Exploratory Data Analysis (EDA)

Panita Thusaranon, Ph.D.

Overview of Data Analytics Process

Step 1

Business Issue Understanding

- Define business objectives
- Gather required information
- Determine appropriate analysis method
- Clarify scope of work
- Identify deliverables

Step 2

Data Understanding

- Collect initial data
- Identify data requirements
- Determine data availability
- Explore data and characteristics

Step 3

Data Preparation

- Gather data from multiple sources
- Cleanse
- Format
- Blend
- Sample

Step 4

Exploratory Analysis and Modeling

- Develop methodology
- Determine important variables
- Build model
- Assess model

Step 5

Validation

- Evaluate results
- Review process
- Determine next steps

Results are valid → proceed to step 6

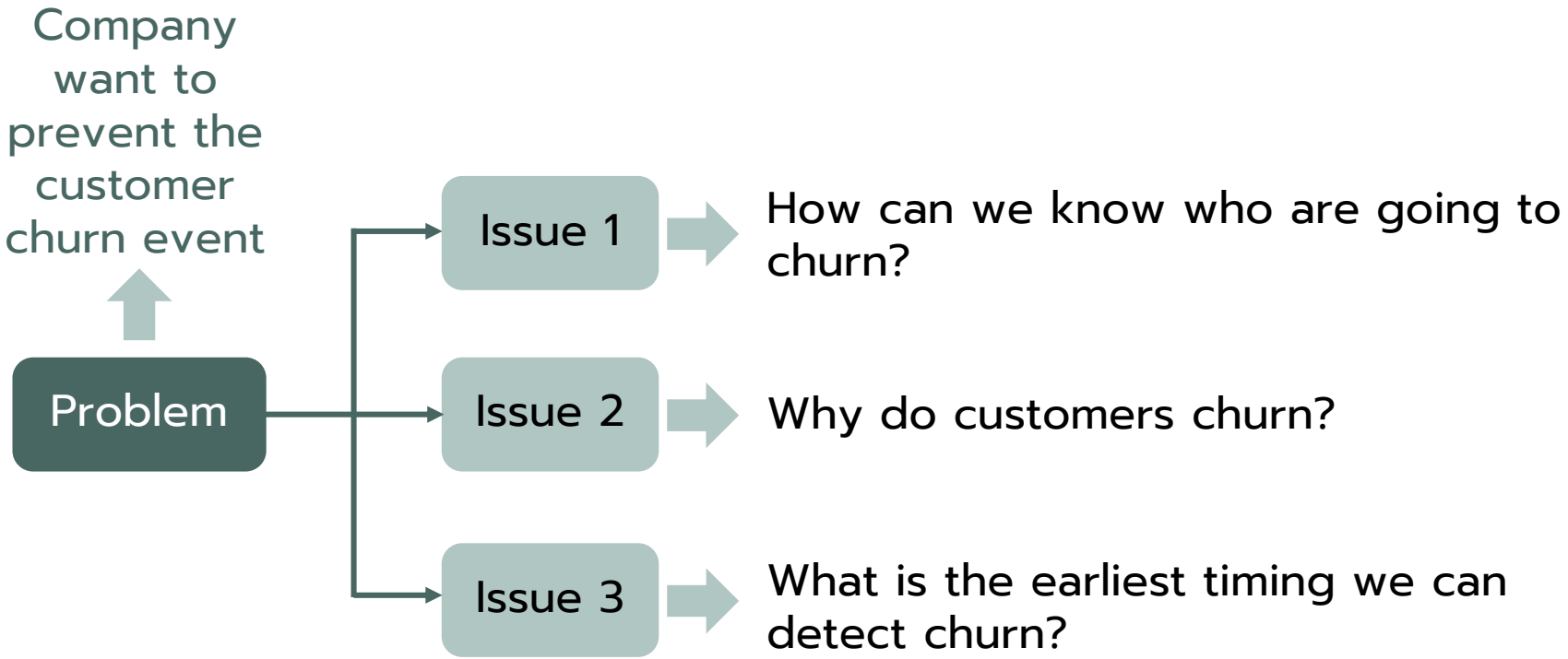
Results are invalid ← revisit steps 1-4

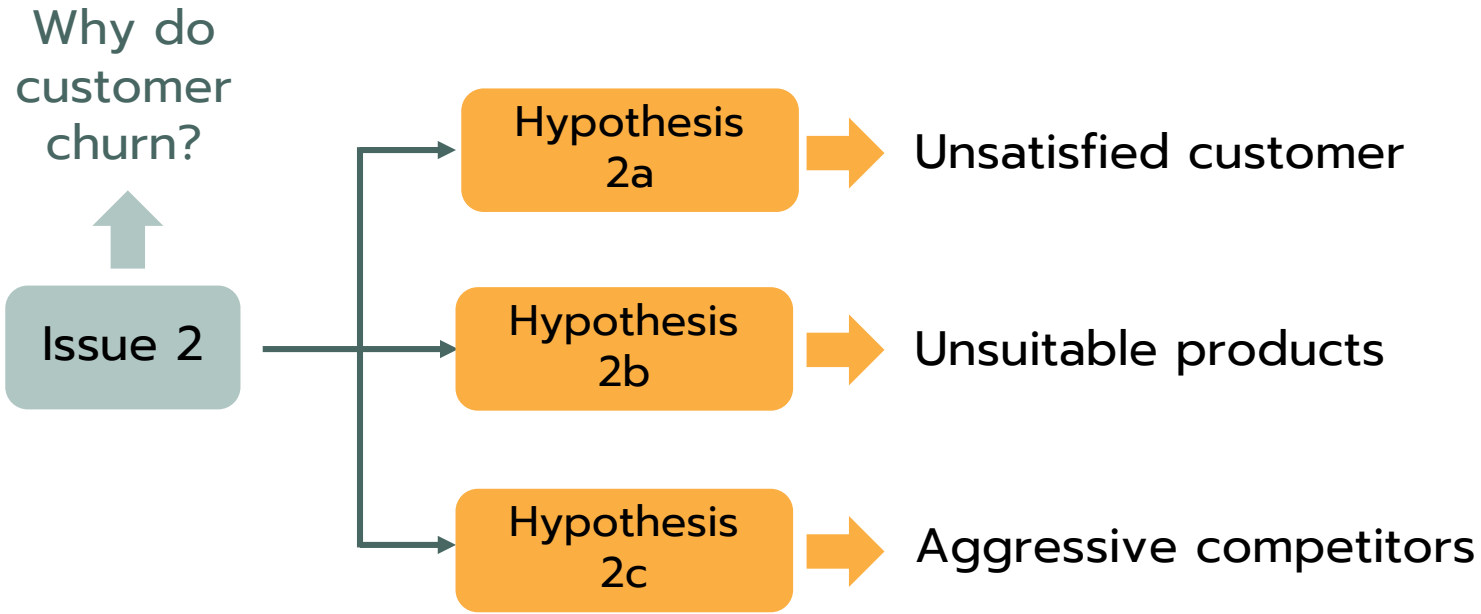
Step 6

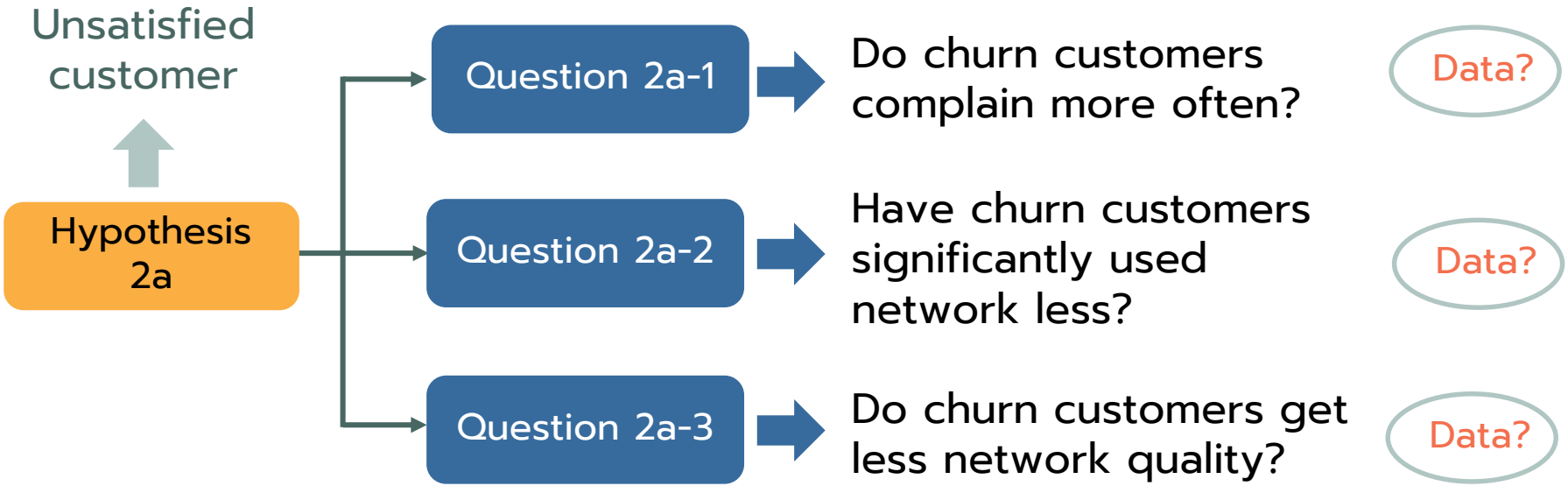
Visualization and Presentation

- Communicate results
- Determine best method to present insights based on analysis and audience
- Craft a compelling story
- Make recommendations

Business Issue Understanding




Why do customer churn?
Issue 2
Hypothesis 2a
Unsatisfied customer
Hypothesis 2b
Unsuitable products
Hypothesis 2c
Aggressive competitors


Unsatisfied customer
Hypothesis 2a
Question 2a-1
Question 2a-2
Question 2a-3
Do churn customers complain more often?
Have churn customers significantly used network less?
Do churn customers get less network quality?
Data?
Data?
Data?

