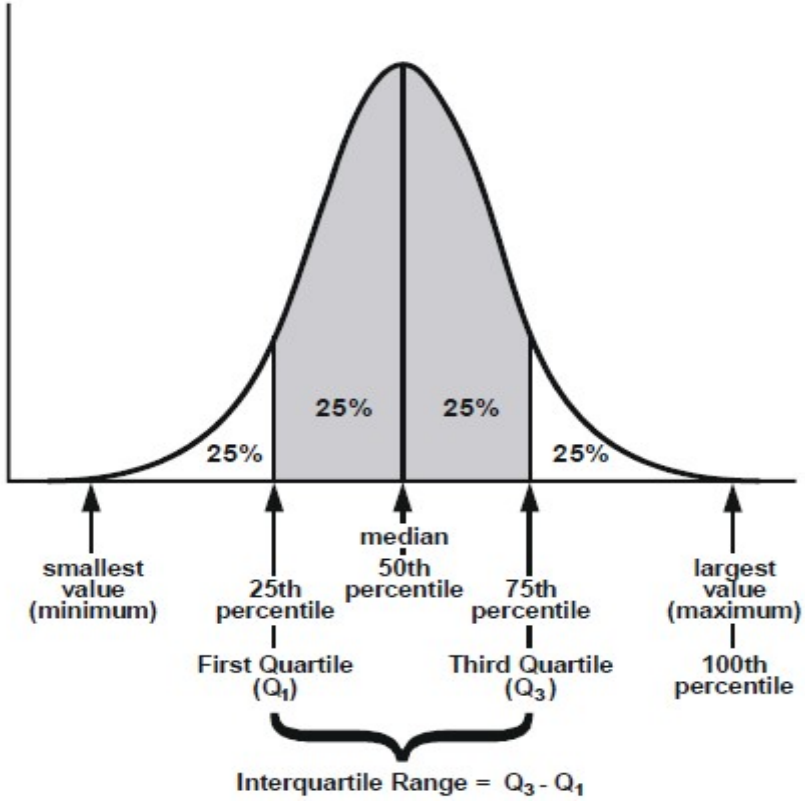


Pie chart

Methods to present “numerical data”

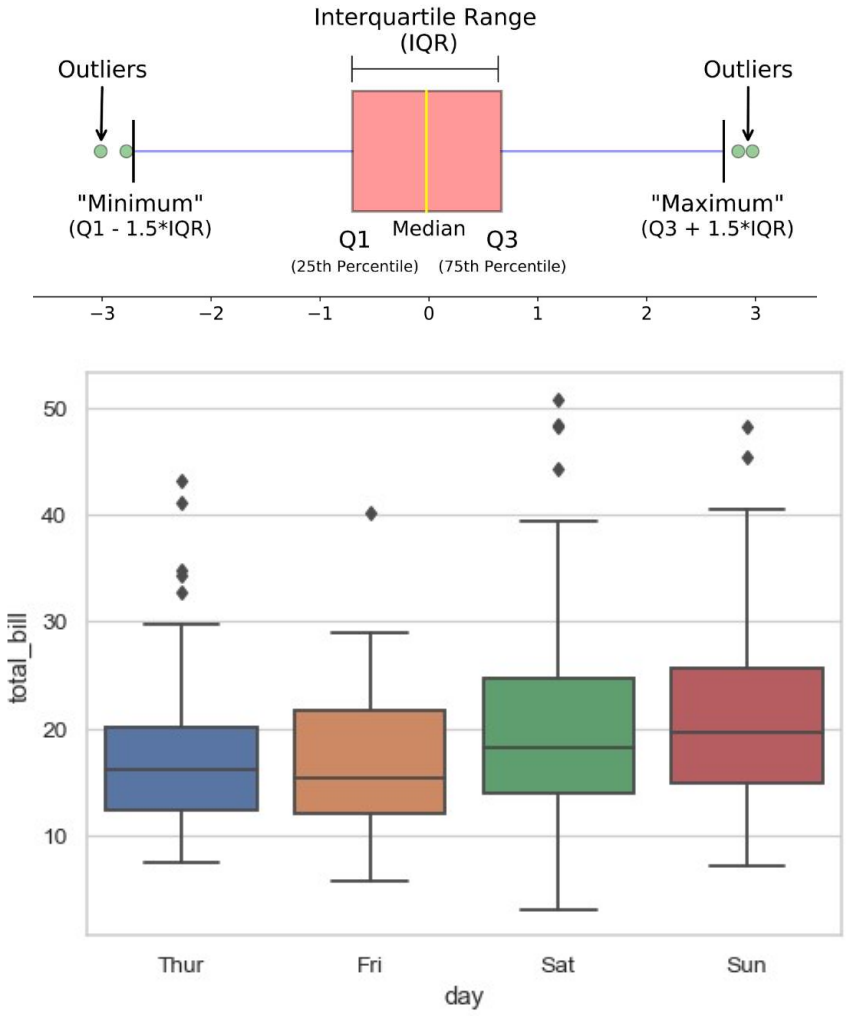
- Summarize
  - Max, min, mean, standard deviation
  - Median, mode, range
  - Histogram
  - InterQuartile Range (IQR) = Q3 - Q1
    - บอกการกระจายตัวของข้อมูลได้ดีกว่า range เพราะไม่ถูกกวนด้วย outliers



