Bachelier de spécialisation en Business Data Analysis

Data mining

Cédric Guilmin

héna llux
HAUTE ÉCOLE DE NAMUR-LIÈGE-LUXEMBOURG

esa
ÉCOLE SUPÉRIEURE DES AFFAIRES

Overview

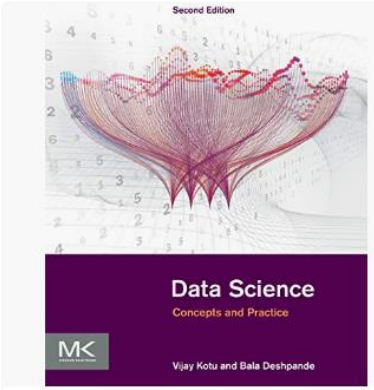
- Introduction
- Tabular data / Data types / data storage
- Descriptive statistics
- Excel
- Python: basics + pandas
- RapidMiner: discovery
- Advanced techniques
  - Unsupervised techniques: clustering, principal components analysis, …
  - Supervised techniques: regression (linear, logistic, …), decision tree, random forest, …
  - Text mining: text classification, topic modelling, regular expressions, …

Introduction

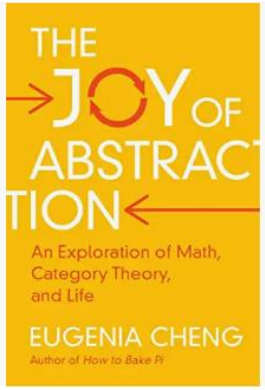
Data is everywhere …

RÉSULTATS

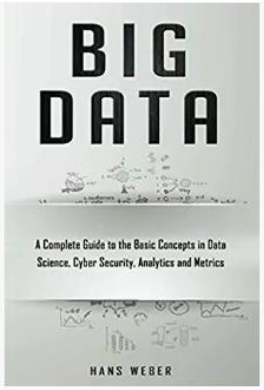
En apprendre plus sur ces résultats.



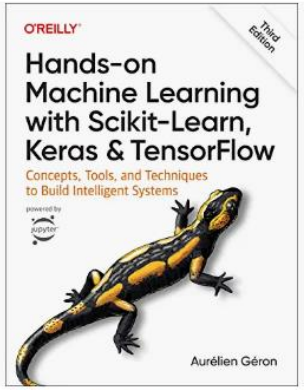
Data Science: Concepts and Practice
Édition en Anglais
de Vijay Kotu
★★★★½ 8
Broché
73,76€
Recevez-le **jeudi 15 septembre**
Livraison à 0,01€ par Amazon
Autres vendeurs sur Amazon
63,03 € (10 offres de produits d'occasion et neufs)



Sponsorisé
The Joy of Abstraction: An Exploration of Math, Category Theory, and Life
Édition en Anglais
de Eugenia Cheng
Relié
27,43€
Livraison à 0,01€ par Amazon
Cet article paraîtra le 13 octobre 2022.



Big Data: A Complete Guide to the Basic Concepts in Data Science, Cyber Security, Analytics and Metrics
Édition en Anglais
de Hans Weber
★★★½☆ 17
Broché
14,72€ PVC: ~~15,77€~~
Recevez-le **jeudi 15 septembre**
Livraison à 0,01€ par Amazon
Autre format: Format Kindle



Hands-on Machine Learning With Scikit-learn, Keras, and Tensorflow: Concepts, Tools, and Techniques to Build...
Édition en Anglais
de Aurélien Géron
Broché
84,42€
Livraison à 0,01€ par Amazon
Cet article paraîtra le 31 octobre 2022.

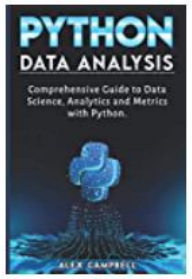
Data is everywhere …

Produits liés à cet article

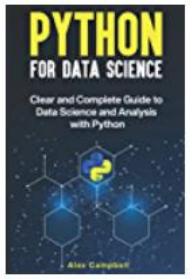
Sponsorisé

Data Visualization Guide: Clear Guide to Data Science and Visualization
Alex Campbell
8
Broché
13,66 € prime

Data Science for Beginners: Comprehensive Guide to Most Important Basics…
Alex Campbell
19
Broché
14,72 € prime



Python Data Analysis: Comprehensive Guide to Data Science, Analytics and Metrics wi...
Alex Campbell
12
Broché
15,77 € prime



Python for Data Science: Clear and Complete Guide to Data Science and Analysis with...
Alex Campbell
6
Broché
15,77 € prime

Python for Data Science: Comprehensive Guide to Data Science with Python
Alex Campbell
10
Broché
15,77 € prime

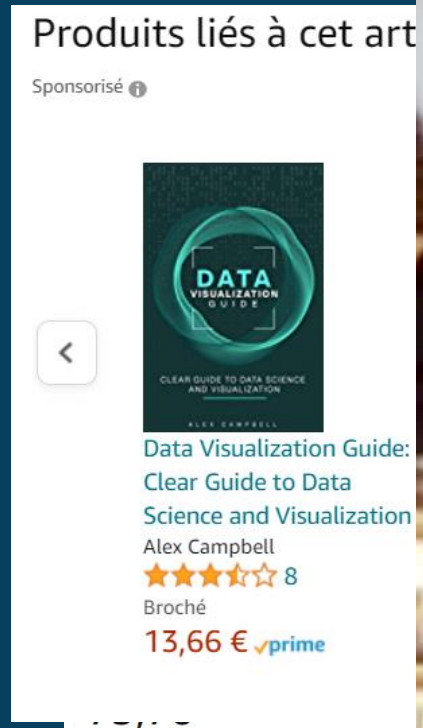
Recevez-le **jeudi 15 septembre**
Livraison à 0,01€ par Amazon
Autres vendeurs sur Amazon
63,03 € (10 offres de produits d'occasion et neufs)

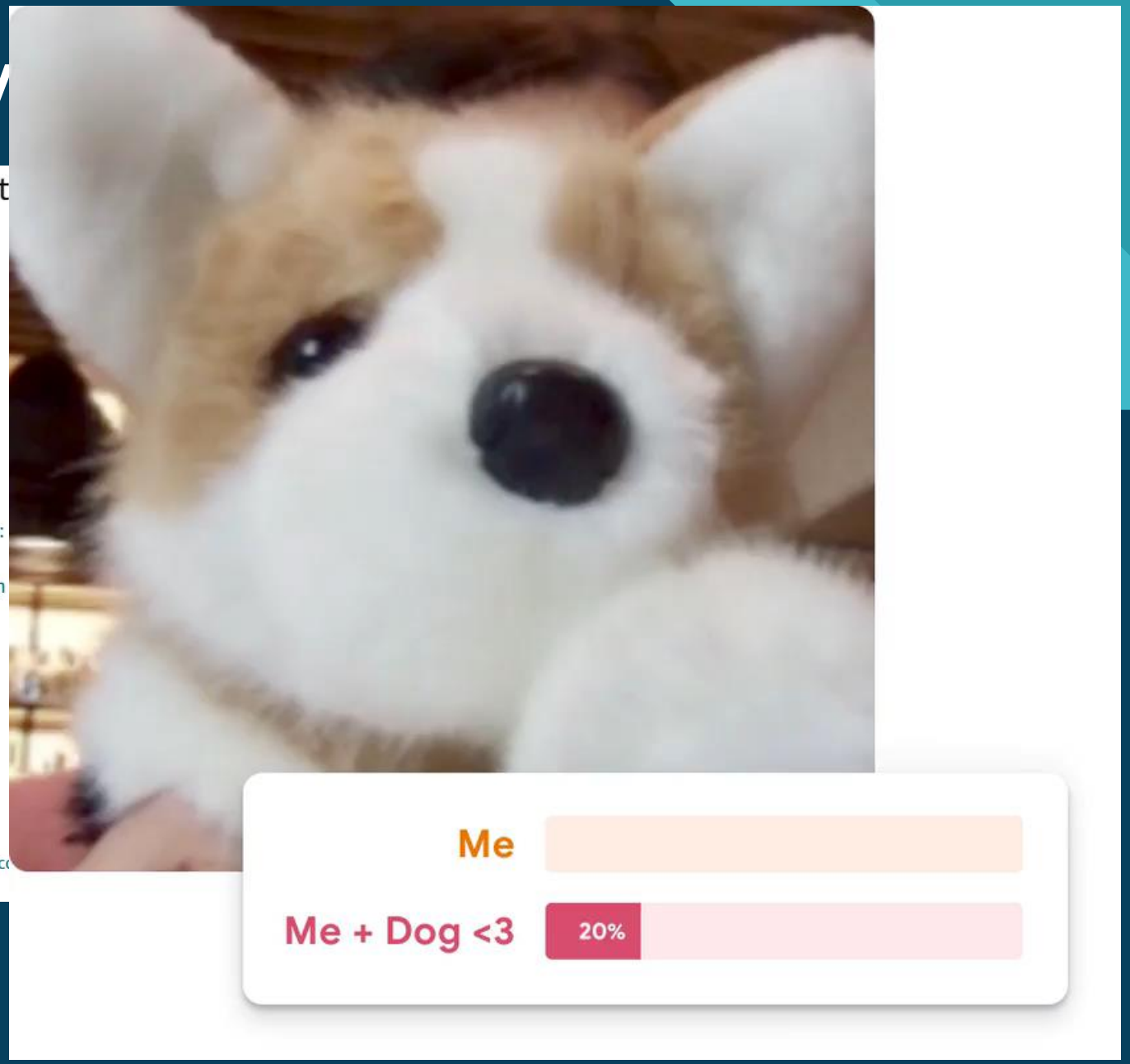
27,43€
Livraison à 0,01€ par Amazon
Cet article paraîtra le 13 octobre 2022.

Broché
14,72€ PVC: ~~15,77€~~
Recevez-le **jeudi 15 septembre**
Livraison à 0,01€ par Amazon
Autre format: Format Kindle

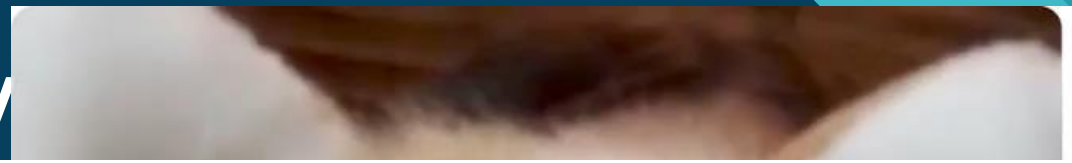
84,42€
Livraison à 0,01€ par Amazon
Cet article paraîtra le 31 octobre 2022.

Data is every





Data is every



Pro

Spons

Published: 25 January 2017

Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun

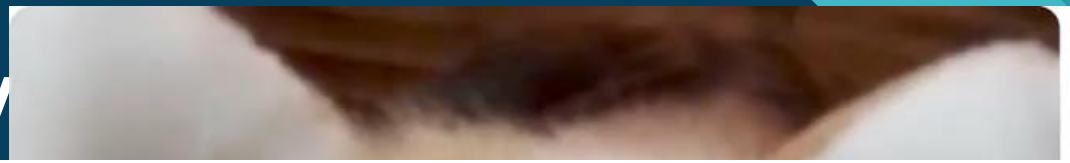
Nature **542**, 115–118 (2017) | Cite this article

A Corrigendum to this article was published on 29 June 2017

Abstract

Skin cancer, the most common human malignancy[1,2,3], is primarily diagnosed visually

Data is every



Published: 25 January 2017

Derma **th**
deep n

Andre Esteva &
Sebastian Th

Nature **542**,

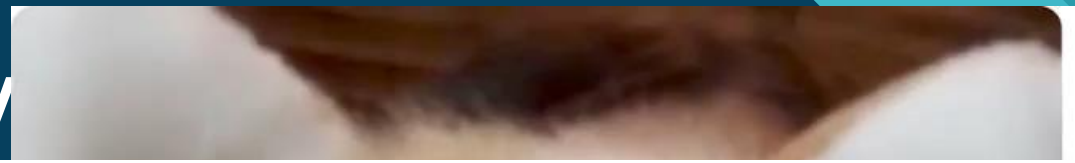
195k Acces

A Co

Abstract

Skin cancer, the most common human malignancy[1,2,3] is primarily diagnosed visually

Data is every



Published: 25 January 2017

US & WORLD / TECH / ARTIFICIAL INTELLIGENCE

French government uses AI to spot undeclared swimming pools – and tax them

Photo by Yang Bo/China News Service via Getty Images

/ The government used machine learning to scan aerial photos of properties

By JAMES VINCENT

Aug 30, 2022, 12:10 PM GMT+2 | 0 Comments

Data is everywhere …

Source: https://hubmeta.com/uncategorized/data-is-the-new-oil/

Data is everywhere …

Source: https://hubmeta.com/uncategorized/data-is-the-new-oil/

Data is everywhere … so the jobs …

Data is everywhere … so the jobs …

Glassdoor 50 Best Jobs In America, 2018					
Ranking	Job	Median Base Salary	Job Score (5.0 scale)	Job Satisfaction (5.0 scale)	Job Openings
1	Data Scientist	$ 110,000	4.8	4.2	4,524
2	DevOps Engineer	$ 105,000	4.6	4	3,369
3	Marketing Manager	$ 85,000	4.6	4	6,436
4	Occupational Therapist	$ 74,000	4.5	4	11,903
5	HR Manager	$ 85,000	4.5	3.9	4,458
6	Electrical Engineer	$ 76,000	4.5	3.9	5,839
7	Strategy Manager	$ 135,000	4.5	4.2	1,195
8	Mobile Developer	$ 90,000	4.5	4.1	1,809
9	Product Manager	$ 113,000	4.4	3.7	7,531
10	Manufacturing Engineer	$ 72,000	4.4	4	4,241
11	Compliance Manager	$ 96,000	4.4	4.3	1,222
12	Finance Manager	$ 116,000	4.4	3.8	2,998
13	Risk Manager	$ 97,000	4.4	4.2	1,209
14	Business Development Manager	$ 75,000	4.4	3.9	4,060
15	Front End Engineer	$ 100,000	4.4	4.2	1,222
16	Site Reliability Engineer	$ 120,000	4.4	4.1	1,064
17	Mechanical Engineer	$ 75,000	4.4	3.8	5,079
18	Analytics Manager	$ 115,000	4.4	3.9	1,381
19	Tax Manager	$ 110,000	4.4	3.7	3,309
20	Creative Manager	$ 110,000	4.3	4.3	824
21	Software Engineer	$ 102,500	4.3	3.6	29,187
22	Hardware Engineer	$ 115,000	4.3	4.2	806
23	Corporate Recruiter	$ 65,000	4.3	4.3	2,330
24	QA Manager	$ 92,000	4.3	3.8	1,741
25	Physician Assistant	$ 104,000	4.3	3.6	5,517
26	Database Administrator	$ 94,000	4.3	3.8	2,370
27	UX Designer	$ 90,000	4.3	3.8	1,963
28	Nursing Manager	$ 84,660	4.3	3.7	4,209
29	Engagement Manager	$ 115,000	4.3	3.7	2,169
30	Solutions Architect	$ 125,000	4.2	3.6	3,325
31	Process Engineer	$ 78,000	4.2	3.8	3,033
32	Reliability Engineer	$ 92,000	4.2	4.3	747
33	Data Engineer	$ 100,000	4.2	3.7	2,816
34	Operations Manager	$ 65,000	4.2	3.8	13,706
35	Speech Language Pathologist	$ 72,000	4.2	3.7	11,573
36	Communications Manager	$ 80,000	4.2	3.9	1,380
37	Audit Manager	$ 100,000	4.2	3.7	1,951
38	Data Analyst	$ 60,000	4.2	3.9	4,729
39	Systems Analyst	$ 75,000	4.2	3.7	2,710
40	Facilities Manager	$ 75,000	4.2	3.8	2,139
41	Strategic Account Manager	$ 85,000	4.2	4.1	808
42	Business Intelligence Developer	$ 86,000	4.1	3.9	882
43	Business Analyst	$ 71,000	4.1	3.6	9,603
44	Accounting Manager	$ 82,000	4.1	3.6	3,273
45	UI Developer	$ 95,000	4.1	3.8	1,004
46	Executive Assistant	$ 55,000	4.1	3.9	4,684
47	Management Consultant	$ 110,000	4.1	3.8	1,024
48	Project Manager	$ 80,000	4.1	3.5	23,274
49	Nurse Practitioner	$ 100,000	4.1	3.5	8,510

Data is everywhere … so the jobs …

BeNeLux (Belgium, Netherlands, Luxembourg):

The most in-demand jobs:

1. Software Engineer
2. Project Manager
3. Account Manager
4. Mechanic/Service Person
5. Business Analyst
6. Salesperson
7. DevOps Engineer
8. Cloud Engineer
9. Product Owner
10. Full Stack Engineer

Jobs that have seen the fastest growing demand:

1. Sales Agent
2. Mechanic/Service Person
3. Cloud Engineer
4. Product Owner
5. Data Engineer

Glassdoor 50 Best Jobs In America, 2018

Ranking	Job	Median Base Salary	Job Score (5.0 scale)	Job Satisfaction (5.0 scale)	Job Openings
1	Data Scientist	$ 110,000	4.8	4.2	4,524
2	DevOps Engineer	$ 105,000	4.6	4	3,369
3	Marketing Manager	$ 85,000	4.6	4	6,436
4	Occupational Therapist	$ 74,000	4.5	4	11,903
5	HR Manager	$ 85,000	4.5	3.9	4,458
6	Electrical Engineer	$ 76,000	4.5	3.9	5,839
7	Strategy Manager	$ 135,000	4.5	4.2	1,195
8	Mobile Developer	$ 90,000	4.5	4.1	1,809
9	Product Manager	$ 113,000	4.4	3.7	7,531
10	Manufacturing Engineer	$ 72,000	4.4	4	4,241
			4.4	4.3	1,222
			4.4	3.8	2,998
			4.4	4.2	1,209
			4.4	3.9	4,060
			4.4	4.2	1,222
			4.4	4.1	1,064
			4.4	3.8	5,079
			4.4	3.9	1,381
			4.4	3.7	3,309
			4.3	4.3	824
			4.3	3.6	29,187
			4.3	4.2	806
			4.3	4.3	2,330
			4.3	3.8	1,741
			4.3	3.6	5,517
			4.3	3.8	2,370
			4.3	3.8	1,963
			4.3	3.7	4,209
			4.3	3.7	2,169
			4.2	3.6	3,325
			4.2	3.8	3,033
			4.2	4.3	747
			4.2	3.7	2,816
			4.2	3.8	13,706
			4.2	3.7	11,573
			4.2	3.9	1,380
			4.2	3.7	1,951
			4.2	3.9	4,729
			4.2	3.7	2,710
			4.2	3.8	2,139
			4.2	4.1	808
			4.1	3.9	882
43	Business Analyst	$ 71,000	4.1	3.6	9,603
44	Accounting Manager	$ 82,000	4.1	3.6	3,273
45	UI Developer	$ 95,000	4.1	3.8	1,004
46	Executive Assistant	$ 55,000	4.1	3.9	4,684
47	Management Consultant	$ 110,000	4.1	3.8	1,024
48	Project Manager	$ 80,000	4.1	3.5	23,274
49	Nurse Practitioner	$ 100,000	4.1	3.5	8,510

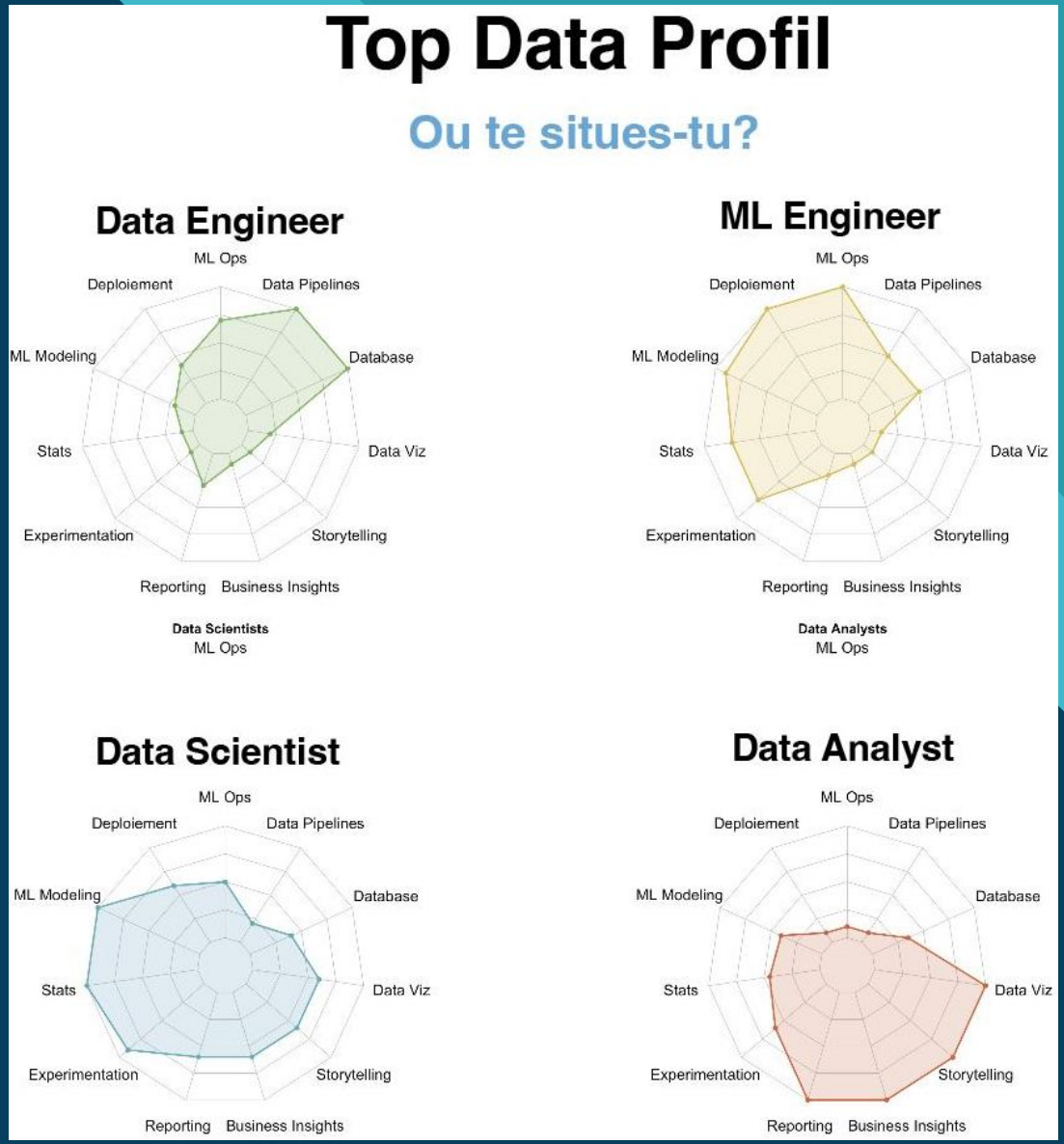
Data is everywhere … so the jobs (for 2022) …

	Job Title	Median Base Salary	Job Satisfaction
#1	Enterprise Architect	$144,997	4.1/5
#2	Full Stack Engineer	$101,794	4.3/5
#3	Data Scientist	$120,000	4.1/5
#4	Devops Engineer	$120,095	4.2/5
#5	Strategy Manager	$140,000	4.2/5
#6	Machine Learning Engineer	$130,489	4.3/5
#7	Data Engineer	$113,960	4.0/5
#8	Software Engineer	$116,638	3.9/5
#9	Java Developer	$107,099	4.1/5
#10	Product Manager	$125,317	4.0/5
#11	Back End Engineer	$112,384	4.2/5
#12	Cloud Engineer	$118,999	4.0/5
#13	HR Manager	$91,502	4.3/5
#14	Business Development Manager	$89,496	4.2/5
#15	Information Security Engineer	$116,919	4.1/5
#16	Physician	$155,400	3.9/5
#17	Corporate Recruiter	$77,700	4.4/5
#18	Salesforce Developer	$98,972	4.2/5
#19	Marketing Manager	$90,748	4.1/5
#20	Consultant	$90,748	3.9/5
#21	Automation Engineer	$86,832	4.1/5
#22	Psychiatrist	$252,385	4.0/5
#23	Sales Manager	$79,962	4.0/5
#24	UX Designer	$97,047	4.0/5
#25	Finance Manager	$114,414	3.9/5
#26	Tax Manager	$125,639	4.0/5
#27	Attorney	$100,831	4.0/5
#28	Dentist	$157,307	3.9/5
#29	Site Reliability Engineer	$137,252	4.0/5
#30	Systems Engineer	$100,831	3.9/5
#31	Electrical Engineer	$86,545	4.0/5
#32	Scrum Master	$109,284	4.1/5
#33	Product Marketing Manager	$125,015	4.1/5
#34	Psychologist	$95,199	3.9/5
#35	Data Analyst	$74,224	4.0/5
#36	Business Analyst	$81,556	3.9/5
#37	QA Engineer	$87,626	4.1/5
#38	Front End Engineer	$81,136	4.0/5
#39	HR Business Partner	$95,431	4.0/5
#40	Project Manager	$86,000	3.8/5
#41	Compliance Officer	$80,000	3.9/5
#42	Program Manager	$81,335	3.9/5
#43	Risk Manager	$102,647	3.9/5
#44	Sales Engineer	$95,809	4.0/5
#45	Solutions Engineer	$100,915	4.1/5
#46	Product Designer	$110,858	4.1/5
#47	Database Architect	$140,000	4.0/5
#48	Realtor	$54,090	4.4/5
#49	Strategic Account Manager	$88,071	4.0/5
#50	Customer Success Manager	$73,702	4.0/5

Data is everywhere … so the jobs …

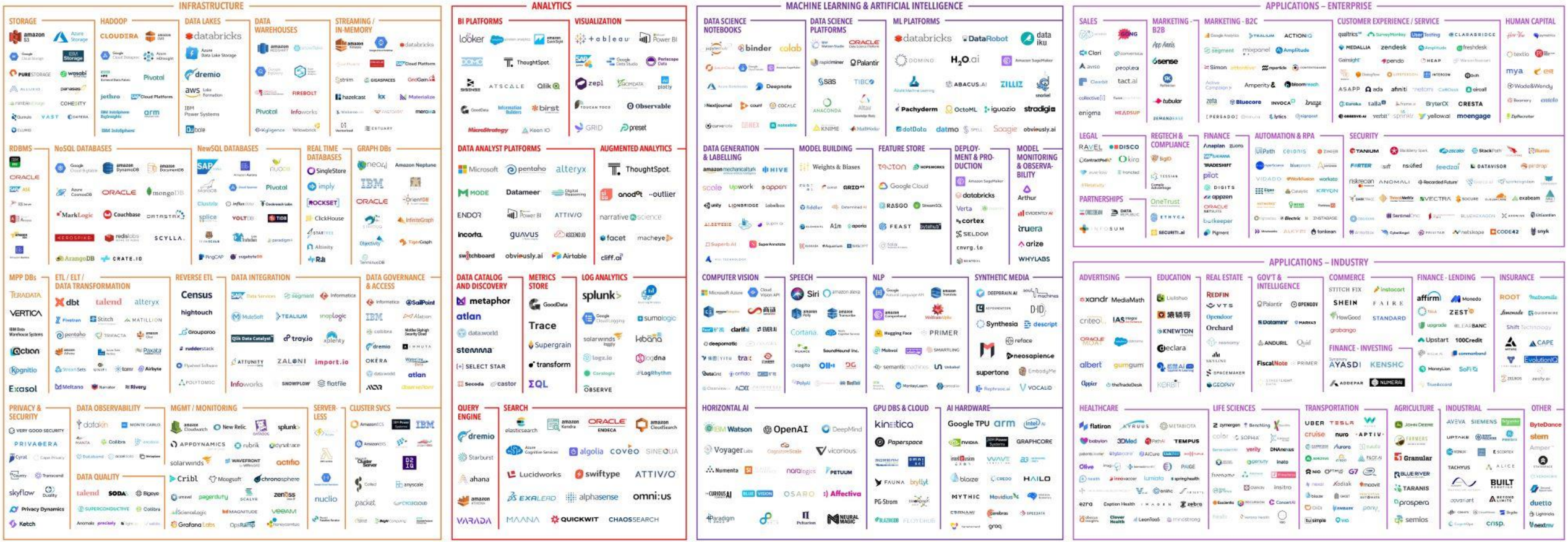
Chart creator: Kevin Rosamont Prombo – LinkedIn

You can assess your own competences and then see what is the closest profile:
https://kevros.shinyapps.io/radar_skills/



Data is everywhere … so the tools …

MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, AND DATA (MAD) LANDSCAPE 2021

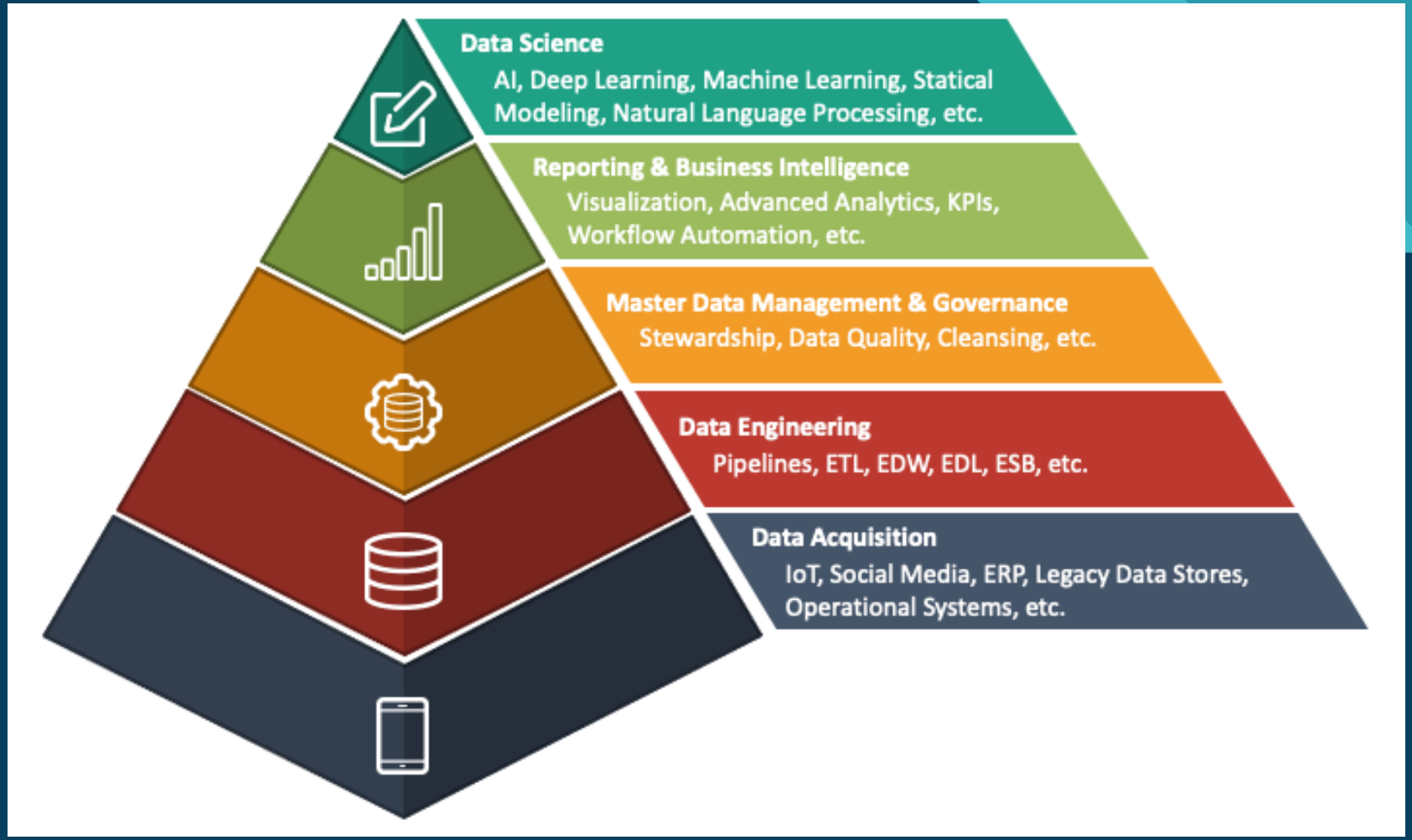


Version 3.0 - November 2021

mattturck.com/data2021

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

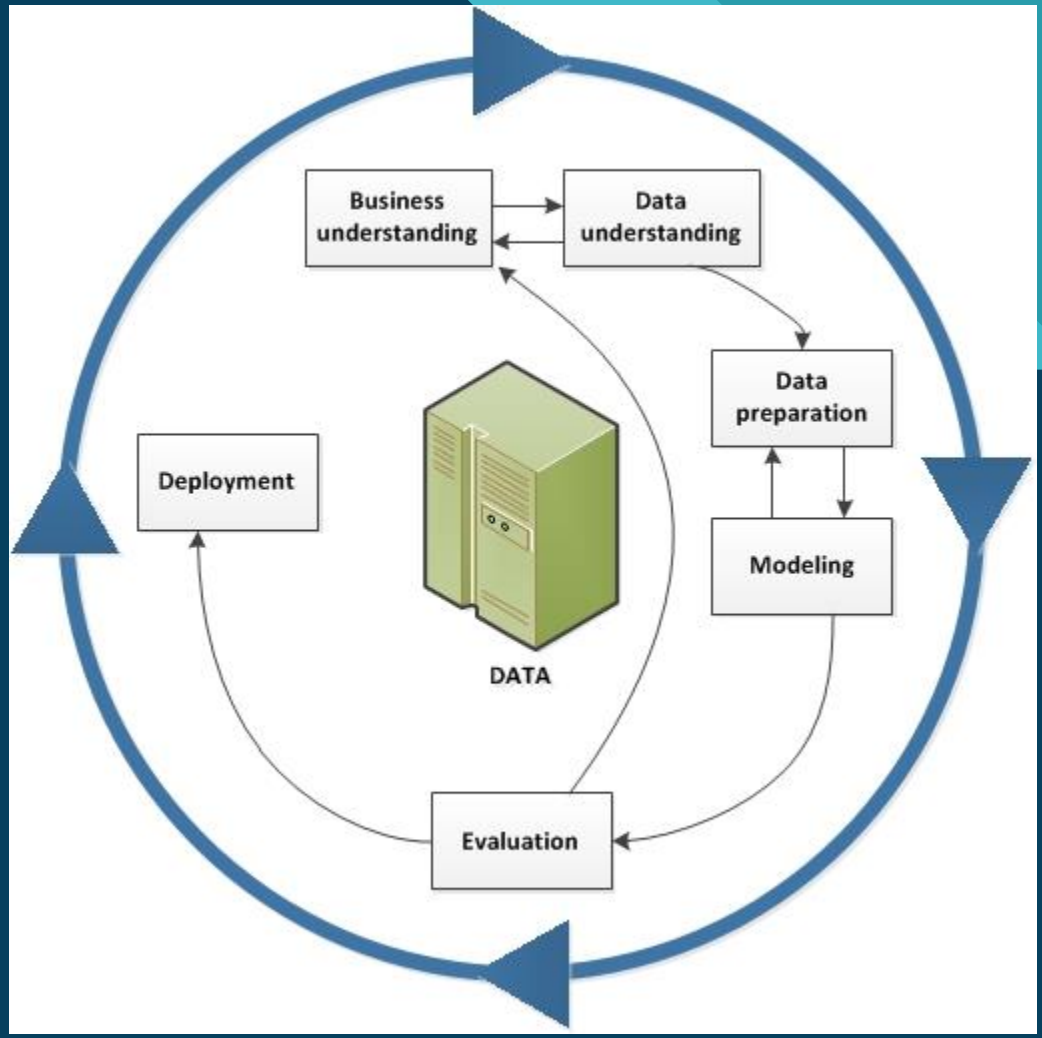
Data foundation as a pyramid



Source: https://www.sketchbubble.com/en/presentation-data-analytics-pyramid.html

CRISP DM

Cross-industry Standard Process for Data Mining



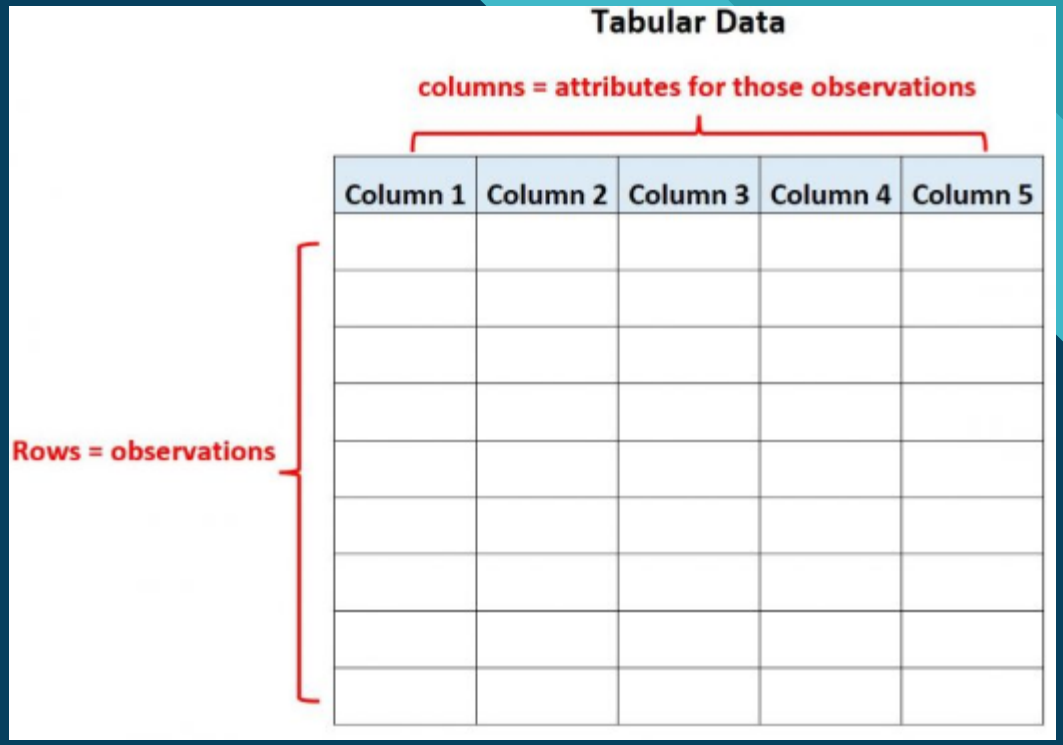
Source: https://www.ibm.com/docs/fr/spss-modeler/saas?topic=dm-crisp-help-overview

Tabular data
Types of variable

Tabular data

Very common data format

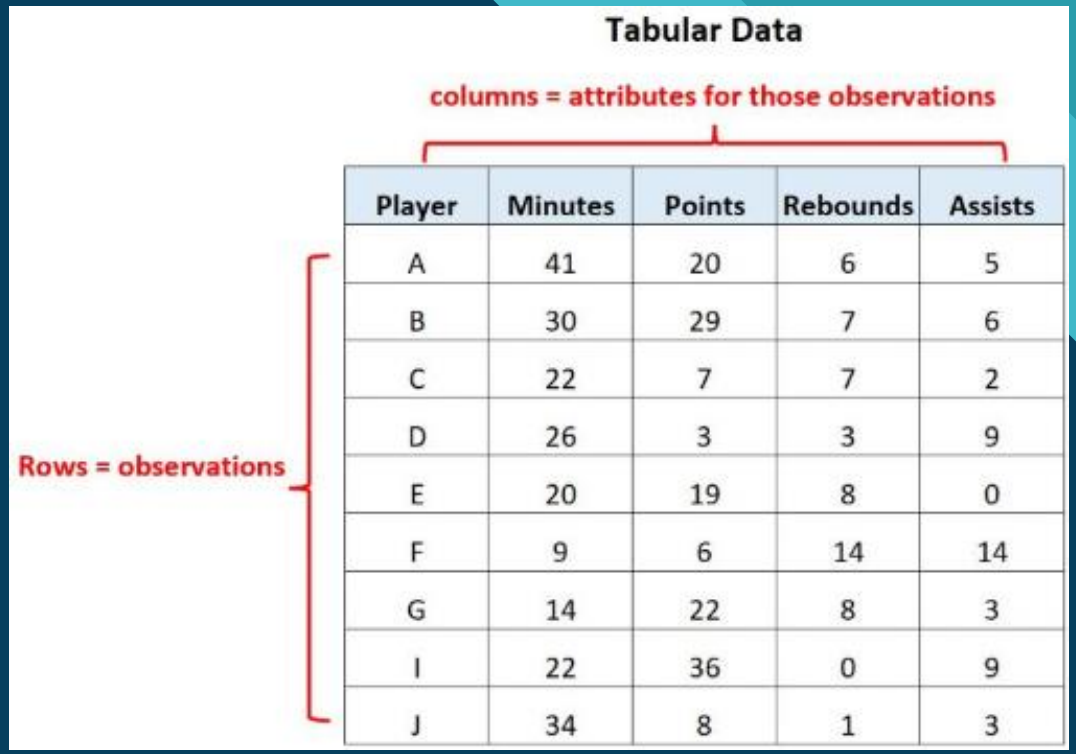
- Oganized by
  - Line = Record = observation = granularity of dataset
    - In other words, what does a record represent in this dataset?
  - Column = attribute = variable = feature (*wording mainly used in machine learning, especially when you create new variables*)



Tabular data

Very common data format

- Oganized by
  - Line = Record = observation = granularity of dataset
    - In other words, what does a record represent in this dataset?
  - Column = attribute = variable = feature (*wording mainly used in machine learning, especially when you create new variables*)



Player	Minutes	Points	Rebounds	Assists
A	41	20	6	5
B	30	29	7	6
C	22	7	7	2
D	26	3	3	9
E	20	19	8	0
F	9	6	14	14
G	14	22	8	3
I	22	36	0	9
J	34	8	1	3

Type of variable

- Categorical variables
  - Nominal. Examples:
    - Sex: Male / female
    - Countries: Belgium, Belarus, France, …
    - Mode of transportation: car, truck, bus, tram, tube, …
  - Ordinal. Examples
    - Appreciation level: Disgusting, Poor, Average, Good, Excellent
    - Rain quantity: low, medium, high

Type of variable

- Numeric variables
  - Continuous variables: infinite number of values
    - Temperature in Celsius
    - Speed of a vehicle
    - Height/Weight of a person
    - Time to answer a question
  - Discrete variable: finite number of values
    - Result at exam: 0, 0.5, 1, 1.5
    - Number of bedrooms in a house

Type of variable

- Numeric variables – continuous variable
  - Interval scale
    - 0 has an arbitrary meaning.
    - Positive or negative.
    - Example: temperature in Celsius (0 means water freeze), ot in Farhenheit
  - Ratio scale
    - 0 means absolute zero (temperature in Kelvin) or is the start of measurement (weight, height always > 0).
    - Always >= 0
    - Ratio means ‘can be used in division’ or ‘in multiplication’
      - We can say that an elephant weights as much as 100 000 mice.
      - >< we cannot say “10°Celsius is twice as warm as 5°Celsius”
    - Examples: sales figures, temperature in Kelvin (cannot be below 0), weight, time, distance, speed (distance / time)

Source: https://www.questionpro.com/blog/ratio-scale-vs-interval-scale/#:~:text=Interval%20scale%20can%20measure%20size,unit%20in%20terms%20of%20another.&text=A%20classic%20example%20of%20an%20interval%20scale%20is%20the%20temperature%20in%20Celsius

Data storage

Data storage

Numerous data storage:

- Plain text
  - csv, tsv or other delimited file types
  - Fixed width
  - json
  - Xml,
  - …

Data storage

Numerous data storage:

- Plain text
  - csv, tsv or other delimited file types
  - Fixed width
  - json
  - Xml,
  - …
- Spreadsheets:
  - Excel:
  - OpenOffice Calc
  - Google sheets
  - …

Data storage

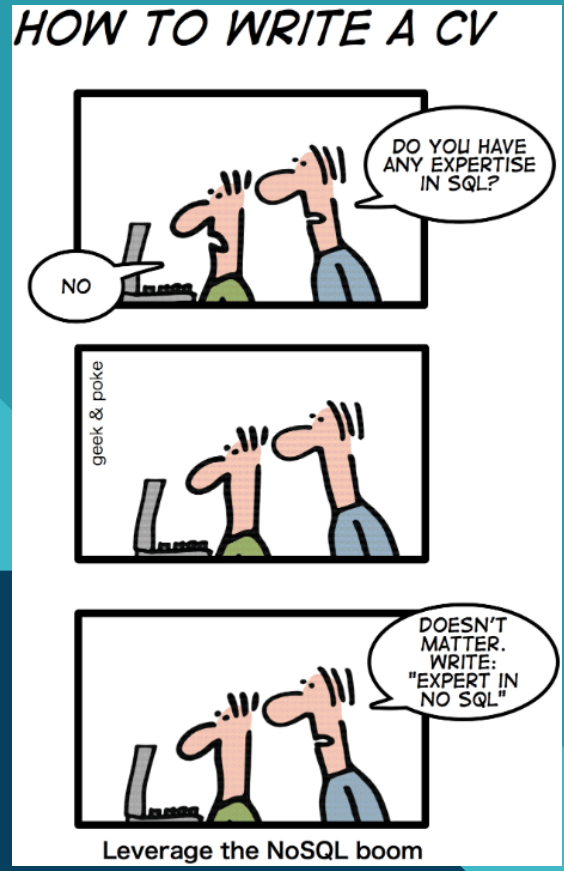
- DBMS = Data Base Management System
  - Relational database: Oracle, MySQL, PostgreSQL, …

Source: https://shopify.engineering/five-common-data-stores-usage

https://spectralops.io/blog/nosql-databases/

Data storage

- DBMS = Data Base Management System
  - Relational database: Oracle, MySQL, PostgreSQL, …
  - NoSQL = Not only SQL:



Source: https://shopify.engineering/five-common-data-stores-usage

https://spectralops.io/blog/nosql-databases/

Data storage

- DBMS = Data Base Management System
  - Relational database: Oracle, MySQL, PostgreSQL, …
  - NoSQL = Not only SQL:
    - columnar/analytical database: IBM Pure Date (Netezza), Greenplum, BigQuery, Amazon Redshift, Azure Synapse Analytics (Azure SQL Data Warehouse), … ;
    - key-value store: Redis, Memcached, RocksDB, Amazon Dynamo DB, …;
    - graph database: Neo4j, …;
    - “search engine” database: ElasticSearch, …;
    - …