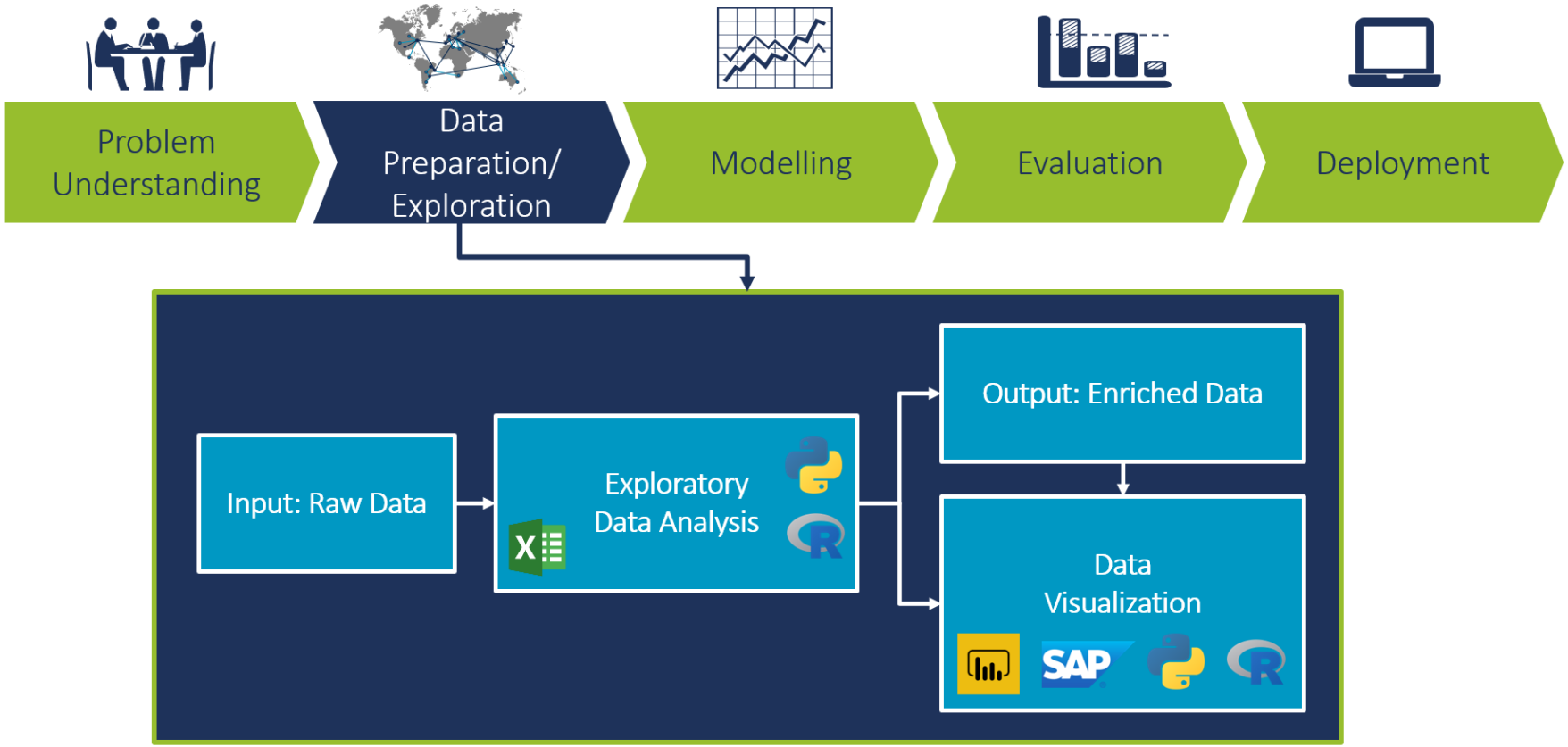
Gathering relevant data and information

- Critical step to support analyses (proving or disproving the hypotheses)
- Know where to dig
- Know how to filter through information
- Know how to verify
- Know how to apply

Question 2a-1

Do churn customers complain more often?

- Customer ID
- Number of contacts
- Sentiment of previous contacts
- Customer status
- Services that each customer has used



https://blog.camelot-group.com/2019/03/exploratory-data-analysis-an-important-step-in-data-science/

What is EDA?

กระบวนการตรวจสอบ สำรวจข้อมูลเบื้องต้น เพื่อตั้งสมมติฐานให้ได้ว่า เราสนใจอยากจะเปรียบเทียบตัวเลขอะไร เปรียบเทียบไปทำไม และอะไรบ้างที่ควรจะถูกนำมาเปรียบเทียบ ก่อนนำไปทดสอบสมมติฐาน วิเคราะห์ข้อมูลอย่างละเอียด หรือสร้างแบบจำลองทางสถิติต่อไป

https://blog.skooldio.com/increase-your-customer-with-exploratory-data-analysis/

ประโยชน์ของการทำ EDA

ช่วยให้เราสร้างความรู้ความเข้าใจพื้นฐานเกี่ยวกับข้อมูลนั้นๆ

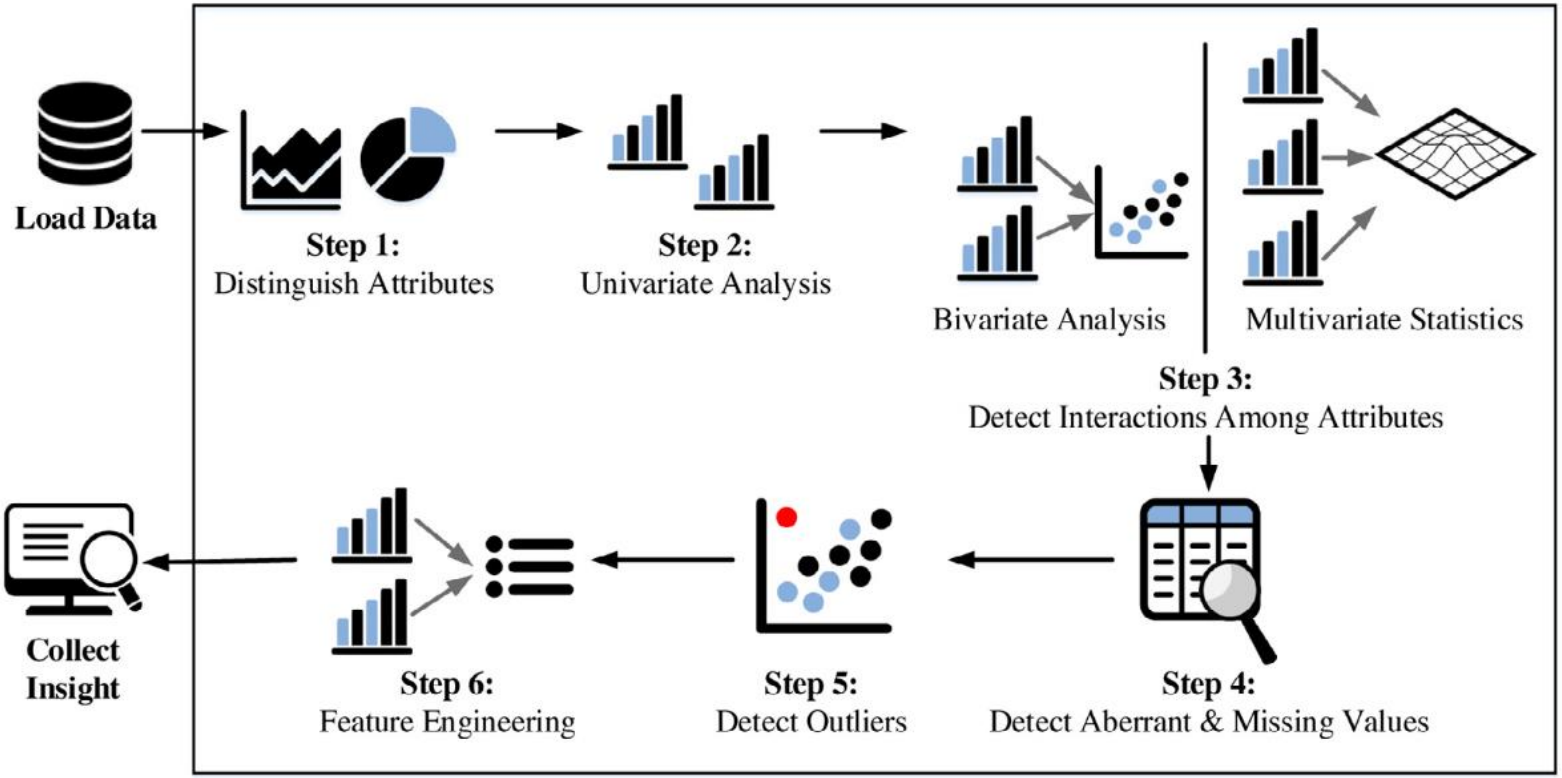
ตรวจสอบสมมติฐานเบื้องต้น ตรวจความผิดพลาดของชุดข้อมูล