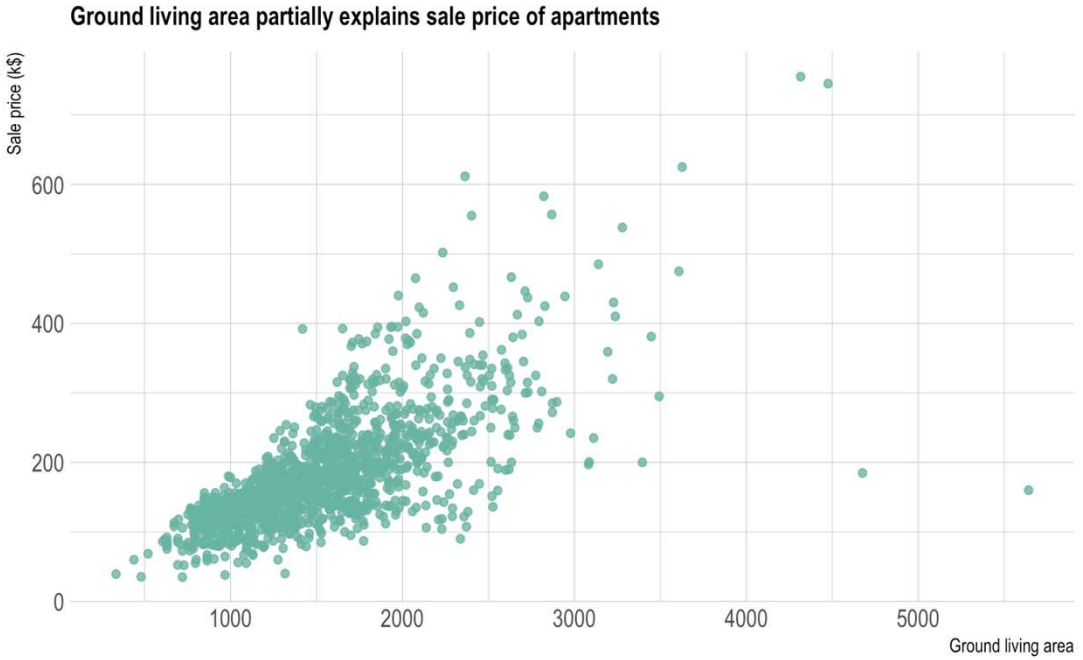


Methods to present “numerical data”