Source: https://shopify.engineering/five-common-data-stores-usage

https://spectralops.io/blog/nosql-databases/

Data storage

- Storage of analytical software:
  - SAS: SAS datasets
  - R: Rdata files, fst files, …
  - …
- Columnar file format (software independent)
  - Parquet file (*optimized for storage, meaning smaller size)*
  - …
- …

Data storage: csv

```
ISO;NAME;ISONAME;region;pop;popUrban
44;BAHAMAS;BAHAMAS;AMR;323063;0.9
84;BELIZE;BELIZE;AMR;269736;0.486
124;CANADA;CANADA;AMR;32268243;0.811
188;COSTA RICA;COSTA RICA;AMR;4327228;0.617
192;CUBA;CUBA;AMR;11269400;0.76
214;DOMINICAN REPUBLIC;DOMINICAN REPUBLIC;AMR;8894907;0.601
222;EL SALVADOR;EL SALVADOR;AMR;6880951;0.601
320;GUATEMALA;GUATEMALA;AMR;12599059;0.472
332;HAITI;HAITI;AMR;8527777;0.388
340;HONDURAS;HONDURAS;AMR;7204723;0.464
388;JAMAICA;JAMAICA;AMR;2650713;0.522
484;MEXICO;MEXICO;AMR;107029360;0.76
558;NICARAGUA;NICARAGUA;AMR;5486685;0.581
591;PANAMA;PANAMA;AMR;3231502;0.578
659;SAINT KITTS/NEVIS;SAINT KITTS AND NEVIS;AMR;42696;0.319
840;UNITED STATES;UNITED STATES;AMR;298212895;0.808
28;ANTIGUA/BARBUDA;ANTIGUA AND BARBUDA;AMR;81485;0.384
32;ARGENTINA;ARGENTINA;AMR;38747148;0.906
```

Data storage: tsv

```
Make    Model   Type    Origin  DriveTrain  MSRP    Invoice EngineSize  Cylinders   Ho
Acura   MDX SUV Asia    All 36945   33337   3.5 6   265 17  23  4451    106 189
Acura   RSX Type S 2dr  Sedan   Asia    Front   23820   21761   2   4   200 24  31  27
Acura   TSX 4dr Sedan   Asia    Front   26990   24647   2.4 4   200 22  29  3230    10
Acura   TL 4dr  Sedan   Asia    Front   33195   30299   3.2 6   270 20  28  3575    10
Acura   3.5 RL 4dr  Sedan   Asia    Front   43755   39014   3.5 6   225 18  24  3880
Acura   3.5 RL w/Navigation 4dr Sedan   Asia    Front   46100   41100   3.5 6   225 18
Acura   NSX coupe 2dr manual S  Sports  Asia    Rear    89765   79978   3.2 6   290 17
Audi    A4 1.8T 4dr Sedan   Europe  Front   25940   23508   1.8 4   170 22  31  3252
Audi    A41.8T convertible 2dr  Sedan   Europe  Front   35940   32506   1.8 4   170 23
Audi    A4 3.0 4dr  Sedan   Europe  Front   31840   28846   3   6   220 20  28  3462
Audi    A4 3.0 Quattro 4dr manual   Sedan   Europe  All 33430   30366   3   6   220 17
```

Source: dataset sashelp.cars from SAS software

Data storage: tsv

```
Make→Model→Type→Origin→DriveTrain→MSRP→Invoice→EngineSize→Cylinders→Ho
Acura→MDX→SUV→Asia→All→36945→33337→3.5→6→265→17→23→4451→106→189LF
Acura→RSX Type S 2dr→Sedan→Asia→Front→23820→21761→2→4→200→24→31→27
Acura→TSX 4dr→Sedan→Asia→Front→26990→24647→2.4→4→200→22→29→3230→10
Acura→TL 4dr→Sedan→Asia→Front→33195→30299→3.2→6→270→20→28→3575→10
Acura→3.5 RL 4dr→Sedan→Asia→Front→43755→39014→3.5→6→225→18→24→3880→
Acura→3.5 RL w/Navigation 4dr→Sedan→Asia→Front→46100→41100→3.5→6→225→18
Acura→NSX coupe 2dr manual S→Sports→Asia→Rear→89765→79978→3.2→6→290→17
Audi→A4 1.8T 4dr→Sedan→Europe→Front→25940→23508→1.8→4→170→22→31→3252→
Audi→A41.8T convertible 2dr→Sedan→Europe→Front→35940→32506→1.8→4→170→23
Audi→A4 3.0 4dr→Sedan→Europe→Front→31840→28846→3→6→220→20→28→3462→
Audi→A4 3.0 Quattro 4dr manual→Sedan→Europe→All→33430→30366→3→6→220→17
```

Source: dataset sashelp.cars from SAS software

Data storage: fixed width

Acura	MDX	SUV	Asia	All	$36,945
Acura	RSX Type S 2dr	Sedan	Asia	Front	$23,820
Acura	TSX 4dr	Sedan	Asia	Front	$26,990
Acura	TL 4dr	Sedan	Asia	Front	$33,195
Acura	3.5 RL 4dr	Sedan	Asia	Front	$43,755
Acura	3.5 RL w/Navigation 4dr	Sedan	Asia	Front	$46,100
Acura	NSX coupe 2dr manual S	Sports	Asia	Rear	$89,765
Audi	A4 1.8T 4dr	Sedan	Europe	Front	$25,940
Audi	A41.8T convertible 2dr	Sedan	Europe	Front	$35,940
Audi	A4 3.0 4dr	Sedan	Europe	Front	$31,840
Audi	A4 3.0 Quattro 4dr manual	Sedan	Europe	All	$33,430
Audi	A4 3.0 Quattro 4dr auto	Sedan	Europe	All	$34,480
Audi	A6 3.0 4dr	Sedan	Europe	Front	$36,640
Audi	A6 3.0 Quattro 4dr	Sedan	Europe	All	$39,640
Audi	A4 3.0 convertible 2dr	Sedan	Europe	Front	$42,490
Audi	A4 3.0 Quattro convertible 2dr	Sedan	Europe	All	$44,240

Source: dataset sashelp.cars from SAS software

Data storage: json


```
[
  {
    "Make": "Acura",
    "Model": " MDX",
    "Type": "SUV",
    "Origin": "Asia",
    "DriveTrain": "All",
    "MSRP": 36945,
    "Invoice": 33337,
    "EngineSize": 3.5,
    "Cylinders": 6,
    "Horsepower": 265,
    "MPG_City": 17,
    "MPG_Highway": 23,
    "Weight": 4451,
    "Wheelbase": 106,
    "Length": 189
  },
  {
    "Make": "Acura",
    "Model": " RSX Type S 2dr",
    "Type": "Sedan",
    "Origin": "Asia",
    "DriveTrain": "Front",
    "MSRP": 23820,
    "Invoice": 21761,
    "EngineSize": 2,
    "Cylinders": 4,
    "Horsepower": 200,
    "MPG_City": 24,
    "MPG_Highway": 31,
    "Weight": 2778,
    "Wheelbase": 101,
    "Length": 172
  },
  {
    "Make": "Acura",
```


Source: dataset sashelp.cars from SAS software

Data storage: json

Sometime the json can be condensed in just 1 line.

This is also fine, (except if someone has modify something in it …)

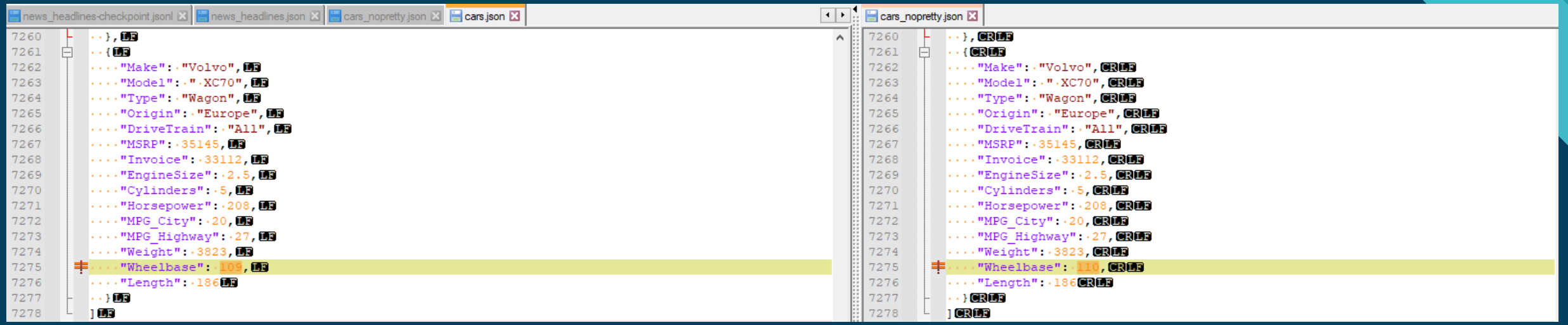

```
[{"Make":"Acura","Model":" MDX","Type":"SUV","Origin":"Asia","DriveTrain":"All","MSRP":36945,"Invoic
```


Source: dataset sashelp.cars from SAS software

Data storage: json

Sometime the json can be condensed in just 1 line.

This is also fine, (except if someone has modified something in it …)

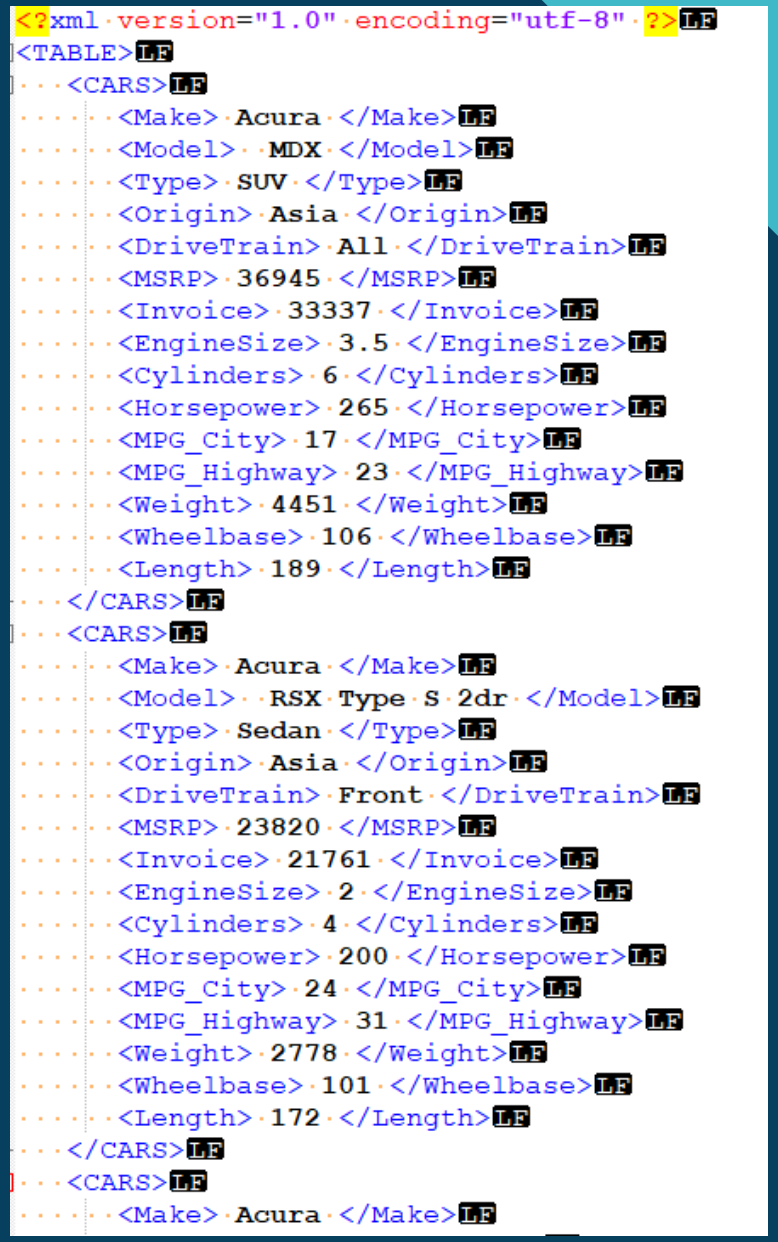


Data storage: jsonl


```
{"Make":"Acura","Model":"·MDX","Type":"SUV","Origin":"Asia","DriveTrain":"All","MSRP":36945,"Invoice":33337,"EngineSize":3.5,"Cylind
{"Make":"Acura","Model":"·RSX·Type·S·2dr","Type":"Sedan","Origin":"Asia","DriveTrain":"Front","MSRP":23820,"Invoice":21761,"EngineSi
{"Make":"Acura","Model":"·TSX·4dr","Type":"Sedan","Origin":"Asia","DriveTrain":"Front","MSRP":26990,"Invoice":24647,"EngineSize":2.4
{"Make":"Acura","Model":"·TL·4dr","Type":"Sedan","Origin":"Asia","DriveTrain":"Front","MSRP":33195,"Invoice":30299,"EngineSize":3.2,
{"Make":"Acura","Model":"·3.5·RL·4dr","Type":"Sedan","Origin":"Asia","DriveTrain":"Front","MSRP":43755,"Invoice":39014,"EngineSize":
{"Make":"Acura","Model":"·3.5·RL·w/Navigation·4dr","Type":"Sedan","Origin":"Asia","DriveTrain":"Front","MSRP":46100,"Invoice":41100,
```


Source: dataset sashelp.cars from SAS software

Data storage: xml



```
<?xml version="1.0" encoding="utf-8" ?>
<TABLE>
  <CARS>
    <Make> Acura </Make>
    <Model>  MDX </Model>
    <Type> SUV </Type>
    <Origin> Asia </Origin>
    <DriveTrain> All </DriveTrain>
    <MSRP> 36945 </MSRP>
    <Invoice> 33337 </Invoice>
    <EngineSize> 3.5 </EngineSize>
    <Cylinders> 6 </Cylinders>
    <Horsepower> 265 </Horsepower>
    <MPG_City> 17 </MPG_City>
    <MPG_Highway> 23 </MPG_Highway>
    <Weight> 4451 </Weight>
    <Wheelbase> 106 </Wheelbase>
    <Length> 189 </Length>
  </CARS>
  <CARS>
    <Make> Acura </Make>
    <Model>  RSX Type S 2dr </Model>
    <Type> Sedan </Type>
    <Origin> Asia </Origin>
    <DriveTrain> Front </DriveTrain>
    <MSRP> 23820 </MSRP>
    <Invoice> 21761 </Invoice>
    <EngineSize> 2 </EngineSize>
    <Cylinders> 4 </Cylinders>
    <Horsepower> 200 </Horsepower>
    <MPG_City> 24 </MPG_City>
    <MPG_Highway> 31 </MPG_Highway>
    <Weight> 2778 </Weight>
    <Wheelbase> 101 </Wheelbase>
    <Length> 172 </Length>
  </CARS>
  <CARS>
    <Make> Acura </Make>
```

Source: dataset sashelp.cars from SAS software

Data storage - plain text

Regarding plain text files, you can inspect them with Notepadd++

Notepad++:

- Free
- Open Source
- Light
- Have a lot of useful features: either built-in or via plugin (Compare, Json Viewer, …)

- Useful to open file containing code (python, R, …)

Data storage: spreadsheets

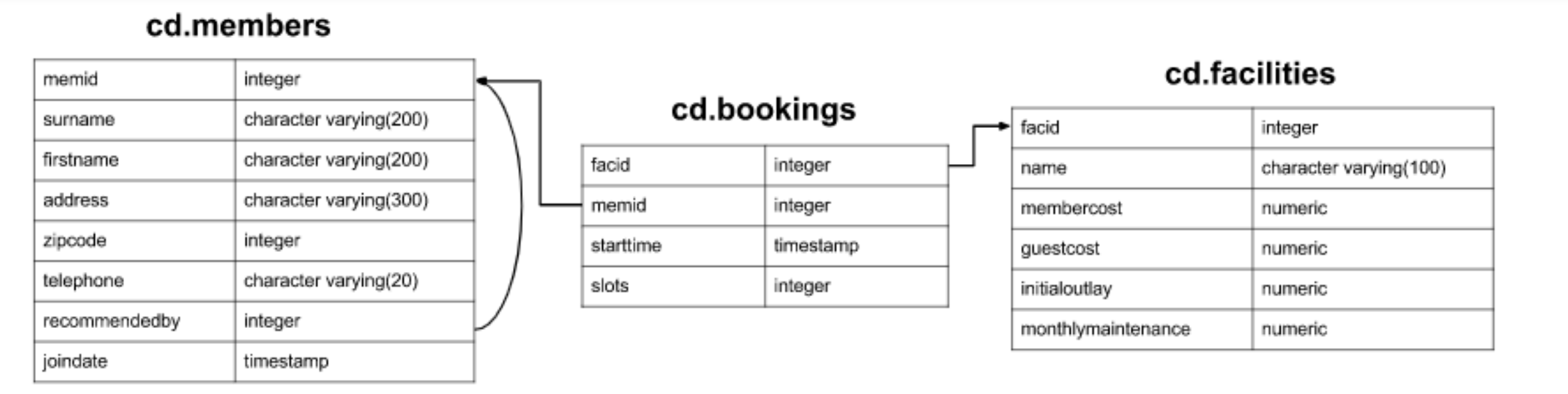
	A	B	C	D	E	F	G	H	I
1	Ticket ID	Product_ID	Unit price	Quantity	Tax applied	Amount without tax	Amount with Tax	Amount without tax (% of tota	Amount with Tax (% of total)
2	10	360	29.6	14	21%	414.4	501.424	6.73%	6.93%
3	19	122	90	7	6%	630	667.8	10.23%	9.23%
4	19	160	8.26	80	21%	660.8	799.568	10.73%	11.05%
5	19	280	65.3	4	21%	261.2	316.052	4.24%	4.37%
6	25	250	120	2	21%	240	290.4	3.90%	4.01%
7	25	124	87.6	18	21%	1576.8	1907.928	25.60%	26.36%
8	30	250	120	7	21%	840	1016.4	13.64%	14.04%
9	30	300	52.4	3	21%	157.2	190.212	2.55%	2.63%
10	30	122	90	8	6%	720	763.2	11.69%	10.55%
11	30	160	8.26	19	12%	156.94	175.7728	2.55%	2.43%
12	35	280	65.3	4	21%	261.2	316.052	4.24%	4.37%
13	40	312	30.2	8	21%	241.6	292.336	3.92%	4.04%

Data storage: relational database

facid	name	membercost	guestcost	initialoutlay	monthlymaintenance
0	Tennis Court 1	5	25	10000	200
1	Tennis Court 2	5	25	8000	200
2	Badminton Court	0	15.5	4000	50
3	Table Tennis	0	5	320	10
4	Massage Room 1	35	80	4000	3000
5	Massage Room 2	35	80	4000	3000
6	Squash Court	3.5	17.5	5000	80
7	Snooker Table	0	5	450	15
8	Pool Table	0	5	400	15

Source: https://pgexercises.com/questions/basic/selectall.html

Data storage: relational database



Source: https://pgexercises.com/questions/basic/selectall.html

Data storage: database

Designed for OLTP or small OLAP workload

OLTP: Online Transactional Processing

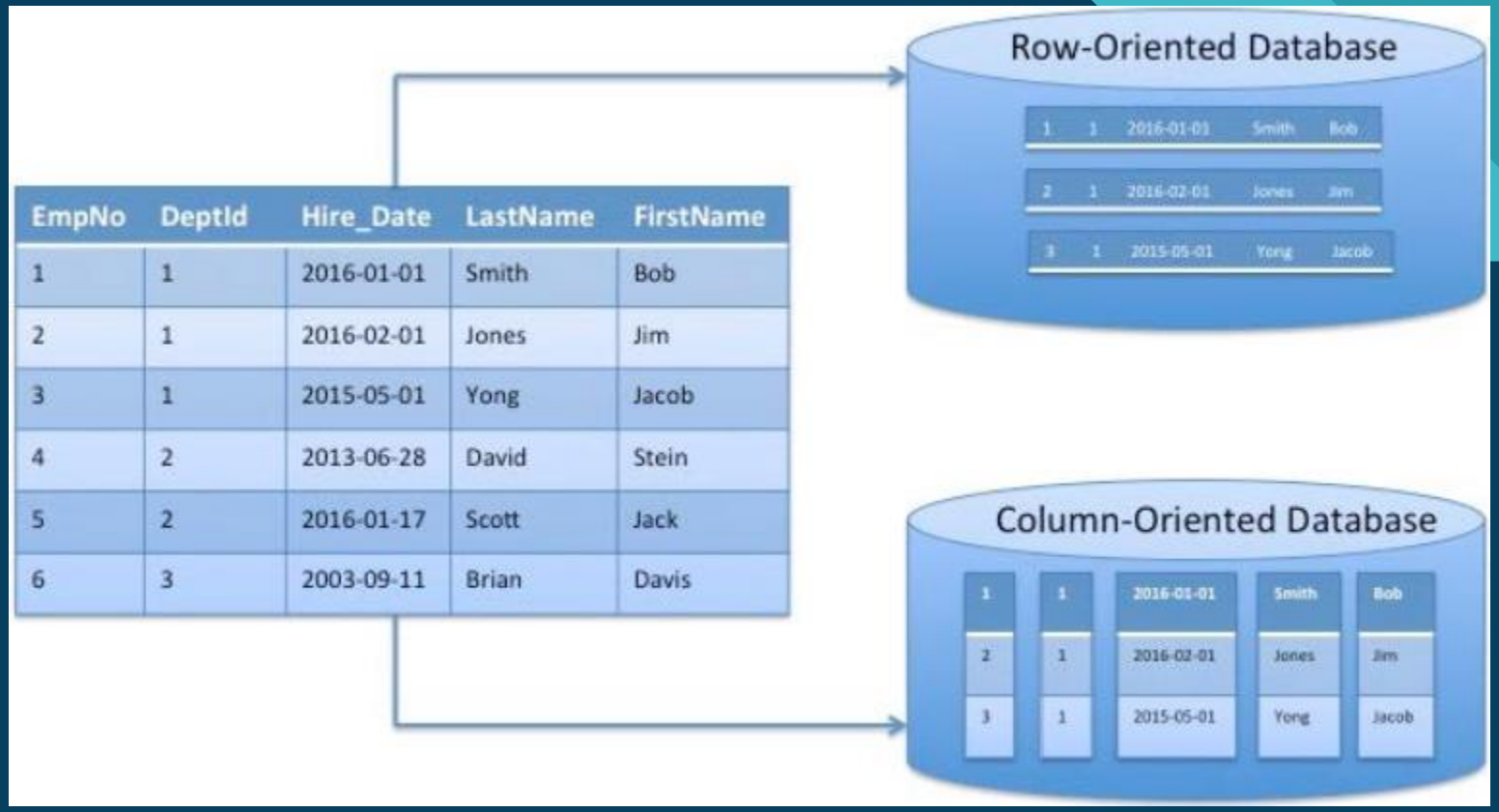
- A lot of small database transaction: usually update, insert or delete records in a databases
- Examples?

OLAP: Online Analytical processing

- Few bigger queries: select a few columns, possibly combined with other tables and then aggregated
- Examples?

Source: https://www.ibm.com/cloud/blog/olap-vs-oltp

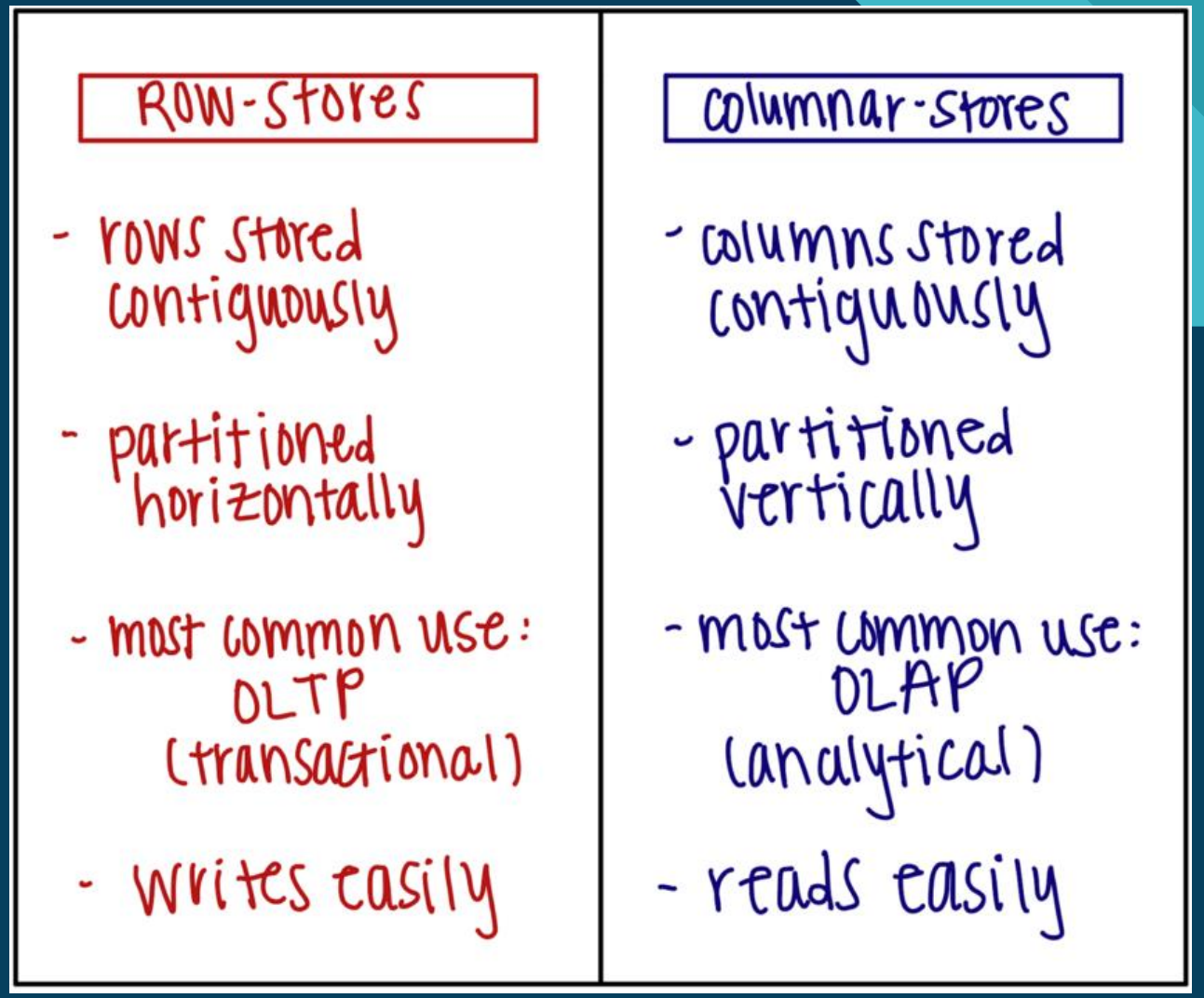
Data storage: row/column oriented database



EmpNo	DeptId	Hire_Date	LastName	FirstName
1	1	2016-01-01	Smith	Bob
2	1	2016-02-01	Jones	Jim
3	1	2015-05-01	Yong	Jacob
4	2	2013-06-28	David	Stein
5	2	2016-01-17	Scott	Jack
6	3	2003-09-11	Brian	Davis

Source: https://docs.aws.amazon.com/whitepapers/latest/data-warehousing-on-aws/data-warehouse-technology-options.html

Data storage: row/column-oriented database



Source: https://medium.com/bluecore-engineering/deciding-between-row-and-columnar-stores-why-we-chose-both-3a675dab4087

Data storage: analytical database

Example: IBM PureData (Netezza)

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
4,9	3	1,4	0,2	setosa
5	3,6	1,4	0,2	setosa
5,1	3,5	1,4	0,2	setosa
4,7	3,2	1,3	0,2	setosa
4,6	3,1	1,5	0,2	setosa

Data storage: analytical database

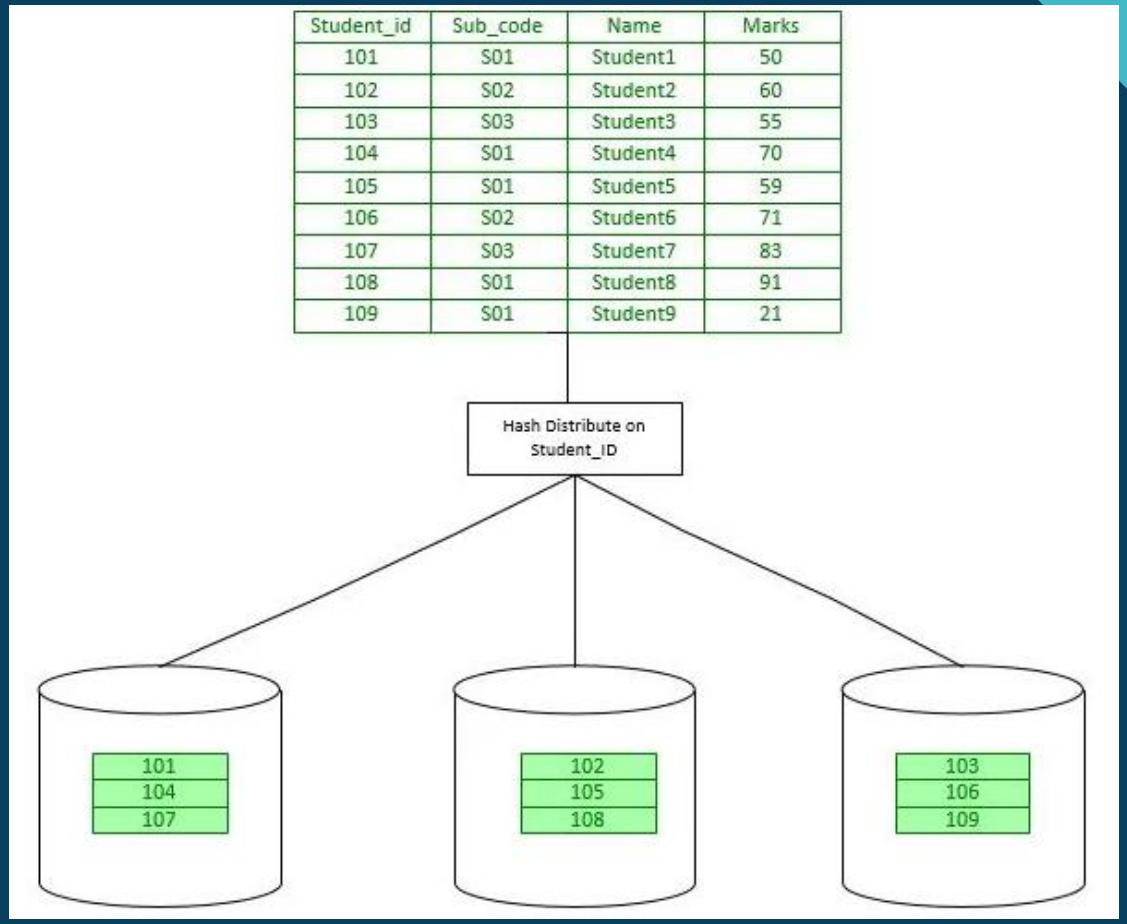
Designed for large OLAP workload

- Most analytical DBMS use columnar storage:
  - Reduce storage due to better compression
  - Faster because only read necessary columns (>< row-based storage)
- Faster because of design
  - Data distributed across multiple nodes/machines (distribution key for Netezza)
  - Reading records with higher chance of meeting the conditions (zone maps for Netezza)

Source: https://www.ibm.com/cloud/blog/olap-vs-oltp

Data storage: analytical database Workload distribution in Netezza

Example: distribution key in IBM Pure Data (Netezza)



Source: https://dwgeek.com/greenplum-table-distribution-best-practices.html/

Data storage: analytical database
Workload distribution in Netezza

Example: distribution key in IBM Pure Data (Netezza)

Distribution key(s):

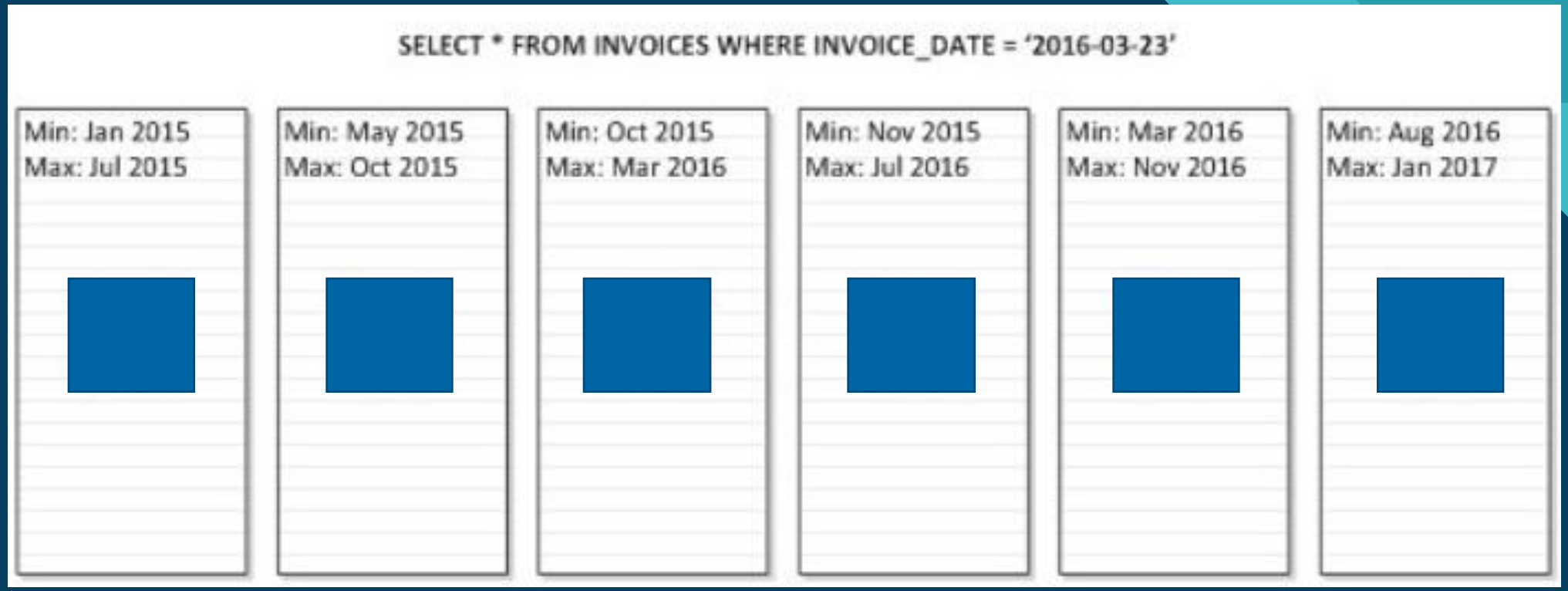
- Specified by database administrator
- 1 to 4 variables maximum
- For each record, value of distribution key(s) is hashed. The result of the hash enables the system to assign the record to a specific node
- Record with same value of distribution key(s) are stored on the same node. Improvement for
  - Computation by group, especially if the grouping variable(s) is(are) distribution key(s) using that variable
  - Joining dataset with the same distribution key(s)

Source: https://www.ibm.com/docs/en/SSULQD_7.2.1/com.ibm.nz.adm.doc/c_sysadm_distribution_keys.html

https://www.ibm.com/docs/en/psfa/7.2.1?topic=keys-criteria-selecting-distribution

Data storage: analytical database
Read necessary records in Netezza

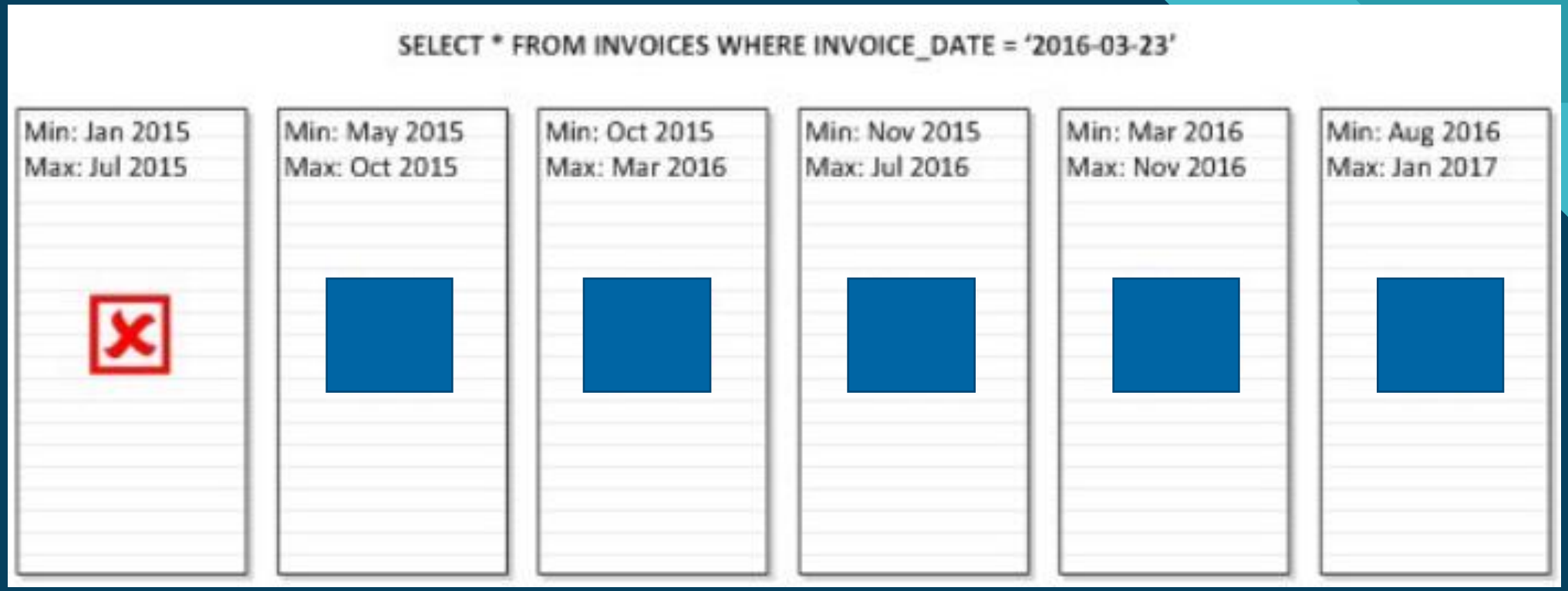
Example: zone maps in IBM Pure Data (Netezza)



Source: https://medium.com/@rleishman/netezza-zone-maps-theyre-the-same-as-indexes-right-fb3249f01cea

Data storage: analytical database
Read necessary records in Netezza

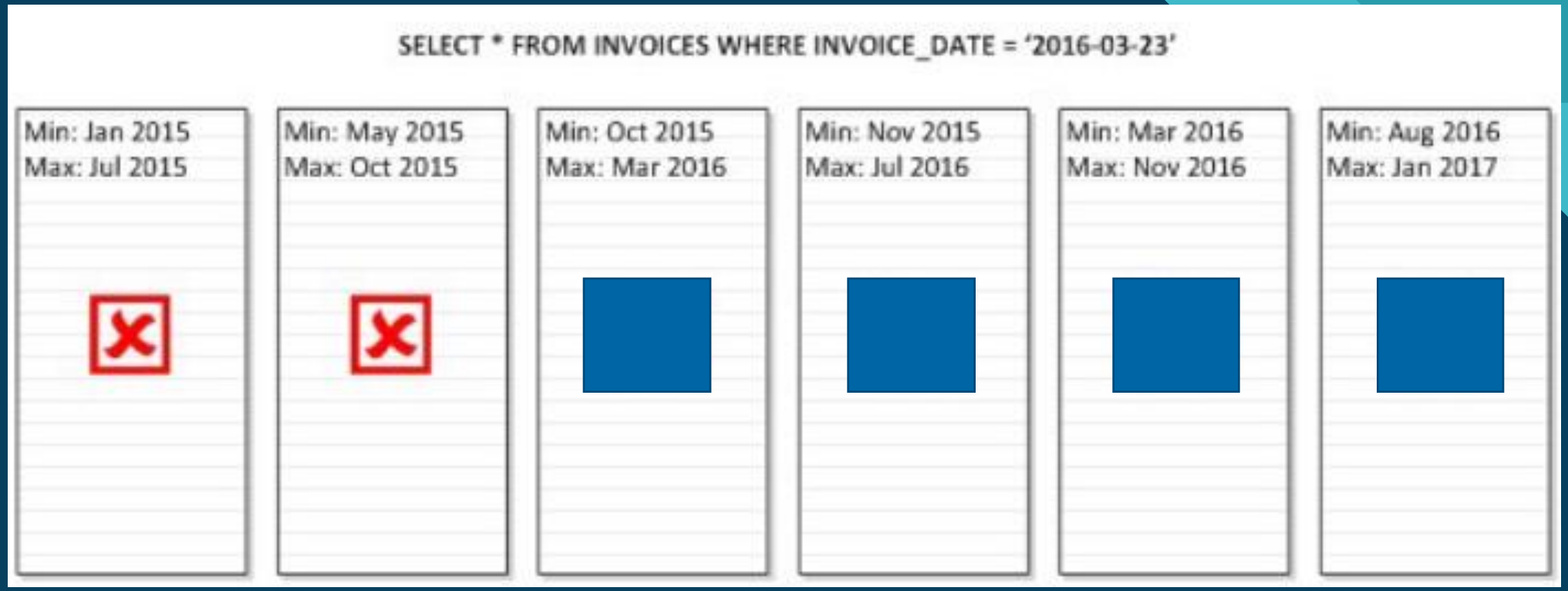
Example: zone maps in IBM Pure Data (Netezza)



Source: https://medium.com/@rleishman/netezza-zone-maps-theyre-the-same-as-indexes-right-fb3249f01cea

Data storage: analytical database
Read necessary records in Netezza

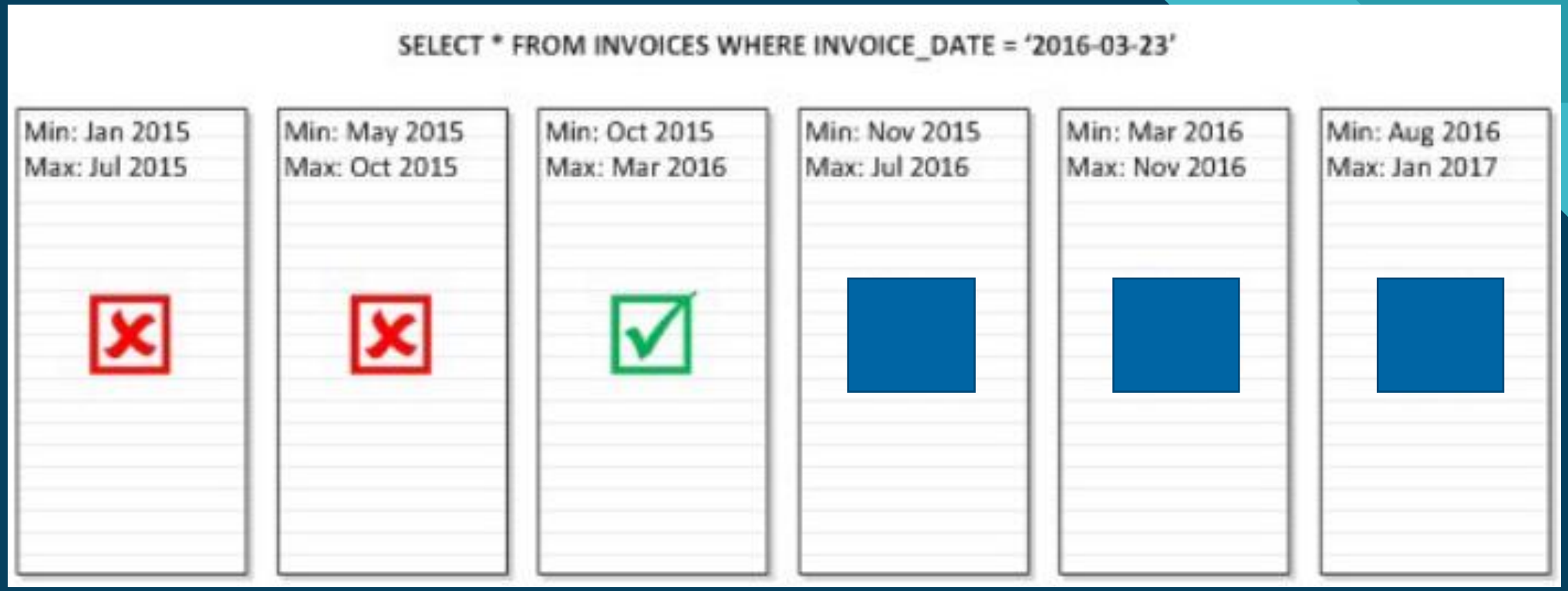
Example: zone maps in IBM Pure Data (Netezza)



Source: https://medium.com/@rleishman/netezza-zone-maps-theyre-the-same-as-indexes-right-fb3249f01cea