ทำให้เห็น outlier หรือค่าที่โดดออกมาจากค่าปกติ เพื่อกันกันความผิดเพี้ยนตอนนำข้อมูลไปคำนวณ

เข้าใจข้อมูล มองเห็น Trends, Patterns หรือ Insights ต่างๆ

https://blog.skooldio.com/increase-your-customer-with-exploratory-data-analysis/

Steps of Data Exploration & Preparation



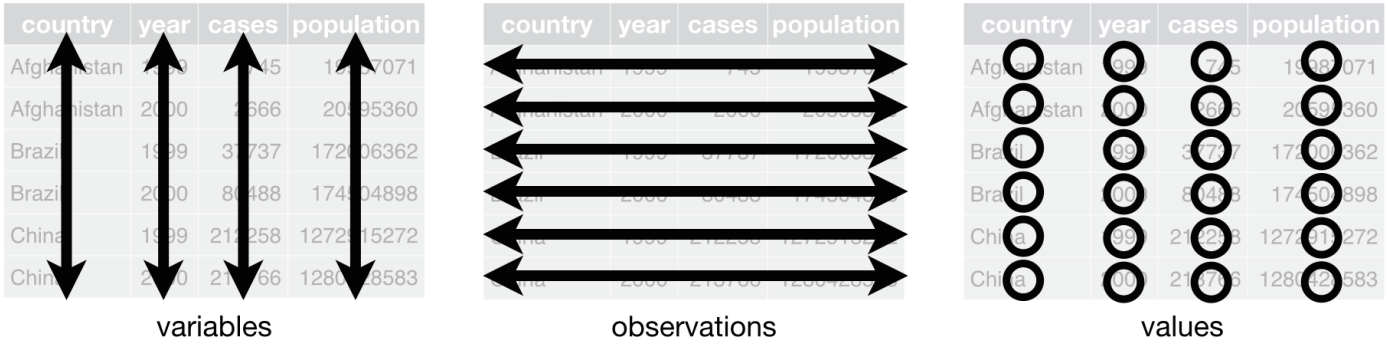
1

Distinguish Attributes

Step 1: Distinguish Attributes

There are three interrelated rules with make a dataset tidy

- Each variable must have its own column
- Each observation must have its own row
- Each value must have its own cell



https://r4ds.had.co.nz/tidy-data

Step 1: Distinguish Attributes

You can represent the same underlying data in multiple ways. The example below shows the same data organised in two different ways. Each dataset shows the same values of four variables country, year, population, and cases, but each dataset organises the values in a different way.

country	year	type	count
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999	cases	212258
China	1999	population	1272915272
China	2000	cases	213766
China	2000	population	1280428583

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

Step 1: Distinguish Attributes

	1999		2000	
country	cases	population	cases	population
Afghanistan	745	19987071	2666	20595360
Brazil	37737	172006362	80488	174504898
China	212258	1272915272	213766	1280428583

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

https://r4ds.had.co.nz/tidy-data

Step 1: Distinguish Attributes

Exploratory data analysis begins with identification of the attributes in a dataset

- **Target/Response/Outcome/Dependent variable** (output): response
- **Predictor/Independent** (input): age, job, gender, marital, education, housing, loan, income, campaign

id	age	job	gender	marital	education	housing	loan	Annual income (k$)	Spending score (1-100)	campaign	response
1	56	housemaid	Female	married	basic.4y	no	no	15	39	A	no
2	57	services	Female	married	high.school	no	no	15	81	A	yes
3	37	services	Male	married	high.school	yes	no	16	6	B	no
4	40	admin	Male	single	basic.6y	no	no	16	77	A	no
5	56	services	Female	married	high.school	no	yes	17	40	B	yes

Step 1: Distinguish Attributes

Data Type of variables

- **Numerical**: age, income, spending score
- **Categorical**: job, gender, marital, education, housing, loan, campaign

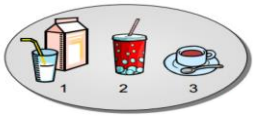
id	age	job	gender	marital	education	housing	loan	Annual income (k$)	Spending score (1-100)	campaign	response
1	56	housemaid	Female	married	basic.4y	no	no	15	39	A	no
2	57	services	Female	married	high.school	no	no	15	81	A	yes
3	37	services	Male	married	high.school	yes	no	16	6	B	no
4	40	admin	Male	single	basic.6y	no	no	16	77	A	no
5	56	services	Female	married	high.school	no	yes	17	40	B	yes

The Four Scales of Measurement

Nominal Scale

Used for naming variables in no particular order

For example, eye colour



Ordinal Scale

Used for variables in ranked order, but the difference between is not determined

For example, #1 happy, #2 neutral, #3 unhappy

Interval Scale

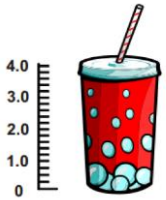
Used for numerical variables with known equal intervals of the same distance

For example, time

Ratio Scale

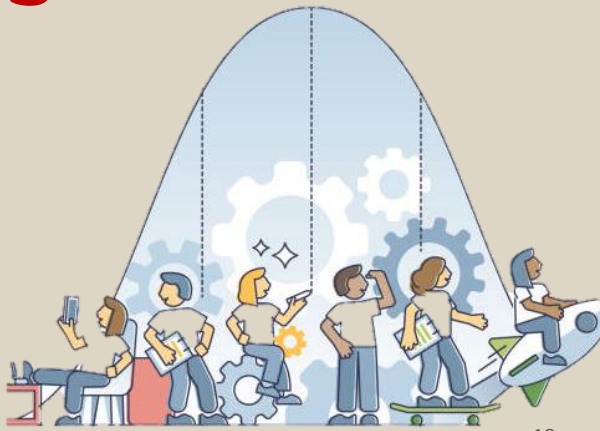
Used for variables on a scale that have measurable intervals

For example, weight



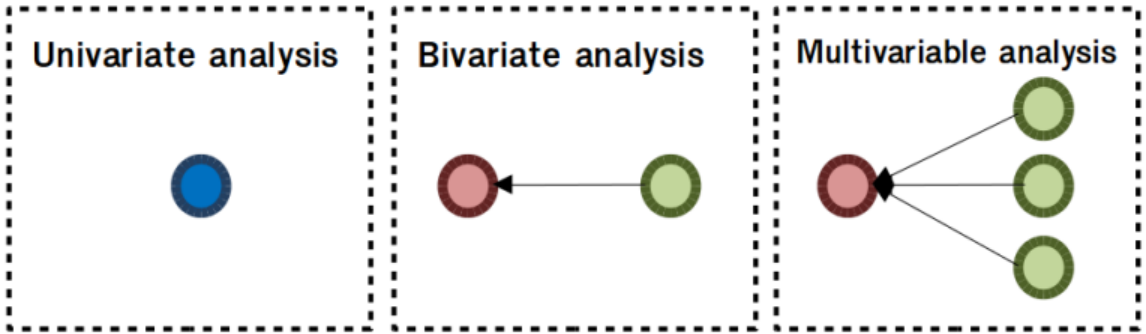
https://studyonline.unsw.edu.au/blog/types-of-data

2 Univariate Analysis
การวิเคราะห์ข้อมูล 1 ตัวแปร



Step 2: Univariate Analysis

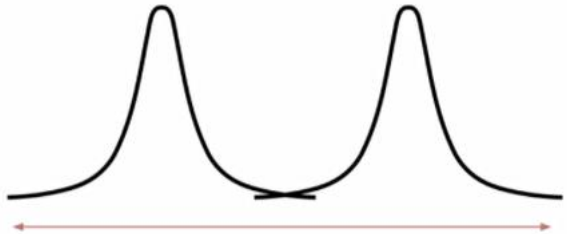
- Explore variable **one by one**
- Perform analysis depends on whether the variable type is continuous or categorical



Step 2: Univariate Analysis

Central Tendency

- **Mean**: the average values
- **Median**: the middle number in sorted data
- **Mode**: the most frequent value

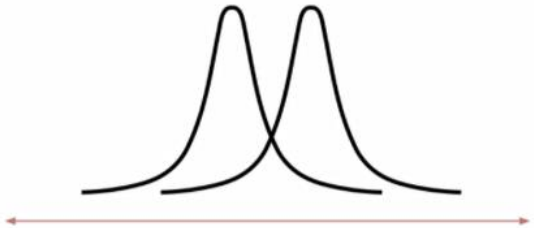


Sorted Data: 2, 3, 3, 3, 5, 7, 8, 8, 10

Mean => (2 + 3 + 3 + 3 + 5 + 7 + 8 + 8 + 10)/9 = 5.44

Median => 2, 3, 3, 3, **5**, 7, 8, 8, 10

Mode => 2, **3**, **3**, **3**, 5, 7, 8, 8, 10



Mean vs. Median

ID	รายได้ (บาท)
1	15,000
8	18,500
2	22,000
4	25,000
9	25,000
3	30,000
7	30,000
6	35,000
10	35,000
5	45,000
11	10,000,000

- ค่ากลางสำหรับข้อมูลที่เป็นตัวเลขมี 2 ค่า คือ
  - ค่าเฉลี่ย (Mean) และ ค่ากลาง หรือ มัธยฐาน (Median)
  - ข้อเสียของ **ค่าเฉลี่ย (Mean) จะอ่อนไหวกับค่าที่เป็น outlier** หรือค่าที่สูงหรือน้อยกว่าปกติ เช่น ในตารางจะได้ ค่าเฉลี่ย (mean) ของรายได้เท่ากับ 934,591 บาท
  - แต่ค่า**มัธยฐาน (Median) จะไม่อ่อนไหวกับค่า Ouliter** ดังนั้น ค่ามัธยฐานของรายได้จะเท่ากับ 30,000 บาท ซึ่งเป็นตัวแทนของรายได้คนส่วนใหญ่มากกว่า

Reference: Cube Analytics Consulting Co., Ltd.