- Boxplot

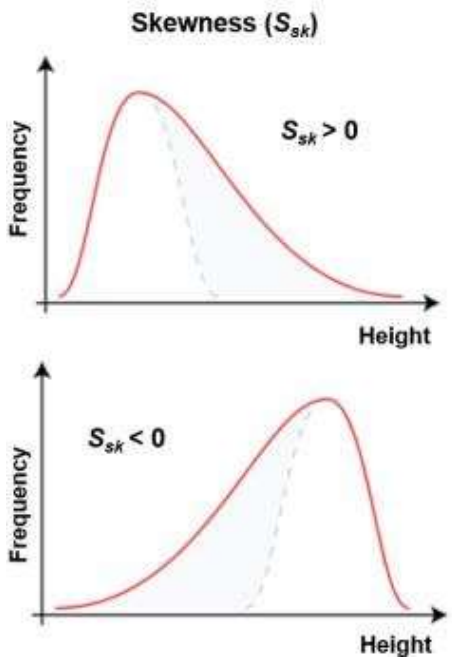


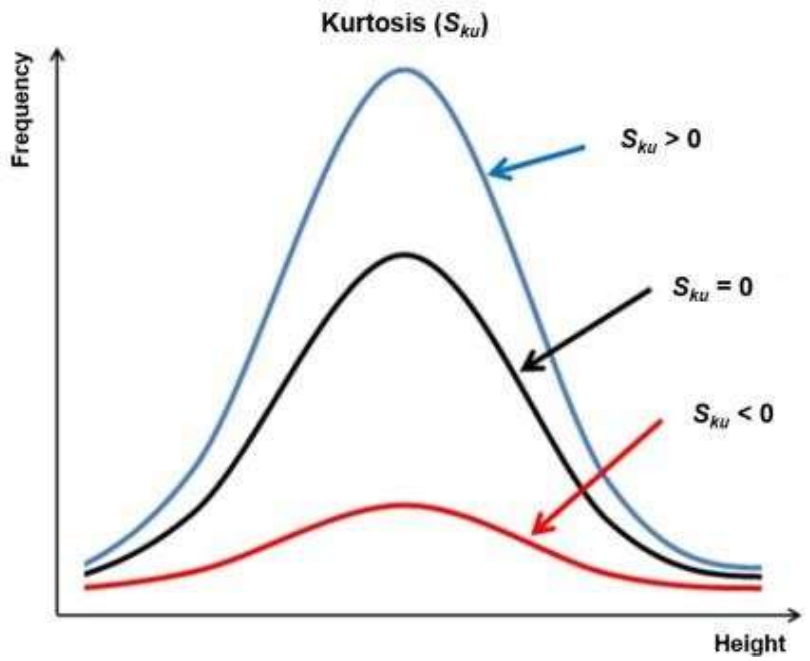
- Scatter plot



Methods to present “numerical data”

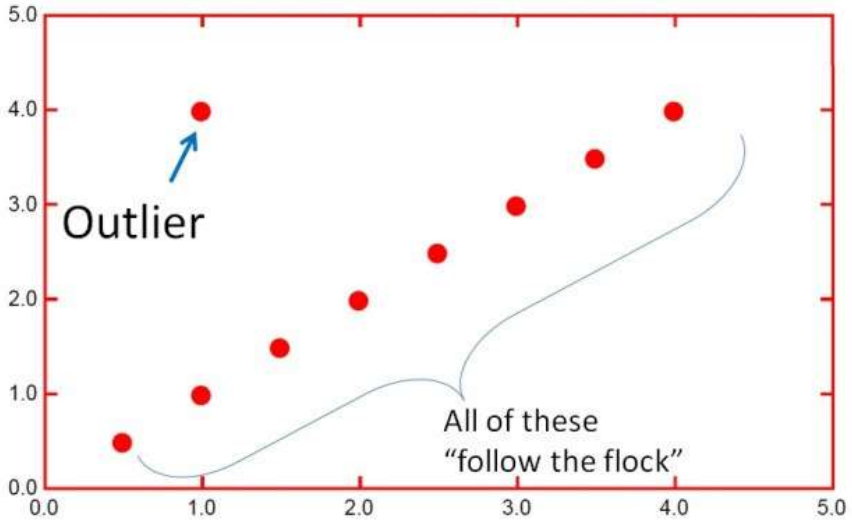
- Skewness
  - a *measure of the asymmetry* of the probability distribution of a random variable about its mean
- Kurtosis
  - a statistical measure used to describe *distribution*. Kurtosis measures extreme values in either tail. Distributions with large kurtosis exhibit tail data exceeding the tails of the normal distribution.