Data storage: analytical database
Read necessary records in Netezza

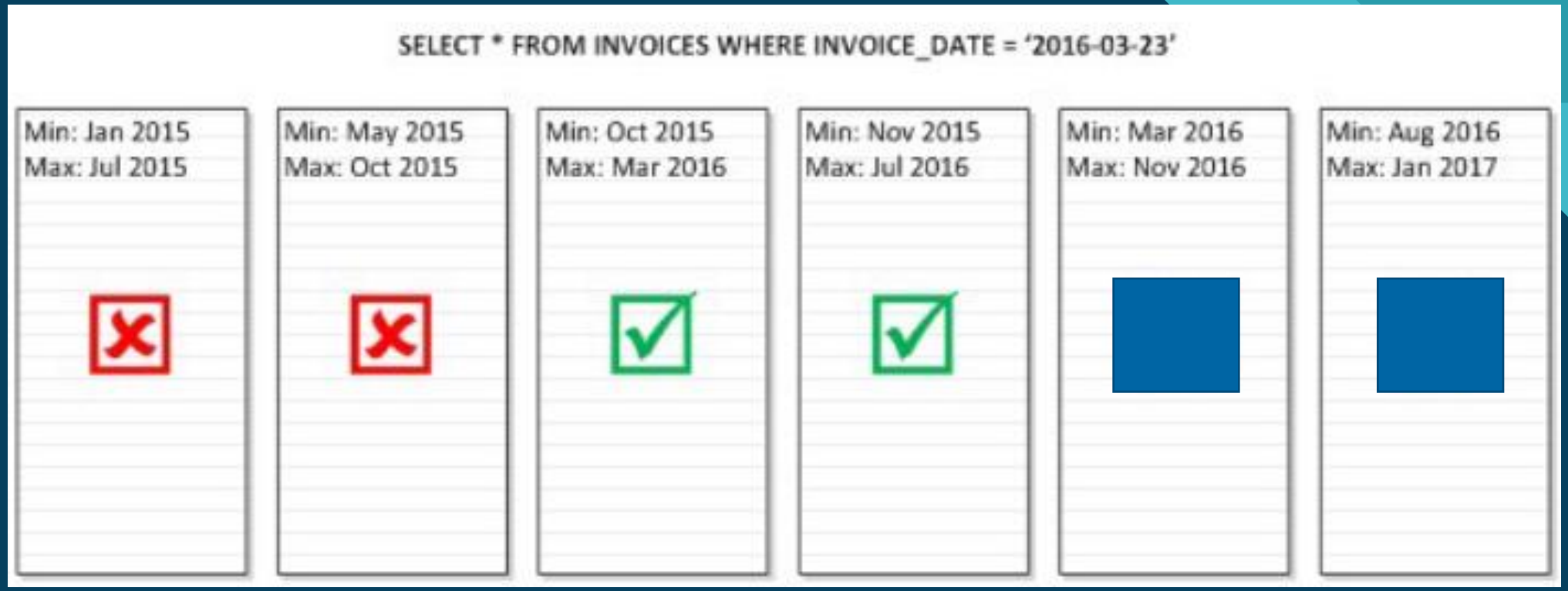
Example: zone maps in IBM Pure Data (Netezza)



Source: https://medium.com/@rleishman/netezza-zone-maps-theyre-the-same-as-indexes-right-fb3249f01cea

Data storage: analytical database
Read necessary records in Netezza

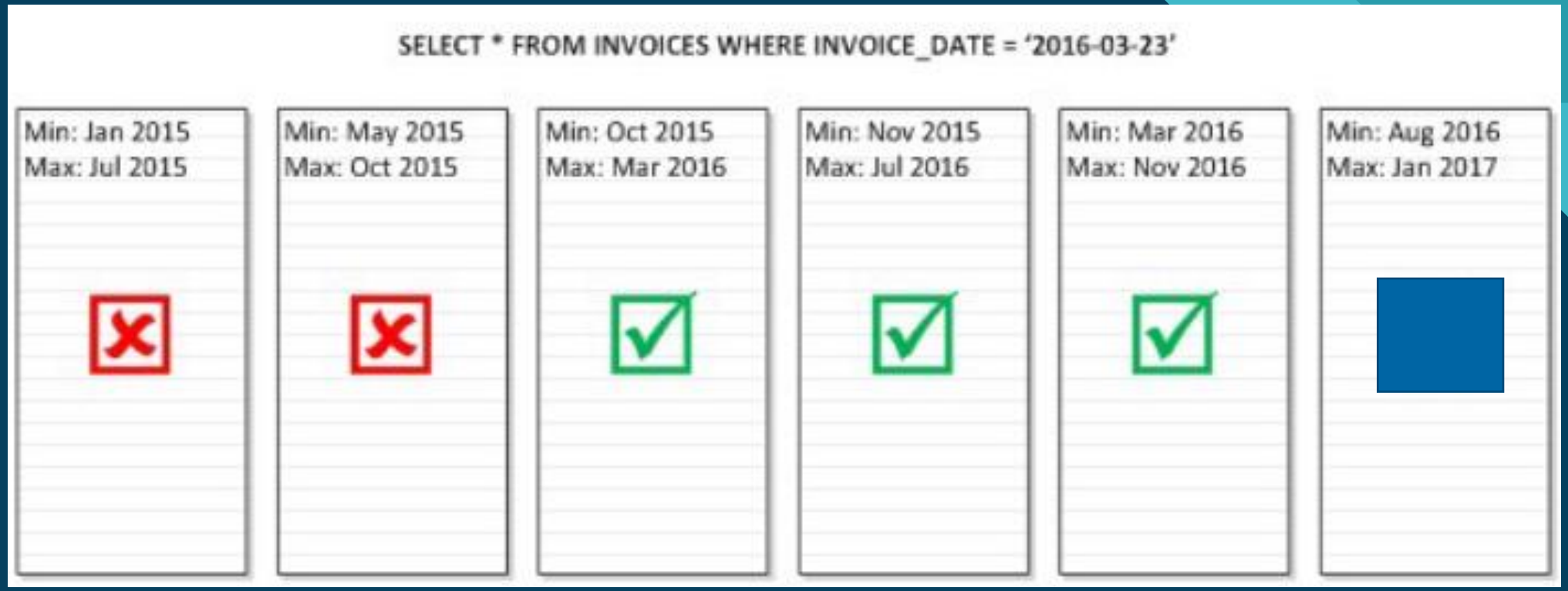
Example: zone maps in IBM Pure Data (Netezza)



Source: https://medium.com/@rleishman/netezza-zone-maps-theyre-the-same-as-indexes-right-fb3249f01cea

Data storage: analytical database
Read necessary records in Netezza

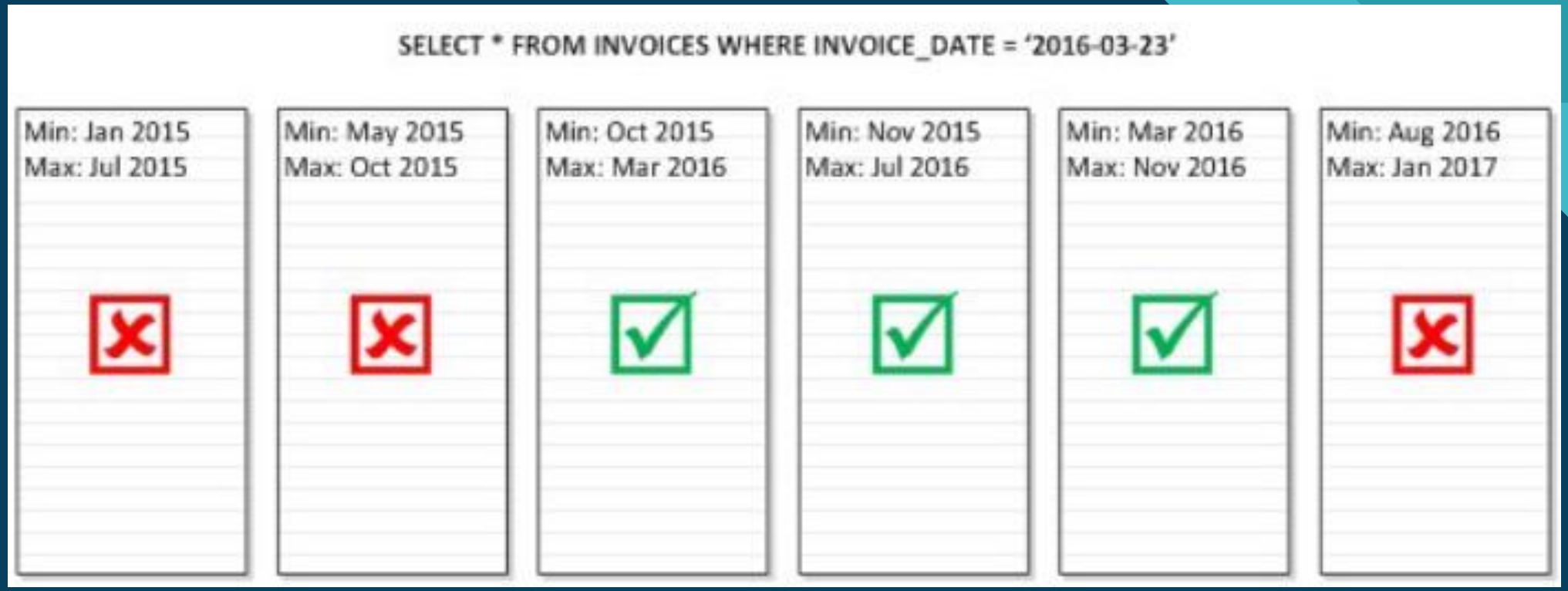
Example: zone maps in IBM Pure Data (Netezza)



Source: https://medium.com/@rleishman/netezza-zone-maps-theyre-the-same-as-indexes-right-fb3249f01cea

Data storage: analytical database
Read necessary records in Netezza

Example: zone maps in IBM Pure Data (Netezza)



Source: https://medium.com/@rleishman/netezza-zone-maps-theyre-the-same-as-indexes-right-fb3249f01cea

Data storage: analytical database
Read necessary records in Netezza

Example: zone maps in IBM Pure Data (Netezza)

Zone maps:

- Summary information about a subset of records
- Some variables types are automatically used for this (not explicitly chosen by database administrator)
- For those variable, minimum and maximum value are stored
- Worked best with value similar when you store them
- Not an index

Source: https://medium.com/@rleishman/netezza-zone-maps-theyre-the-same-as-indexes-right-fb3249f01cea

Data storage: analytical database Index vs zone maps

Similarities:

- Speed up the time to find records matching a specific criteria

Dissimilarities:

Index (relational DBMS)	Zone maps (Netezza)
List all distinct values for the whole dataset	Store only minimum and maximum value for a slice of the dataset
For each distinct value, list all records number to read	Indicate if a block (slice) of data can be trustly ignored.
More expensive to update	Cheap to update

Source: https://medium.com/@rleishman/netezza-zone-maps-theyre-the-same-as-indexes-right-fb3249f01cea

Descriptive statistics

Here is an example

Variable	*Sample*
Age in years: mean (SD)	47.3 (17.4)
Gender (%)	
Women	65.7
Men	34.3
Occupation (%)	
In paid work	37.2
Retired	25.5
Unable to work due to illness/disability	10.4
Looking after home/family	9.1
Unemployed and looking for work	7.8
Living situation (%)	
Own their own home	48.7
Renting from a housing association	30.0
Renting from a private landlord	12.6
Living with parents/family	7.0
Neighbourhood (%)	
Thickley (most deprived)	38.9
Byerley	29.9
Sunnydale (least deprived)	31.4
Years lived in Shildon: mean (SD)	32.9 (20.1)
Longstanding illness, disability or infirmity (%)	46.2

Note: $n = 233$

Source: https://uniskills.library.curtin.edu.au/numeracy/statistics/interpreting/

Type of descriptive statistics

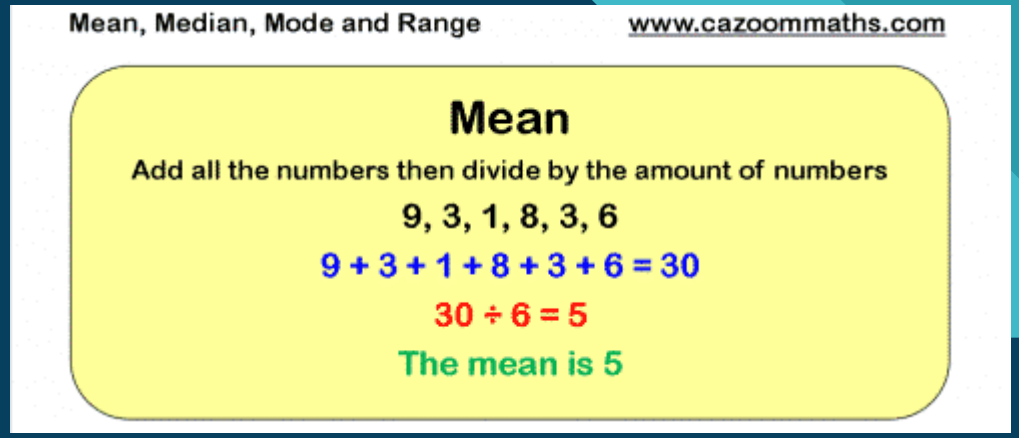
- Measure of frequency
- Measures of central tendency
- Measures of variation
- Measures of association

Measures of frequency

- Count / percentage
- Useful for
  - categorical variables
  - and discrete variables (if it doesn't take too many distinct values and it makes sense for your story)

Living situation (%)	
Own their own home	48.7
Renting from a housing association	30.0
Renting from a private landlord	12.6
Living with parents/family	7.0

Measures of central tendency

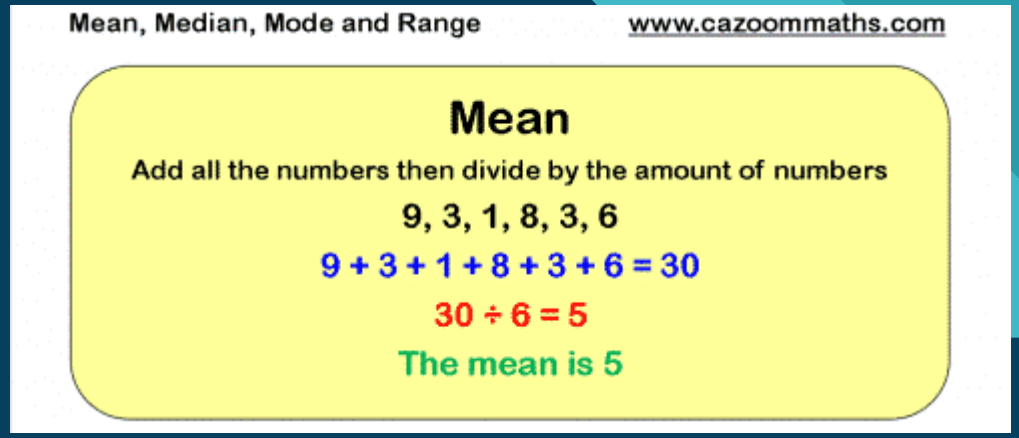


Source: https://danielmiessler.com/blog/difference-median-mean/

Measures of central tendency

Mean formula:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

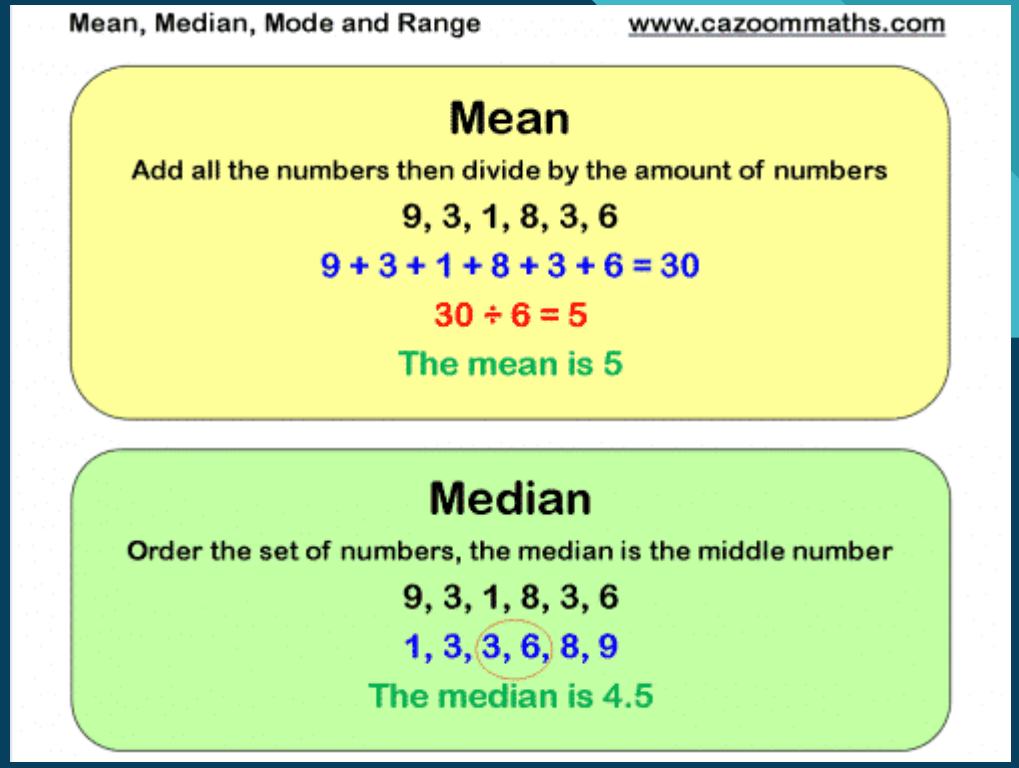


Source: https://danielmiessler.com/blog/difference-median-mean/

Measures of central tendency

Mean formula:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

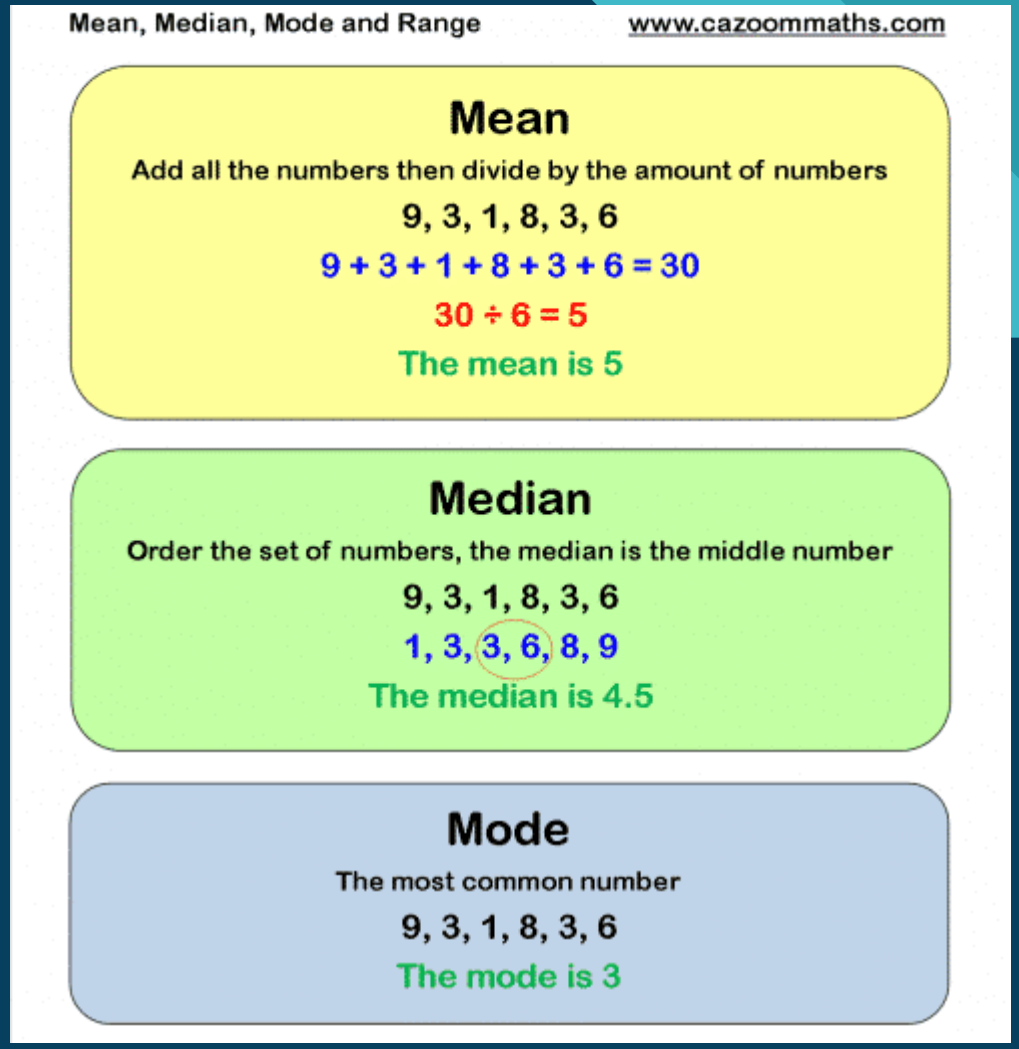


Source: https://danielmiessler.com/blog/difference-median-mean/

Measures of central tendency

Mean formula:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$



Source: https://danielmiessler.com/blog/difference-median-mean/

Measures of central tendency

- Mean
  - For numeric variables only
- Median
  - For numeric variables and ordinal variables
    - For ordinal, if there is an even number of elements, you can't take the middle points, so you take the very next value as being the median
- Mode
  - For numeric variables (if not too many distinct values) and categorical variables

Measures of central tendency

- Weighted mean
  - For numeric variables only
  - You need a 2nd variable that represents the weight of each observation

Measures of central tendency

- Weighted mean
  - For numeric variables only
  - You need a 2nd variable that represents the weight of each observation

  - Formula for weighted mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i \, x_i}{\sum_{i=1}^{n} w_i}$$

Measures of central tendency

- Weighted mean
  - For numeric variables only
  - You need a 2nd variable that represents the weight of each observation

  - Formula for weighted mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i \, x_i}{\sum_{i=1}^{n} w_i}$$

- Weighted percentile …

Measures of central tendency - weighted mean

Shareholder	Vote	Weight	Vote * Weight
John	1	2.5	2.5
Brenda	0	1	0
Leo	0	0.5	0
Paul	0	1	0
Matthew	1	1.2	1.2
Marc	0	1.4	0
Sophie	1	0.8	0.8
Aaron	1	0.6	0.6
Joe	0	0.4	0
Joey	1	2.4	2.4
Alice	0	1.4	0
Lien	0	1	0

Mean 41.67%

Weighted mean 52.82%

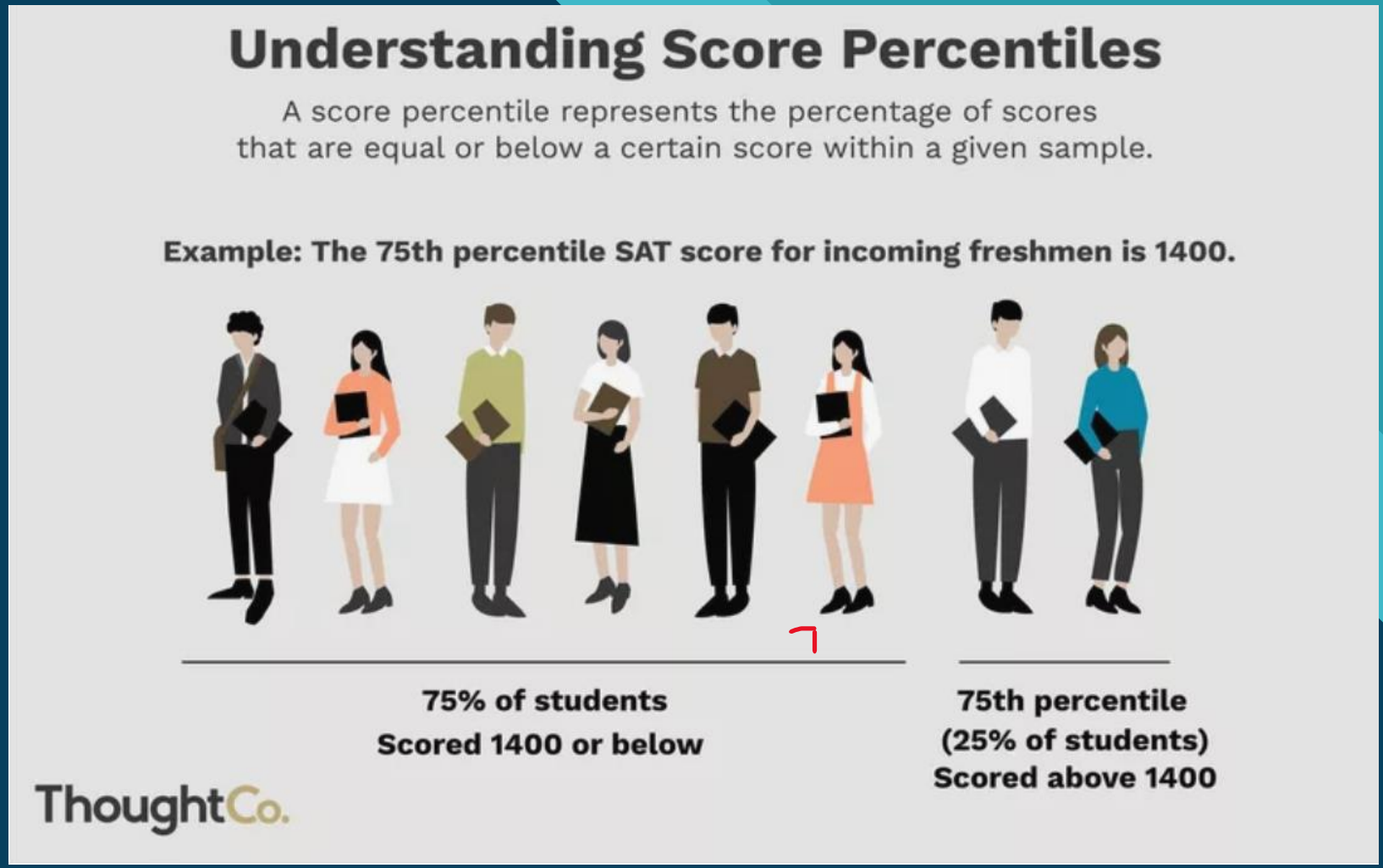
=> **Doing a simple mean is not correct in this case !!!**

Decomposed approach for weighted mean

1) create new column where you multiple Vote by Weight
2) sum this new column 7.5
3) sum column representing Weight 14.2
4) Divide 2) by 3) 52.82%

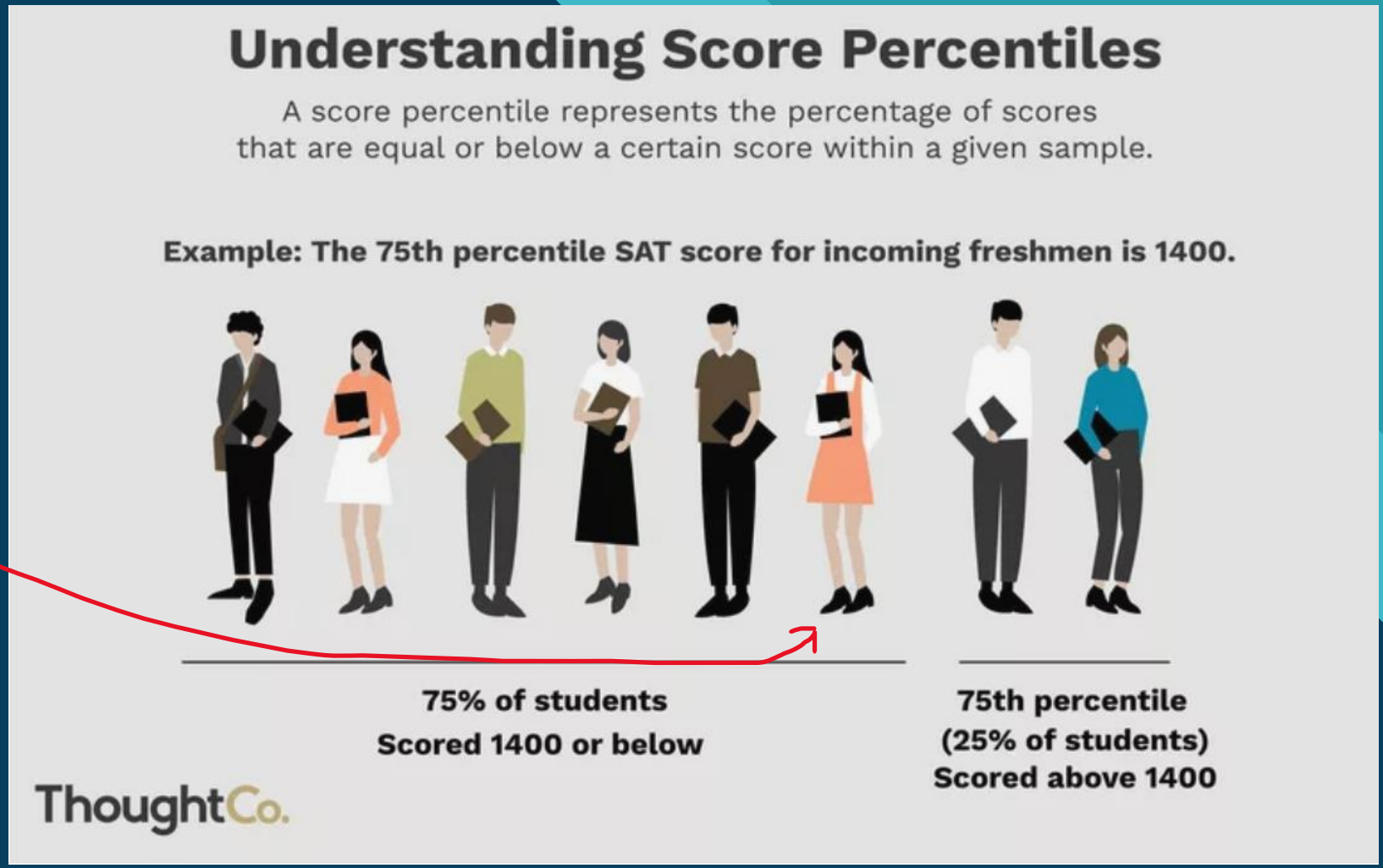
See file: Mean vs weighted mean.xlsx and solve it by yourself

Percentile



Source: https://www.thoughtco.com/how-to-understand-score-percentiles-3211610

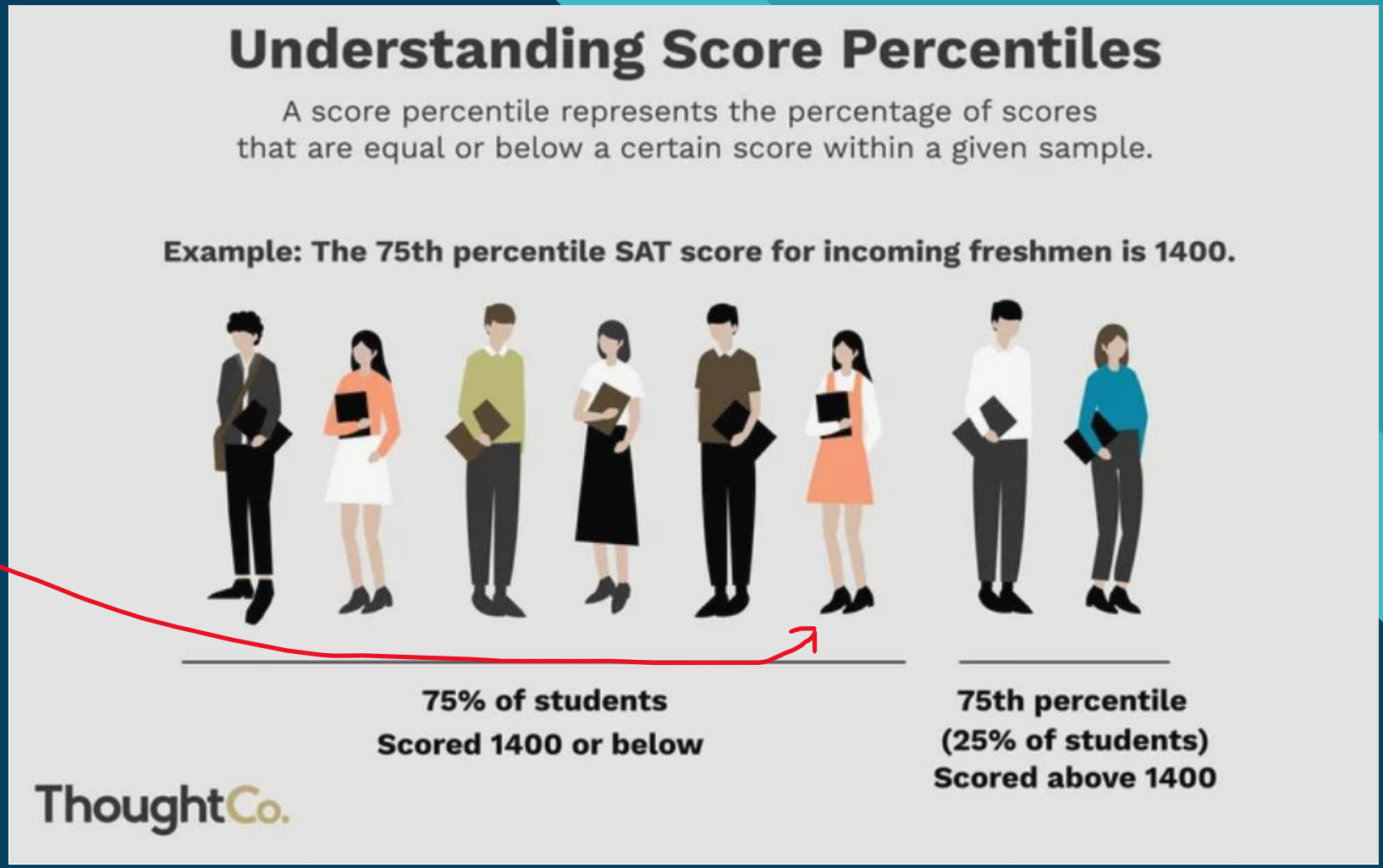
Percentile



Source: https://www.thoughtco.com/how-to-understand-score-percentiles-3211610

Percentile

This student ranks at percentile 75%

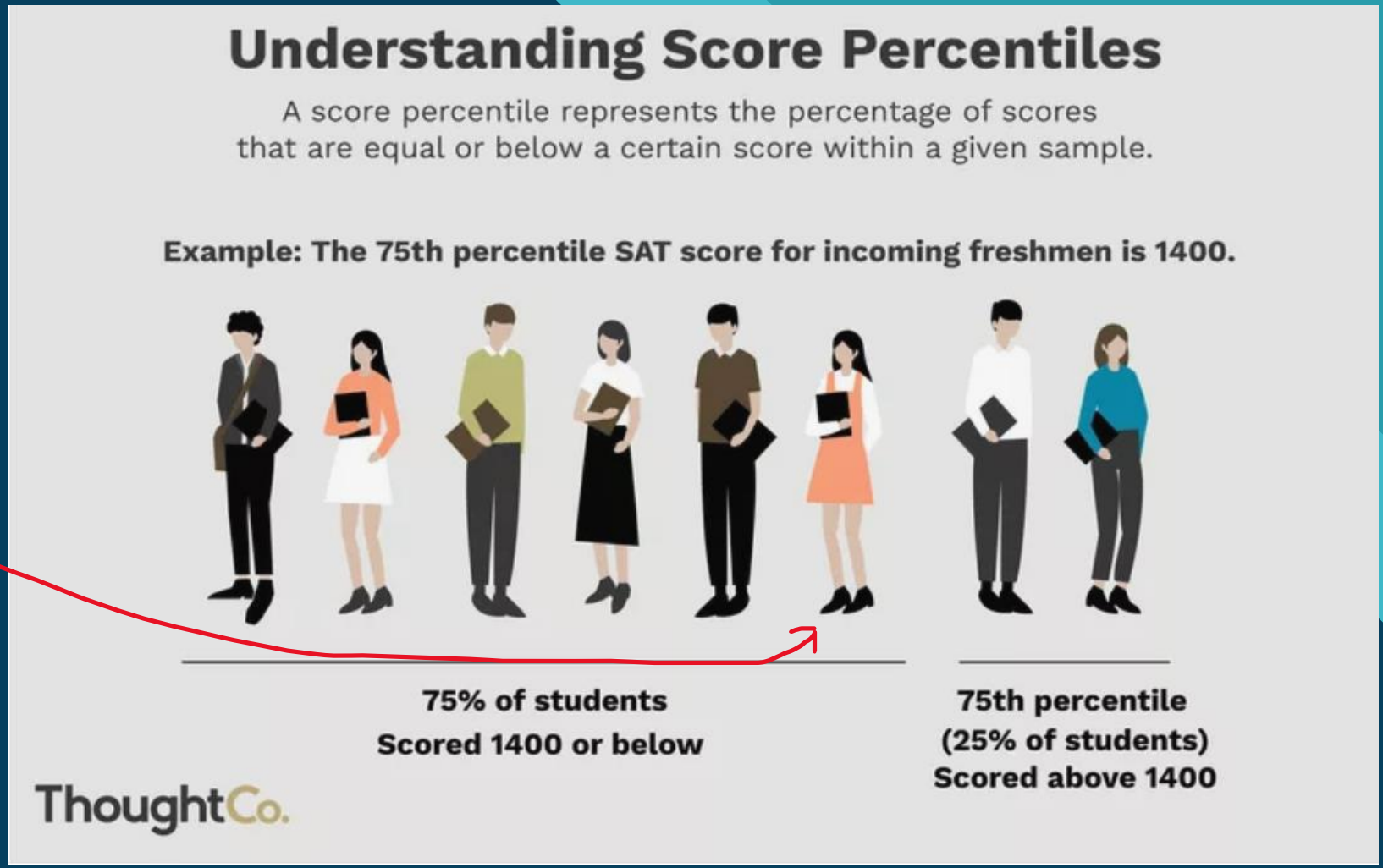


Source: https://www.thoughtco.com/how-to-understand-score-percentiles-3211610

Percentile

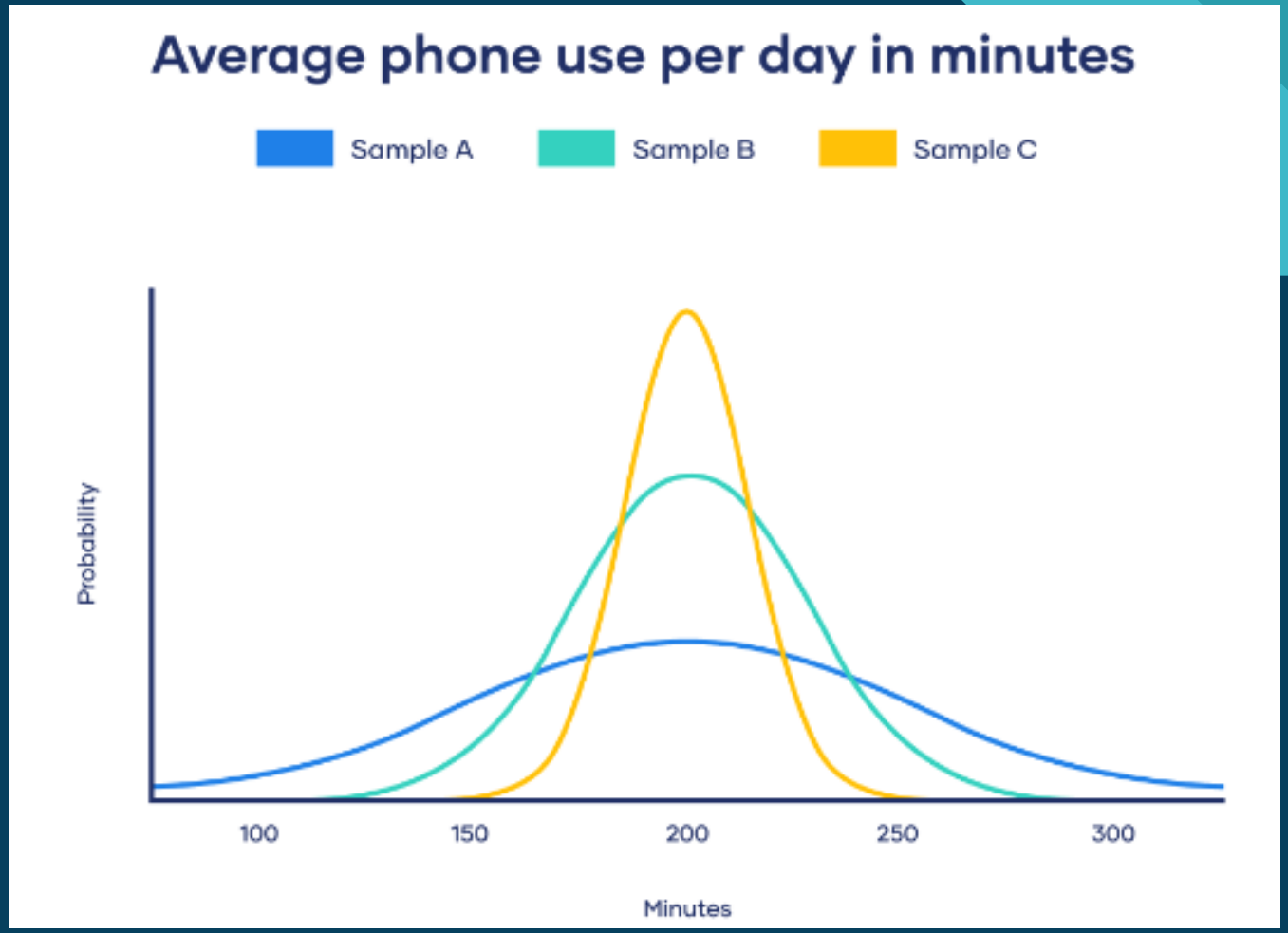
This student ranks at percentile 75%

Which student is at percentile 87.5%?