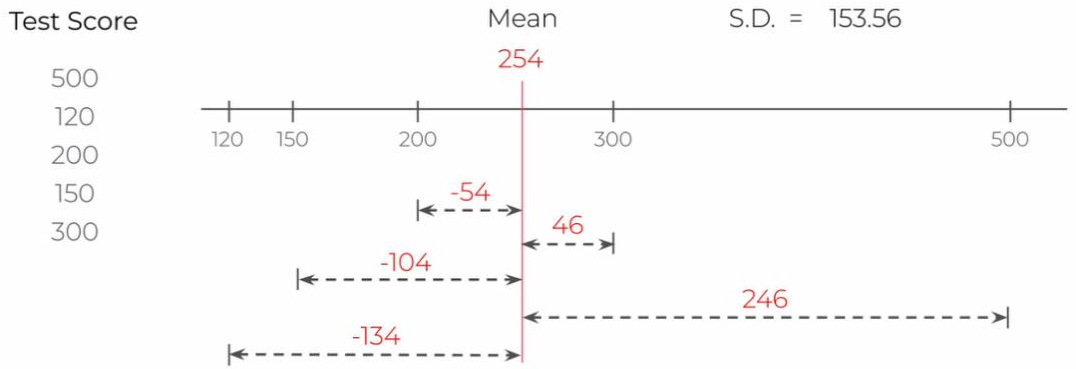
https://www.facebook.com/datacube.th

Step 2: Univariate Analysis

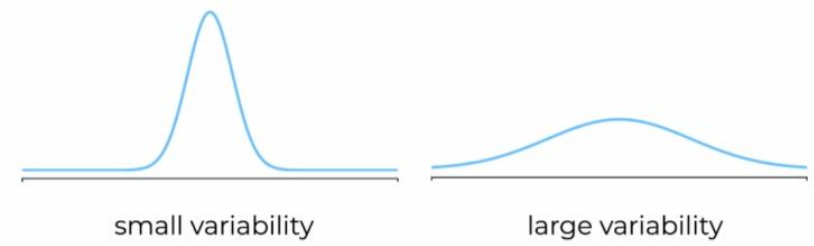
Spread

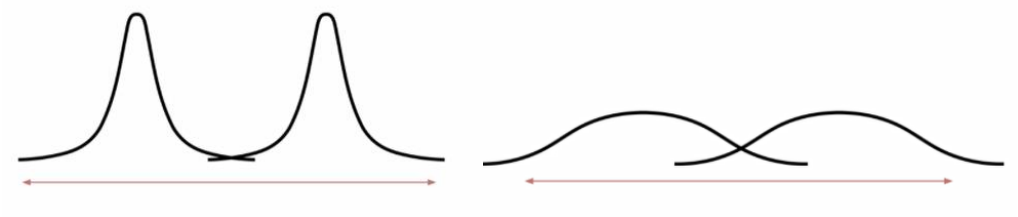
- **Range**: max - min
- **Variance** (s^2):

$$s^2 = \frac{\sum (X - \bar{X})^2}{N - 1}$$



- **Standard deviation (SD)**: square root of the variance

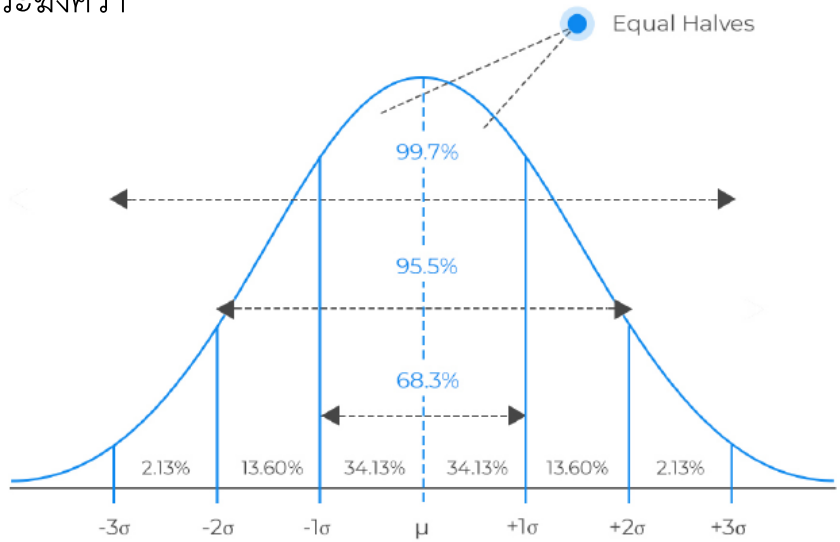




Step 2: Univariate Analysis

Distribution

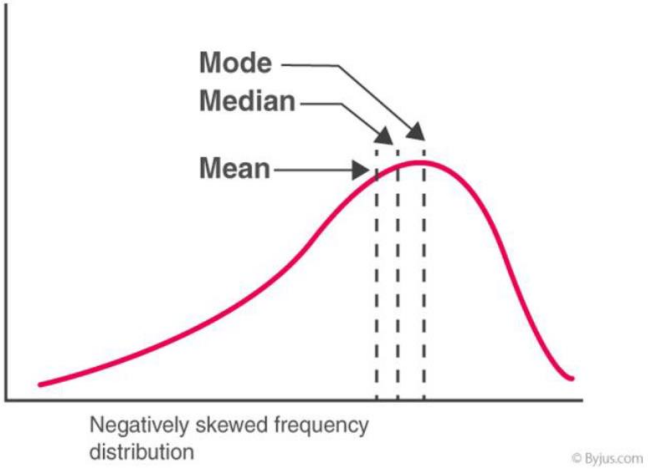
- สามารถ plot การกระจายตัวของข้อมูลในรูปแบบของ curve ได้โดยทั่วไปเราอยากให้ข้อมูลอยู่ในรูปแบบของ Normal Distribution หรือรูประฆังคว่ำ



Step 2: Univariate Analysis

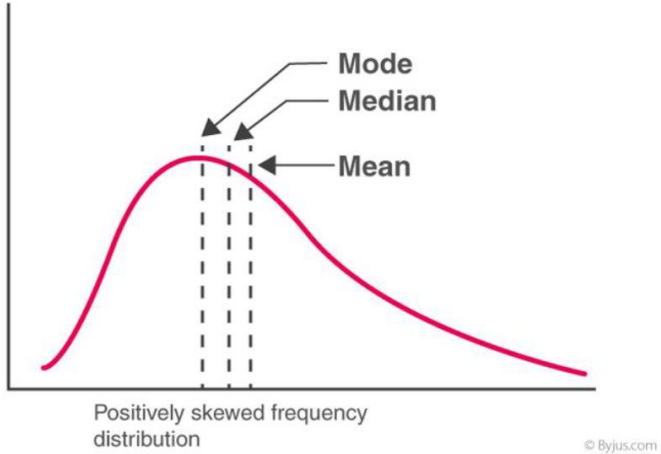
Skewness

- ค่าวัด **ความสมมาตรของข้อมูล** (ค่าความเบ้ของข้อมูล)



การแจกแจงข้อมูลแบบเบ้ซ้าย

Negative Skewed Distribution



การแจกแจงข้อมูลแบบเบ้ขวา

Positive Skewed Distribution

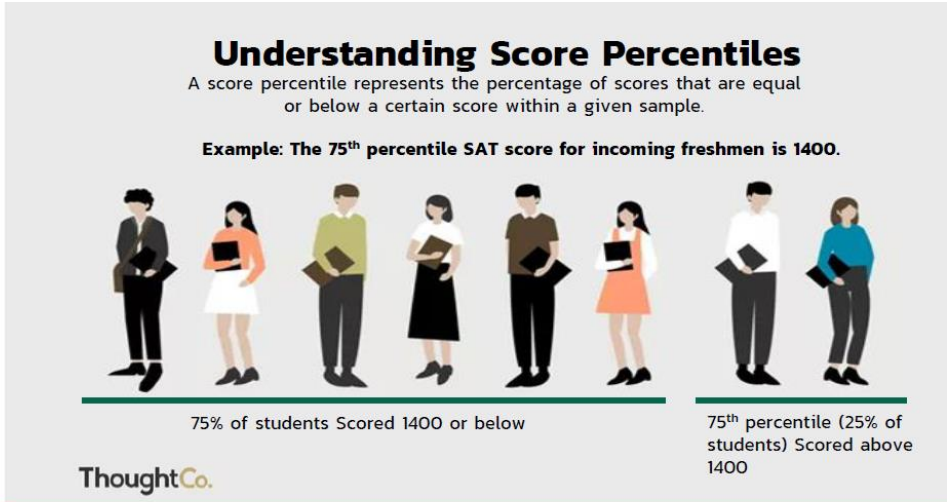
Step 2: Univariate Analysis

Percentile

- A measure indicating the value below which a given percentage of observations in a group of observations falls