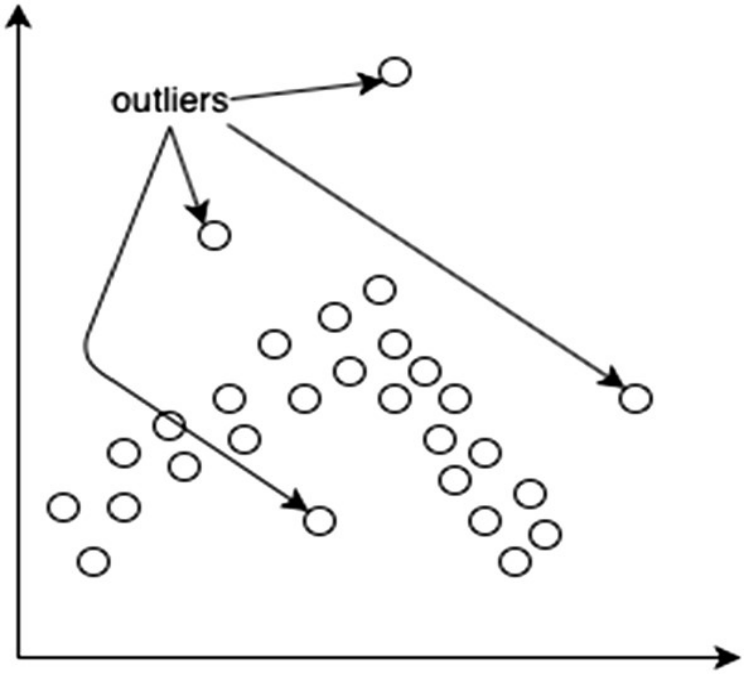


Numerical data – outliers

- In statistics, an *outlier* is a data point that differs significantly from other observations. An outlier may be due to variability in the measurement, or it may indicate experimental error.



Never mind what the axes mean...