Source: https://www.thoughtco.com/how-to-understand-score-percentiles-3211610

Measures of variation



Source: https://www.scribbr.com/statistics/variability/#:~:text=Range%3A%20the%20difference%20between%20the,squared%20distances%20from%20the%20mean

Measures of variation

- Range: max value – min value
  - Useful for numeric variables

Source: https://study.com/academy/lesson/measures-of-variability-range-variance-standard-deviation.html

Measures of variation

- Range: max value – min value
  - Useful for numeric variables

$$Range = Max - Min$$

Source: https://study.com/academy/lesson/measures-of-variability-range-variance-standard-deviation.html

Measures of variation

- Range: max value – min value
  - Useful for numeric variables

$$Range = Max - Min$$

- Variance
  - Useful for numeric variables

$$V = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Source: https://study.com/academy/lesson/measures-of-variability-range-variance-standard-deviation.html

Measures of variation

- Range: max value – min value
  - Useful for numeric variables

$$Range = Max - Min$$

- Variance
  - Useful for numeric variables

$$V = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Standard deviation
  - Useful for numeric variables

$$SD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

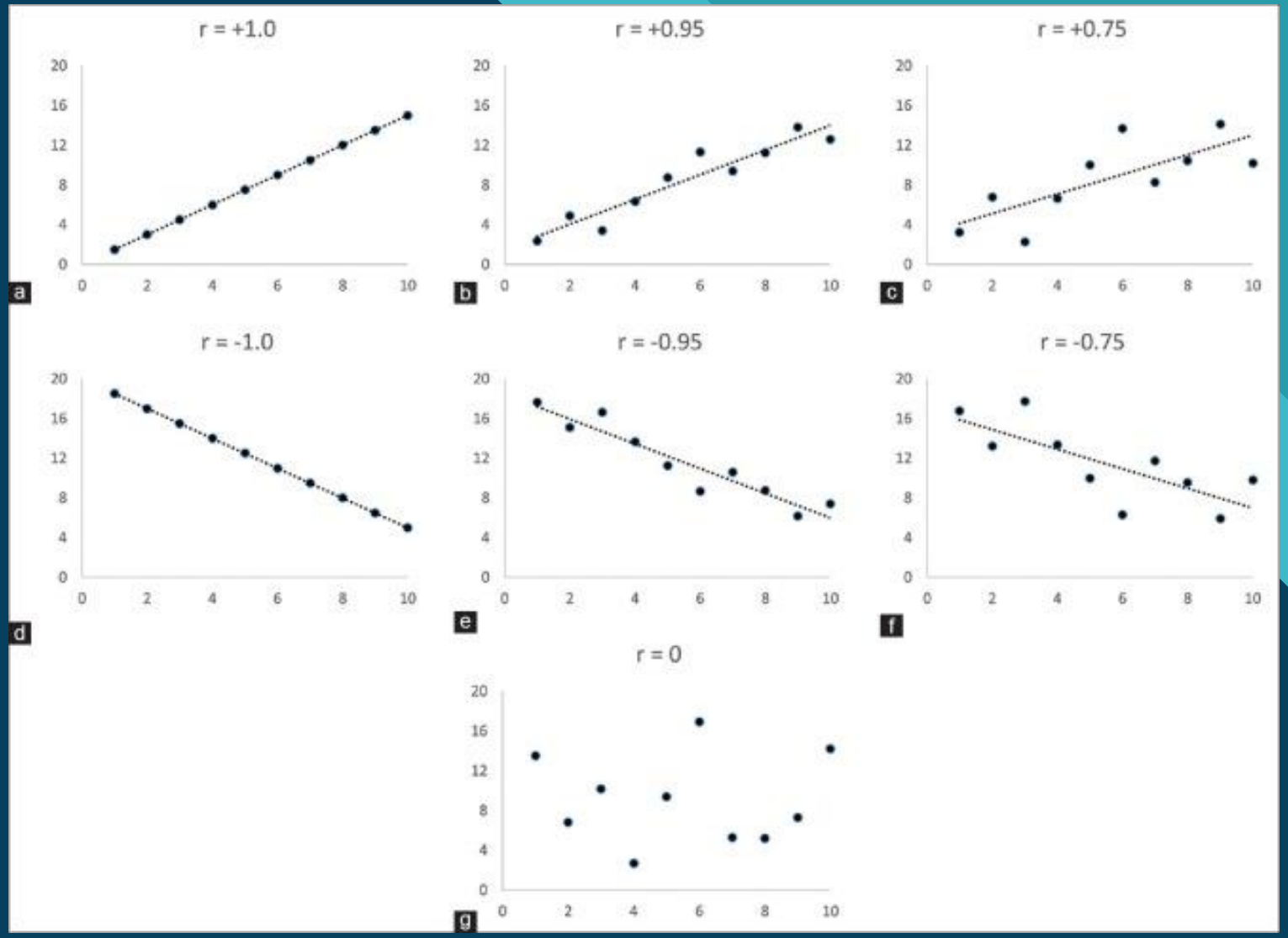
Source: https://study.com/academy/lesson/measures-of-variability-range-variance-standard-deviation.html

Measures of association

- Pearson's correlation
  - Most common way for measuring association between 2 numeric variables
  - Measure linear association between 2 variables
  - Always between -1 and 1
    - -1: perfect negative association
    - 0: no linear association
    - 1: perfect positive association
  - Correlation ≠ causation

Measures of association

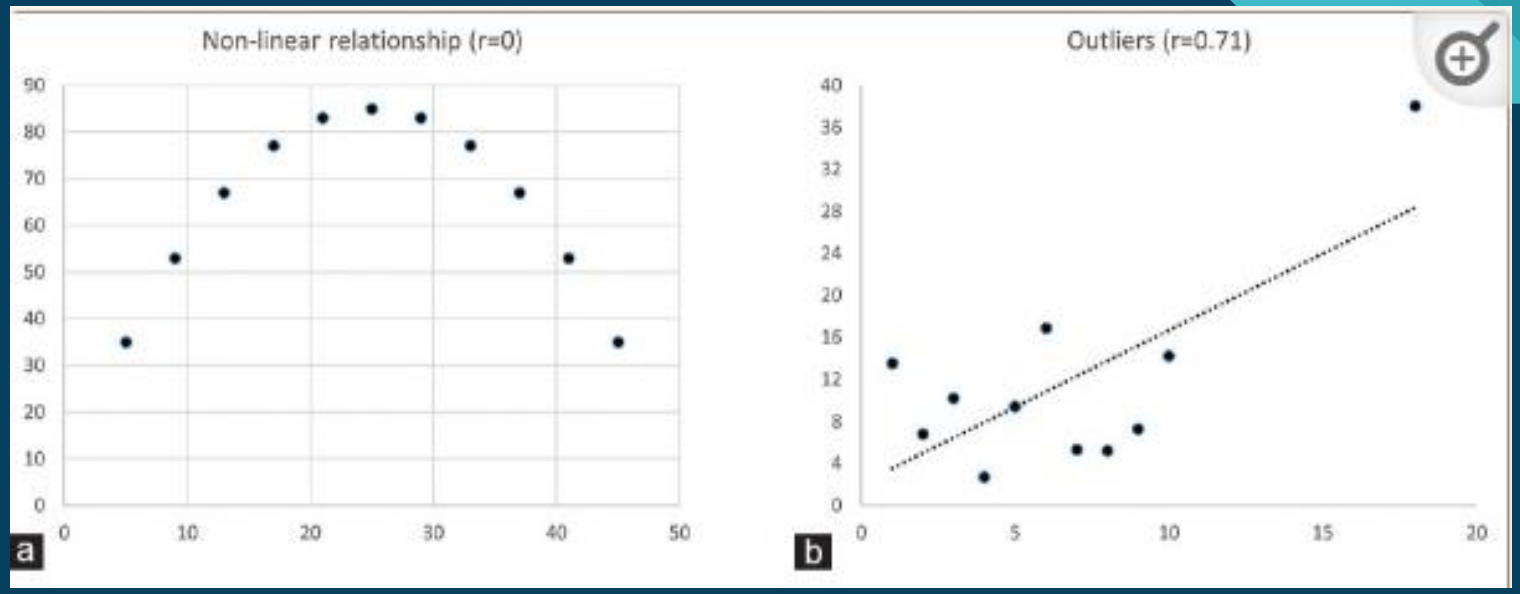
- Pearson's correlation – Examples



Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5079093/

Measures of association

- Pearson's correlation – Limitations



Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5079093/

https://en.wikipedia.org/wiki/Simpson%27s_paradox

Measures of association

- Correlation ≠ causation

Source:
https://www.biostat.jhsph.edu/courses/bio621/misc/Chocolate%20consumption%20cognitive%20function%20and%20nobel%20laurates%20(NEJM).pdf

Measures

- Correlation ≠

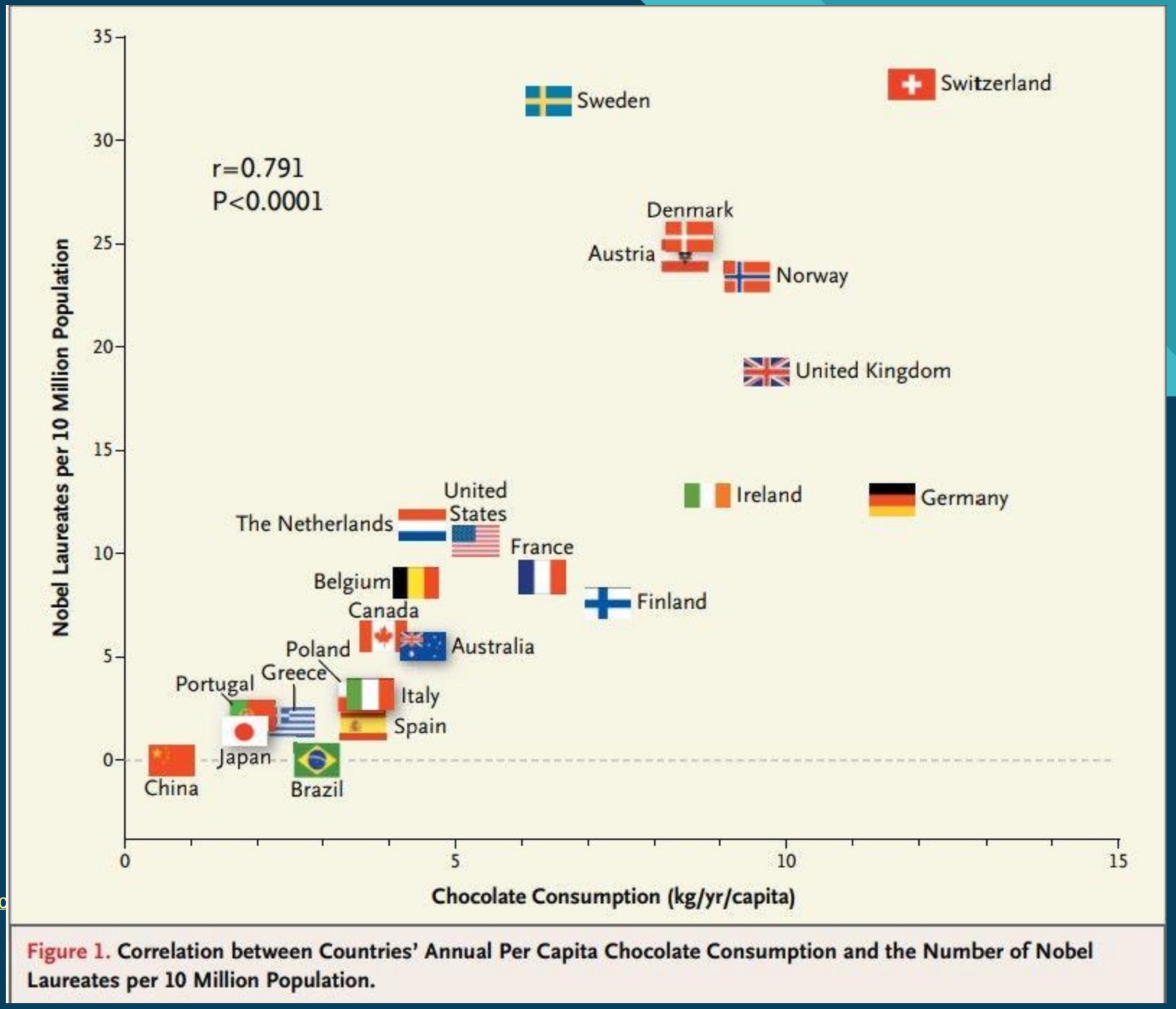


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Source: https://www.biostat.jhsph.edu/c pdf

Simpson's paradox

- University of Berkley – Admission rates

Department	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
Total	**4526**	**39%**	**2691**	**45%**	**1835**	**30%**

Legend:

greater percentage of successful applicants than the other gender

greater number of applicants than the other gender

bold - the two 'most applied for' departments for each gender

Source: https://en.wikipedia.org/wiki/Simpson%27s_paradox

https://www.linkedin.com/posts/awadelrahman_paradox-simpsons-datascience-activity-6979074241333735424-vaH8/?utm_source=share&utm_medium=member_ios

Simpson's paradox

- University of Berkley – Admission rates

Department	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
A	933	64%	**825**	62%	108	82%
B	585	63%	**560**	63%	25	68%
C	918	35%	325	37%	**593**	34%
D	792	34%	417	33%	375	35%
E	584	25%	191	28%	**393**	24%
F	714	6%	373	6%	341	7%
Total	**4526**	**39%**	**2691**	**45%**	**1835**	**30%**

Legend:

greater percentage of successful applicants than the other gender

greater number of applicants than the other gender

bold - the two 'most applied for' departments for each gender

Source: https://en.wikipedia.org/wiki/Simpson%27s_paradox

https://www.linkedin.com/posts/awadelrahman_paradox-simpsons-datascience-activity-6979074241333735424-vaH8/?utm_source=share&utm_medium=member_ios

Simpson's paradox

- University of Be
  rates

	Women	
	Applicants	Admitted
	108	82%
	25	68%
	593	34%
	375	35%
	393	24%
	341	7%
	1835	**30%**

nder

Source: https://en.wikipedia.org/wik

https://www.linkedin.com/posts/awadelrahman_paradox-simpsons-datascience-activity-6979074241333735424-vaH8/?utm_source=share&utm_medium=member_ios

Discover Excel

Discover Excel

- Freeze columns + data view
- Conditional formatting
- Formulas, cell references, Locking cell references, formula extension
- Pivot table
- Import data
- VLOOKUP
- Analysis toolkit: descriptive statistics, correlation matrix
  - Configuration: https://www.ablebits.com/office-addins-blog/linear-regression-analysis-excel/#linear-regression-Excel-Analysis-ToolPak

- And much more … (not covered in this course)

Exercice with Excel

01-countries_iso.xlsx: tab ‘Warm-up’

- Let’s work together

01-countries_iso.xlsx: tab ‘countries iso’

- On your own or by 2
- How data were brought to excel from the website?
- Fill each column with the right formula

02-countries_ue.txt

- Read instruction in 02-countries_ue-readme.txt

Discover Python

A word about programming languages

Jul 2023	Jul 2022	Change	Programming Language	Ratings	Change
1	1		Python	13.42%	-0.01%
2	2		C	11.56%	-1.57%
3	4	↑	C++	10.80%	+0.79%
4	3	↓	Java	10.50%	-1.09%
5	5		C#	6.87%	+1.21%
6	7	↑	JavaScript	3.11%	+1.34%
7	6	↓	Visual Basic	2.90%	-2.07%
8	9	↑	SQL	1.48%	-0.16%
9	11	↑	PHP	1.41%	+0.21%
10	20	⇈	MATLAB	1.26%	+0.53%

Source: https://www.tiobe.com/tiobe-index/

A word about programming languages

In my opinion, most relevant programming language for a (business) data analyst:

(the following are also the starting point for any aspiring data scientist)

A word about programming languages

In my opinion, most relevant programming language for a (business) data analyst:

(the following are also the starting point for any aspiring data scientist)

- SQL
  - this is a must have nowadays and whatever your role is in the data area

A word about programming languages

In my opinion, most relevant programming language for a (business) data analyst:

(the following are also the starting point for any aspiring data scientist)

- SQL
  - this is a must have nowadays and whatever your role is in the data area
- Python
  - Most popular programming language: free, wide adoption/community, multi purpose, …

A word about programming languages

In my opinion, most relevant programming language for a (business) data analyst:

(the following are also the starting point for any aspiring data scientist)

- SQL
  - this is a must have nowadays and whatever your role is in the data area
- Python
  - Most popular programming language: free, wide adoption/community, multi purpose, …
- (optional) SAS and/or R: can be more important depending of your sector / usage of your data / specific techniques to apply

Discover Python

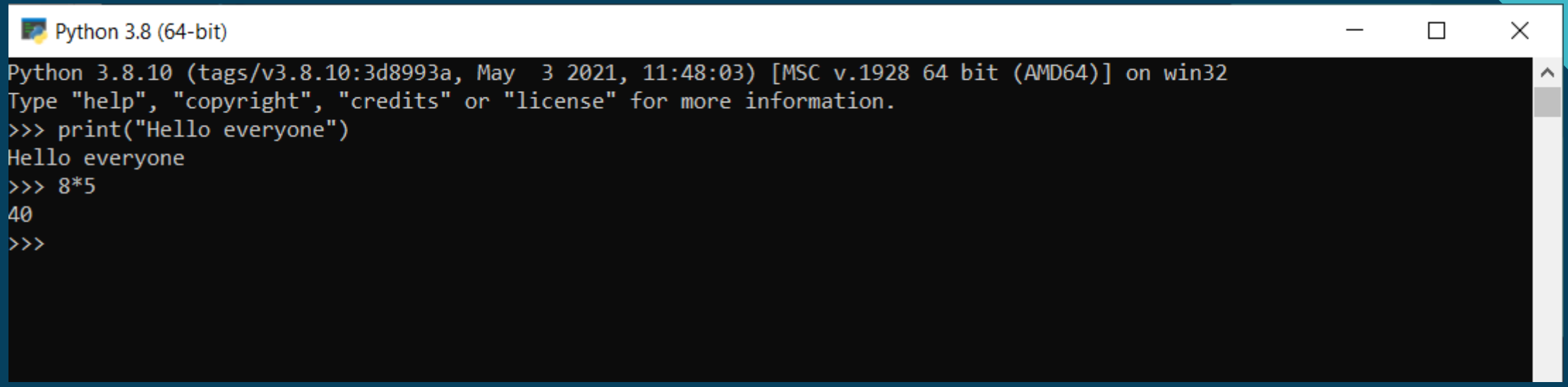
- Basic of the language
  - Number + math operations
  - String
  - List
  - Dictionary
  - Looping
- Compute descriptive statistics
- Pandas

- And much more (not covered in this course)

Discover Python

Environment where you can learn Python:

- Locally
  - Python installed locally without IDE

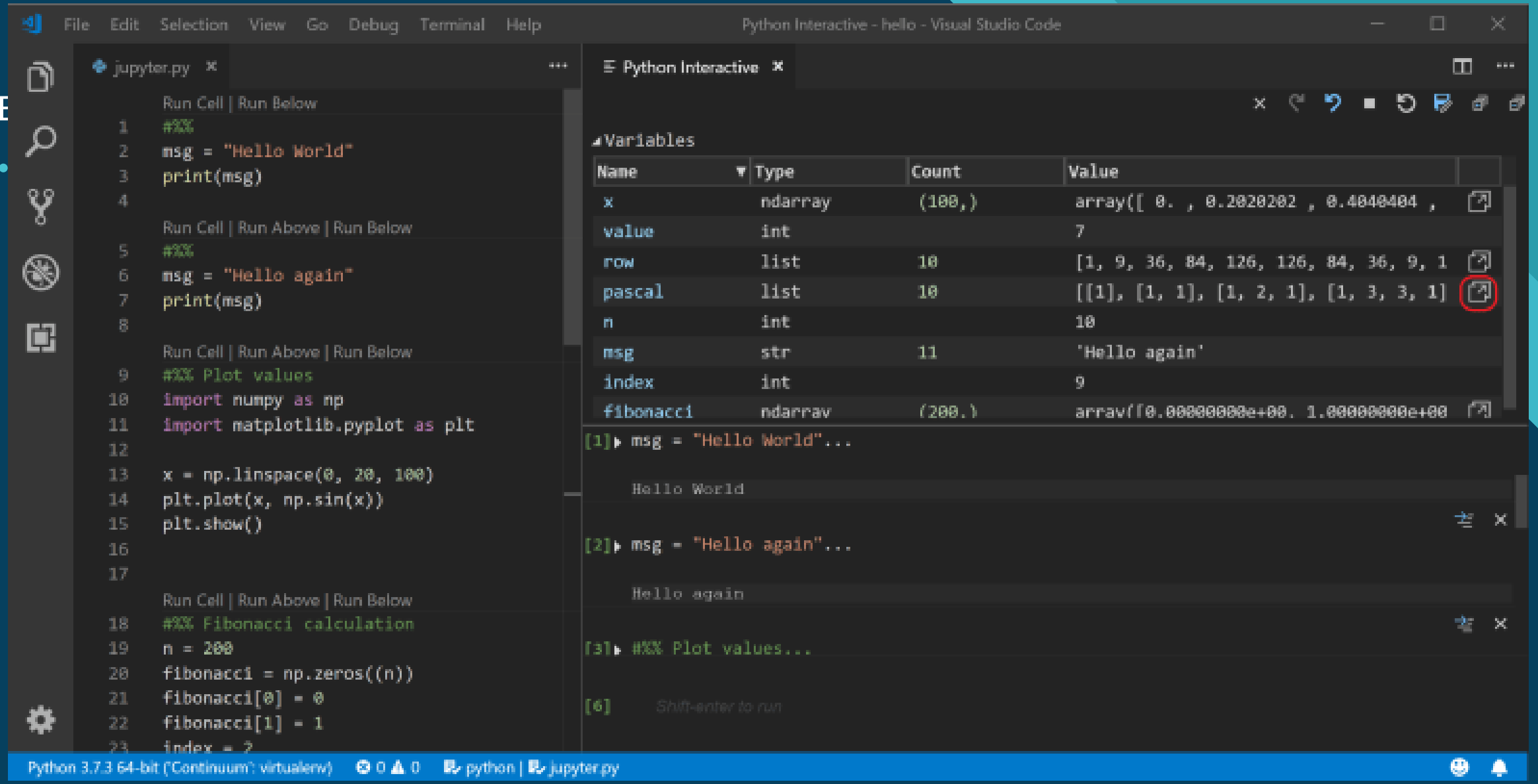


Discover Python

Environment where you can learn Python:

- Locally
  - Python installed locally with IDE

Discover Python



Discover Python

- Locally
  - Anaconda + virtual environment
    - + Spyder

Discover Python

- Loca
  - A

Spyder (Python 3.9)

File Edit Search Source Run Debug Consoles Projects Tools View Help

C:\Users\cguilmin\OneDrive - Business & Decision Europe\Documents\Nerdlandfestival\People-Counting-in-Real-Time

...ve - Business & Decision Europe\Documents\Nerdlandfestival\People-Counting-in-Real-Time\cut_video.py

Run.py × cut_video.py × testing.py × inference on images.py × untitled0.py* ×

```python
# -*- coding: utf-8 -*-
"""
Created on Fri Apr 15 15:33:14 2022

@author: cguilmin
"""

import cv2

file = ".\\videos\\example_01.mp4"
seconds = [(33,38)]



cap = cv2.VideoCapture(file)
ret, frame = cap.read()
h, w, _ = frame.shape

fps = cap.get(cv2.CAP_PROP_FPS)
parts = [(part1 * fps, part2 * fps) for part1, part2 in seconds]

fourcc = cv2.VideoWriter_fourcc(*"XVID")
writers = [cv2.VideoWriter(f"part{start}-{end}.mp4", fourcc, fps, (w, h)) \
           for start, end in seconds]

f = 0
while ret:
    f += 1
    for i, part in enumerate(parts):
        start, end = part
        if start <= f <= end:
            writers[i].write(frame)
    ret, frame = cap.read()

for writer in writers:
    writer.release()

cap.release()
```

Name	Type	Size	Value
cap	VideoCapture	1	VideoCapture object of cv2 module
end	float	1	1140.0
f	int	1	1283
file	str	23	.\videos\example_01.mp4
fourcc	int	1	1145656920
fps	float	1	30.0
frame	NoneType	1	NoneType object
h	int	1	300
i	int	1	0
part	tuple	2	(990.0, 1140.0)
parts	list	1	[(990.0, 1140.0)]
ret	bool	1	False
seconds	list	1	[(33, 38)]

Help Variable Explorer Plots Files

Console 1/A ×

```
Python 3.9.12 (main, Apr  4 2022, 05:22:27) [MSC v.1916 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 8.2.0 -- An enhanced Interactive Python.

In [1]: runfile('C:/Users/cguilmin/OneDrive - Business & Decision Europe/Documents/
Nerdlandfestival/People-Counting-in-Real-Time/cut_video.py', wdir='C:/Users/cguilmin/
OneDrive - Business & Decision Europe/Documents/Nerdlandfestival/People-Counting-in-
Real-Time', current_namespace=True)

In [2]:
```

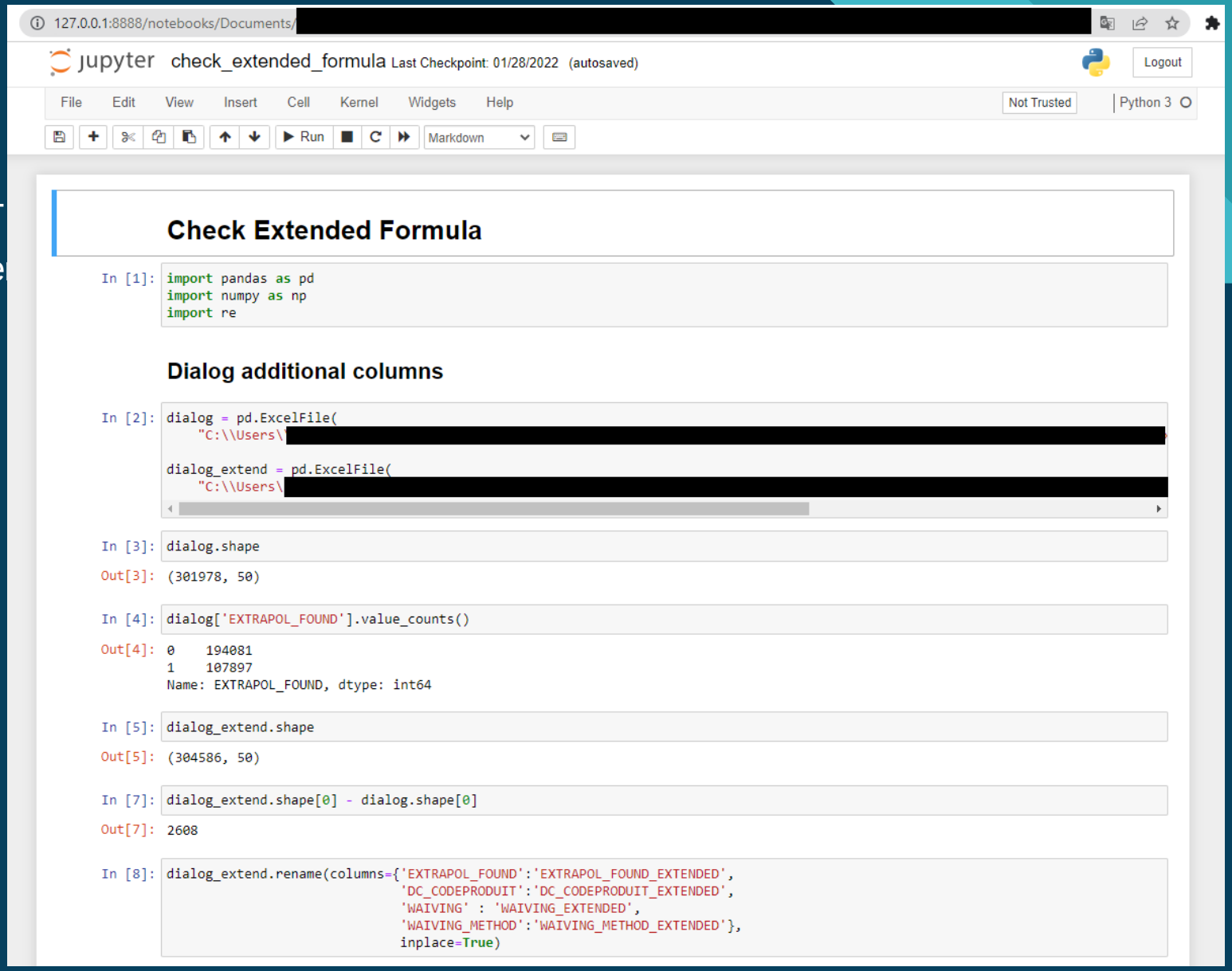
IPython console History

Discover Python

- Locally
  - Anaconda + virtual environment
    - + Jupyter

Discover Python

- Locally
  - Anaconda +
    - + Jupyte



Discover Python

Environment where you can learn Python:

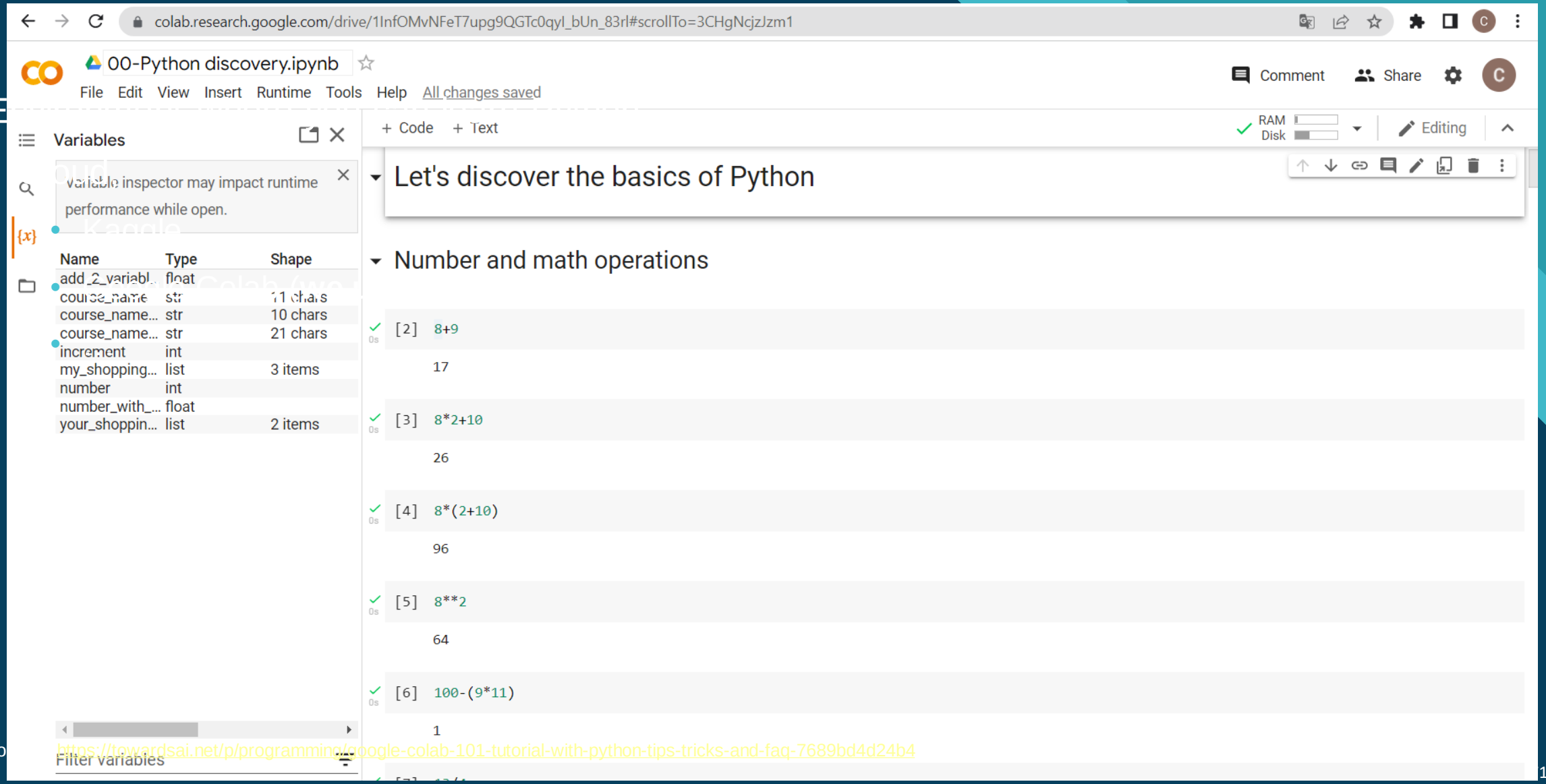
- Cloud,
  - Kaggle
  - Google Colab (**we use this one in this course**)
  - …

Source : https://towardsai.net/p/programming/google-colab-101-tutorial-with-python-tips-tricks-and-faq-7689bd4d24b4

Discover Python

E

-



So

https://towardsai.net/p/programming/google-colab-101-tutorial-with-python-tips-tricks-and-faq-7689bd4d24b4

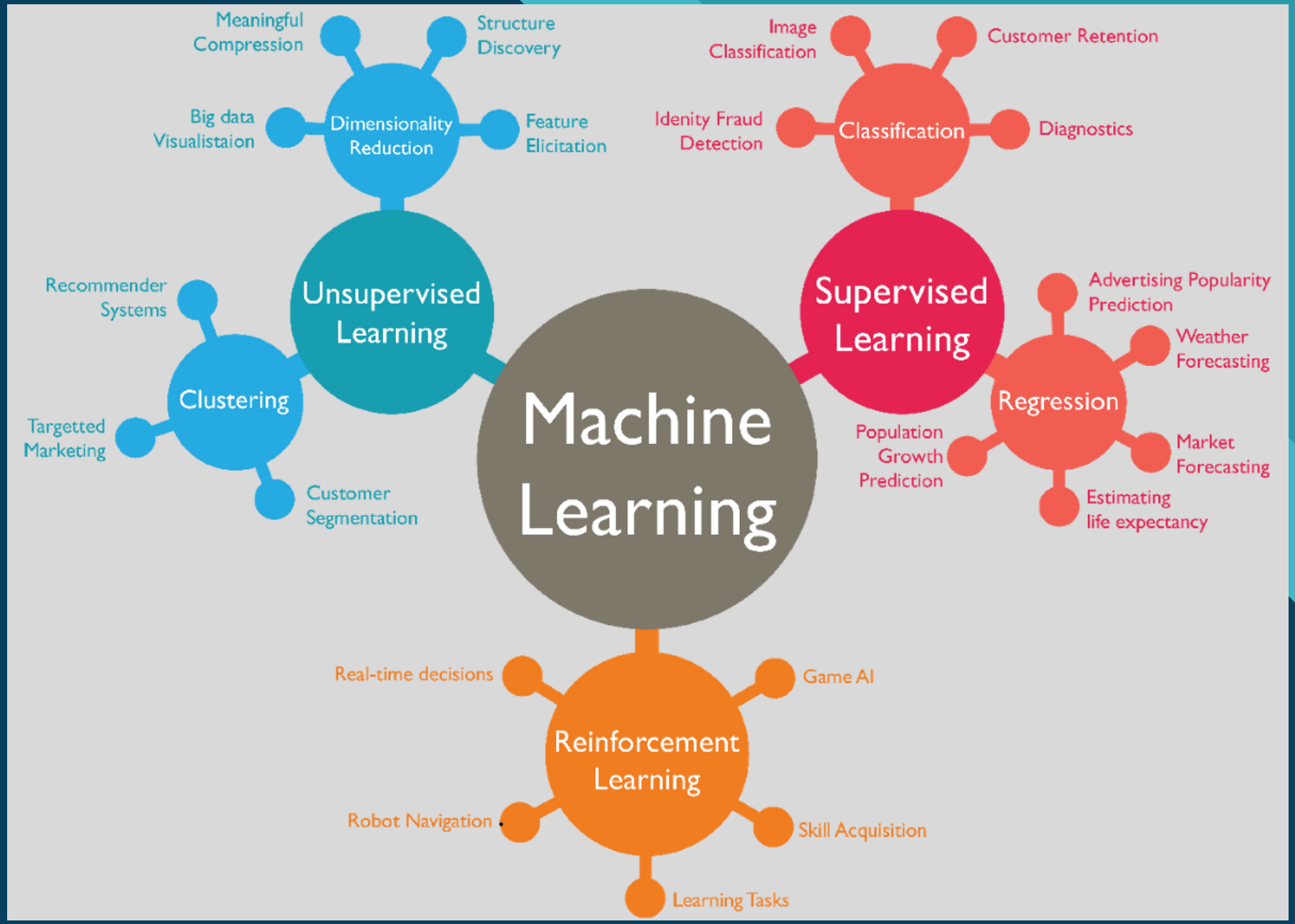
Discover Python

Exercice on Google Colab

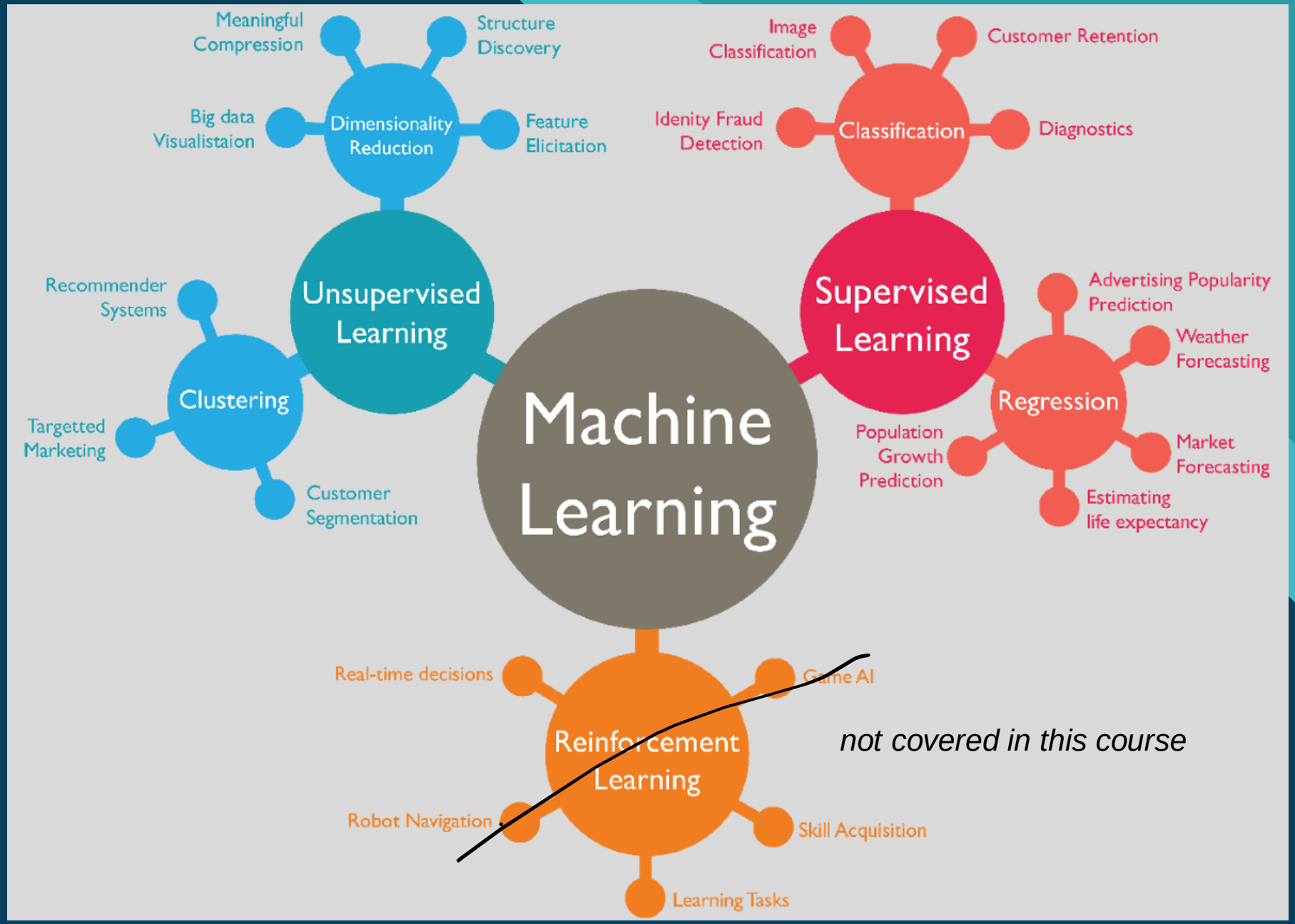
- 00-Python discovery: discovery the basic of python programming language
  - https://github.com/cedricguilmin/datamining/blob/main/00_Python_discovery.ipynb
- 10-Pandas discovery.ipynb: discover the fascinating pandas library, the ‘go-to’ library when you want to manipulate and extract datasets

Advanced techniques

Some vocabulary



Some vocabulary



Ressources

- Mostly 1 reference book used
  - "Data Science - Concepts and Pratice", V. Kotu and B. Deshpande, 2018, 2nd Edition, ISBN: 9780128147610.

- http://www.introdatascience.com/
  - Overview of each chapter
  - Slides about main concepts
  - RapidMiner flows + data

Tools

To perform these advanced techniques, we will use RapidMiner Studio

- Education licence can be found here: https://rapidminer.com/platform/educational/

Tools

To

RAPIDMINER

Platform Solutions Why RapidMiner Resource Center About Us Account Request a Demo

Home > Data Science Platform > RapidMiner for Academics

RapidMiner for Academics

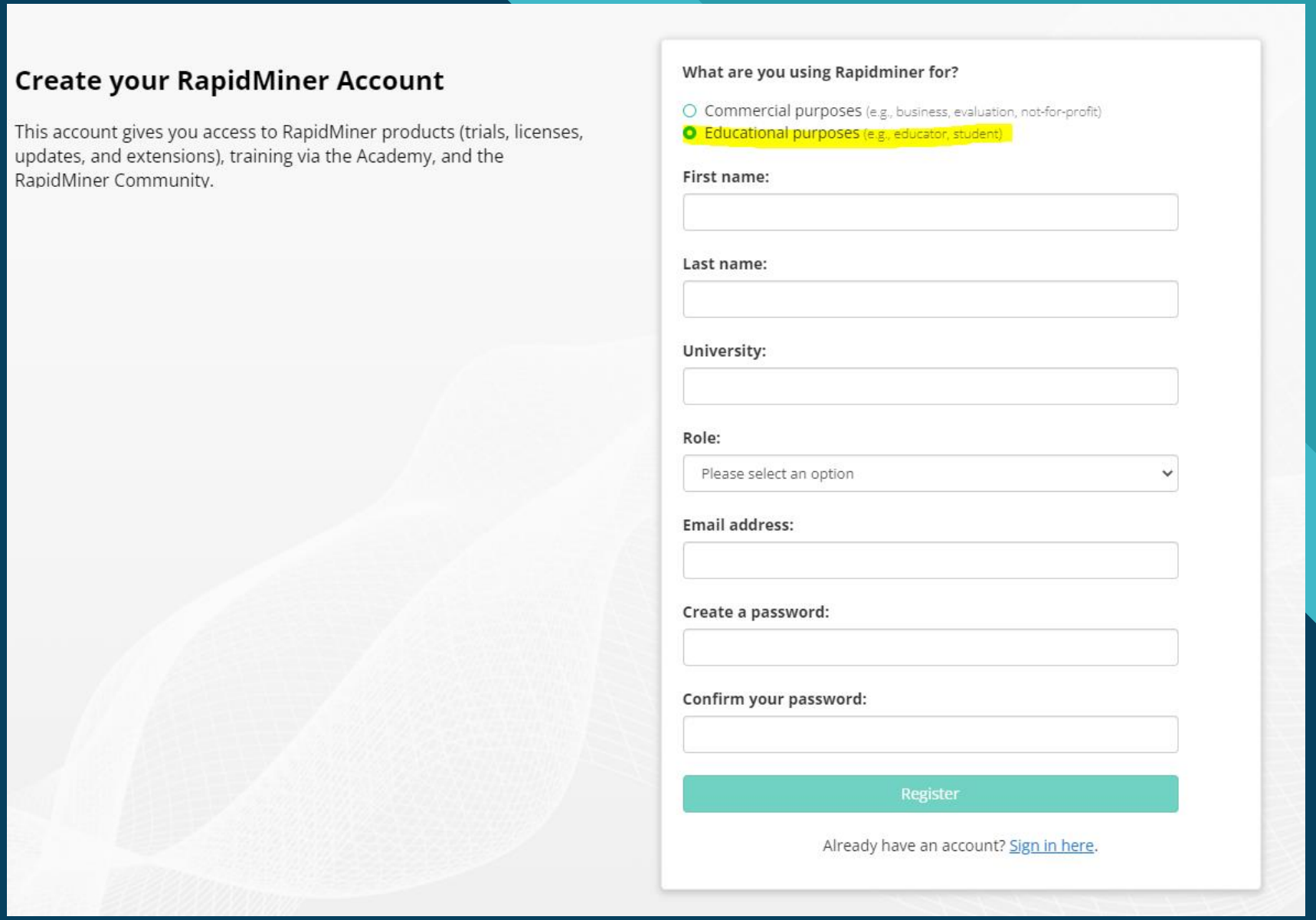
RapidMiner Studio is available for academics looking for an end-to-end data science platform for instructional or research purposes.

Download Studio

MANNING SCHOOL OF BUSINESS UMASS LOWELL NYU SCHOOL OF PROFESSIONAL STUDIES Middlesex University London DEAKIN UNIVERSITY WKU Baruch COLLEGE ZICKLIN SCHOOL OF BUSINESS

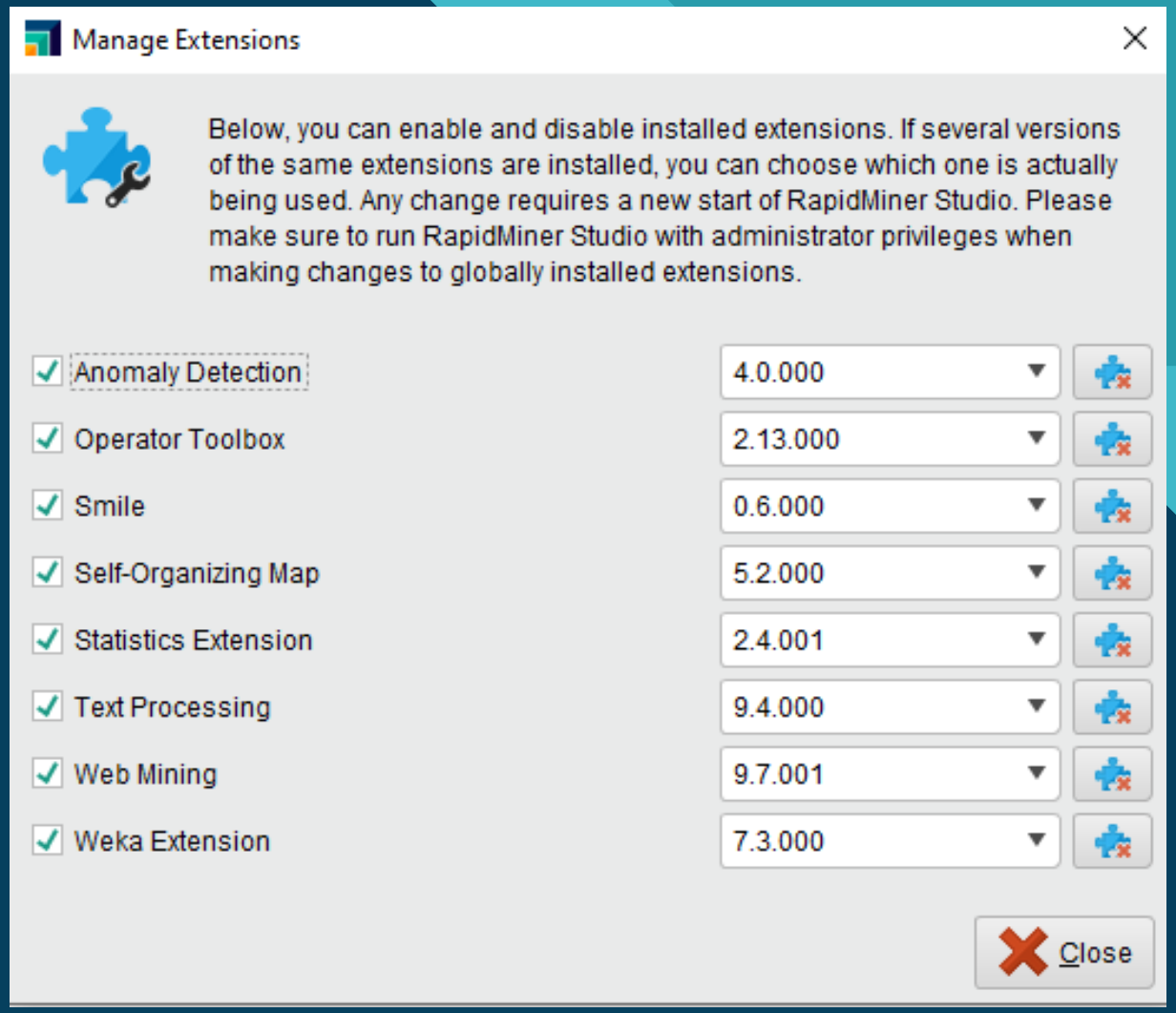
Tools

Make sure to select ‘Educational purposes’

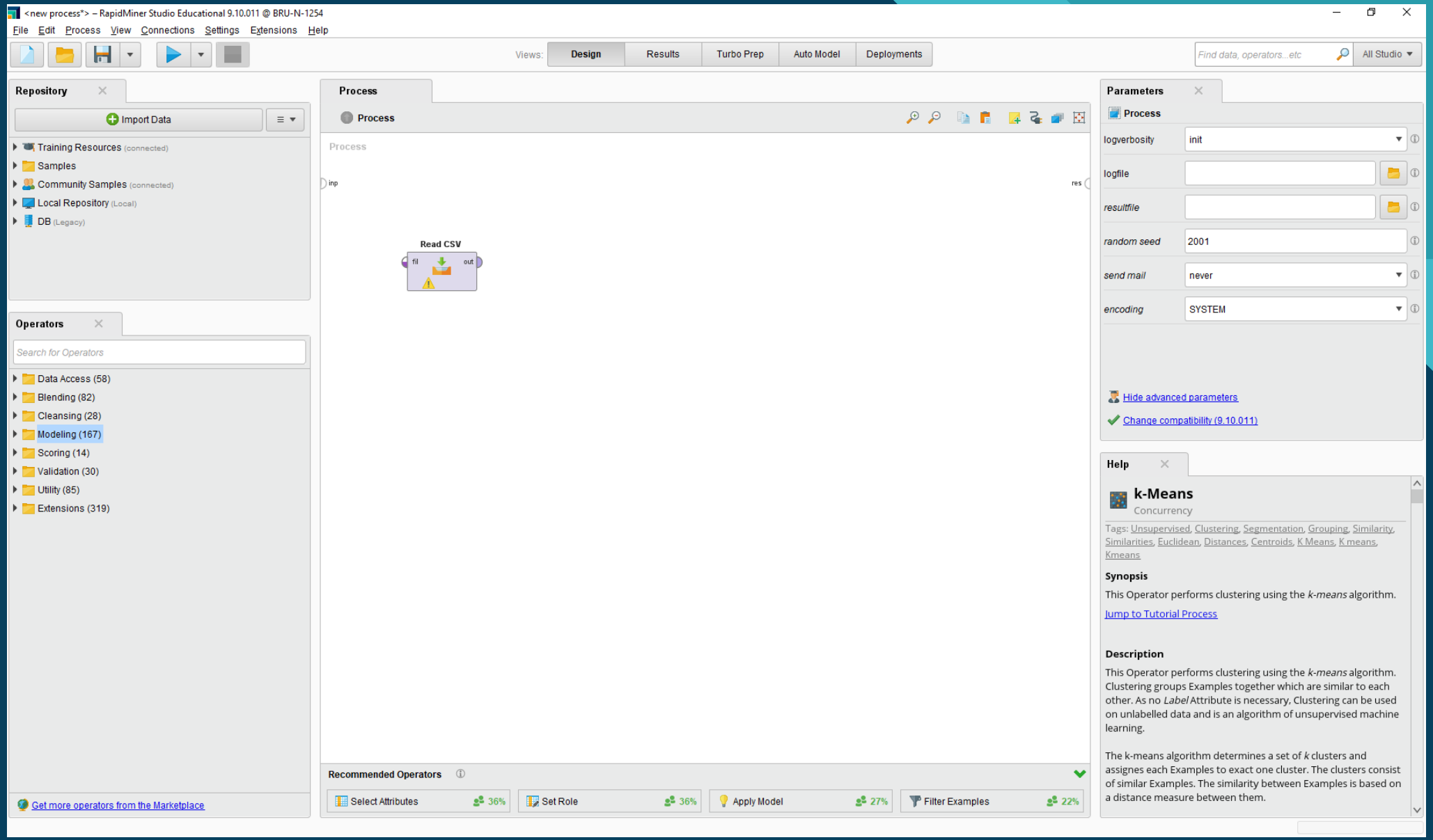


Tools

Once installed, you can install the following extensions to supercharge your RapidMiner!