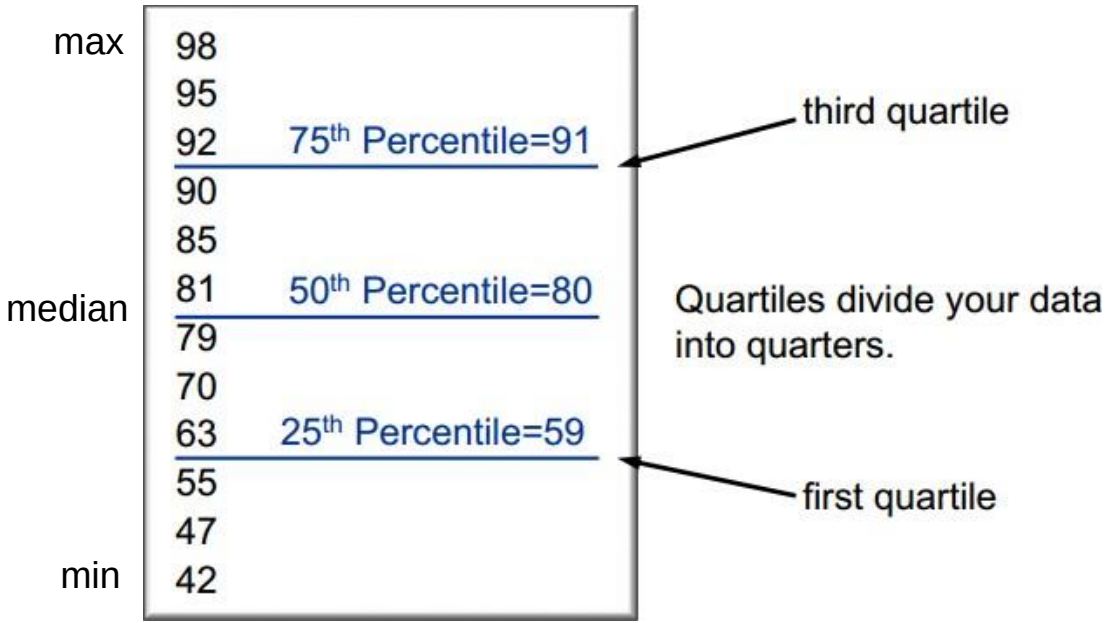
ตัวอย่าง

เปอร์เซ็นไทล์ (Percentile)



เปอร์เซ็นไทล์ที่ 75% ของคะแนนสอบ
สำหรับการจัดสอบทั่วประเทศ
เท่ากับ 1,400 คะแนน

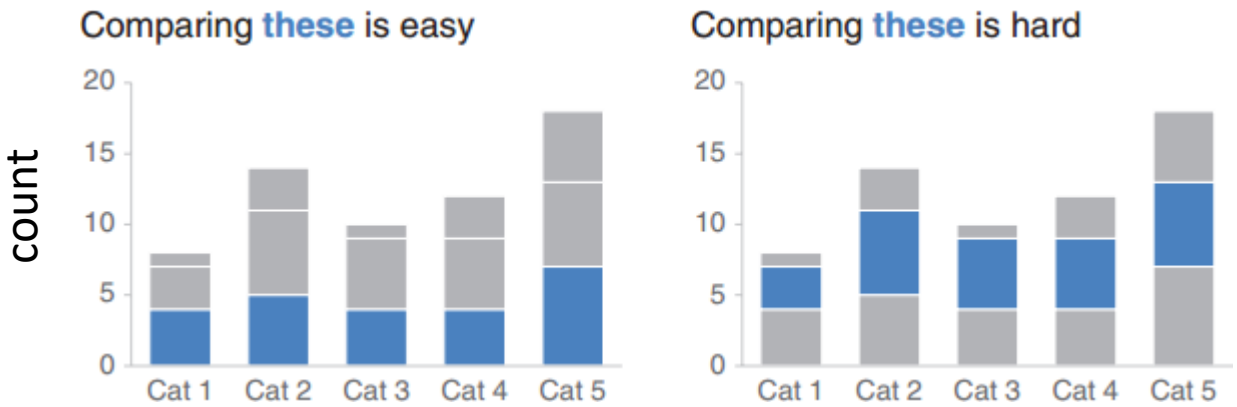
https://youtu.be/bzCupPVvn3Q

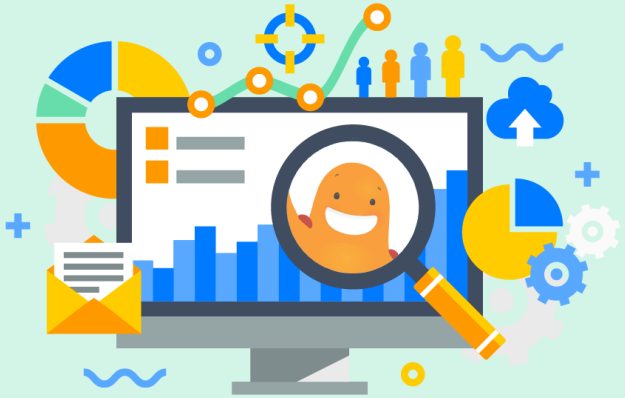

max
98
95
92
75th Percentile=91
90
85
median
81
50th Percentile=80
79
70
63
25th Percentile=59
55
47
min
42
third quartile
Quartiles divide your data into quarters.
first quartile

Step 2: Univariate Analysis

Categorical variables:

- Count / Count %





3 Bivariate Analysis
การวิเคราะห์ข้อมูล 2 ตัวแปร

Step 3: Bivariate Analysis

- Find out the relationship between **two variables** (x and y)
- If **highly** associated
  - both two variables are predictors => either variable x or y can be used in further analysis
  - One predictor and one target variable => the target variable is dominated by this predictor, no need to construct complex models.

Step 3: Bivariate Analysis

- Perform bi-variate analysis for any combination of categorical and continuous variables
  - Continuous & Continuous
  - Categorical & Categorical
  - Categorical & Continuous

credit_limit	total_spending
100000.00	3214.47
1000000.00	120754.11
200000.00	15000.00
200000.00	23149.50
500000.00	31145.03
150000.00	23012.75
50000.00	40012.83
...	...

Step 3: Bivariate Analysis

Numerical & Numerical

- **Correlation**: When two variables **move** together, we say they are **correlated**.

credit_limit	total_spending
100000.00	3214.47
1000000.00	120754.11
200000.00	15000.00
200000.00	23149.50
500000.00	31145.03
150000.00	23012.75
50000.00	40012.83
...	...

Step 3: Bivariate Analysis

Numerical & Numerical

- Pearson product moment Correlation coefficient (r)
- Measure of the linear correlation between 2 variables (The degree to which 2 variables move in relation to each other)

$$r = \frac{\text{Covariance (X,Y)}}{\text{SQRT(Variance(X)*Variance (Y))}}$$

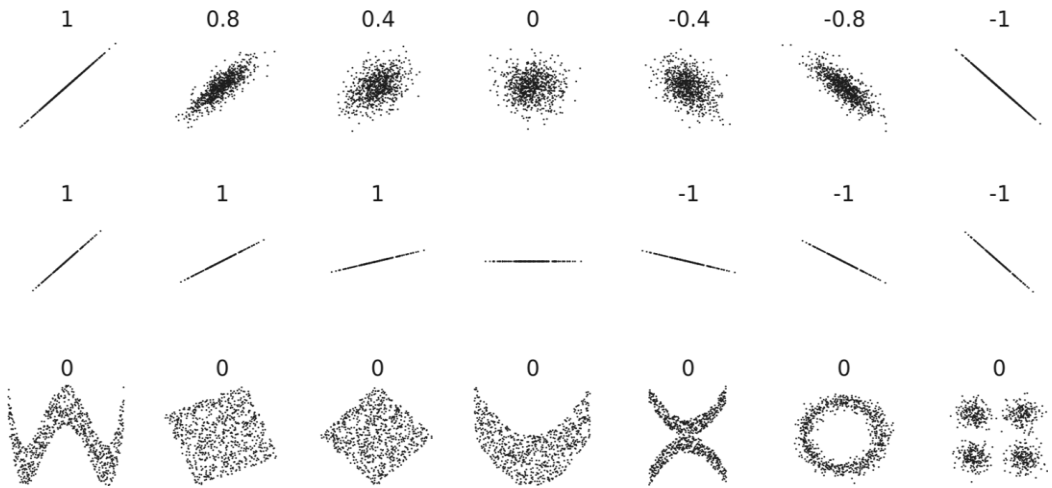
$$\text{Covariance (X,Y)} = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{n-1}$$

$$\text{Variance(X)} = \frac{\Sigma(X - \overline{X})^2}{n-1} \qquad \text{Variance(Y)} = \frac{\Sigma(Y - \overline{Y})^2}{n-1}$$