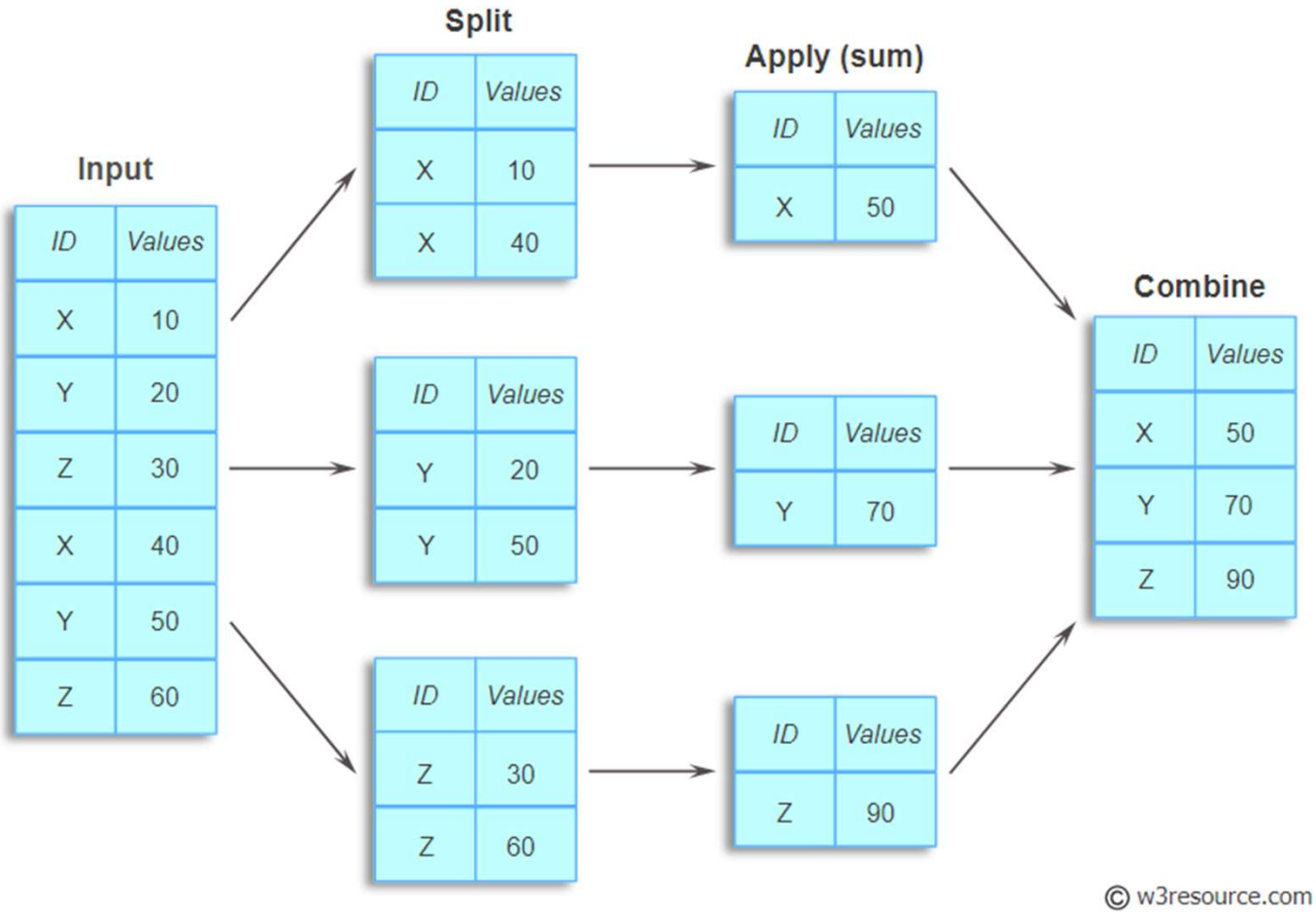


Data exploration.ipynb

- Basic data descriptive
  - df.shape()
  - df.info()
  - df.describe()
  - df.head()
  - df.tail()
  - df.columns
- Data access/filtering
  - .iloc[]
- Basic statistics
  - min, max, mean, std
  - skewness, kurtosis
- Multivariate statistics
  - Correlation, histogram
- sort() function
- apply() function
- groupby() function

Pandas's groupby function

- Split-Apply-Combine strategy



Example of groupby function – iris dataset

	species	sepal_length	sepal_width	petal_length	petal_width
0	setosa	5.1	3.5	1.4	0.2
1	setosa	4.9	3.0	1.4	0.2
2	setosa	4.7	3.2	1.3	0.2
3	setosa	4.6	3.1	1.5	0.2
4	setosa	5.0	3.6	1.4	0.2
50	versicolor	7.0	3.2	4.7	1.4
51	versicolor	6.4	3.2	4.5	1.5
52	versicolor	6.9	3.1	4.9	1.5
53	versicolor	5.5	2.3	4.0	1.3
54	versicolor	6.5	2.8	4.6	1.5
100	virginica	6.3	3.3	6.0	2.5
101	virginica	5.8	2.7	5.1	1.9
102	virginica	7.1	3.0	5.9	2.1
103	virginica	6.3	2.9	5.6	1.8
104	virginica	6.5	3.0	5.8	2.2

SUM

SUM

SUM

species	sepal_length	sepal_width	petal_length	petal_width
setosa	24.3	16.4	7.0	1.0
versicolor	32.3	14.6	22.7	7.2
virginica	32.0	14.9	28.4	10.5

Case study 1 – Adult dataset

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	age	workclass	fnlwgt	education	ucation-nu	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	urs-per-we	native-country	salary
2	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
3	50	Self-emp-not-	83311	Bachelors	13	Married-civ-spo	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
4	38	Private	215646	HS-grad	9	Divorced	Handlers-clean	Not-in-family	White	Male	0	0	40	United-States	<=50K
5	53	Private	234721	11th	7	Married-civ-spo	Handlers-clean	Husband	Black	Male	0	0	40	United-States	<=50K
6	28	Private	338409	Bachelors	13	Married-civ-spo	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
7	37	Private	284582	Masters	14	Married-civ-spo	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
8	49	Private	160187	9th	5	Married-spouse	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
9	52	Self-emp-not-	209642	HS-grad	9	Married-civ-spo	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
10	31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
11	42	Private	159449	Bachelors	13	Married-civ-spo	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
12	37	Private	280464	Some-college	10	Married-civ-spo	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
13	30	State-gov	141297	Bachelors	13	Married-civ-spo	Prof-specialty	Husband	Asian-Pac-	Male	0	0	40	India	>50K
14	23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
15	32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
16	40	Private	121772	Assoc-voc	11	Married-civ-spo	Craft-repair	Husband	Asian-Pac-	Male	0	0	40	?	>50K
17	34	Private	245487	7th-8th	4	Married-civ-spo	Transport-mov	Husband	Amer-Indi	Male	0	0	45	Mexico	<=50K
18	25	Self-emp-not-	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
19	32	Private	186824	HS-grad	9	Never-married	Machine-op-in	Unmarried	White	Male	0	0	40	United-States	<=50K
20	38	Private	28887	11th	7	Married-civ-spo	Sales	Husband	White	Male	0	0	50	United-States	<=50K

Case study 1 – Adult dataset

Assignment #1

Exploratory data analysis with Pandas

*In this task you should use Pandas to answer a few questions about the Adult dataset. *