Tools


<new process*> – RapidMiner Studio Educational 9.10.011 @ BRU-N-1254
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc All Studio
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
DB (Legacy)
Operators
Search for Operators
Data Access (58)
Blending (82)
Cleansing (28)
Modeling (167)
Scoring (14)
Validation (30)
Utility (85)
Extensions (319)
Get more operators from the Marketplace
Process
Process
Process
inp
res
Read CSV
fil
out
Recommended Operators
Select Attributes 36%
Set Role 36%
Apply Model 27%
Filter Examples 22%
Parameters
Process
logverbosity init
logfile
resultfile
random seed 2001
send mail never
encoding SYSTEM
Hide advanced parameters
Change compatibility (9.10.011)
Help
k-Means
Concurrency
Tags: Unsupervised, Clustering, Segmentation, Grouping, Similarity, Similarities, Euclidean, Distances, Centroids, K Means, K means, Kmeans
Synopsis
This Operator performs clustering using the k-means algorithm.
Jump to Tutorial Process
Description
This Operator performs clustering using the k-means algorithm. Clustering groups Examples together which are similar to each other. As no Label Attribute is necessary, Clustering can be used on unlabelled data and is an algorithm of unsupervised machine learning.
The k-means algorithm determines a set of k clusters and assignes each Examples to exact one cluster. The clusters consist of similar Examples. The similarity between Examples is based on a distance measure between them.

Unsupervised techniques

Clustering

A word about clustering, aka cluster analysis / data segmentation

- Unsupervised techniques such that:
  - Similar observations within the same group
  - Dissimilar observations in different group

- Techniques types
  - Hierarchical clustering
  - Centroid based clustering (aka prototype based clustering)
  - Density based clustering
  - …

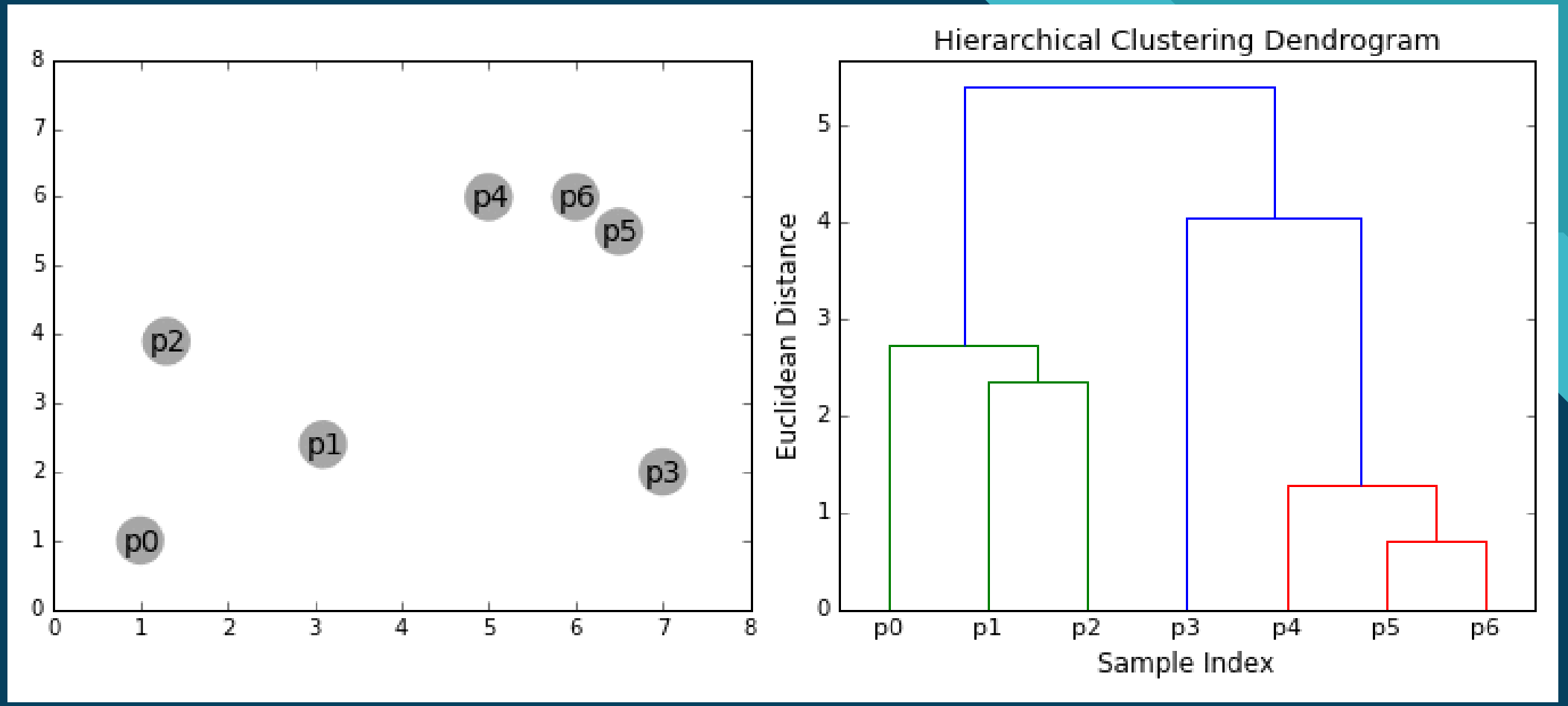
Clustering

- Applications:
  - Find group of similar customer to know who they are, how they interact with us, …
  - Propose similar product in a webshop
  - Find similar people in a social network
  - Find outlier(s) in data
  - Data summarization / reduction
  - …

Hierachical clustering

Source: https://dashee87.github.io/data%20science/general/Clustering-with-Scikit-with-GIFs/

Hierachical clustering



Source: https://dashee87.github.io/data%20science/general/Clustering-with-Scikit-with-GIFs/

Clustering

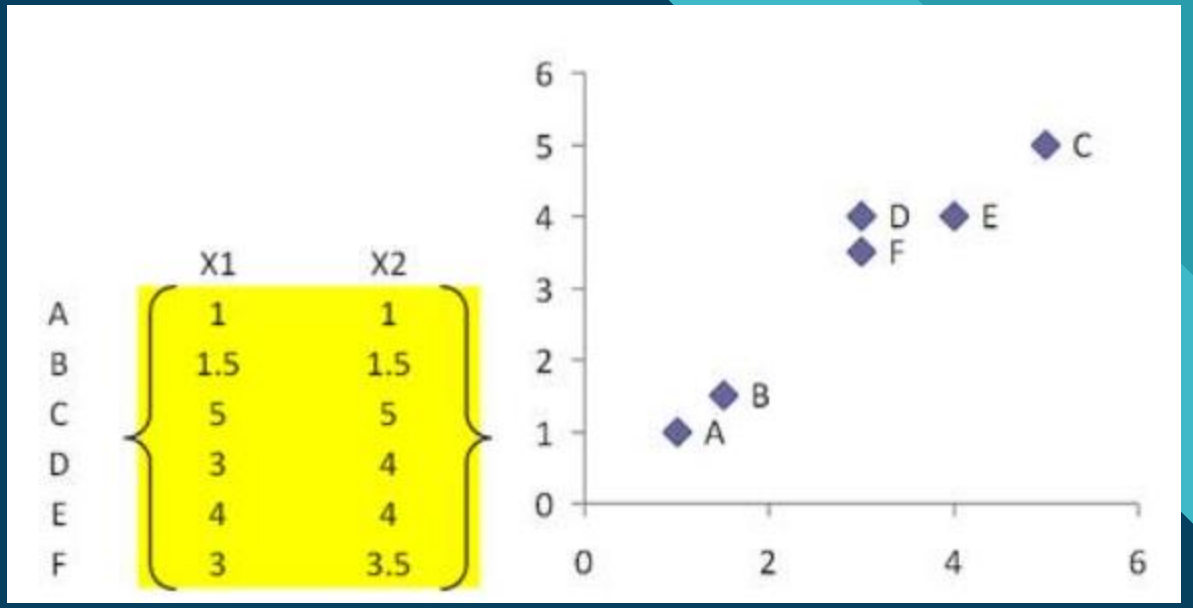
How to measure distance between 2 observations?

Euclidian distance

$$d_{euc}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Source: https://www.datanovia.com/en/lessons/clustering-distance-measures/#:~:text=The%20classification%20of%20observations%20into,a%20dissimilarity%20or%20distance%20matrix.

Clustering



Distance matrix

Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

Source: https://people.revoledu.com/kardi/tutorial/Clustering/Distance%20Matrix.htm

https://people.revoledu.com/kardi/tutorial/Clustering/Distance%20Matrix.htm

Clustering

Type of variables impact the measure we should use

- Only numerical variable => Euclidian distance, Manhattan distance, Minkowski distance, …
- Only nominal variables => simple matching, …
  - Only binary data: Hamming distance, Jaccard distance, …
- Mixed variables types
  - https://arxiv.org/pdf/cs/9701101.pdf

Clustering

Some measure can/must be standardized before applying more complex techniques (clustering, LOESS, principal composent analysis, regression, …)

Most commun way of standardization of a variable

$$x_i \Rightarrow value\ of\ variable\ x\ for\ observation\ i$$

$$mean_x \Rightarrow mean\ of\ variable\ x\ for\ all\ observations$$

$$sd_x \Rightarrow standard\ deviation\ of\ variable\ x\ for\ all\ obervations$$

Source: https://www.datanovia.com/en/lessons/clustering-distance-measures/#:~:text=The%20classification%20of%20observations%20into,a%20dissimilarity%20or%20distance%20matrix.

Other standardization/scaling techniques here: https://en.wikipedia.org/wiki/Feature_scaling

Clustering

Some measure can/must be standardized before applying more complex techniques (clustering, LOESS, principal composent analysis, regression, …)

Most commun way of standardization of a variable

$$x_i \Rightarrow value\ of\ variable\ x\ for\ observation\ i$$

$$mean_x \Rightarrow mean\ of\ variable\ x\ for\ all\ observations$$

$$sd_x \Rightarrow standard\ deviation\ of\ variable\ x\ for\ all\ obervations$$

$$x_i' = \frac{x_i - mean_x}{sd_x}$$

Source: https://www.datanovia.com/en/lessons/clustering-distance-measures/#:~:text=The%20classification%20of%20observations%20into,a%20dissimilarity%20or%20distance%20matrix.

Other standardization/scaling techniques here: https://en.wikipedia.org/wiki/Feature_scaling

Hierachical clustering

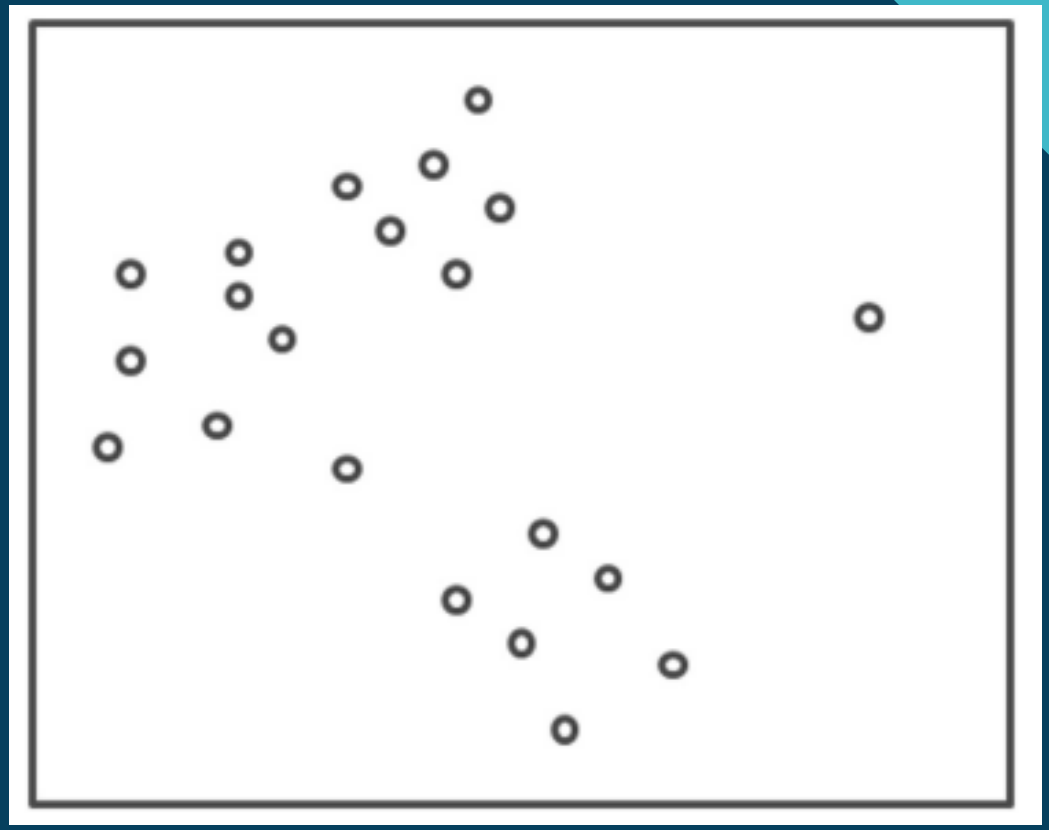
Pros

- No need to specify a priori the number of cluster we need. Based on dendogram, we can choose which number of clusters is relevant for our analysis
- Easy to understand
- Always the same result because no random initialization

Cons

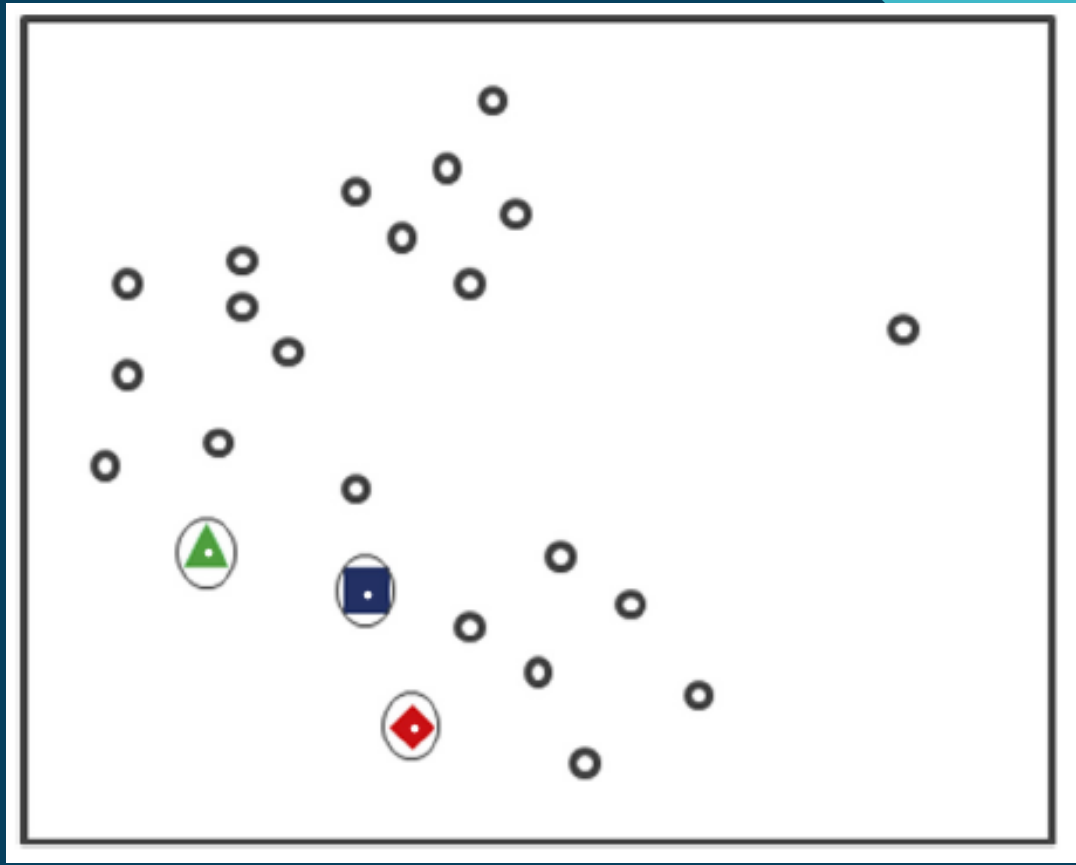
- Can be slow (for datasets with high number of records) because it has to compute the distance matrix

Centroid based clustering: k-means



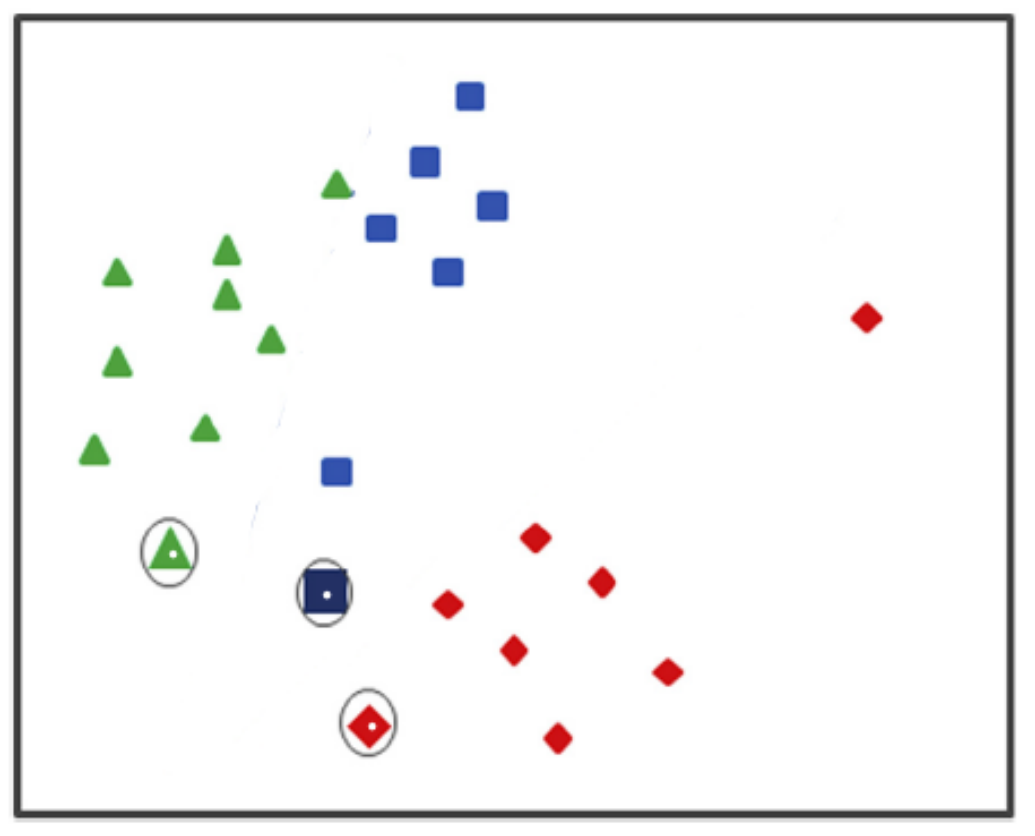
Source: "Data Science - Concepts and Pratice", V. Kotu and B. Deshpande, 2018, 2nd Edition, ISBN: 9780128147610.

Centroid based clustering: k-means



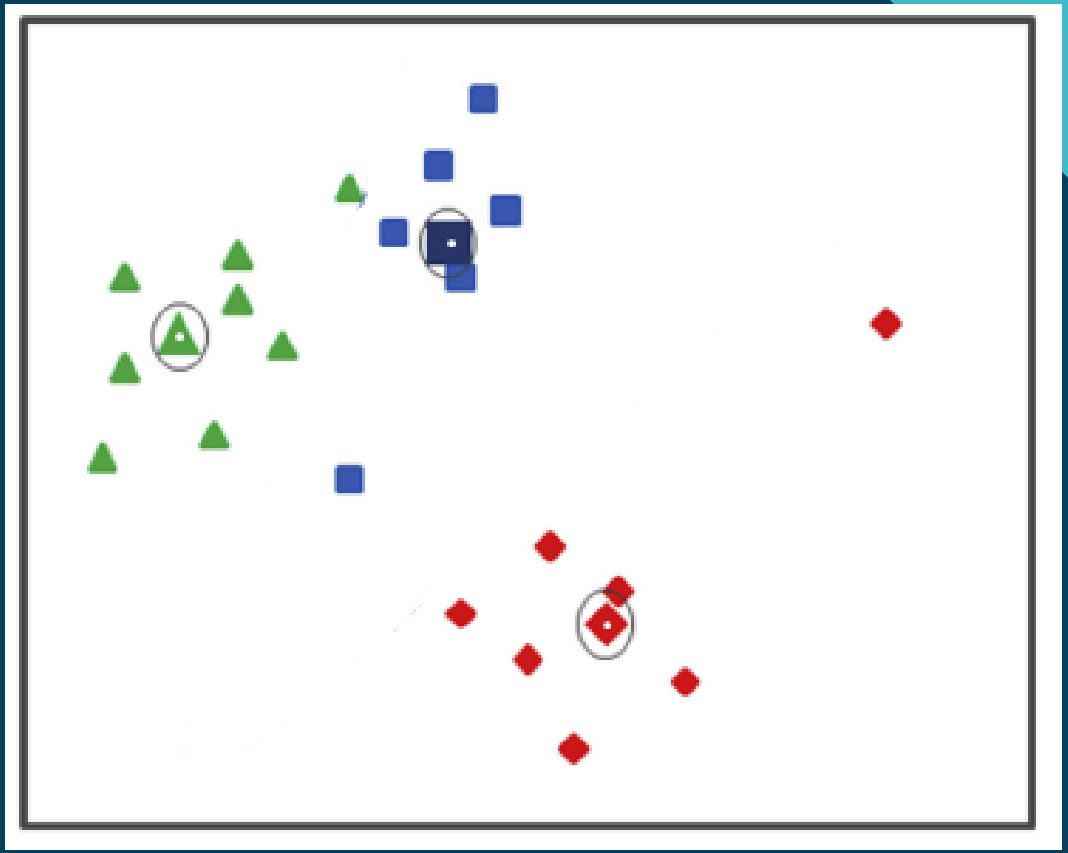
Source: "Data Science - Concepts and Pratice", V. Kotu and B. Deshpande, 2018, 2nd Edition, ISBN: 9780128147610.

Centroid based clustering: k-means



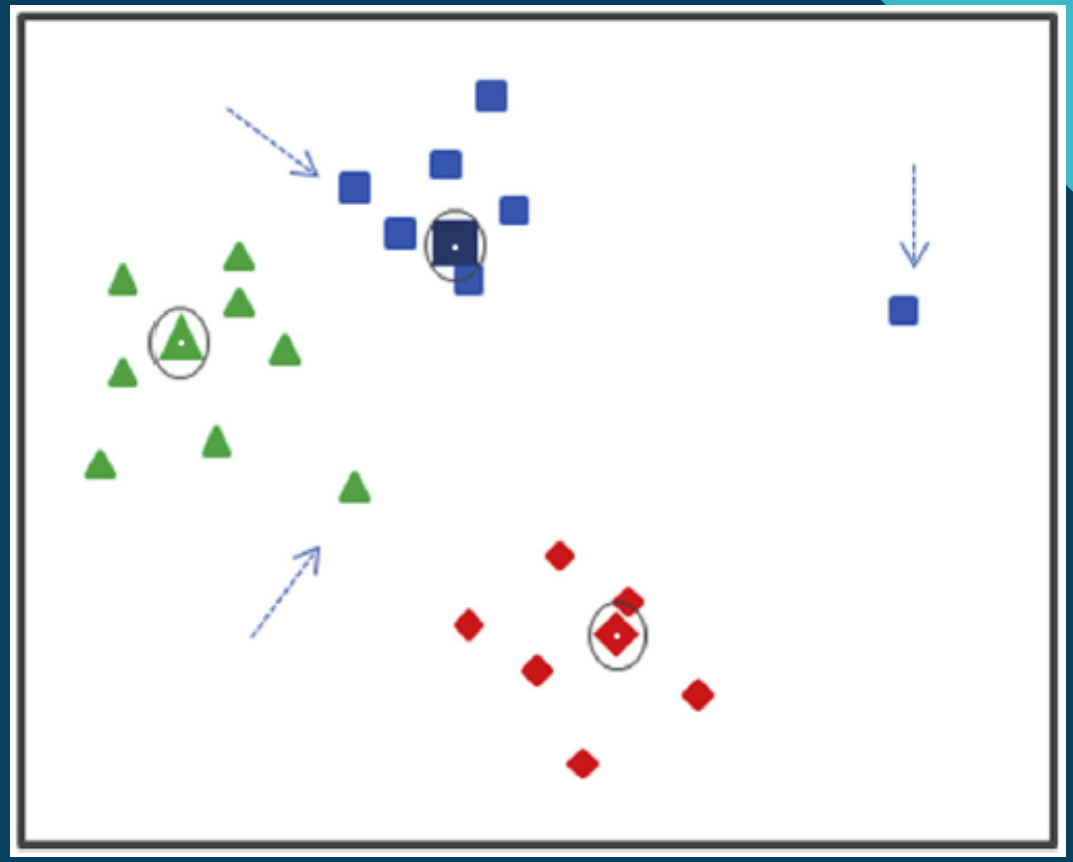
Source: "Data Science - Concepts and Pratice", V. Kotu and B. Deshpande, 2018, 2nd Edition, ISBN: 9780128147610.

Centroid based clustering: k-means



Source: "Data Science - Concepts and Pratice", V. Kotu and B. Deshpande, 2018, 2nd Edition, ISBN: 9780128147610.

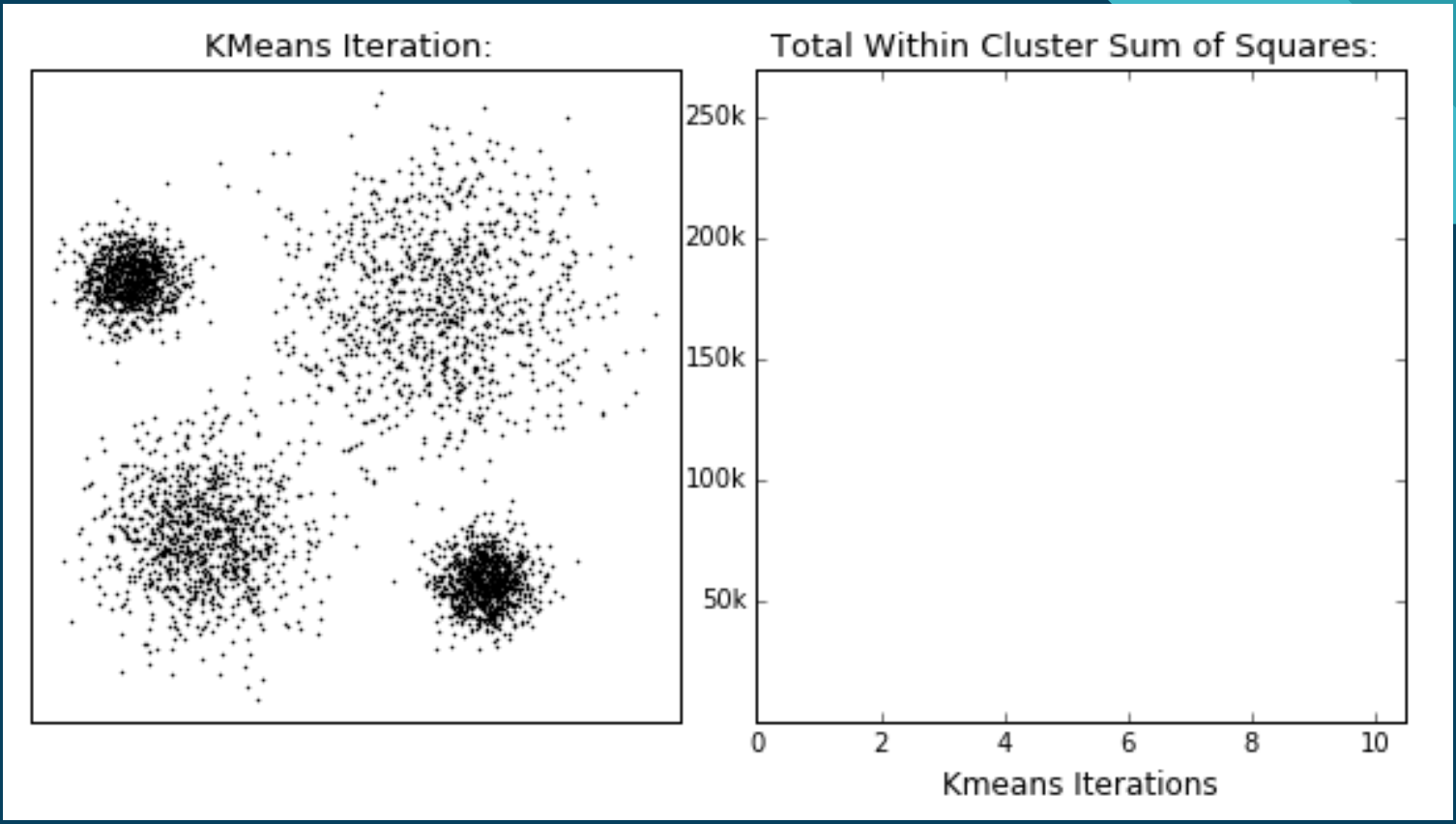
Centroid based clustering: k-means



Till we reach convergence …

Source: "Data Science - Concepts and Pratice", V. Kotu and B. Deshpande, 2018, 2nd Edition, ISBN: 9780128147610

Centroid based clustering: k-means



Source: https://dashee87.github.io/data%20science/general/Clustering-with-Scikit-with-GIFs/

Centroid based clustering: k-means

Pros

- Quicker because only distance to compute between centroid and points

Cons

- Specify a priori the number of cluster
- Not always providing the same result because of random choice for the starting centroid
    - Can be run multiple times

Ressource about k-means and implementation in Python: https://neptune.ai/blog/k-means-clustering

Centroid based clustering: k-means

How to choose the relevant number of cluster? = Which k value should I use?

General principle:

1) Iterate over the multiple value of k: for instance from 2 to 10, or 20, or …
    1) For each iteration with a specific value of k, record a specific metrics
2) Plot that specific metrics inr egards to the different value of k

Ressource: https://towardsdatascience.com/are-you-still-using-the-elbow-method-5d271b3063bd

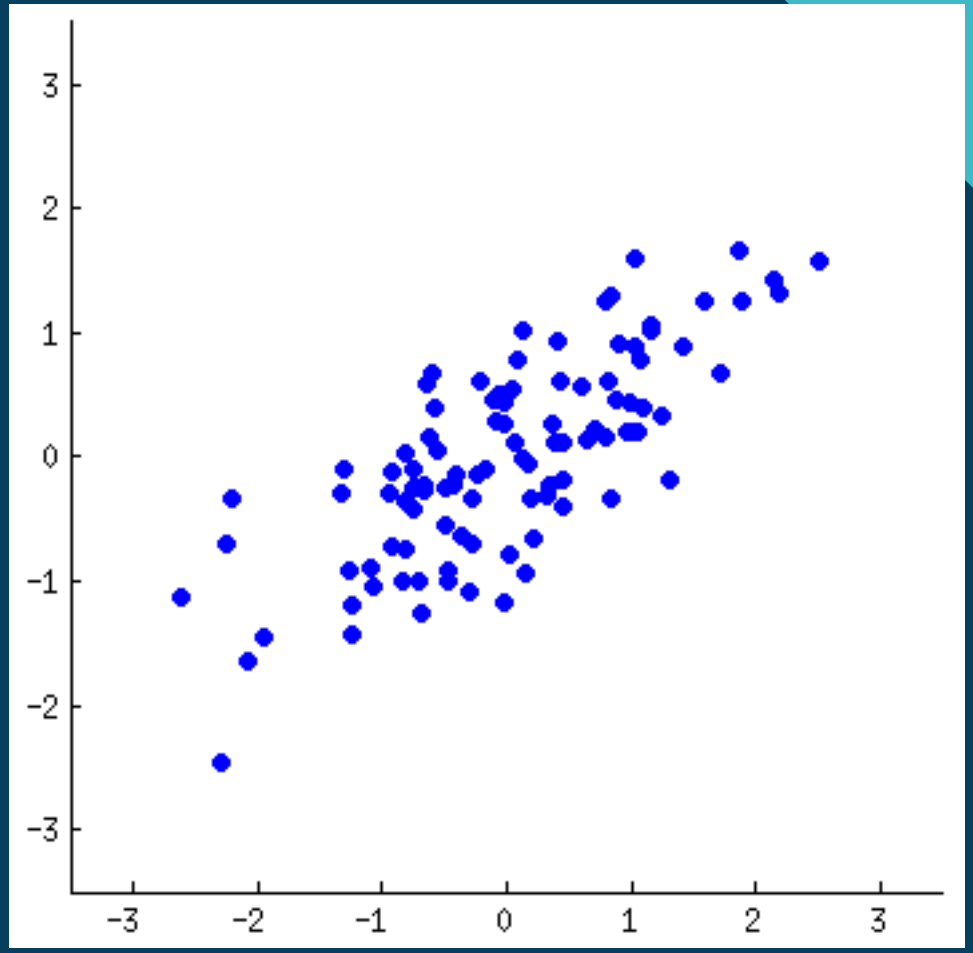
Dimensionality reduction

A word about dimensionality reduction

- Techniques to reduce the number of variable in your dataset by keeping as much information as possible:
  - Can be used as a preprocessing step before clustering
    - Either by using the most informative new components (meaning the new variables capturing the most information)
    - Identify most important original variables

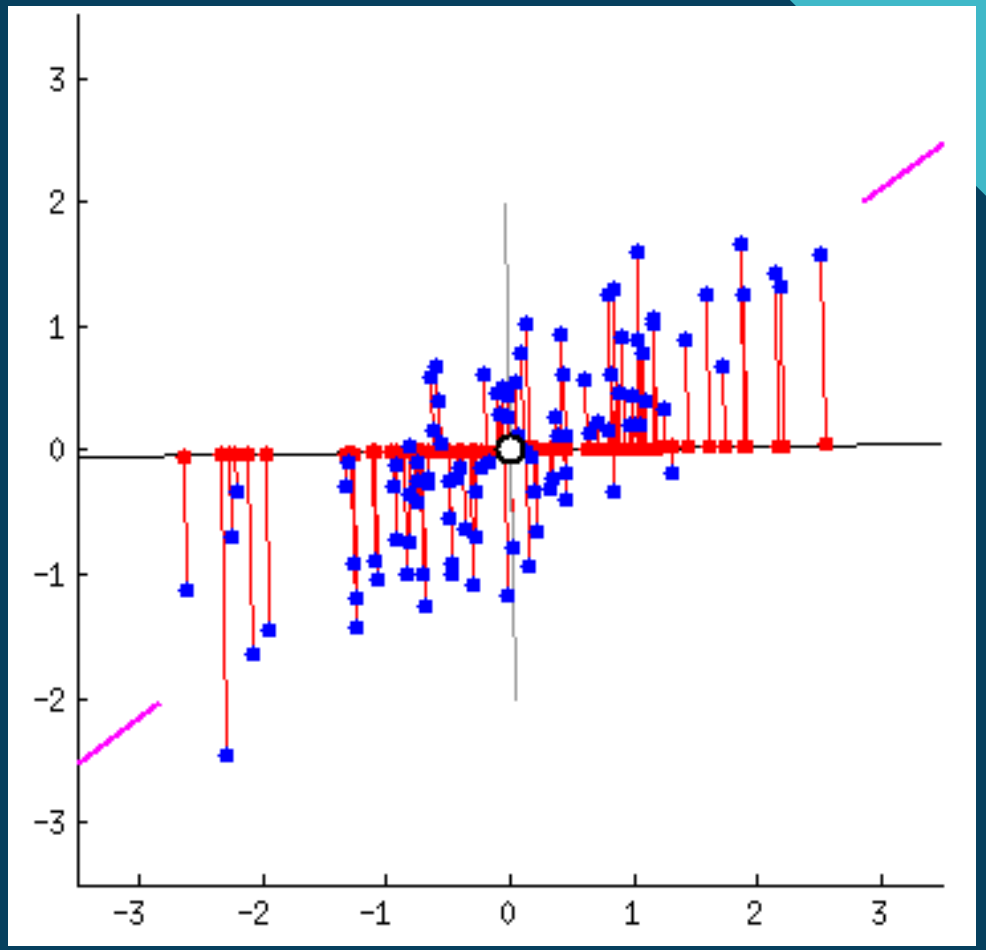
- Techniques
  - Principal Components Analysis
  - t-SNE

Dimensionality reduction



Source: https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues

Dimensionality reduction



Source: https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues

Dimensionality reduction

1) Original dataset 02-countries_ue.txt

We use those 8 numeric variables to do the dimensionality reduction.

Pays	PIB	Dep_pub	Rev_net	Esp_vie	Chomage	Education	Gini	Securite
BE	32700	17187.500	20.800	80.300	8.300	30	26.600	96.300
BG	4800	1790.800	15.100	73.800	10.300	20.700	33.200	18.300
CZ	14300	6261.600	16.600	77.700	7.300	15.300	24.900	31.700
DK	42500	24562.800	26.700	79.300	7.500	29.500	26.900	88.900
DE	30500	14503.300	17.200	80.500	7.100	24.300	29.300	74
EE	10700	4348.600	20.900	76	16.900	34.100	31.300	36.100
IE	35000	23073.500	10.200	81	13.700	33	33.200	23.100
GR	19600	10091.100	6.900	80.600	12.600	19.600	32.900	34.200
ES	22800	10410.600	14.900	82.300	20.100	25.800	33.900	50.800
FR	29900	16901.100	18.800	81.900	9.700	24.400	29.800	56.100
IT	25700	12930.800	17.600	81.800	8.400	12.100	31.200	43.600
CY	20700	9569.400	19.200	81.500	6.400	30.900	29.100	8.700
LV	8100	3757	12.800	73.700	19.800	25.300	36.100	25.200
LT	8400	3428.800	12.800	73.500	17.800	28.100	36.900	22.900
LU	78600	33738.500	19	80.800	4.600	31.600	27.900	64.500
HU	9700	4802.900	16.100	74.700	11.200	19	24.100	39.300
MT	15000	6368.800	16.600	81.400	6.900	12.500	28.400	28.800
NL	35400	18133.600	25.100	81	4.500	28.700	25.500	74.400

Dimensionality reduction

2) Normalized dataset

Pays	PIB	Dep_pub	Rev_net	Esp_vie	Chomage	Education	Gini	Securite
BE	0.582	0.669	0.772	0.466	-0.404	0.820	-0.752	1.467
BG	-1.203	-1.264	-0.401	-1.709	0.044	-0.419	0.935	-0.988
CZ	-0.596	-0.703	-0.092	-0.404	-0.628	-1.138	-1.186	-0.566
DK	1.209	1.595	1.987	0.131	-0.583	0.753	-0.675	1.234
DE	0.441	0.332	0.031	0.533	-0.673	0.061	-0.062	0.765
EE	-0.826	-0.943	0.793	-0.973	1.523	1.366	0.450	-0.428
IE	0.729	1.408	-1.410	0.700	0.806	1.220	0.935	-0.837
GR	-0.256	-0.222	-2.089	0.566	0.559	-0.565	0.858	-0.487
ES	-0.052	-0.182	-0.442	1.135	2.240	0.261	1.114	0.035
FR	0.403	0.633	0.361	1.001	-0.090	0.074	0.066	0.202
IT	0.134	0.135	0.114	0.968	-0.382	-1.565	0.424	-0.191
CY	-0.186	-0.287	0.443	0.867	-0.830	0.940	-0.113	-1.290
LV	-0.992	-1.017	-0.875	-1.742	2.172	0.194	1.676	-0.771
LT	-0.973	-1.058	-0.875	-1.809	1.724	0.567	1.881	-0.843
LU	3.519	2.747	0.402	0.633	-1.233	1.033	-0.419	0.466
HU	-0.890	-0.886	-0.195	-1.408	0.246	-0.645	-1.390	-0.327
MT	-0.551	-0.689	-0.092	0.834	-0.718	-1.511	-0.292	-0.657
NL	0.755	0.788	1.658	0.700	-1.255	0.647	-1.033	0.778

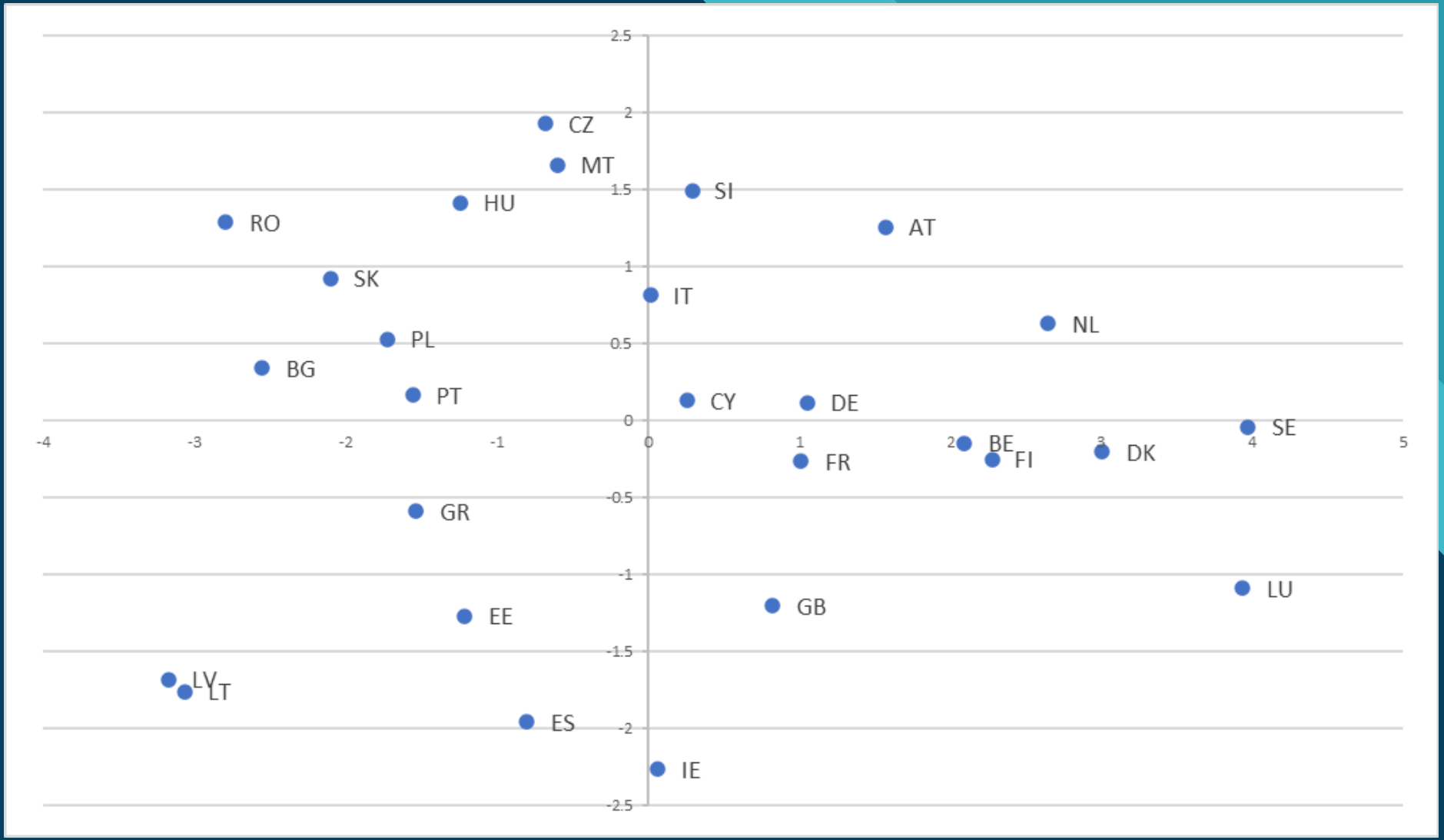
Dimensionality reduction

3) Dataset after PCA

Pays	pc_1	pc_2	pc_3	pc_4	pc_5	pc_6	pc_7	pc_8
BE	2.094	-0.144	0.645	0.281	0.224	-0.006	-0.451	-0.003
BG	-2.551	0.343	0.657	-0.732	0.009	0.748	-0.022	-0.016
CZ	-0.677	1.929	0.093	-0.167	-0.102	-0.455	-0.019	0.036
DK	3.008	-0.200	0.849	-0.420	0.449	0.265	0.703	-0.328
DE	1.055	0.119	-0.355	0.187	0.185	0.360	-0.417	0.105
EE	-1.212	-1.266	1.998	0.011	-0.654	-0.104	0.450	0.119
IE	0.068	-2.264	-1.518	-0.554	-0.436	-0.709	-0.236	-0.464
GR	-1.533	-0.586	-1.671	0.507	0.139	-0.337	-0.730	0.028
ES	-0.803	-1.958	-0.398	1.623	-0.042	-0.519	0.517	0.191
FR	1.015	-0.266	-0.540	0.436	-0.212	0.065	0.318	-0.146
IT	0.016	0.813	-1.315	0.725	0.237	0.489	0.763	-0.042
CY	0.258	0.133	-0.215	-0.393	-1.986	0.207	0.178	0.097
LV	-3.168	-1.679	0.863	-0.122	0.554	-0.016	0.244	0.000
LT	-3.065	-1.759	0.871	-0.451	0.272	0.349	0.013	0.046
LU	3.934	-1.086	-1.539	-1.902	0.636	-0.338	0.384	0.410
HU	-1.244	1.417	1.154	-0.372	0.322	-0.962	-0.241	-0.035
MT	-0.600	1.661	-0.932	0.669	-0.469	0.186	0.321	0.091
NL	2.647	0.629	0.537	-0.183	-0.457	0.314	0.252	-0.026