Step 3: Bivariate Analysis

Numerical & Numerical

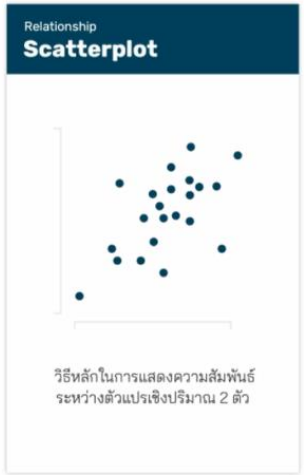
- Correlation values
  - 1 = Positive linear correlation
  - 0 = no linear correlation
  - -1 = Negative linear correlation



Step 3: Bivariate Analysis

Numerical & Numerical

- Visualizing Relationship: Scatter Plot



Step 3: Bivariate Analysis

Categorical & Categorical

- Two-way table/contingency table
  - Values => Count of observations in each combination of row and column categories

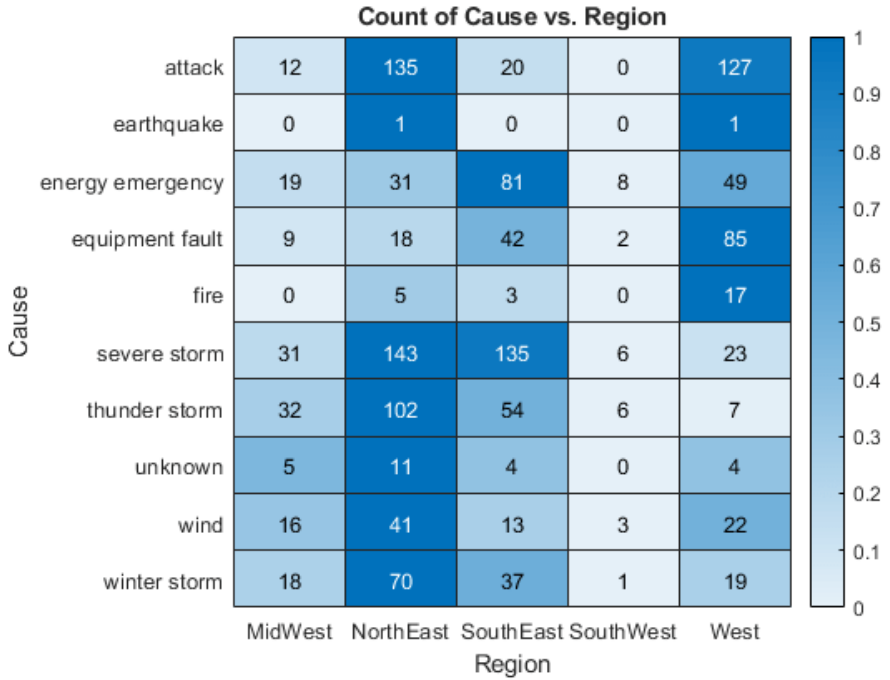
merchant_type	card_not_present
Retail	false
Transportation	true
Restaurants	true
Transportation	false
Retail	true
Utility	false
Retail	false
...	...

merchant_type	card_not_present	card_present
Retail	5102	938
Transportation	632	485
Restaurants	2100	3822
Utility	496	329

Step 3: Bivariate Analysis

Categorical & Categorical

- Visualizing Relationship: Heat Map



Step 3: Bivariate Analysis

Categorical & Numerical

Segment	Country	Gross Sales	Discounts	Sales	COGS	Profit
Government	United States of America	$ 8,001.00	$ -	$ 8,001.00	$ 5,715.00	$ 2,286.00
Midmarket	United States of America	$ 9,225.00	$ -	$ 9,225.00	$ 6,150.00	$ 3,075.00
Government	France	$ 27,615.00	$ 276.15	$ 27,338.85	$ 19,725.00	$ 7,613.85
Midmarket	France	$ 34,440.00	$ 344.40	$ 34,095.60	$ 22,960.00	$ 11,135.60
Government	France	$ 7,210.00	$ 72.10	$ 7,137.90	$ 5,150.00	$ 1,987.90
Government	France	$ 4,473.00	$ 44.73	$ 4,428.27	$ 3,195.00	$ 1,233.27
Government	Canada	$ 9,282.00	$ 92.82	$ 9,189.18	$ 6,630.00	$ 2,559.18
Channel Partners	United States of America	$ 22,296.00	$ 222.96	$ 22,073.04	$ 5,574.00	$ 16,499.04
Government	Mexico	$ 423,500.00	$ 4,235.00	$ 419,265.00	$ 314,600.00	$ 104,665.00
Government	United States of America	$ 17,703.00	$ 177.03	$ 17,525.97	$ 12,645.00	$ 4,880.97
Channel Partners	Canada	$ 17,340.00	$ 173.40	$ 17,166.60	$ 4,335.00	$ 12,831.60
Enterprise	United States of America	$ 41,250.00	$ 412.50	$ 40,837.50	$ 39,600.00	$ 1,237.50
Channel Partners	France	$ 32,052.00	$ 320.52	$ 31,731.48	$ 8,013.00	$ 23,718.48
Channel Partners	Germany	$ 9,192.00	$ 91.92	$ 9,100.08	$ 2,298.00	$ 6,802.08
Small Business	Mexico	$ 148,200.00	$ 1,482.00	$ 146,718.00	$ 123,500.00	$ 23,218.00
Government	Mexico	$ 488,950.00	$ 4,889.50	$ 484,060.50	$ 363,220.00	$ 120,840.50
Government	France	$ 754,250.00	$ 7,542.50	$ 746,707.50	$ 560,300.00	$ 186,407.50

Step 3: Bivariate Analysis

Categorical & Numerical

- Split data into smaller groups based on the categorical variable (dimension) and aggregate the numerical variable (values) for each group

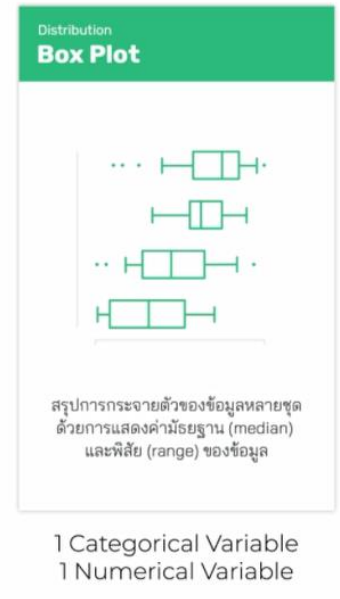
Row Labels	Sum of Gross Sales
Canada	26,932,163.50
France	26,081,674.50
Germany	24,921,467.50
Mexico	22,726,935.00
United States of America	27,269,358.00
Grand Total	**127,931,598.50**

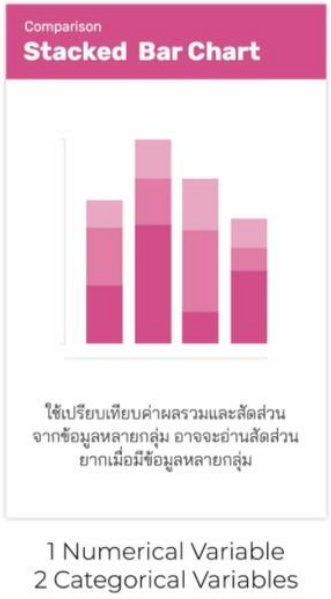
Row Labels	Sum of Gross Sales	Sum of Profit	Max of Discounts
Canada	26,932,163.50	3,529,228.89	119,756.00
France	26,081,674.50	3,781,020.78	111,375.00
Germany	24,921,467.50	3,680,388.82	106,512.00
Mexico	22,726,935.00	2,907,523.11	149,677.50
United States of America	27,269,358.00	2,995,540.67	125,820.00
Grand Total	**127,931,598.50**	**16,893,702.26**	**149,677.50**

Step 3: Bivariate Analysis

Categorical & Numerical

- Visualizing Relationship:





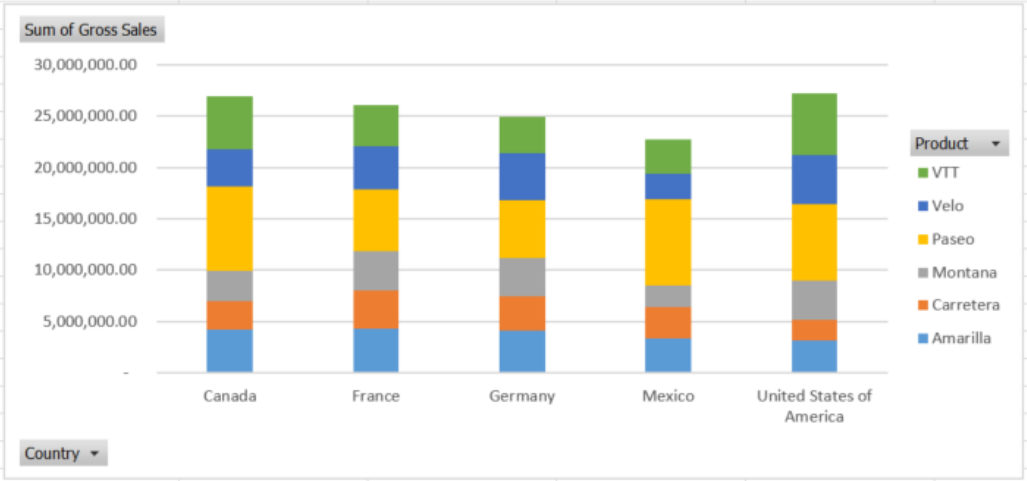
Step 3: Bivariate Analysis

Correlation VS Causation

- **Causation**: When changes in one variable (x) affect changes in another variable (y), we say that X causes Y.
    - Examples: Education -> Higher wages
- **Correlation**: When two variables move together, we say they are correlated.
- Correlation does not necessarily imply causality