0. Load adult.csv dataset into this notebook.

```
In [1]: # You code here
```

1. How many men and women (*sex* feature) are represented in this dataset?

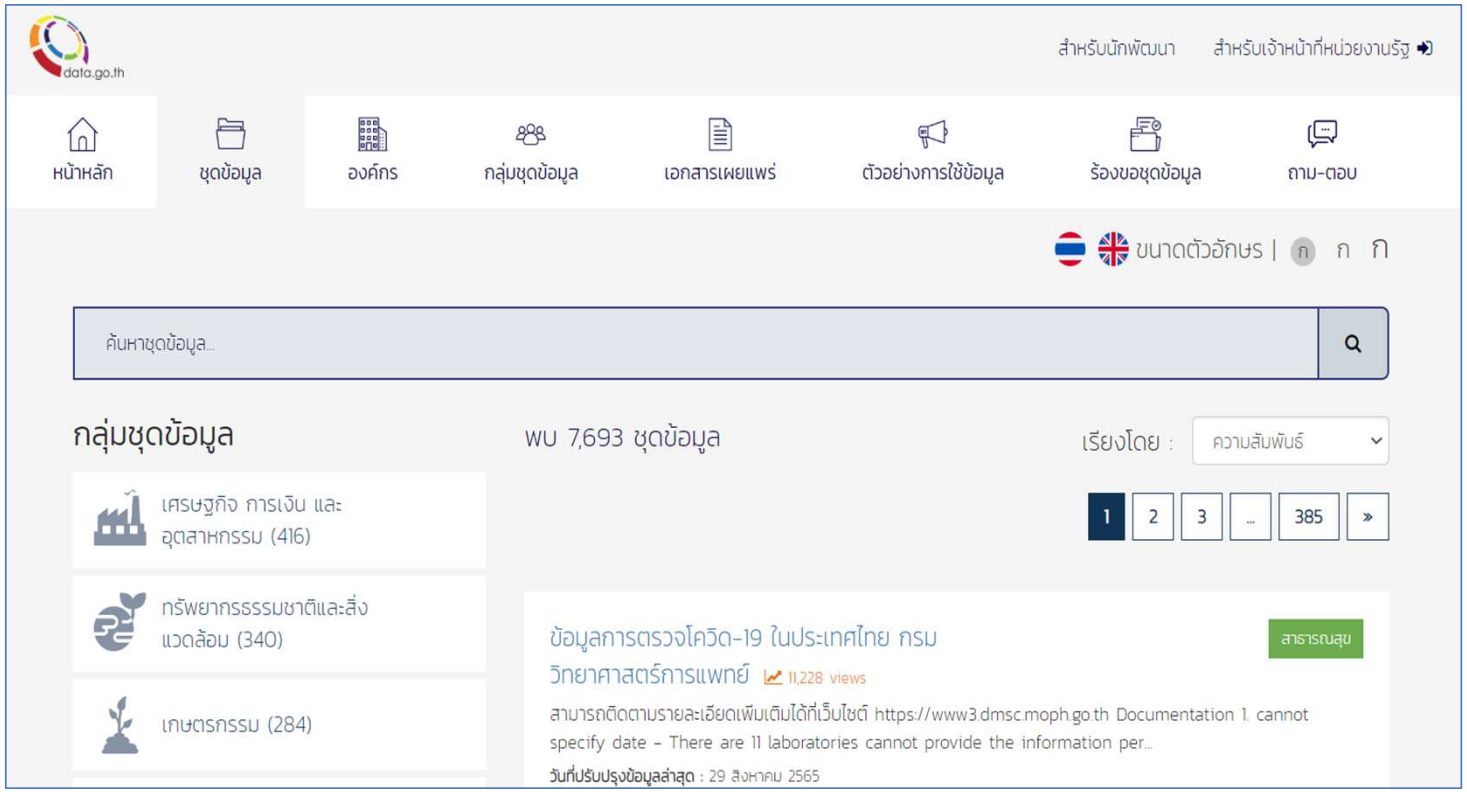
```
In [ ]: # You code here
```

2. What is the average age (*age* feature) of women?

```
In [ ]: # You code here
```

Case study 2 – Thai government data

- https://data.go.th/dataset



- ให้ไปสำรวจข้อมูลที่น่าสนใจ นำมาวิเคราะห์ต่อขั้นต่อไป
- ในขั้นแรกนี้ ให้นำเข้ามาแล้วสำรวจเบื้องต้น

More about groupby funciton

- https://www.analyticsvidhya.com/blog/2020/03/groupby-pandas-aggregating-data-python/
- https://pandas.pydata.org/pandas-docs/version/0.23/generated/pandas.core.groupby.DataFrameGroupBy.agg.html
- https://www.w3resource.com/python-exercises/pandas/groupby/index.php