Dimensionality reduction

4) Show the 2 first dimensions (done via Excel)



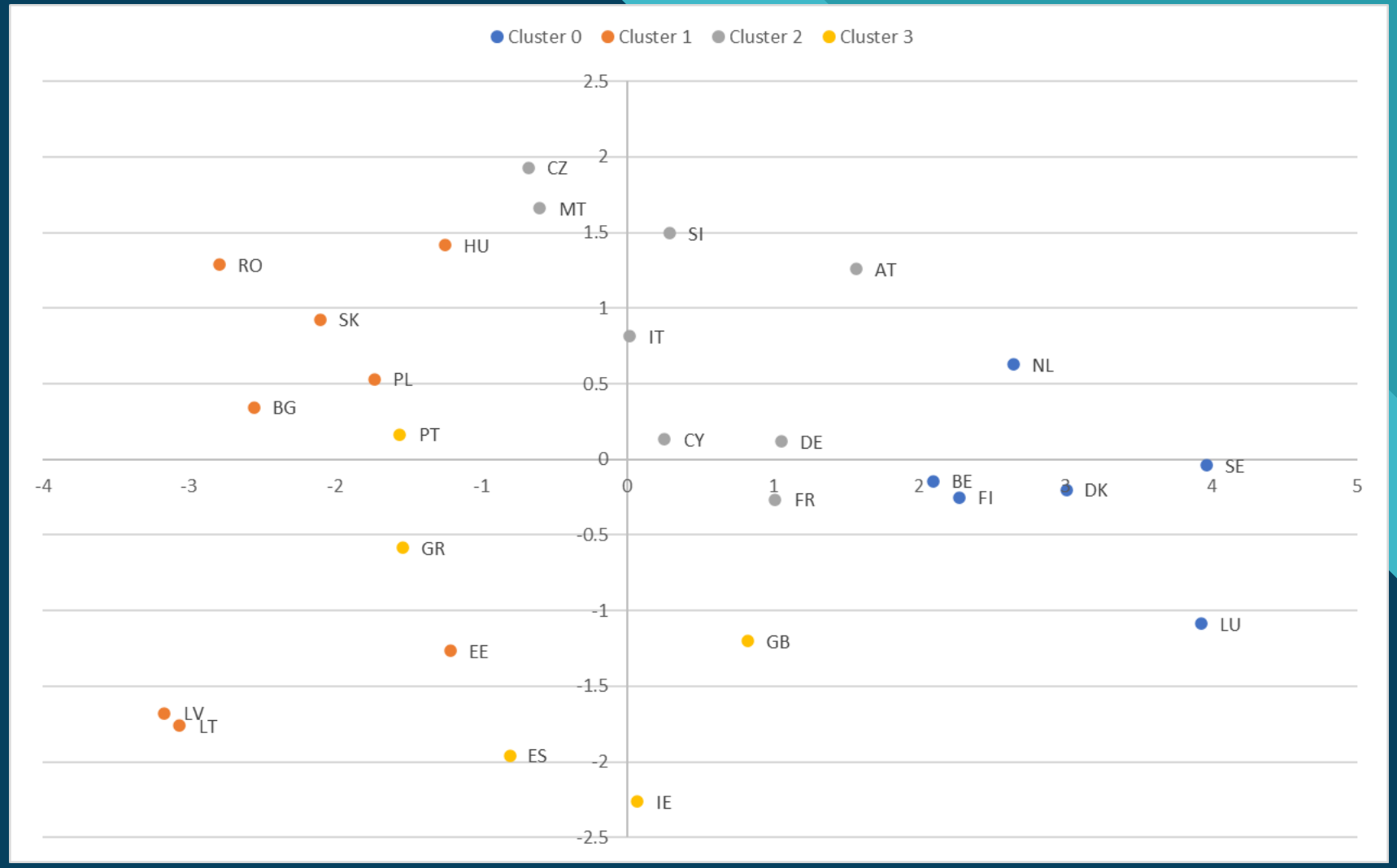
Dimensionality reduction

4) Variance per PC and cumulative variance

Component	Proportion of Variance	Cumulative Variance
PC 1	0.542	0.542
PC 2	0.173	0.714
PC 3	0.124	0.838
PC 4	0.059	0.897
PC 5	0.040	0.937
PC 6	0.039	0.975
PC 7	0.022	0.997
PC 8	0.003	1.000

Dimensionality reduction + Clustering

5) Clustering on the first 5 principal components (done via RapidMiner) and display of cluster in the first 2 dimensions (done via Excel)



Summary exercice

Cereal data

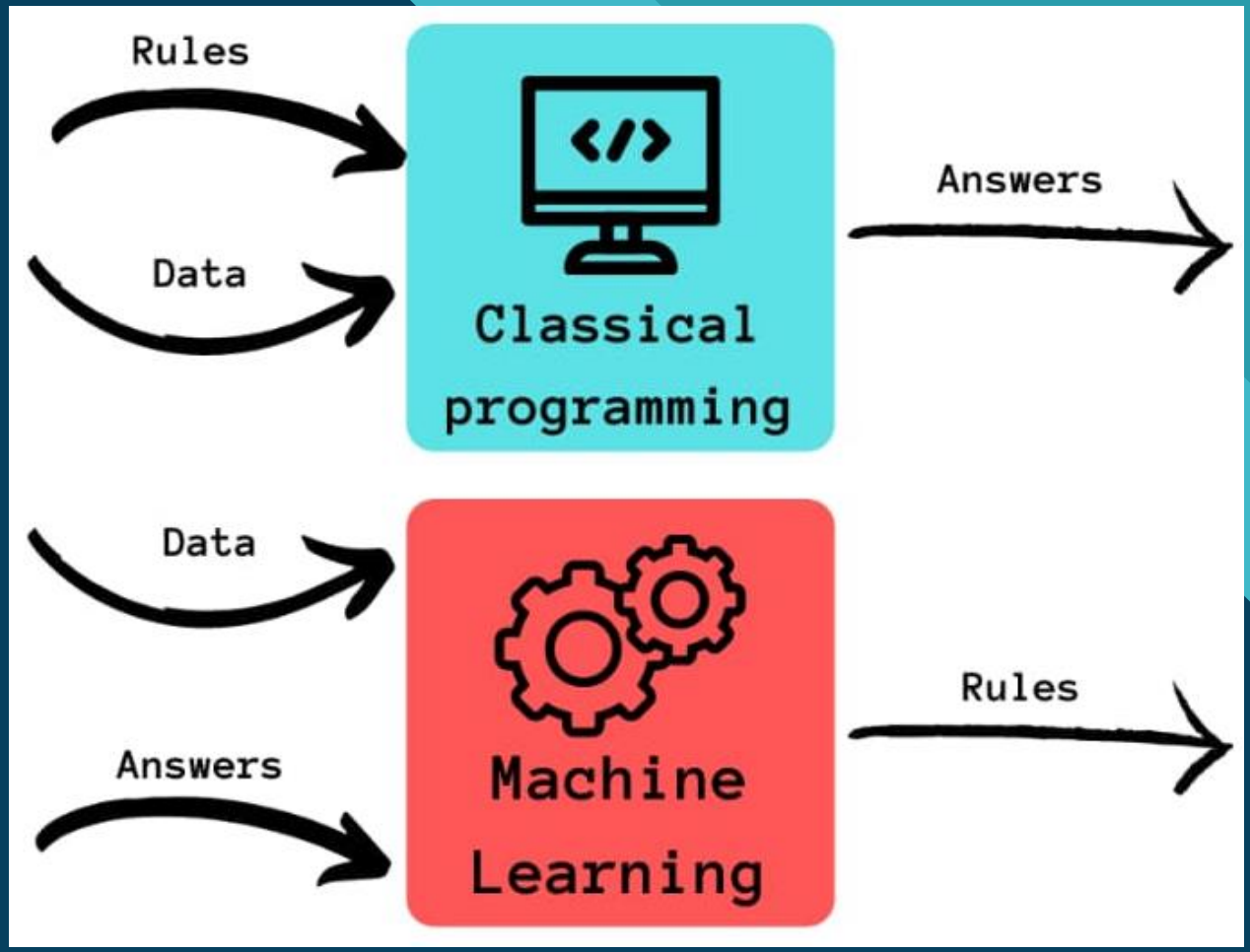
As a data analyst, run some analysis to give an overview of cereals proposed in your supermarket.

- 10-cereal-readme.txt: context and meta data explained
- 10-cereal.csv: data
- 10-cereal_manufacturer.txt: data about complete name for manufacturer

Supervised techniques

Introduction

- Traditional vs machine learning approach

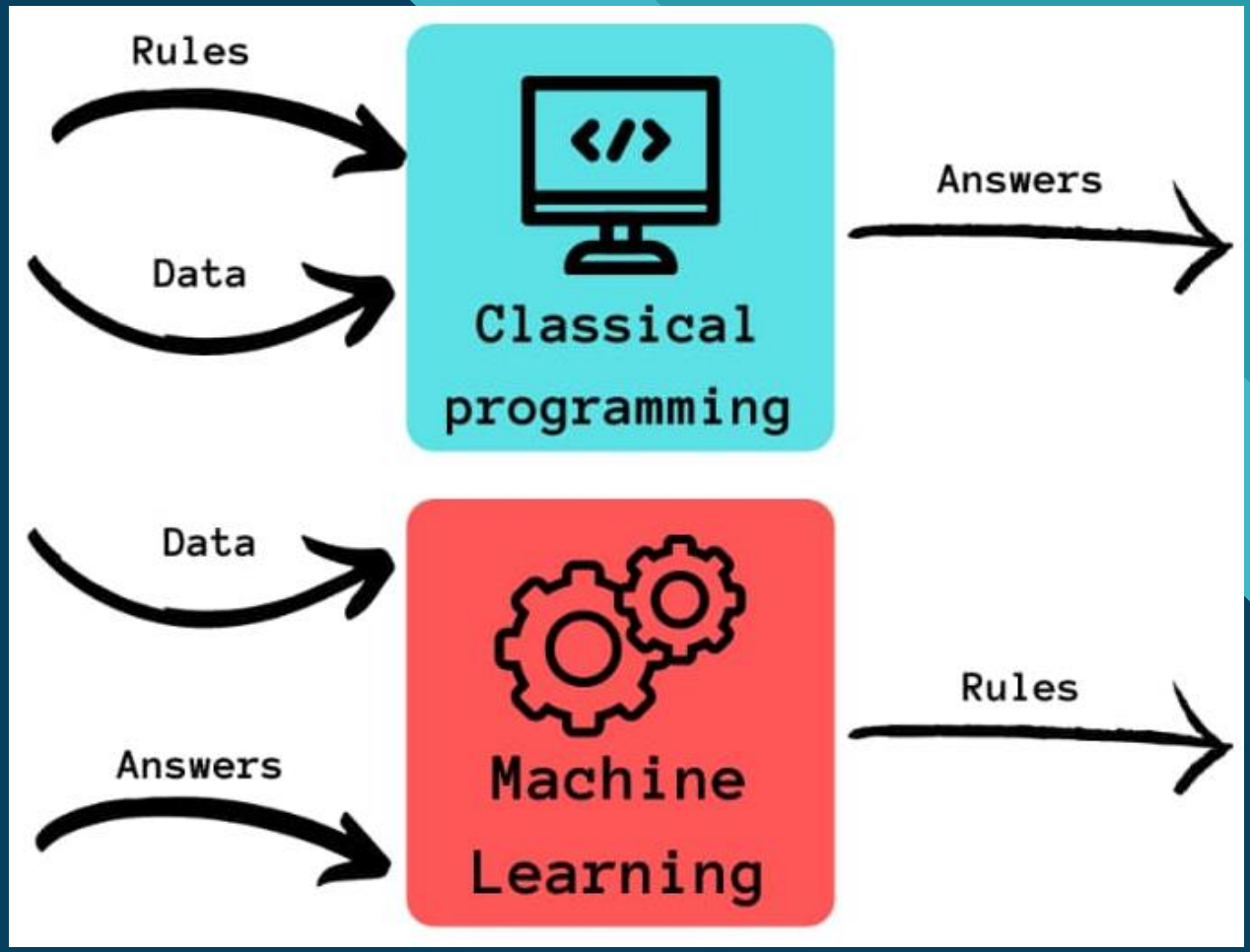


Source: https://inlocrobotics.com/en/evolution-of-artificial-intelligence/

https://futurice.com/blog/differences-between-machine-learning-and-software-engineering

Introduction

- Traditional vs machine learning approach



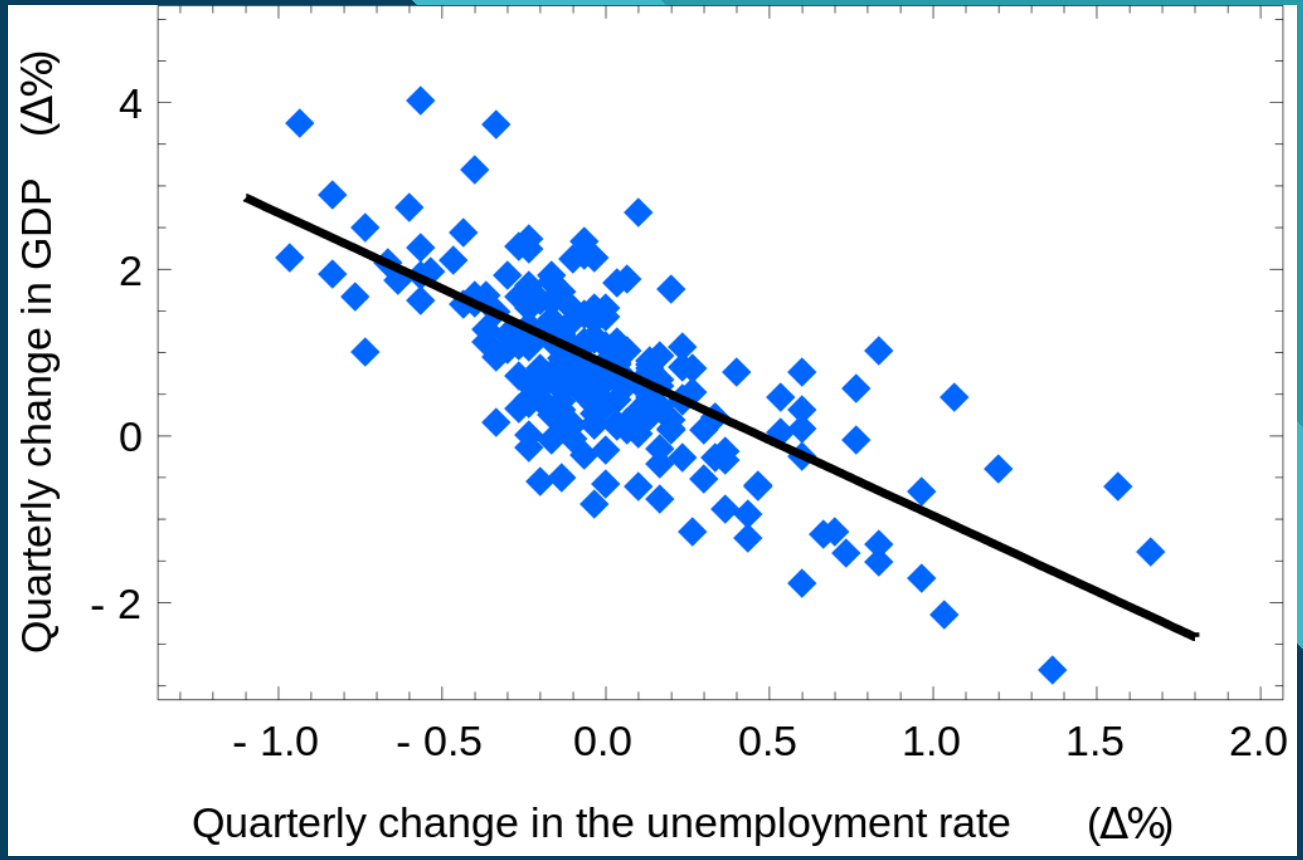
Supervised = we give the answers in input

Source: https://inlocrobotics.com/en/evolution-of-artificial-intelligence/

https://futurice.com/blog/differences-between-machine-learning-and-software-engineering

Linear regression

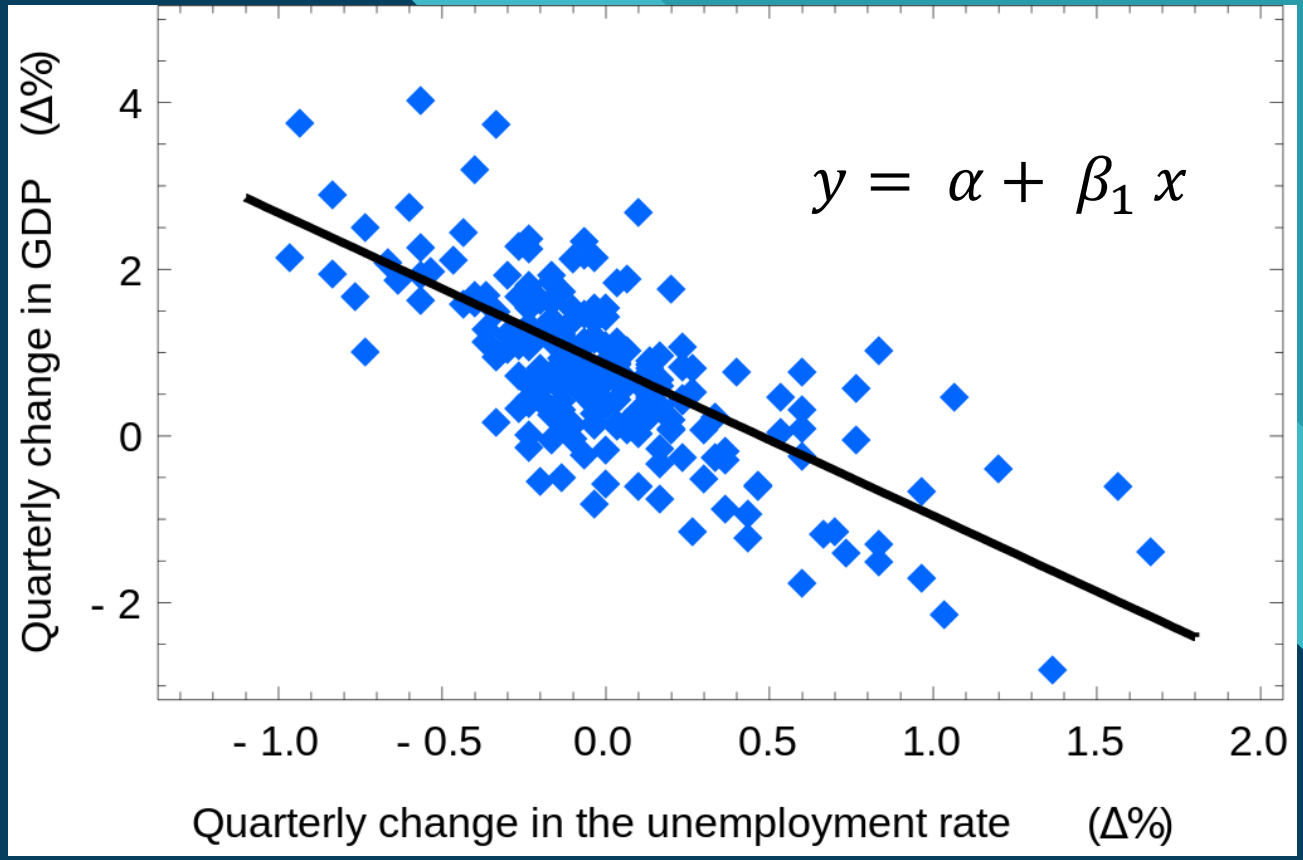
What is linear regression?



Source: https://en.wikipedia.org/wiki/Simple_linear_regression

Linear regression

What is linear regression?

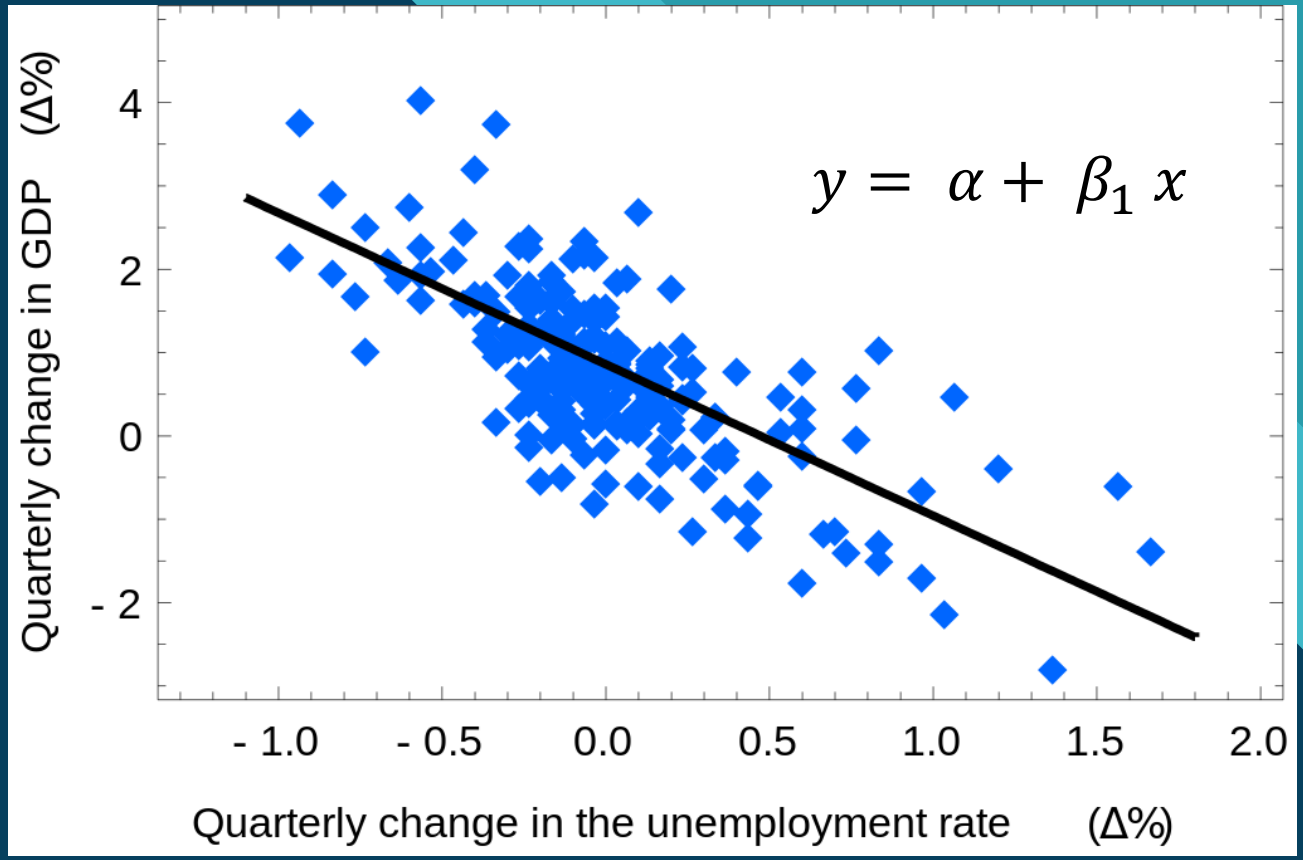


Source: https://en.wikipedia.org/wiki/Simple_linear_regression

Linear regression

What is linear regression?

- Target variable = dependent variable

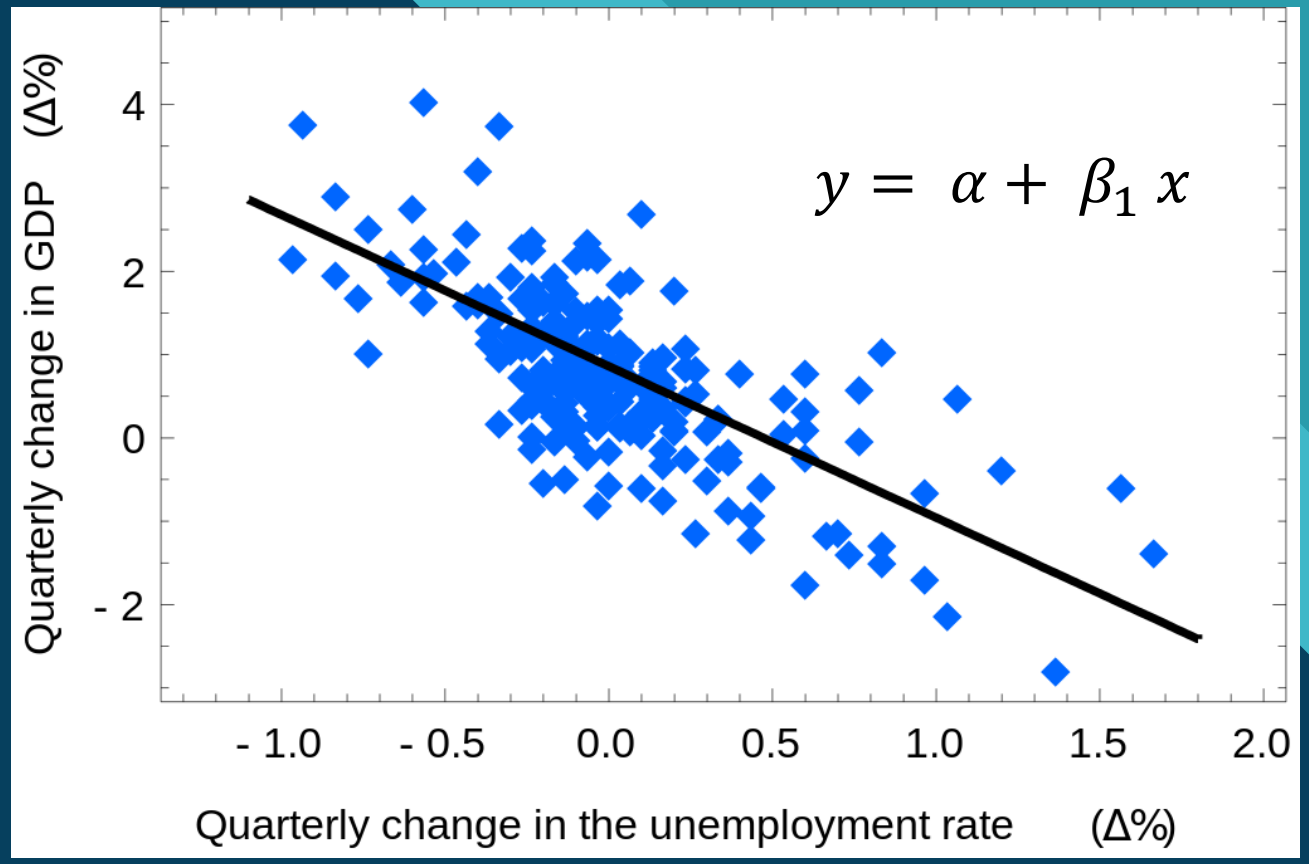


Source: https://en.wikipedia.org/wiki/Simple_linear_regression

Linear regression

What is linear regression?

- Target variable = dependent variable
  - Target variable must be numeric

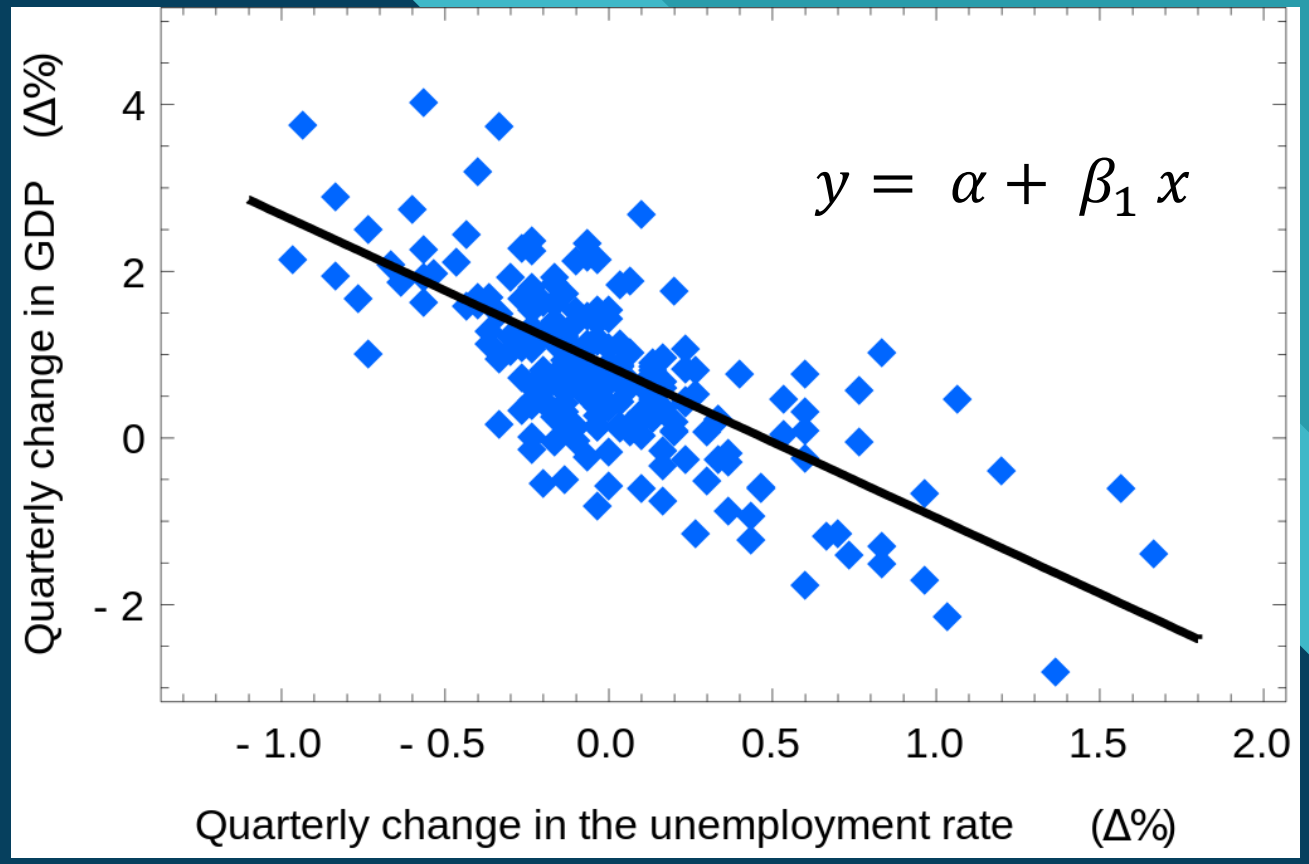


Source: https://en.wikipedia.org/wiki/Simple_linear_regression

Linear regression

What is linear regression?

- Target variable = dependent variable
  - Target variable must be numeric
- Explanatory variable = independent variable

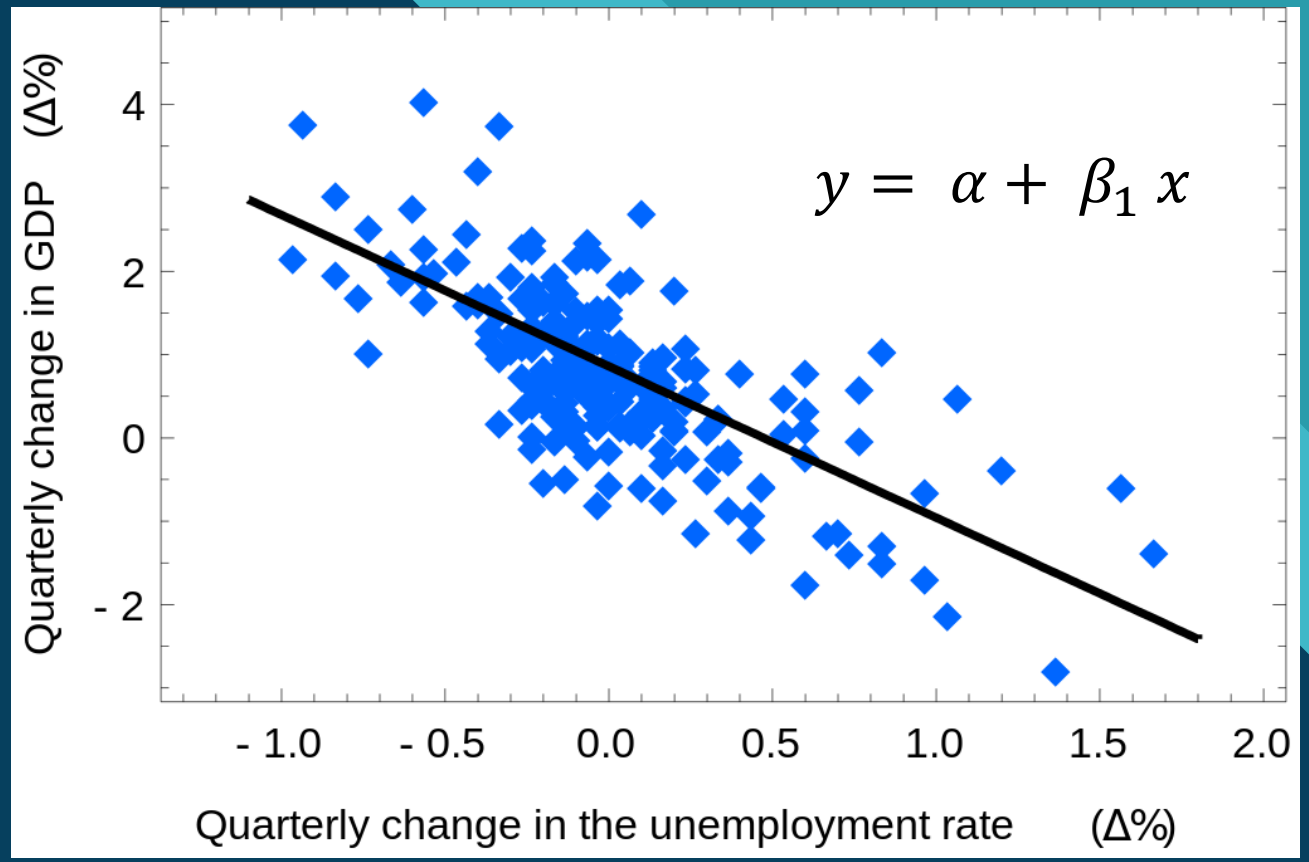


Source: https://en.wikipedia.org/wiki/Simple_linear_regression

Linear regression

What is linear regression?

- Target variable = dependent variable
  - Target variable must be numeric
- Explanatory variable = independent variable
  - Explanatory variable must be numeric

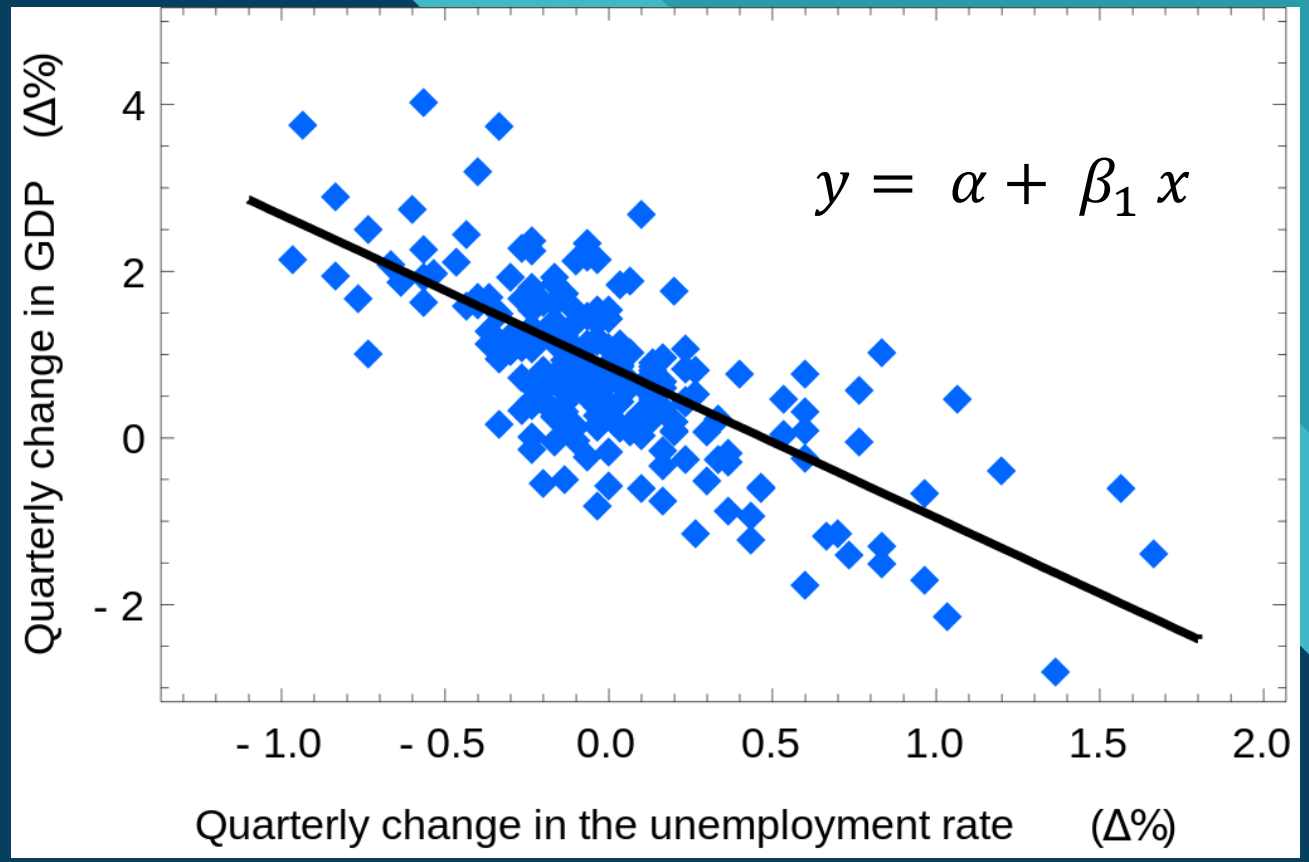


Source: https://en.wikipedia.org/wiki/Simple_linear_regression

Linear regression

What is linear regression?

- Target variable = dependent variable
  - Target variable must be numeric
- Explanatory variable = independent variable
  - Explanatory variable must be numeric
    - categorical variables => one-hot encoding



Source: https://en.wikipedia.org/wiki/Simple_linear_regression

One-hot encoding

Some algorithms (linear regression, logistic regression, …) needs that categorical variables are preprocessed before being used in the algorithm

Type
AA
AB
CD
AA

Onehot encoding →

Type	AA_Onehot	AB_Onehot	CD_Onehot
AA	1	0	0
AB	0	1	0
CD	0	0	1
AA	0	0	0

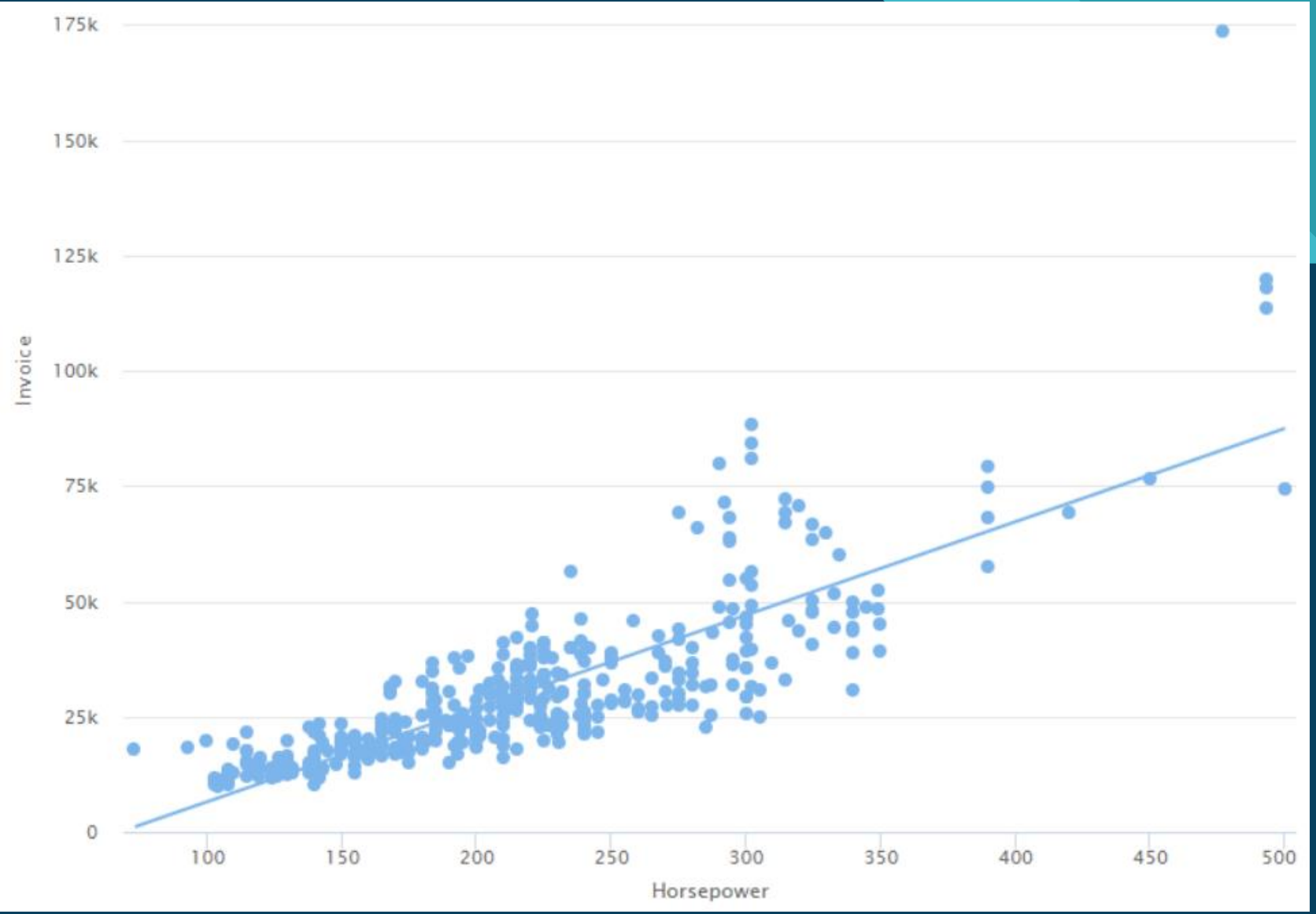
Source: https://www.educative.io/blog/one-hot-encoding

Linear regression

Cars dataset

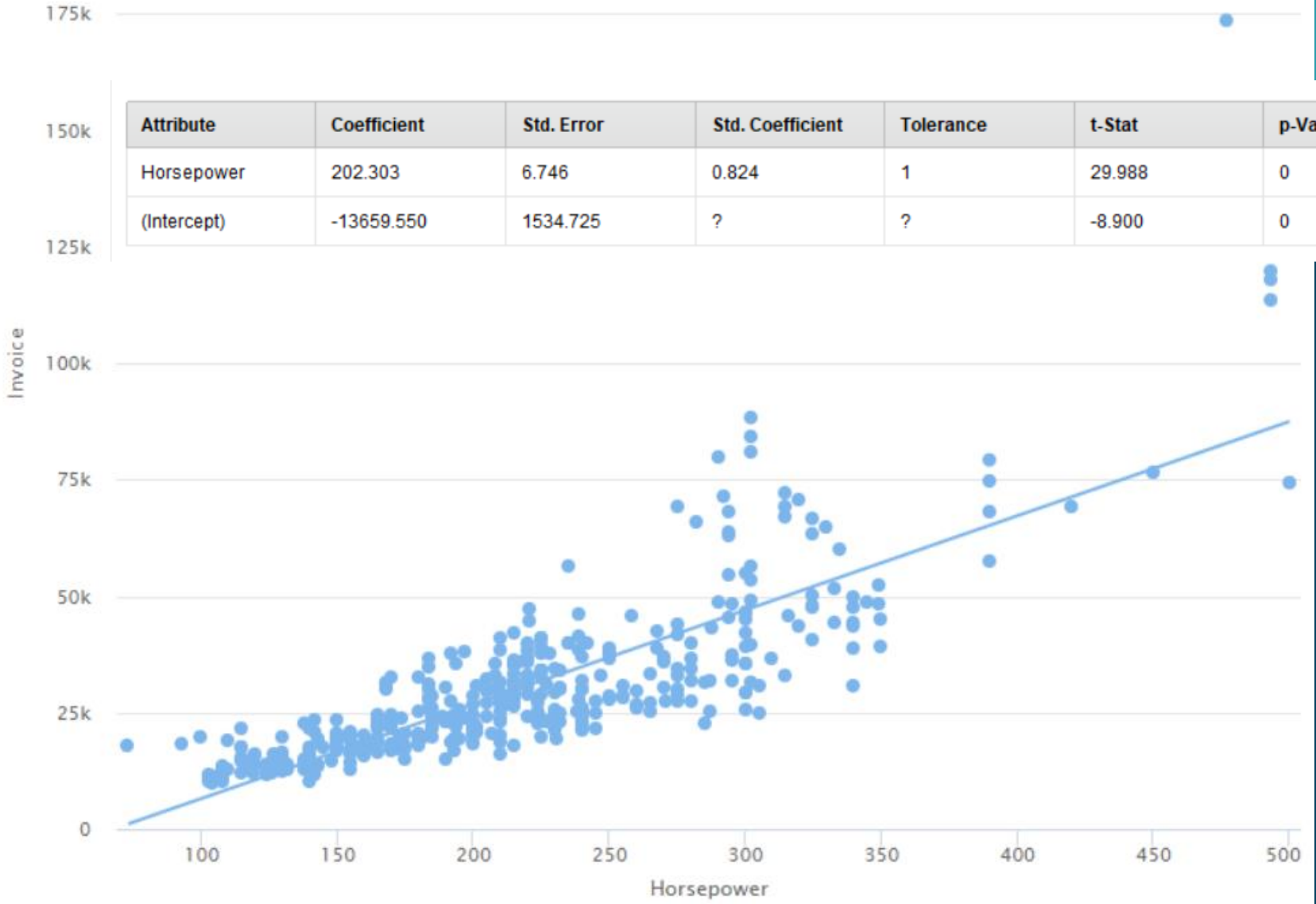
Row No.	Invoice	Origin	Horsepower	ï»¿Make	Model	Type	DriveTrain	MSRP	EngineSize
1	33337	Asia	265	Acura	MDX	SUV	All	36945	3.500
2	21761	Asia	200	Acura	RSX Type S 2...	Sedan	Front	23820	2
3	24647	Asia	200	Acura	TSX 4dr	Sedan	Front	26990	2.400
4	30299	Asia	270	Acura	TL 4dr	Sedan	Front	33195	3.200
5	39014	Asia	225	Acura	3.5 RL 4dr	Sedan	Front	43755	3.500
6	41100	Asia	225	Acura	3.5 RL w/Navi...	Sedan	Front	46100	3.500
7	79978	Asia	290	Acura	NSX coupe 2...	Sports	Rear	89765	3.200
8	23508	Europe	170	Audi	A4 1.8T 4dr	Sedan	Front	25940	1.800
9	32506	Europe	170	Audi	A41.8T conve...	Sedan	Front	35940	1.800
10	28846	Europe	220	Audi	A4 3.0 4dr	Sedan	Front	31840	3
11	30366	Europe	220	Audi	A4 3.0 Quattr...	Sedan	All	33430	3
12	31388	Europe	220	Audi	A4 3.0 Quattr...	Sedan	All	34480	3
13	33129	Europe	220	Audi	A6 3.0 4dr	Sedan	Front	36640	3
14	35992	Europe	220	Audi	A6 3.0 Quattr...	Sedan	All	39640	3
15	38325	Europe	220	Audi	A4 3.0 conver...	Sedan	Front	42490	3