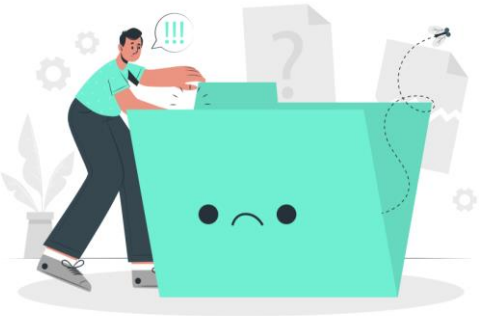
Step 3: Multivariable Analysis

Sum of Gross Sales	Products						
Country	Amarilla	Carretera	Montana	Paseo	Velo	VTT	Grand Total
Canada	4,164,683.50	2,825,853.50	2,982,114.50	8,172,612.00	3,660,387.00	5,126,513.00	26,932,163.50
France	4,318,664.00	3,687,829.50	3,843,216.00	5,984,767.00	4,244,434.50	4,002,763.50	26,081,674.50
Germany	4,123,204.50	3,306,376.00	3,798,355.00	5,555,838.00	4,637,903.00	3,499,791.00	24,921,467.50
Mexico	3,303,196.00	3,112,011.00	2,052,575.00	8,432,206.00	2,510,373.00	3,316,574.00	22,726,935.00
United States of America	3,127,531.50	2,005,450.50	3,873,574.00	7,466,239.00	4,773,671.00	6,022,892.00	27,269,358.00
Grand Total	**19,037,279.50**	**14,937,520.50**	**16,549,834.50**	**35,611,662.00**	**19,826,768.50**	**21,968,533.50**	**127,931,598.50**



4

Missing Values Treatment



Step 4: Missing Values Treatment

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55		Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57		Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	2	1	50%
M	4	2	50%
Missing	2	2	100%

playing cricket by males is higher than females

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55	F	Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57	F	Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	4	3	75%
M	4	2	50%

playing cricket by females is higher than males

Step 4: Missing Values Treatment

Incomplete (Missing) Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred

Step 4: Missing Values Treatment

Disguised Missing Data?

- Disguised missing data is the missing data entries that are not explicitly represented as such, but instead appear as potentially valid data values
  - Information about "State" is missing >> "Alabama" is used as disguise

Step 4: Missing Values Treatment

Disguised Missing Data?

Wrong conclusion

Unreasonable results

Step 4: Missing Values Treatment

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification) - not effective when the % of missing values per attribute varies considerably
- Fill in missing values
  - Manually
  - Using a global constant
  - Using a measure of central tendency for the attribute, such as mean, median, or mode
  - Using the central tendency of the class
  - Using the most probable value

Step 4: Missing Values Treatment

Which are the methods to treat missing values ?

- Mean/Mode/Median Imputation:
  - **Generalized Imputation**: In this case, we calculate the mean or median for all non missing values of that variable then replace missing value with mean or median. Like in above table, variable "Manpower" is missing so we take average of all non missing values of "Manpower" (28.33) and then replace missing value with it.

Gender	Manpower	Sales
M	25	343
F	.	280
M	33	332
M	.	272
F	25	.
M	29	326
	26	259
M	32	297

Step 4: Missing Values Treatment

Which are the methods to treat missing values ?

- Mean/Mode/Median Imputation:
  - **Similar case Imputation**: In this case, we calculate average for gender "Male" (29.75) and "Female" (25) individually of non missing values then replace the missing value based on gender. For "Male", we will replace missing values of manpower with 29.75 and for "Female" with 25.

Gender	Manpower	Sales
M	25	343
F	.	280
M	33	332
M	.	272
F	25	.
M	29	326
	26	259
M	32	297

Step 4: Missing Values Treatment

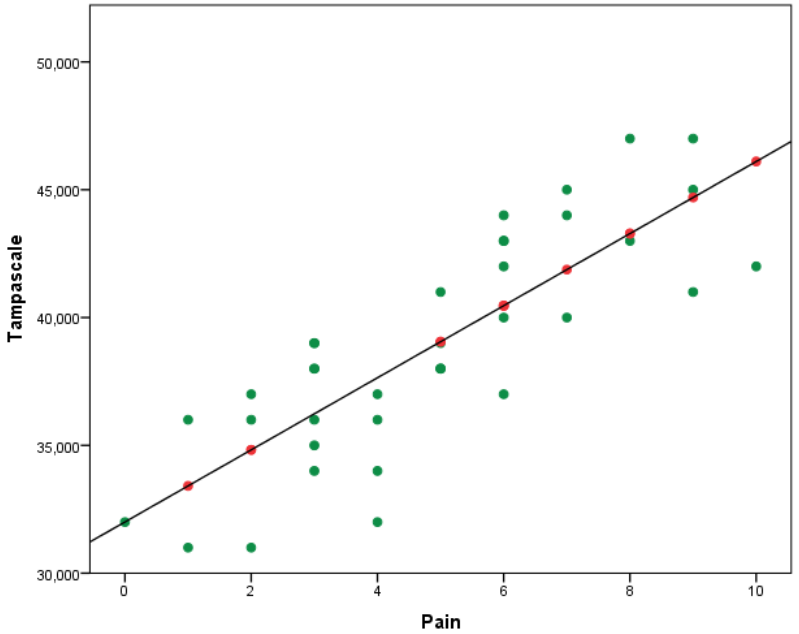
Which are the methods to treat missing values ?

- Regression Imputation:

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Extensions Window Help

20 : Tampa_ImpRegr...

	ID	Pain	Tampascale	Disability	Radiation	Tampa_ImpRegress
1	1	9	45	20	1	.
2	2	6	.	10	0	40,465
3	3	1	36	1	0	.
4	4	5	38	14	0	.
5	5	6	44	14	1	.
6	6	7	.	11	1	41,875
7	7	8	43	18	0	.
8	8	6	43	11	1	.
9	9	2	.	11	1	34,825
10	10	4	36	3	0	.
11	11	5	38	16	1	.
12	12	9	47	14	0	.
13	13	0	32	3	1	.
14	14	6	.	12	0	40,465
15	15	3	34	13	0	.



https://bookdown.org/mwheymans/bookmi/single-missing-data-imputation.html

Step 4: Missing Values Treatment

Which are the methods to treat missing values ?

- **List wise deletion**, we delete observations where any of the variable is missing. Simplicity is one of the major advantage of this method, but this method reduces the power of model because it reduces the sample size.

List wise deletion

Gender	Manpower	Sales
M	25	343
F	.	280
M	33	332
M	.	272
F	25	.
M	29	326
	26	259
M	32	297

Step 4: Missing Values Treatment

Which are the methods to treat missing values ?

- **Pair wise deletion**, we perform analysis with all cases in which the variables of interest are present. Advantage of this method is, it keeps as many cases available for analysis. One of the disadvantage of this method, it uses different sample size for different variables.

Pair wise deletion

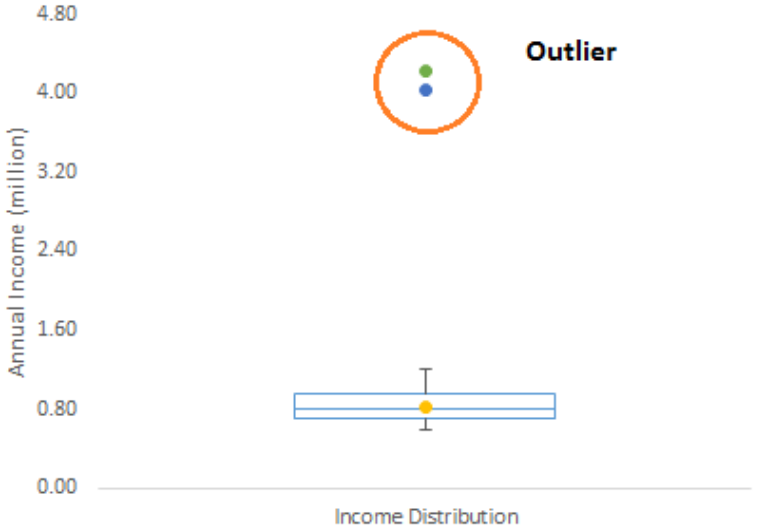
Gender	Manpower	Sales
M	25	343
F	.	280
M	33	332
M	.	272
F	25	.
M	29	326
	26	259
M	32	297

5

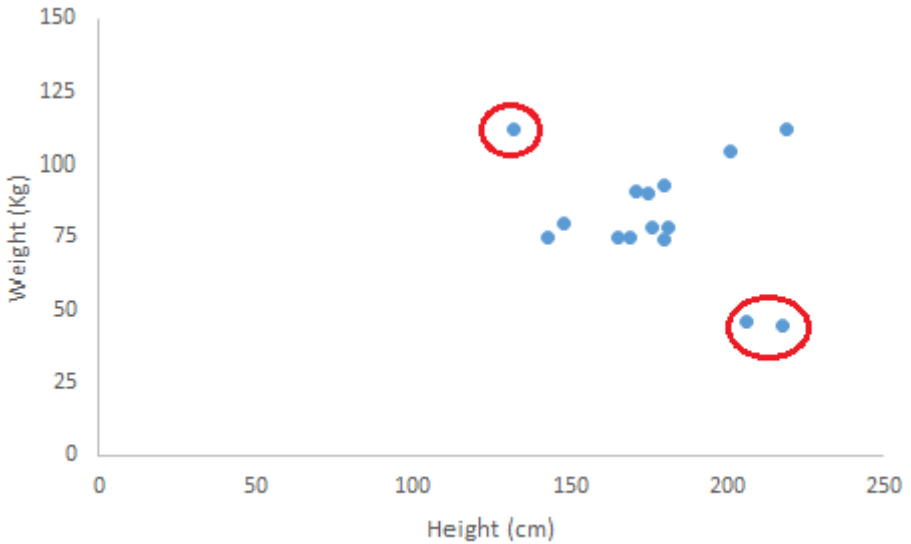
Outlier Treatment

Step 5: Outlier Treatment

Univariate



Multivariate



Step 5: Outlier Treatment

What is the impact of Outliers on a dataset?