Linear regression



Linear regression

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
Horsepower	202.303	6.746	0.824	1	29.988	0	****
(Intercept)	-13659.550	1534.725	?	?	-8.900	0	****



Linear regression

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
Horsepower	202.303	6.746	0.824	1	29.988	0	****
(Intercept)	-13659.550	1534.725	?	?	-8.900	0	****

Regression line is the following:

Linear regression

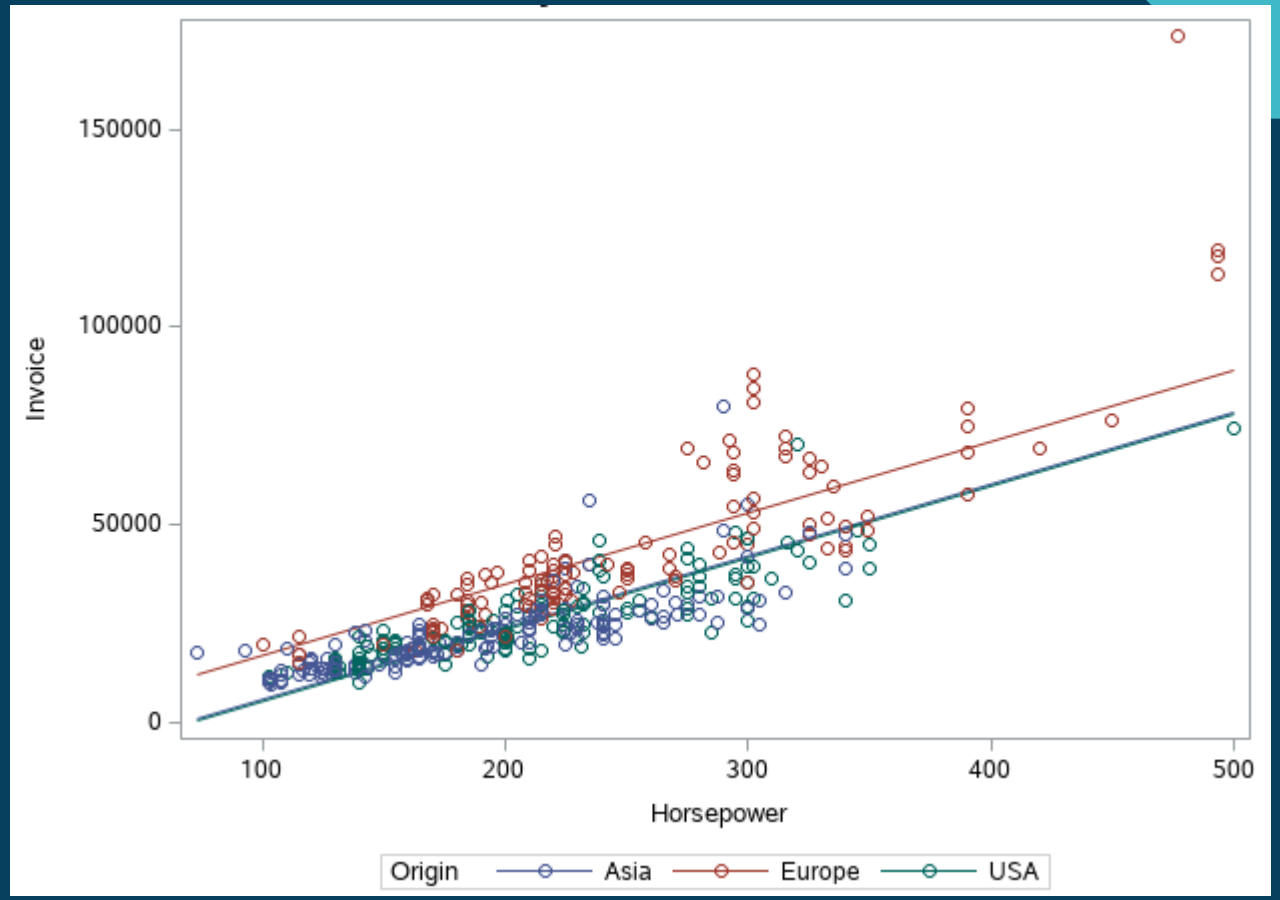
Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
Horsepower	202.303	6.746	0.824	1	29.988	0	****
(Intercept)	-13659.550	1534.725	?	?	-8.900	0	****

Regression line is the following:

$$Invoice = -13659.55 + 202.303\ Horsepower$$

Linear regression

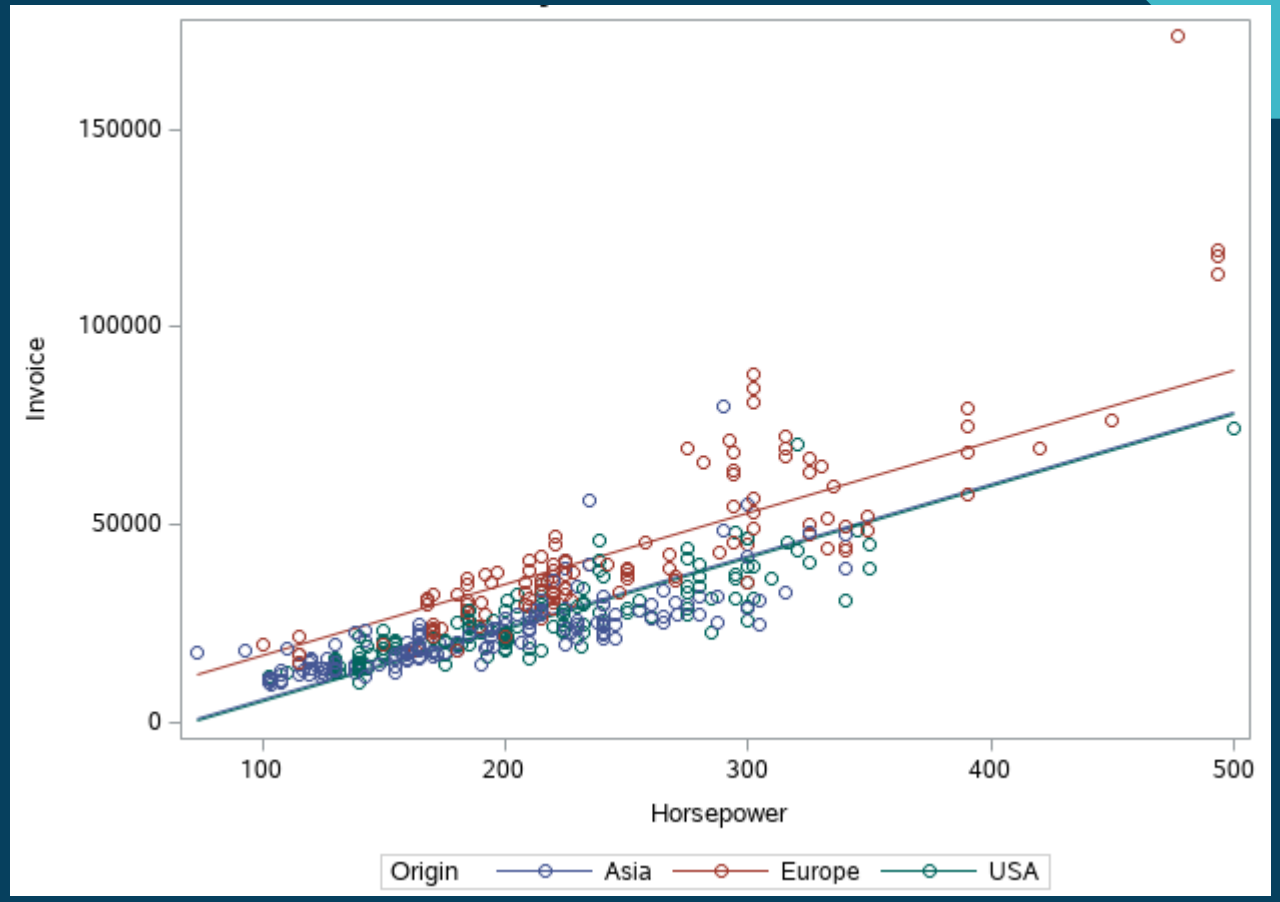
Include intercept for each different Origin (this graph is not done with RapidMiner)



Linear regression

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
Origin = Asia	-10743.094	1129.606	-0.294	0.971	-9.510	0	****
Origin = USA	-11390.396	1106.336	-0.307	0.998	-10.296	0	****
Horsepower	180.577	6.331	0.735	0.888	28.520	0	****
(Intercept)	-1091.166	1782.320	?	?	-0.612	0.541	

Include intercept for each different Origin (this graph is not done with RapidMiner)



Linear regression

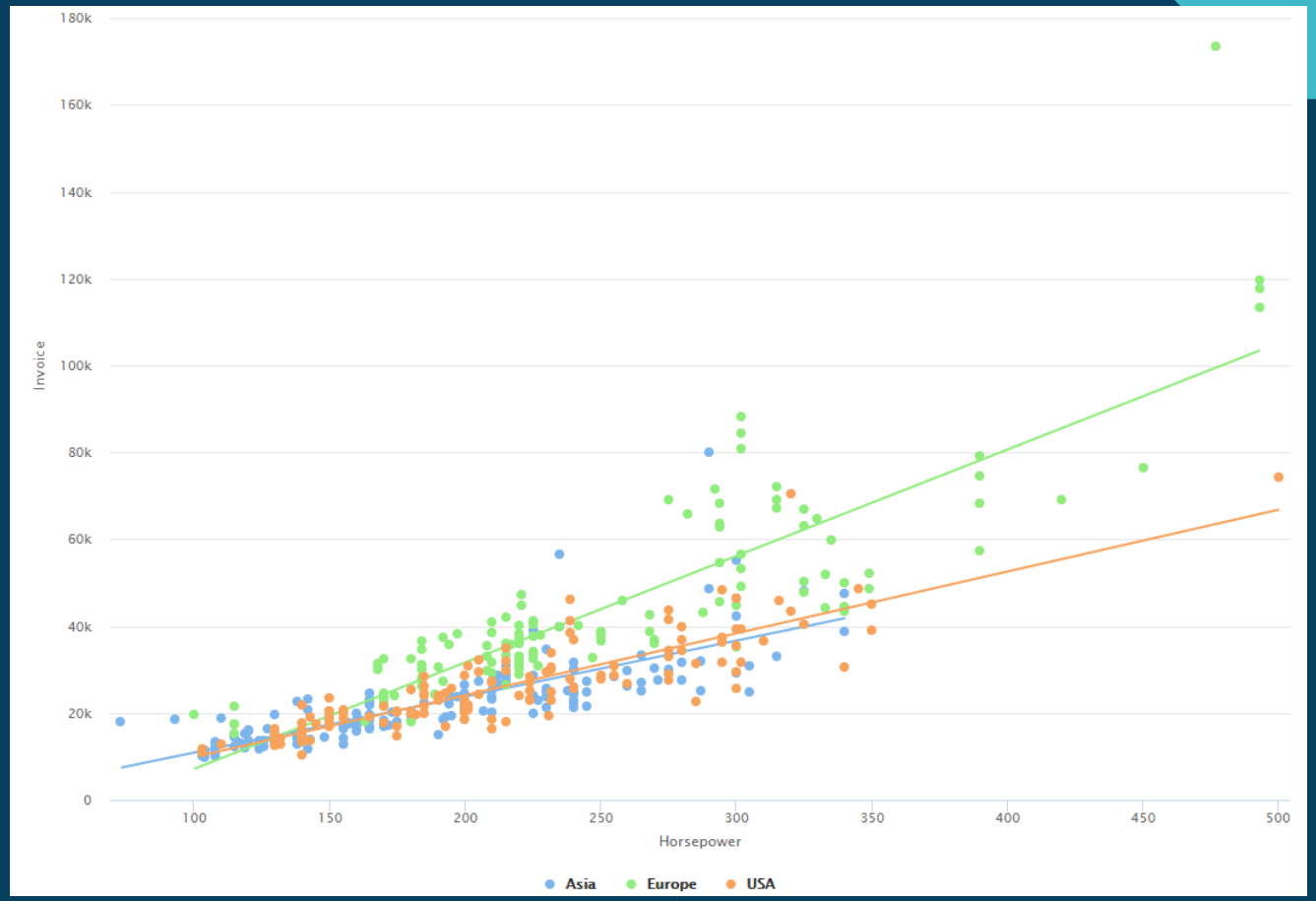
Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
Origin = Asia	-10743.094	1129.606	-0.294	0.971	-9.510	0	****
Origin = USA	-11390.396	1106.336	-0.307	0.998	-10.296	0	****
Horsepower	180.577	6.331	0.735	0.888	28.520	0	****
(Intercept)	-1091.166	1782.320	?	?	-0.612	0.541	

Regression line is the following

$$Invoice = -1091.166 + 180.577\ Horsepower \Rightarrow Europe\ cars$$
$$Invoice = -12481.562 + 180.577\ Horsepower \Rightarrow USA\ cars$$
$$Invoice = -11834.26 + 180.577\ Horsepower \Rightarrow Asia\ cars$$

Linear regression

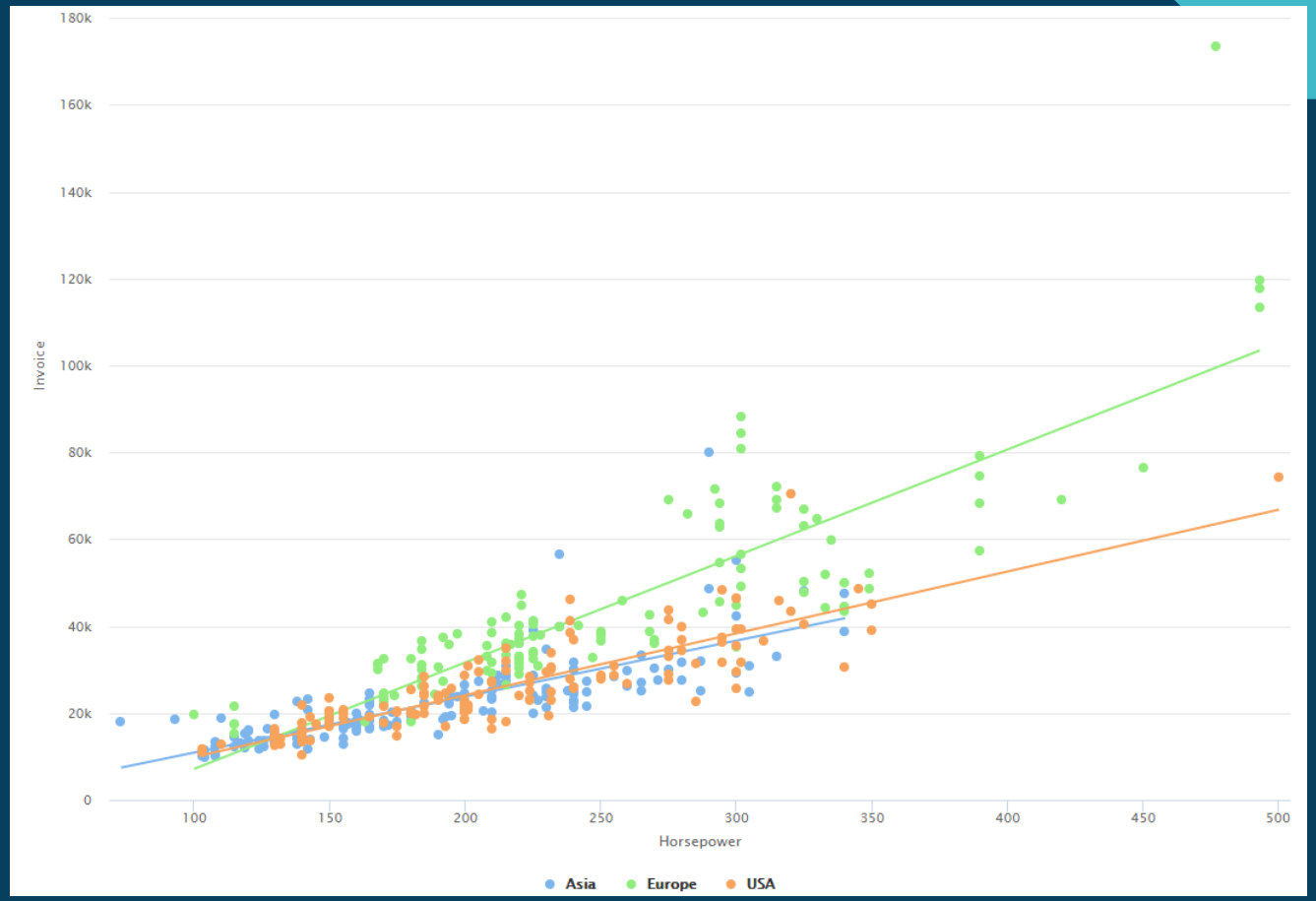
Include intercept for each different Origin but also a specific slope



Linear regression

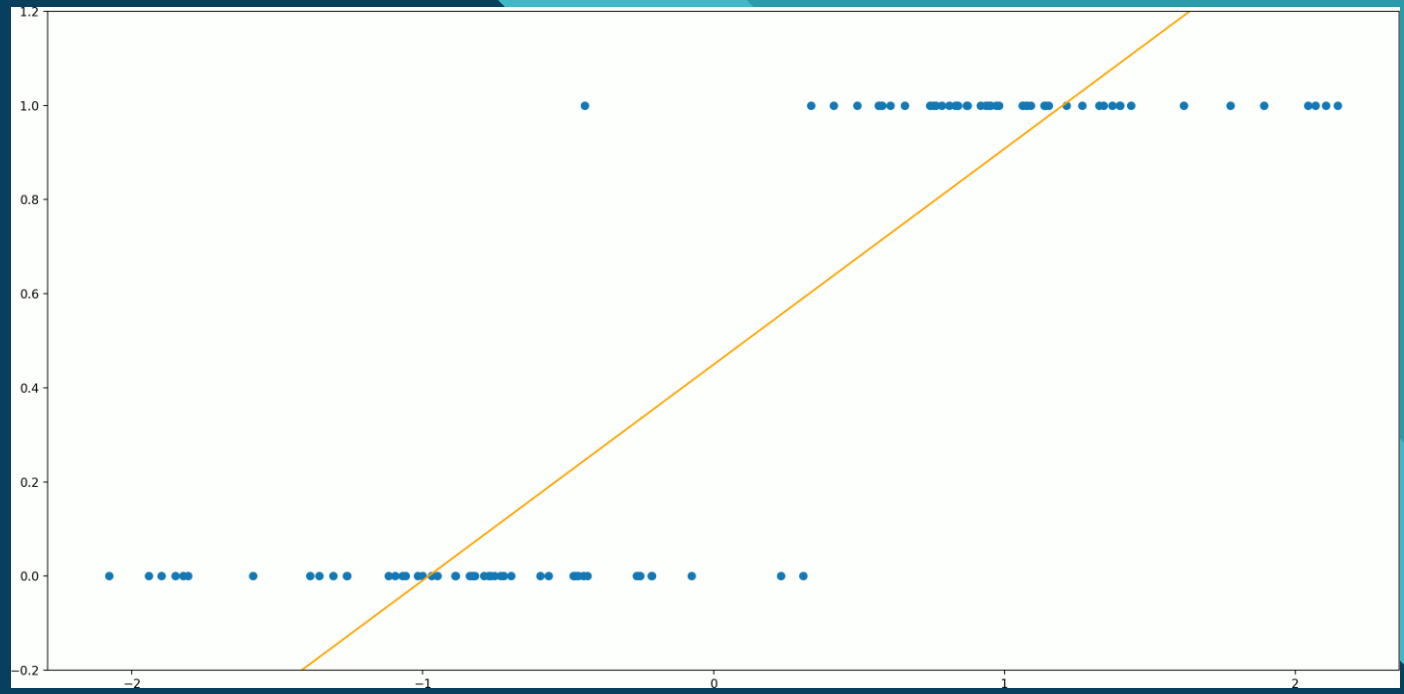
Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
Horsepower	245.105	9.028	0.998	0.506	27.148	0	****
Origin_Asia	15374.982	3219.659	0.421	0.845	4.775	0.000	****
Origin_USA	12996.661	3329.966	0.350	0.955	3.903	0.000	****
Horsepower_Asia	-116.252	14.091	-0.652	0.969	-8.250	0.000	****
Horsepower_USA	-102.742	13.812	-0.628	1.000	-7.439	0.000	****
(Intercept)	-17345.439	2387.261	?	?	-7.266	0.000	****

Include intercept for each different Origin but also a specific slope



Logistic regression

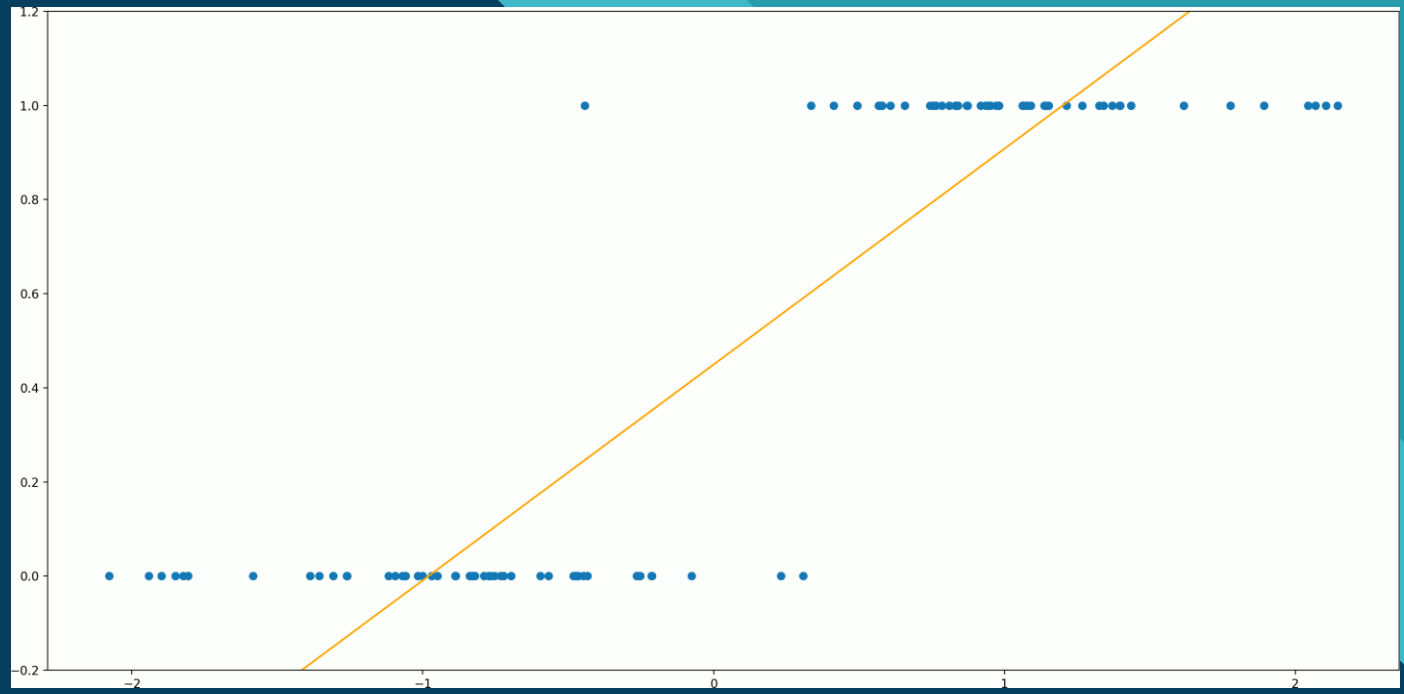
What is logistic regression?



Source:

Logistic regression

What is logistic regression?

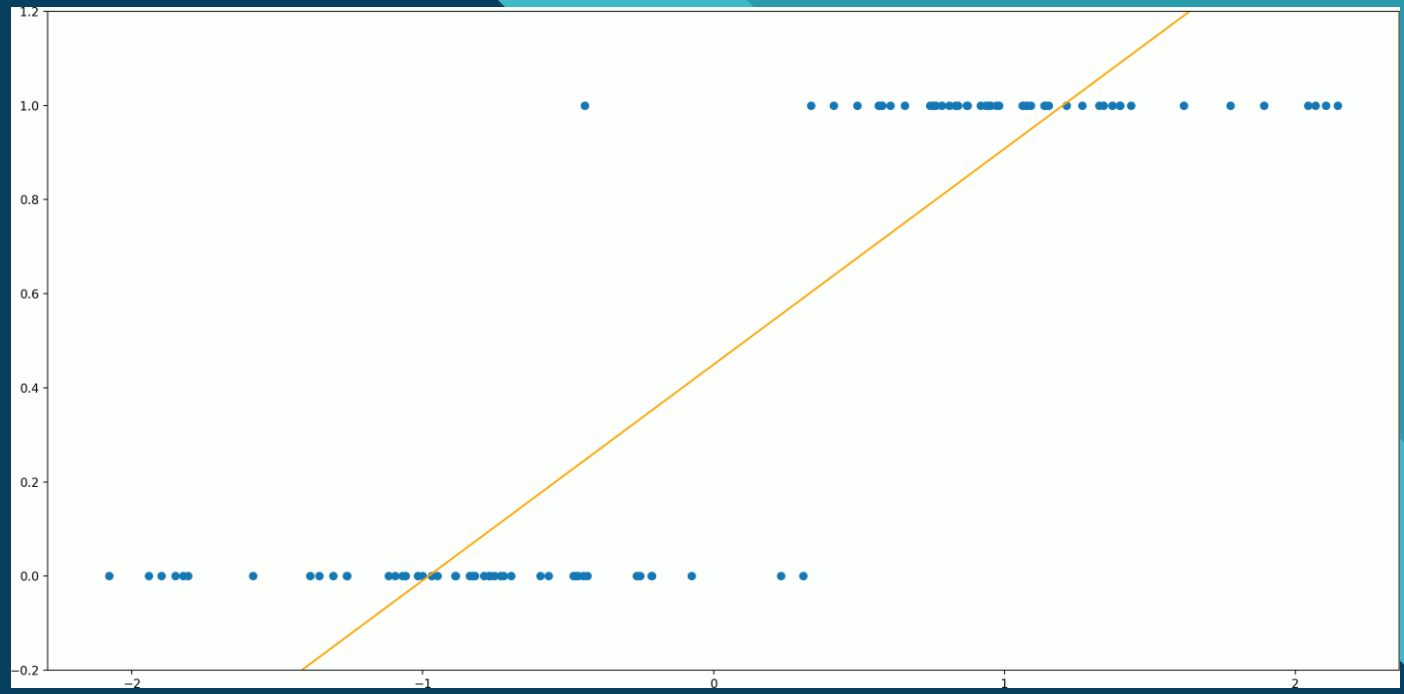


Source:

Logistic regression

What is logistic regression?

- Target variable = dependent variable

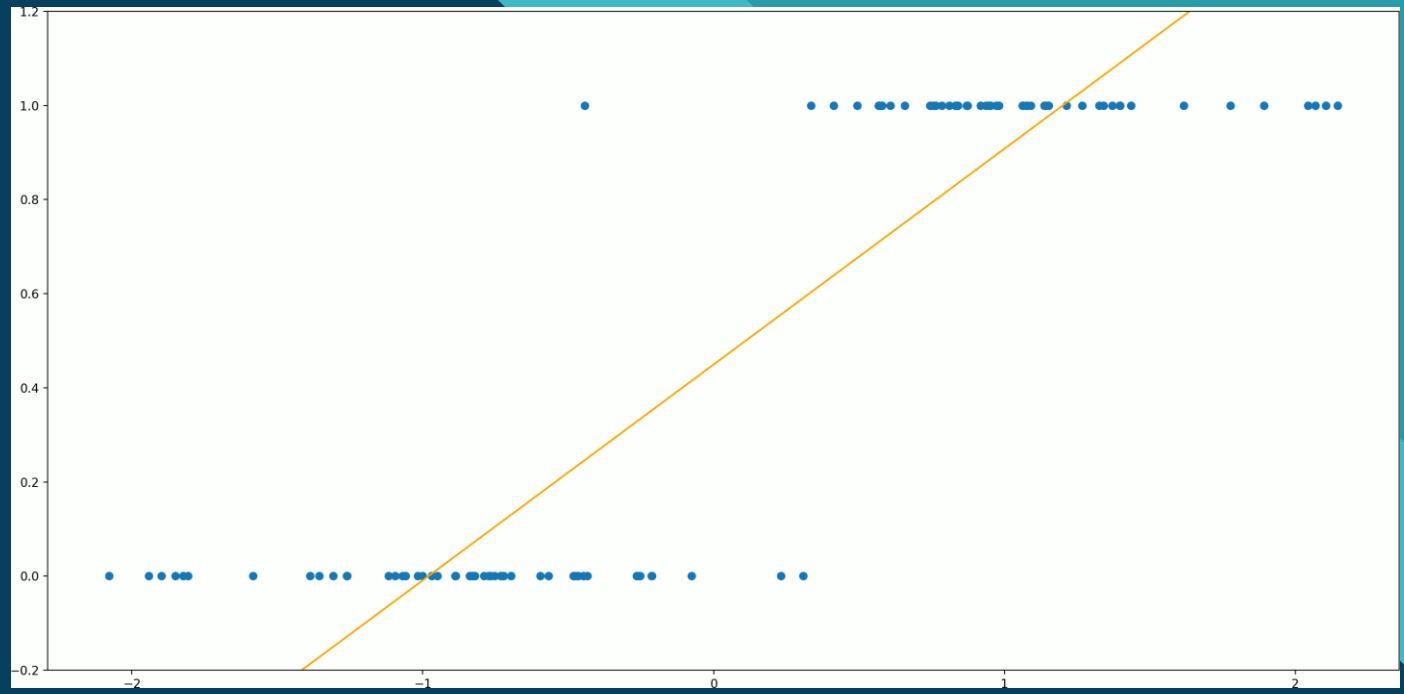


Source:

Logistic regression

What is logistic regression?

- Target variable = dependent variable
  - Target variable must be binary (fraud/no fraud, cancer/no cancer, buy/not nuy, …)

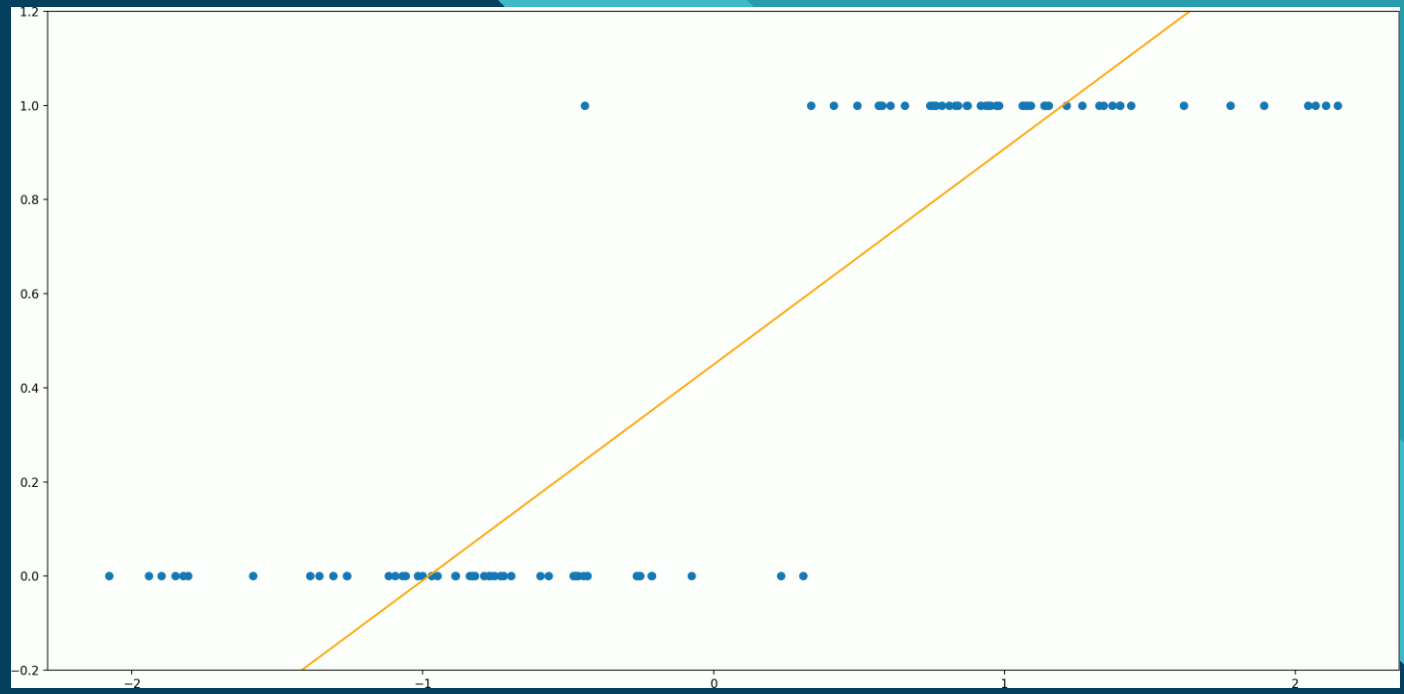


Source:

Logistic regression

What is logistic regression?

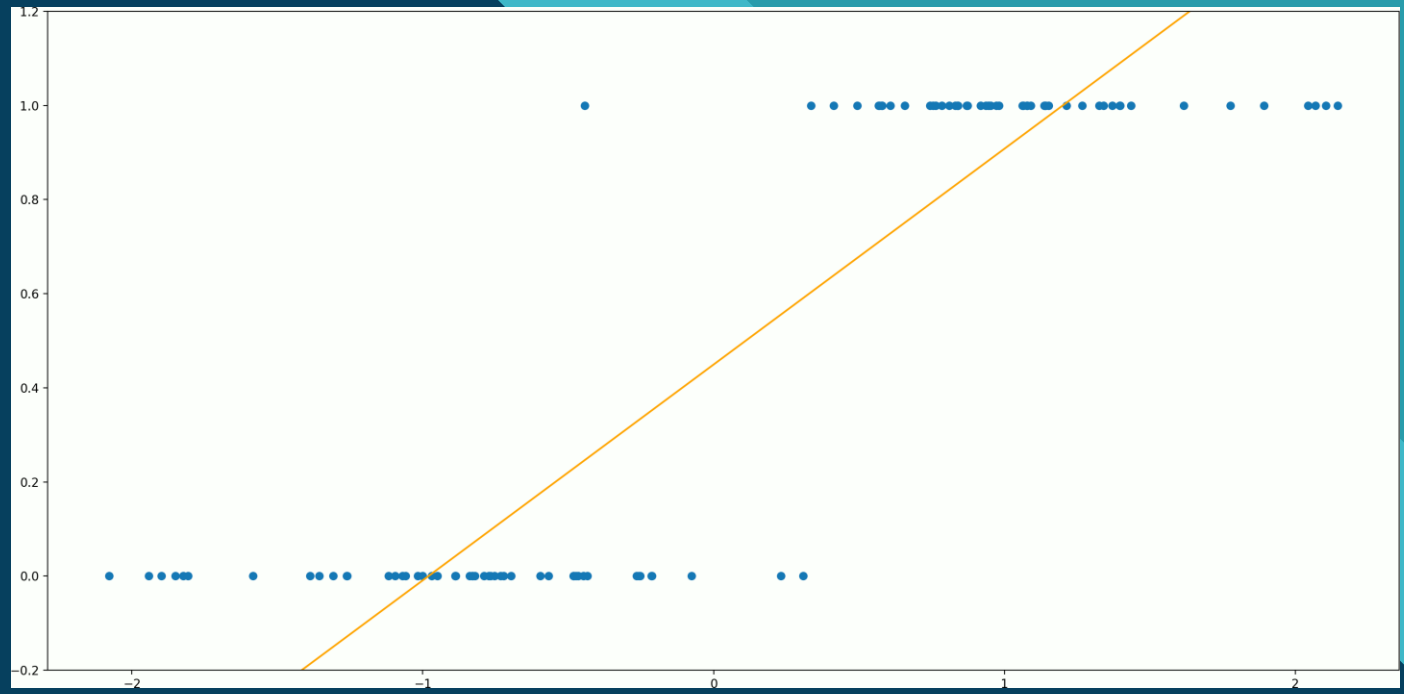
- Target variable = dependent variable
  - Target variable must be binary (fraud/no fraud, cancer/no cancer, buy/not nuy, …)
- Explanatory variable = independent variable



Source:

Logistic regression

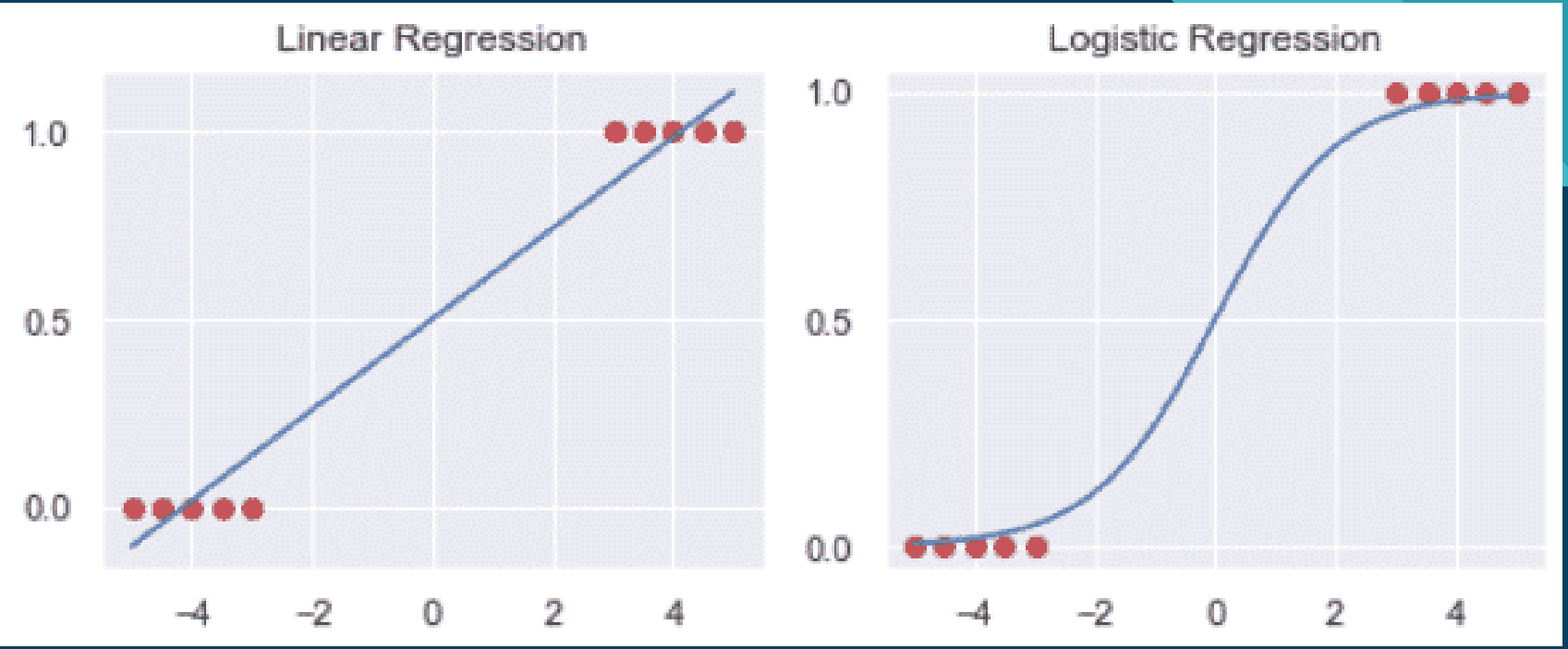
What is logistic regression?



- Target variable = dependent variable
  - Target variable must be binary (fraud/no fraud, cancer/no cancer, buy/not nuy, …)
- Explanatory variable = independent variable
  - Explanatory variable must be numeric (categorical had to be transformed via one-hot encoding)

Source:

Linear vs logistic regression



Source: https://www.jcchouinard.com/logistic-regression/

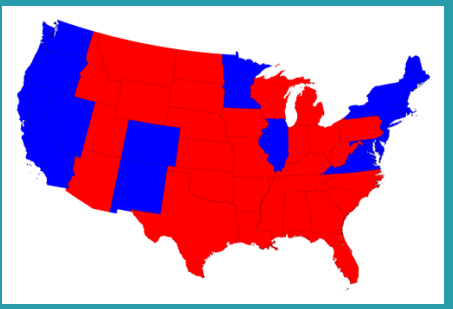
Decision tree

- Supervised technique
- Can be
  - Classification tree: target variable is categorical
  - Regression tree: target variable is numeric
- Explanatory variable can be either numeric or categorical

A lot of other variants/improvements exists: random forests, gradient boosting trees, …

Source:

Decision tree



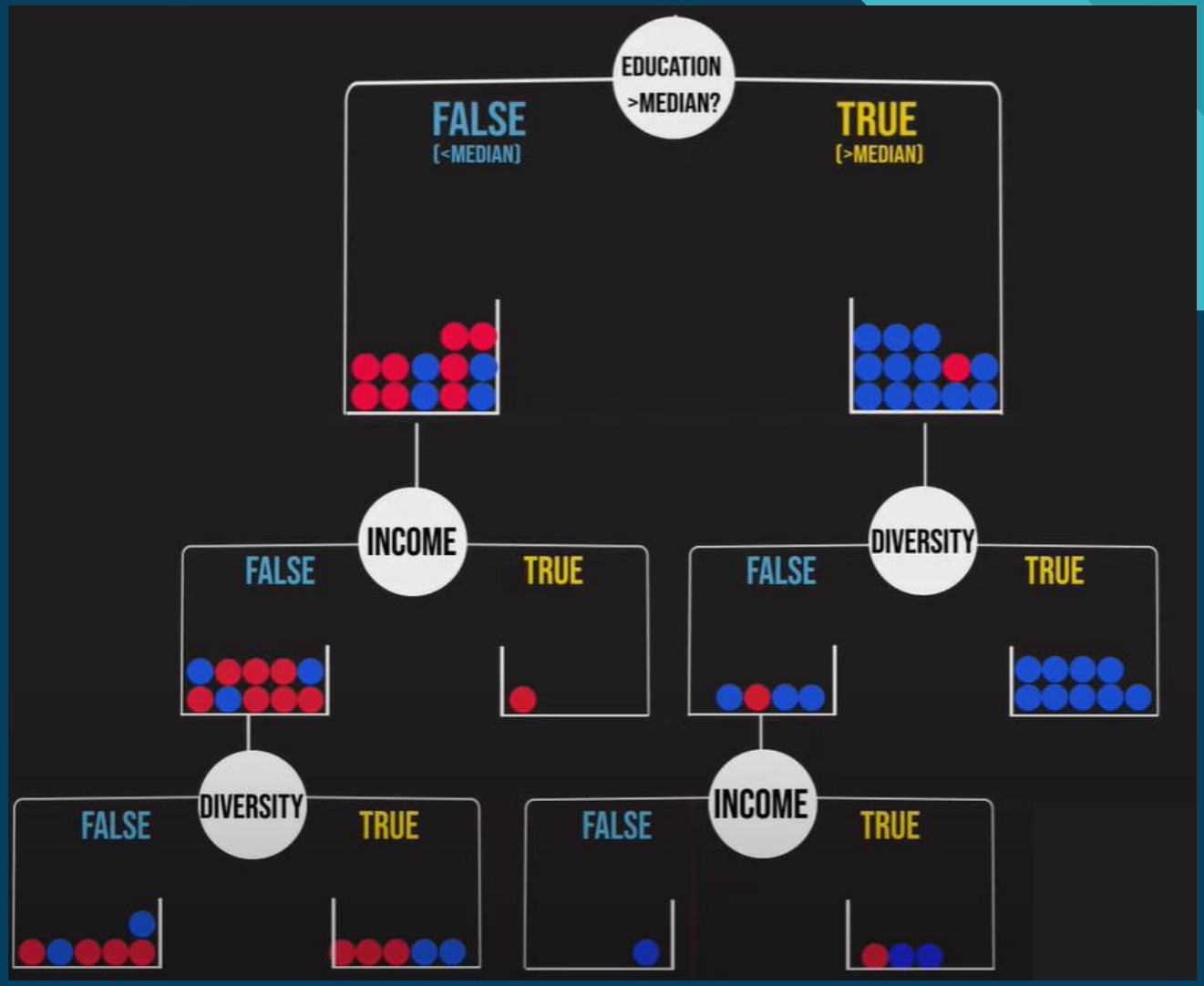
State	Median_income	%_Bachelors_degree_or_higher	%_Diversity	Political_Leaning
Illinois	70145	0.33	0.61	0
Montana	57679	0.31	0.86	1
Nevada	61864	0.24	0.49	0
New Jersey	74176	0.38	0.55	0
Mississippi	42781	0.21	0.57	1
Indiana	59892	0.25	0.79	1
Maine	58663	0.30	0.93	0
Tennessee	56060	0.26	0.74	1
West Virginia	50573	0.20	0.92	1
Missouri	61726	0.28	0.80	1
Florida	54644	0.29	0.53	1
Connecticut	72812	0.38	0.66	0
Colorado	73034	0.39	0.68	0
Hawaii	80108	0.32	0.21	0
Oregon	69165	0.32	0.75	0

Source: https://demcastusa.com/2019/11/11/land-doesnt-vote-people-do-this-electoral-map-tells-the-real-story/

https://www.youtube.com/watch?v=zs6yHVtxyv8

https://colab.research.google.com/drive/1Z1aiXravQxrUKJbk1ciTy0DZ_yhBopZS?usp=sharing#scrollTo=c-xBX3WBttPs

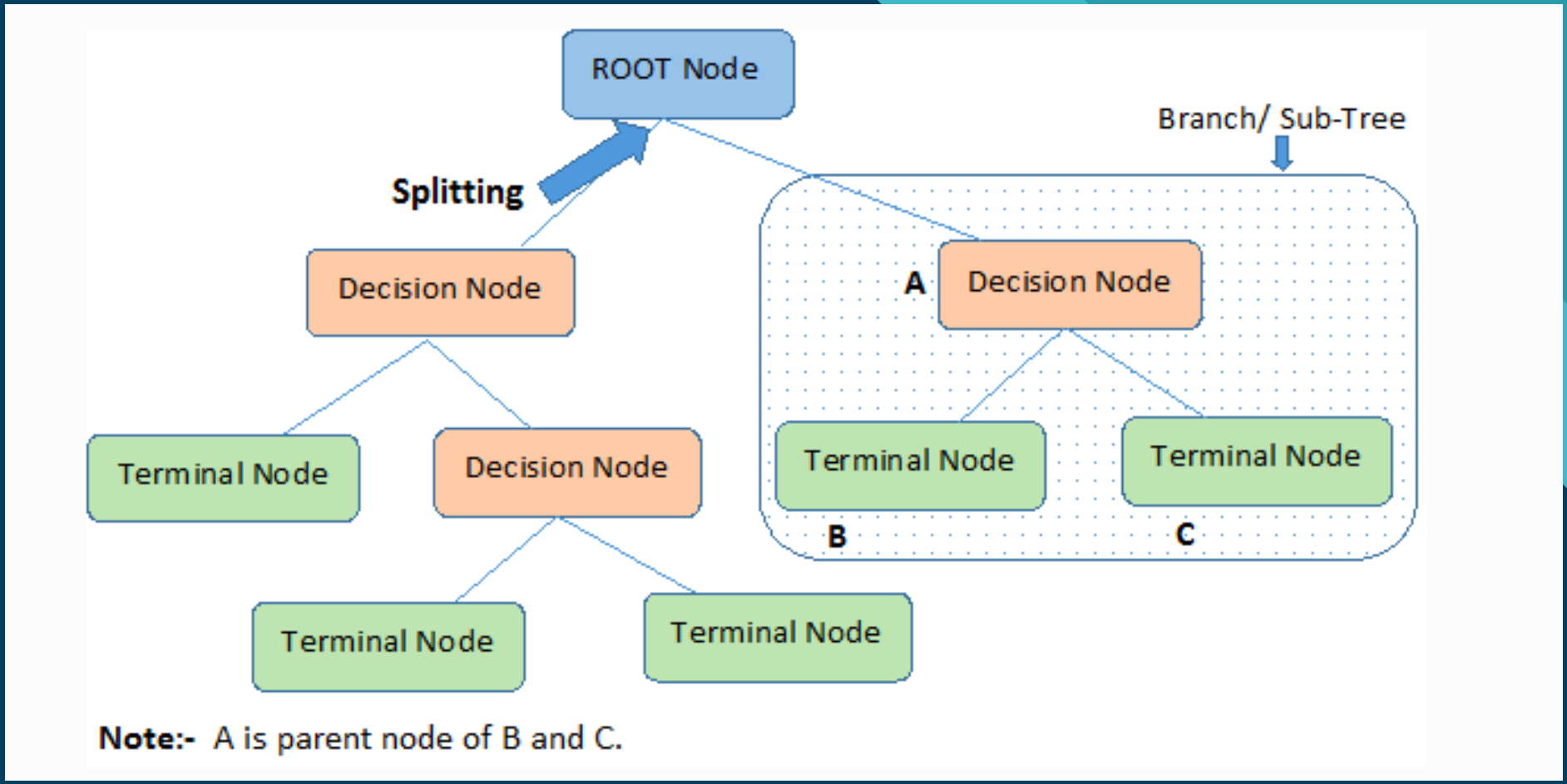
Decision tree



Source: https://www.youtube.com/watch?v=zs6yHVtxyv8

Decision tree

Terminology



Source: https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

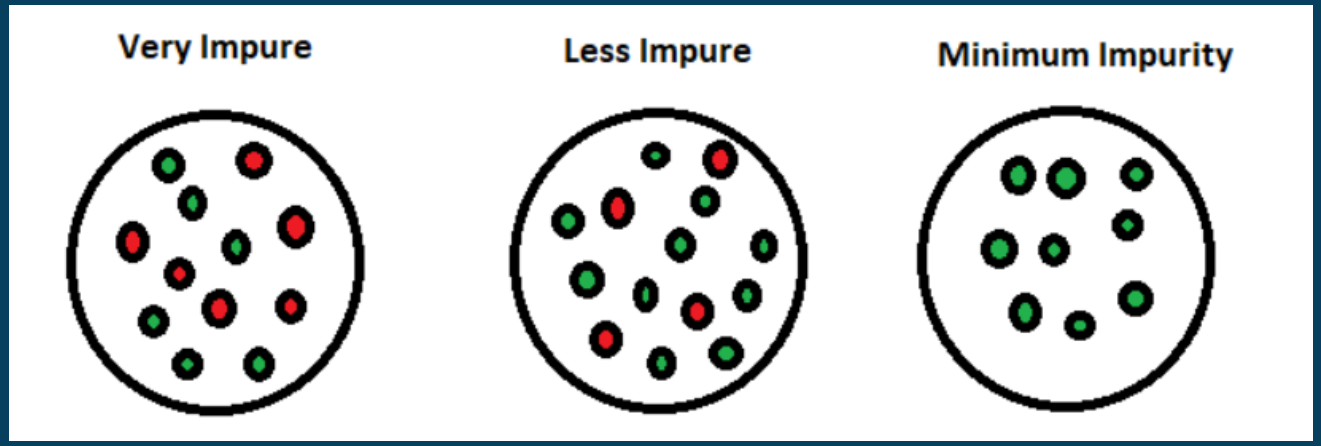
https://www.youtube.com/watch?v=JcI5E2Ng6r4

Decision tree

How the split is done? At each potential split, we measure the entropy

Entropy = impurity = measure of dissimilarity of target variable

- High entropy = observations are not very similar => bad splitting choice
- Low entropy = observation are more similar => good splitting choice

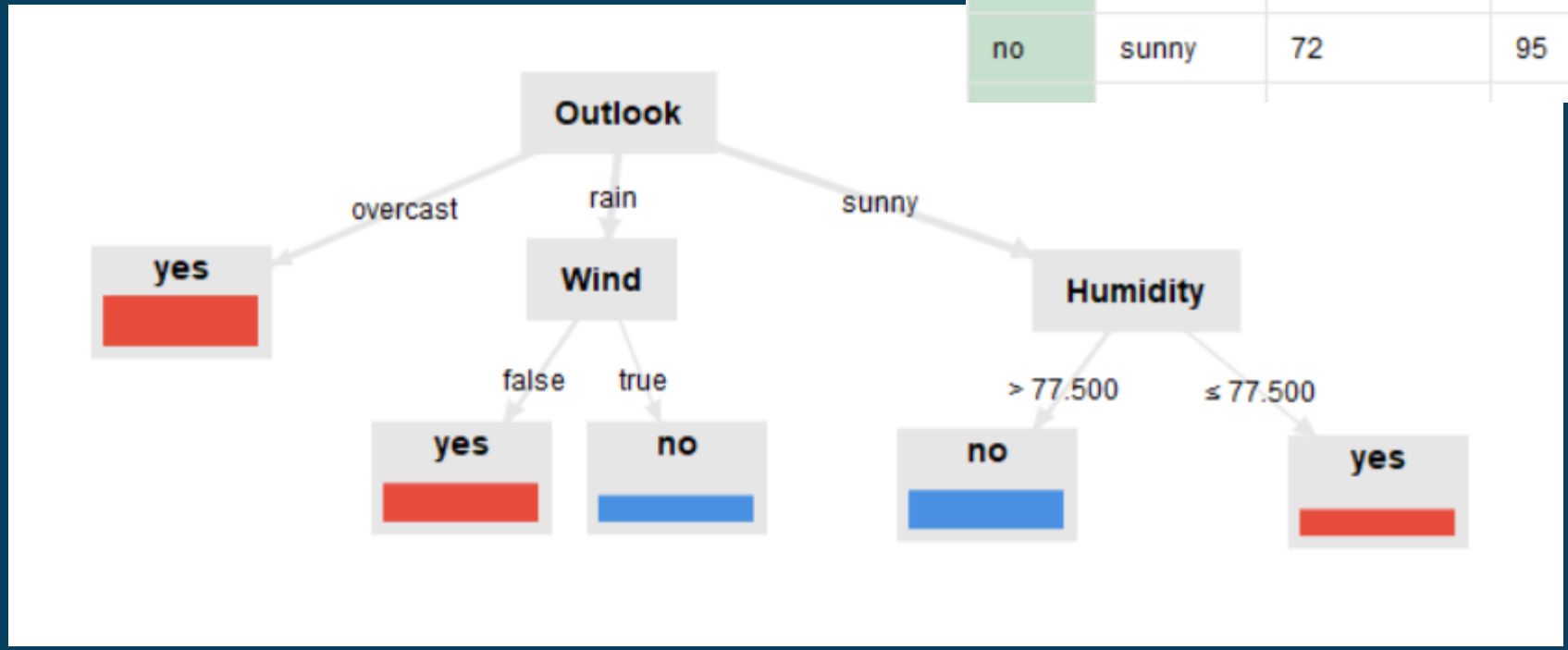


Source: https://www.section.io/engineering-education/entropy-information-gain-machine-learning/#:~:text=Entropy%20is%20an%20information%20theory,tree%20chooses%20to%20split%20data.

Decision tree - golf dataset

Example: play or not to play golf

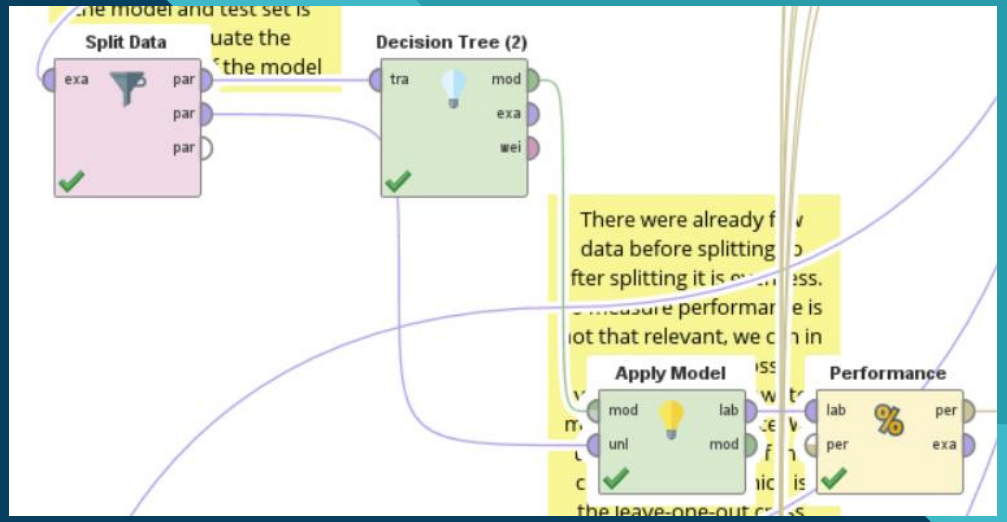
Play	Outlook	Temperature	Humidity	Wind
no	sunny	85	85	false
no	sunny	80	90	true
yes	overcast	83	78	false
yes	rain	70	96	false
yes	rain	68	80	false
no	rain	65	70	true
yes	overcast	64	65	true
no	sunny	72	95	false



Decision tree - golf dataset

Evaluate performance:

we must first split into training and test set



accuracy: 75.00%

	true no	true yes	class precision
pred. no	1	1	50.00%
pred. yes	0	2	100.00%
class recall	100.00%	66.67%	

Model performance evaluation

To **fairly** evaluate if the model performance, i.e. we look the performance of the model on unseen data.

We proceed as follows:

1) Randomly split the data into 2 part
    1) Training set (usually 70-80% initial dataset): part of the data used in the learning
    2) Test set (usually 20-30%): part of the data being scored with the learned model.

With bad luck, you could have a test set that is not representative (especially if you don't have much data) and so being very optimistic (or pessimistic) about the true performance of the model. We can solve if by repeating the above process multiple times: this is called cross-validation. This method is covered in the slides a bit later.

Model performance evaluation - metrics

How to measure the performance of my model?

- Classification tasks (categorical target variable) => confusion matrix
  - From confusion matrix, a lot of metrics can be used => Accuracy, Precision, Recall, F-measure, …

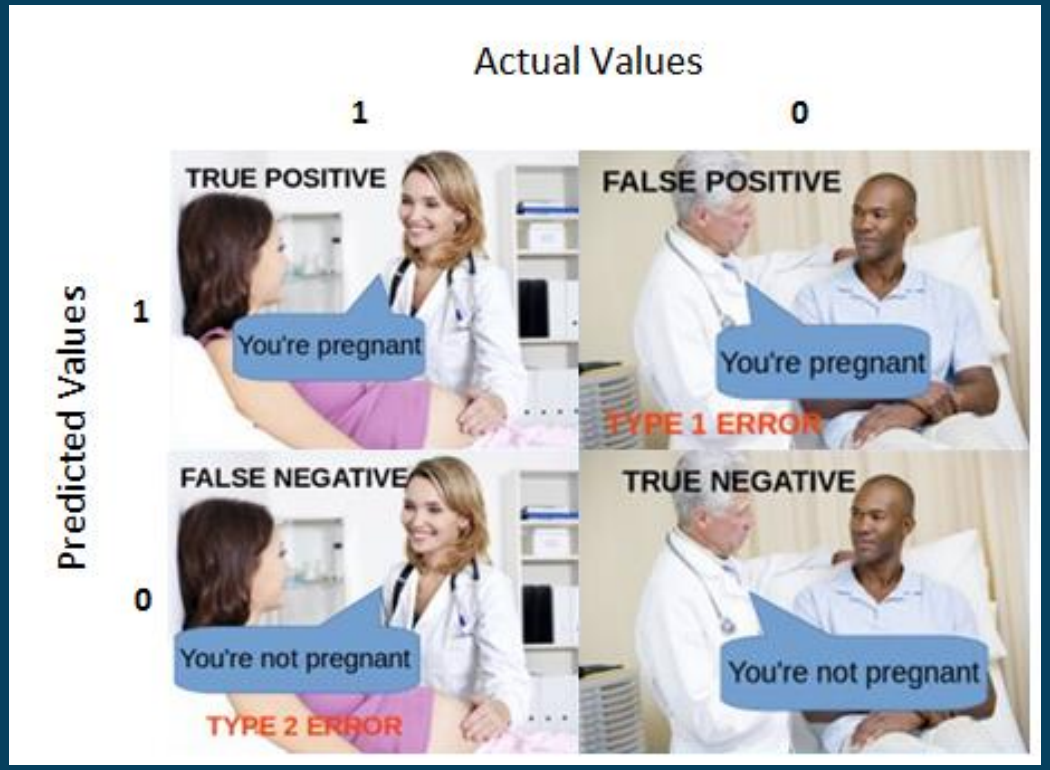
- Regression tasks (numerical target variable) => MSE, RMSE, MAE, …

Source: https://neptune.ai/blog/performance-metrics-in-machine-learning-complete-guide

Confusion matrix

Useful for classification tasks

Example of a confusion matrix for a binary classification task:

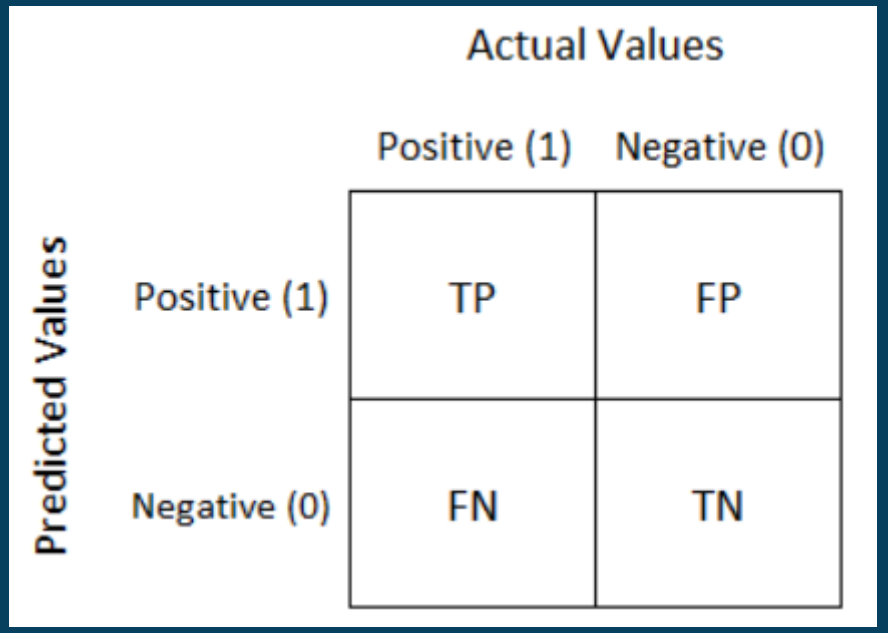


Source: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

Confusion matrix

Useful for classification tasks

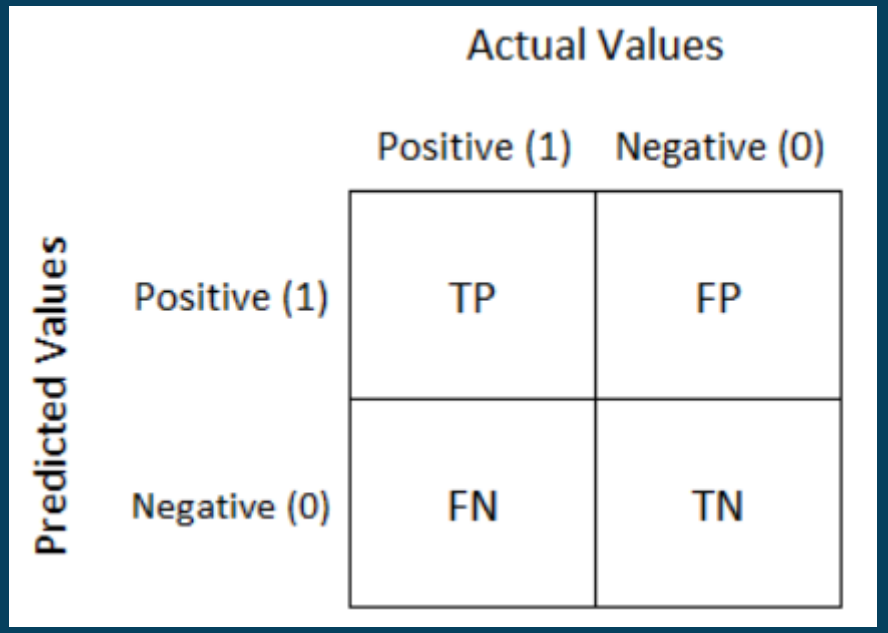
Example of a confusion matrix for a binary classification task:



Confusion matrix

Useful for classification tasks

Example of a confusion matrix for a binary classification task:



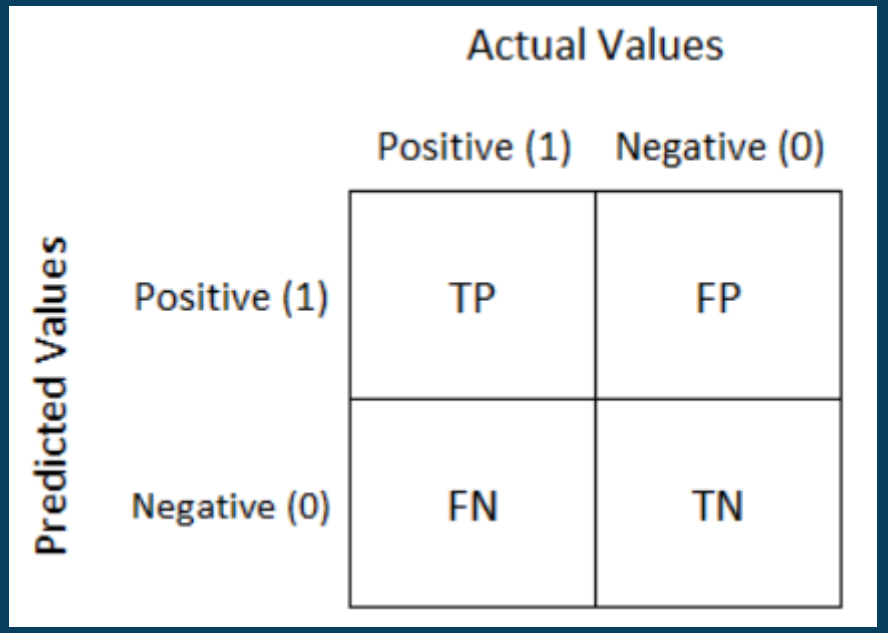
Accuracy =

$$\frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Confusion matrix

Useful for classification tasks

Example of a confusion matrix for a binary classification task:



Accuracy =

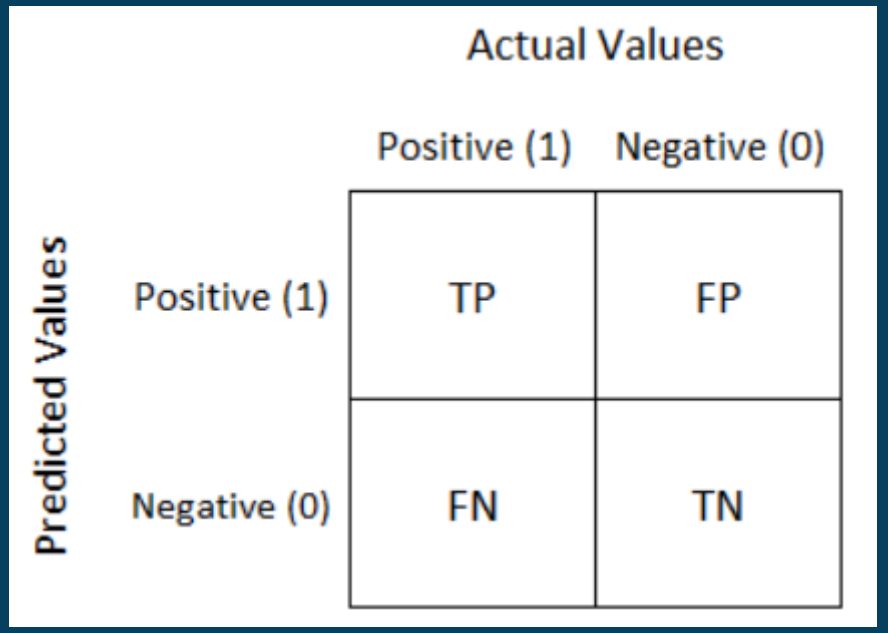
$$\frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

$$\textit{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

Confusion matrix

Useful for classification tasks

Example of a confusion matrix for a binary classification task:



Accuracy =

$$\frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

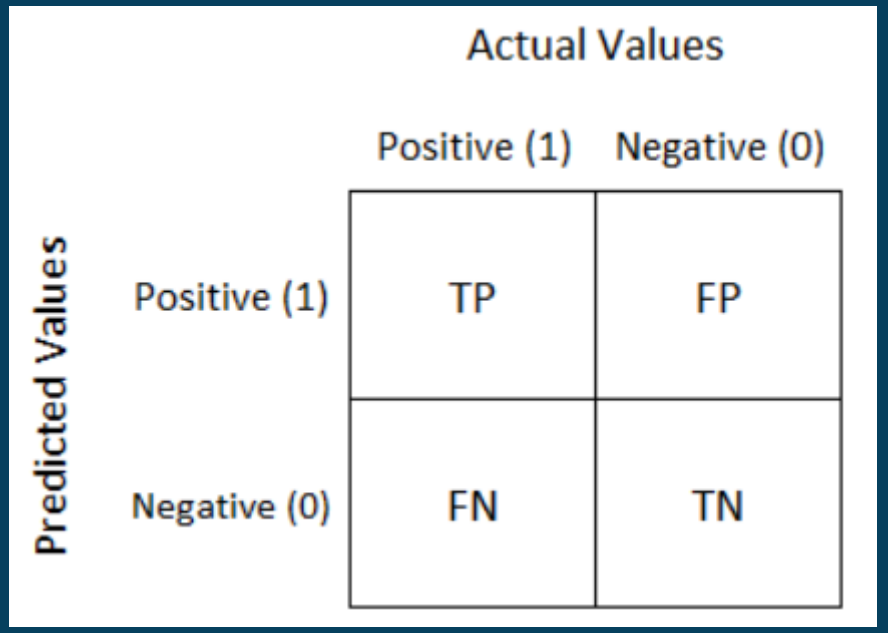
$$\textit{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$\textit{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Confusion matrix

Useful for classification tasks

Example of a confusion matrix for a binary classification task:



Accuracy =

$$\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$$

$$\textit{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$

$$\textit{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

$$\textit{F - measure} = \frac{2*\mathrm{Recall}*\mathrm{Precision}}{\mathrm{Recall}+\mathrm{Precision}}$$

Confusion matrix

Real word examples

A. Sensitivity and Specificity

356 clinical case samples (including symptomatic and asymptomatic cases) which include 138 confirmed as COVID-19 positive and 218 confirmed as COVID-19 negative by PCR assay, were obtained for testing, and then compared the test results between Wondfo 2019-nCoV Antigen Test (Lateral Flow Method) and the PCR results. The results are shown below.

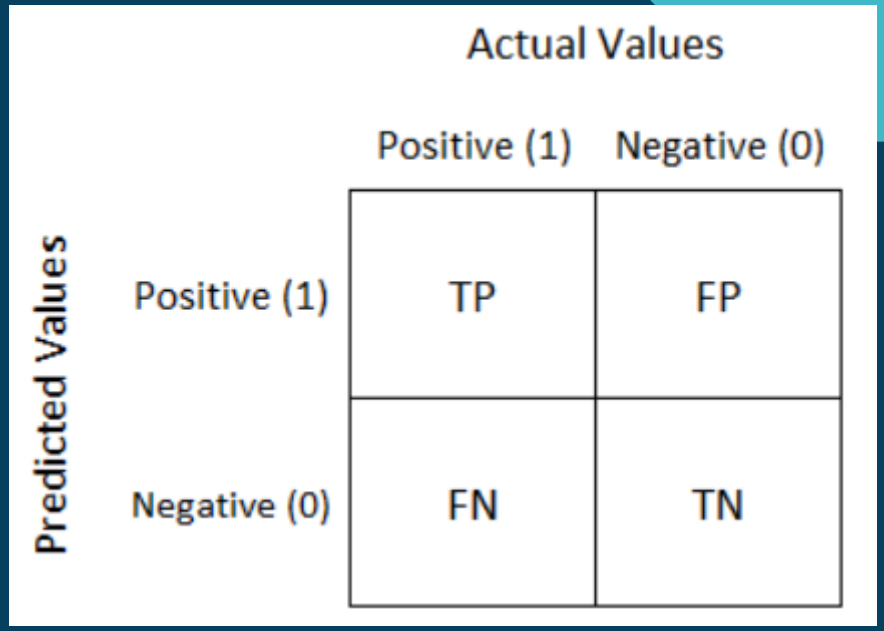
Reagent		PCR		Total
		Positive	Negative	
Wondfo 2019-nCoV Antigen Test (Lateral Flow Method)	Positive	135	2	137
	Negative	3	216	219
Total		138	218	356

Sensitivity: 97.83% (95%CI: 93.78%~99.55%)
Specificity: 99.08% (95%CI: 96.73%~99.89%)
Total agreement: 98.60% (95%CI: 96.75%~99.54%)

Discover by yourself what are sensitivity, specificity and total agreement.

Confusion matrix

- You analyze medical data to detect cancer. Which metric do you look at: precision or recall?
- You want to detect customer that will go to competitors. You want to phone them but doing useless phone calls are costly and bad for customer experience. Precision or recall?
- A vendor sells a model of churn detection model. Accuracy ratio of that model is 98.5%. The training dataset is based on people switching bank during the 90s, which was quite uncommon at that time. You trust him or not? Explain.



$$\textbf{Precision} = \frac{TP}{TP + FP}$$

$$\textbf{Recall} = \frac{TP}{TP + FN}$$

$$\textbf{F - measure} = \frac{2*Recall*Precision}{Recall + Precision}$$

Source: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

Decision tree

Pre-prunning techniques

= When do we stop building the tree?

- All records in that node have the same target value
- Maximal depth of the tree reached
- Not a minimal information gain threshold (in other words, computed entropy was not below a given threshold)
- Low number of observation in the current nodes
- …

Decision tree

Post-pruning techniques

= Once built, how do we remove some branches of the tree that does not generalize ?

- Cut branches as long as the misclassification error (for classification tasks) (or MSE (for regressions tasks)) decrease

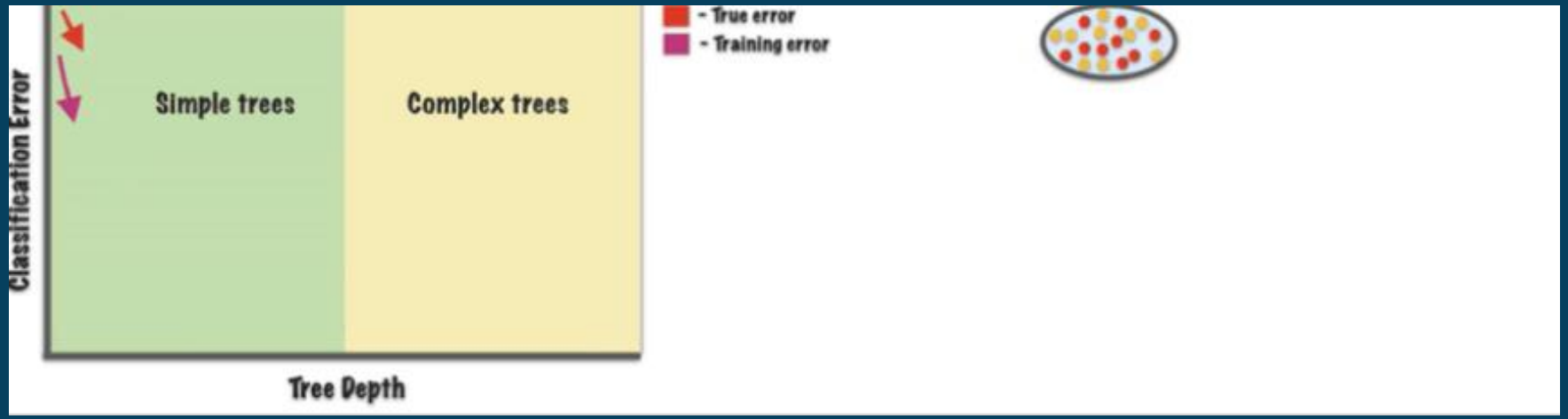
Source: https://towardsdatascience.com/an-exhaustive-guide-to-classification-using-decision-trees-8d472e77223f

Decision tree

Post-pruning techniques

= Once built, how do we remove some branches of the tree that does not generalize ?

- Cut branches as long as the misclassification error (for classification tasks) (or MSE (for regressions tasks)) decrease



Source: https://towardsdatascience.com/an-exhaustive-guide-to-classification-using-decision-trees-8d472e77223f

Cross-validation

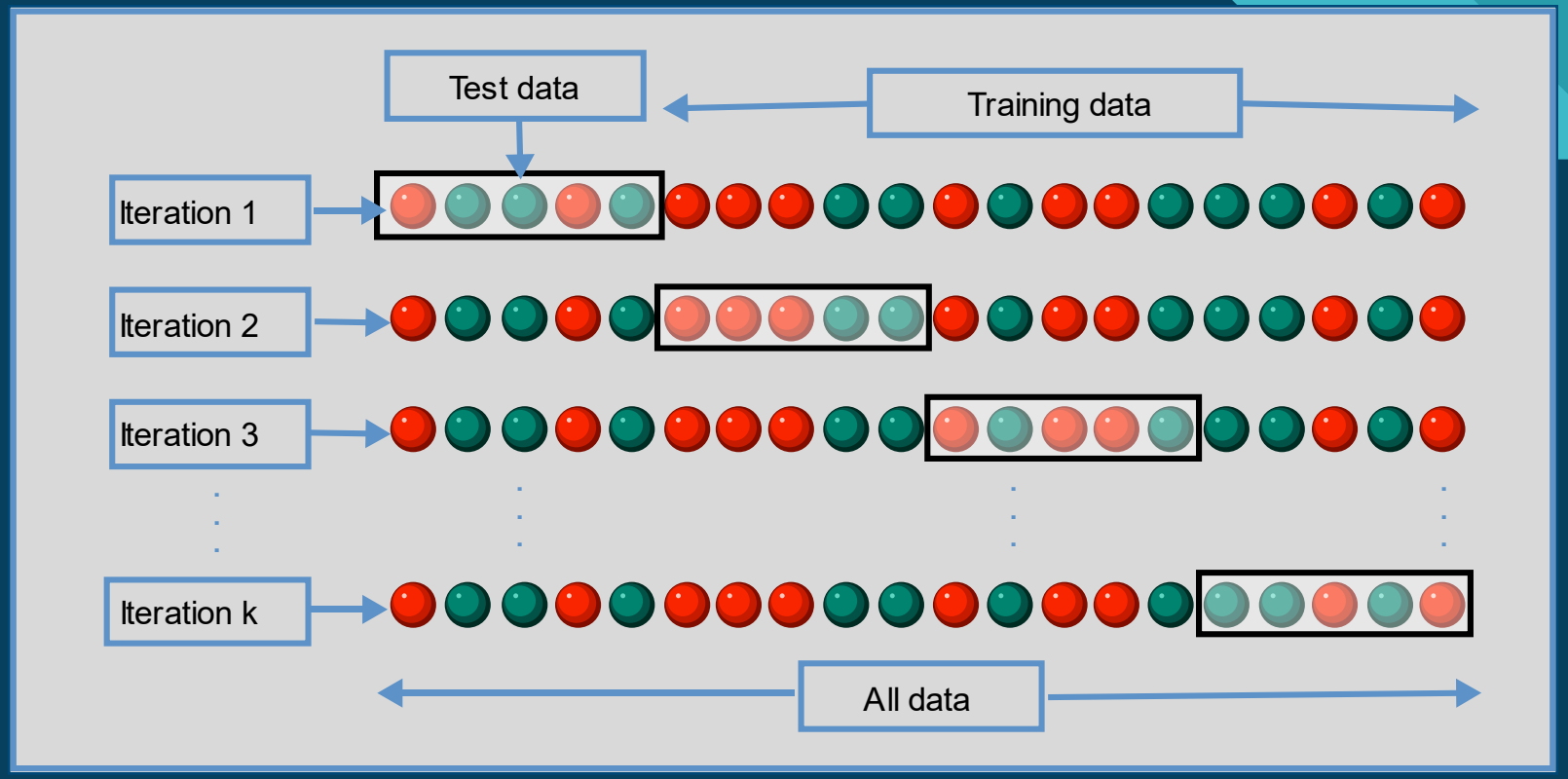
How to evaluate performance of the algorithm without being so dependent of the split between training and test set?

Use cross-validation to evaluate if a model is generalizing correctly

Used for supervised learning (for classification or regression tasks)

K-fold cross-validation

How does it work?



Source: https://en.wikipedia.org/wiki/Cross-validation_(statistics)

Decision tree – golf dataset – cross-validation

Which model do we choose based on the results (via cross-validation) of 2 different models?

For golf dataset, we use leave-one out cross validation because really few observations so we really need all of them (except one) to learn something.

accuracy: 78.57% +/- 42.58% (micro average: 78.57%)

	true no	true yes	class precision
pred. no	3	1	75.00%
pred. yes	2	8	80.00%
class recall	60.00%	88.89%	

accuracy: 78.57% +/- 42.58% (micro average: 78.57%)

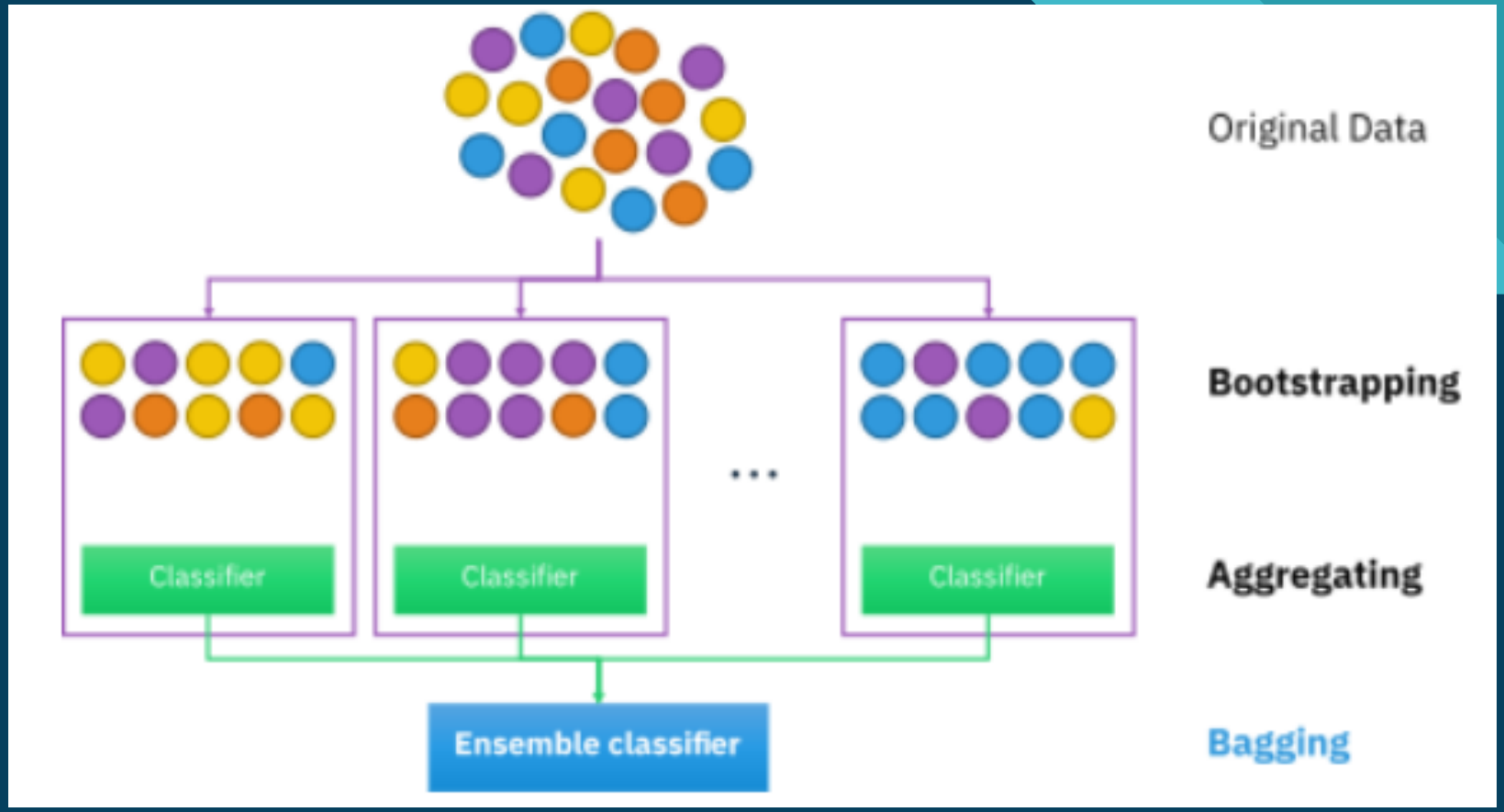
	true no	true yes	class precision
pred. no	4	2	66.67%
pred. yes	1	7	87.50%
class recall	80.00%	77.78%	

Random forest

Wisdom of the Crowds

Source: https://ai-pool.com/a/s/random-forests-understanding

Random forest



Wisdom of the Crowds

Source: https://ai-pool.com/a/s/random-forests-understanding

Random forest

Pros

- Generalize better well => less overfitting
- Can easily determine importance of each explanatory variables

Cons

- Time consuming (but tasks can be parallelized)
- More difficult to visualize

Source: https://www.ibm.com/cloud/learn/random-forest

Text mining

Text mining

Techniques meant for analyzing textual data:

- Text classification (supervised learning)
  - detecting the subject of a mail and forward it to the correct department
  - Spam filtering, …
- Sentiment analysis (supervised and/or rules)
  - evaluate if the customer feedback is rather positive or negative
- Named entity recognition (supervised and/or rules)
  - Extract amount, customer information, date from an invoice, …
- Topic modelling (unsupervised, +- clustering)
  - Find similar documents, …

Regular expressions

Language to check if a string contains/follows a given pattern using regular expressions

Aka “Regex”

Examples:

- Validate the format of a phone number, national register number, account number, date, email, …
- Extract an amount from an invoice,

Regular expressions

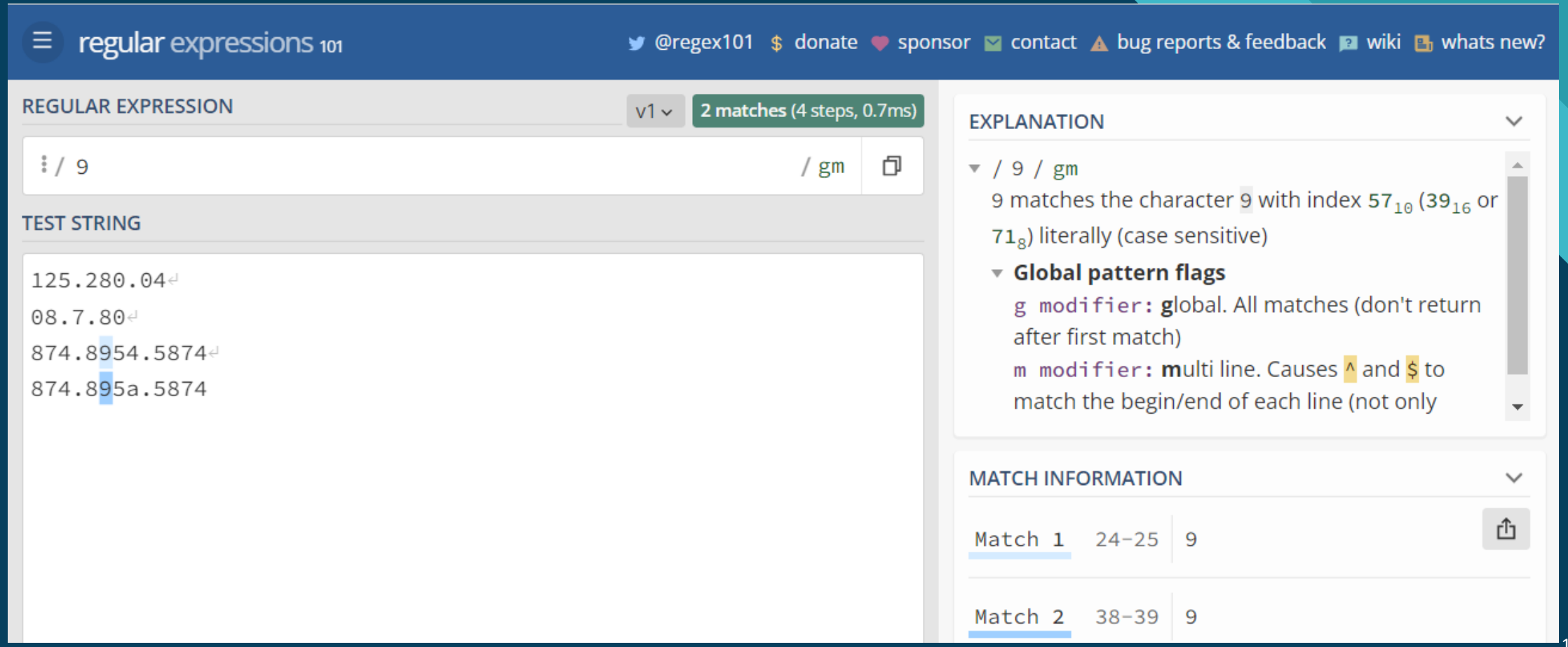
Very good website to develop his regular expressions skill!

Free + interactive + explanation:

- https://regex101.com/

Regular expressions

https://regex101.com/



Regular expressions

Let's use https://regex101.com/r/ji0u5M/1 to do the exercice!

Build the following regular expressions

1. Contains digit 9
2. Contains at least 4 consecutive digits
3. Start with digit 8
4. End with digit 74
5. Has a length between 7 and 10
6. Contains at least 2 digit between the first . and second .
7. Does not start with digit 0 or 1
8. Contains any letter
9. Follows this structure: at least 2 digits, followed with ., between 1 and 3 digits, followed with ., between 2 and 3 digits

This is it!