Without Outlier	With Outlier
4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7	4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7,300
Mean = 5.45	Mean = 30.00
Median = 5.00	Median = 5.50
Mode = 5.00	Mode = 5.00
Standard Deviation = 1.04	Standard Deviation = 85.03

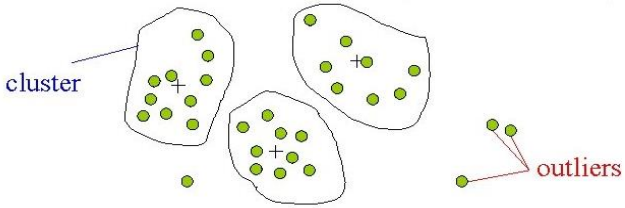
Step 5: Outlier Treatment

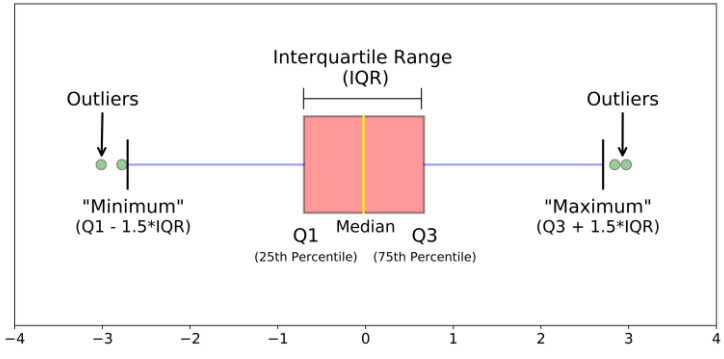
- What causes Outliers?

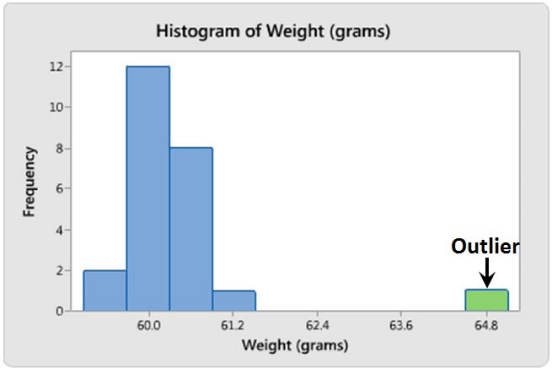
Step 5: Outlier Treatment

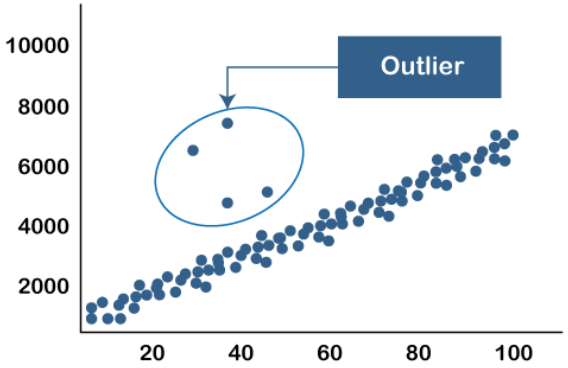
How to detect Outliers?

- Most commonly used method to detect outliers is visualization, like Box-plot, Histogram, Scatter Plot









Step 5: Outlier Treatment

Step 5: Outlier Treatment

- Binning
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.

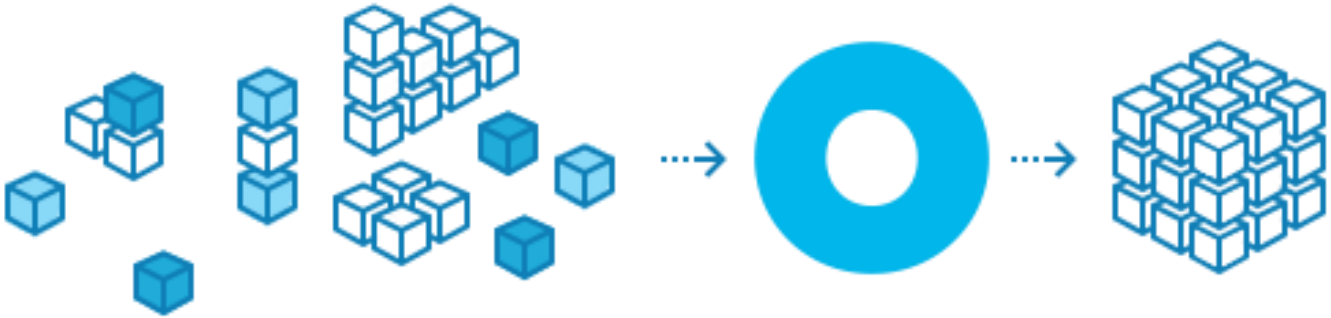
Sorted data for price (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:
Bin1: 4, 8, 15
Bin2: 21, 21, 24
Bin3: 25, 28, 34

Smoothing by bin means:
Bin1: 9, 9, 9
Bin2: 22, 22, 22
Bin3: 29, 29, 29

Smoothing by bin Boundaries:
Bin1: 4, 4, 15
Bin2: 21, 21, 24
Bin3: 25, 25, 34

6 Data Transform and Feature Engineering

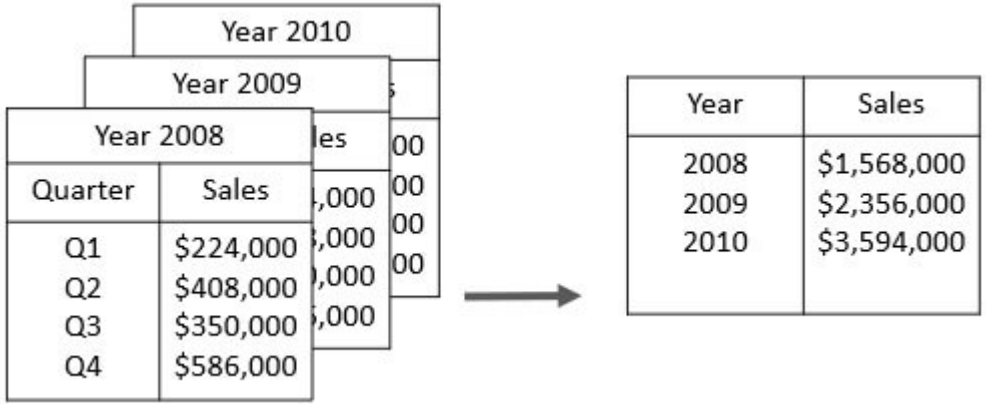


Step 6: Data Transformation and Feature Engineering

- A function that maps the entire set of values of a given attribute to a new set of replacement values
- Methods
  - Smoothing: Remove noise from data
  - Attribute/feature construction:
    - New attributes constructed from the given ones
  - Aggregation: Summarization, data cube construction
  - Normalization: Scaled to fall within a smaller, specified range such as -1.0 to 1.0, or 0.0 to 1.0
    - min-max normalization, z-score normalization, normalization by decimal scaling
  - Discretization:
    - Divide the range of a continuous attribute into intervals
    - Concept hierarchy climbing

Step 6: Data Transformation and Feature Engineering

- Aggregation



Step 6: Data Transformation and Feature Engineering

- Min-max normalization: performs a linear transformation on the original data

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- Example
  - Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0].
  - Then \$73,000 is mapped to

$$\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$$

Step 6: Data Transformation and Feature Engineering

- **Creating derived variables:** This refers to creating new variables from existing variable(s) using set of functions or different methods.

Emp_Code	Gender	Date	New_Day	New_Month	New_Year
A001	Male	21-Sep-11	21	9	2011
A002	Female	27-Feb-13	27	2	2013
A003	Female	14-Nov-12	14	11	2012
A004	Male	07-Apr-13	7	4	2013
A005	Female	21-Jan-11	21	1	2011
A006	Male	26-Apr-13	26	4	2013
A007	Male	15-Mar-12	15	3	2012

Step 6: Data Transformation and Feature Engineering

- **Creating dummy variables / Categorical encoding** : Convert categorical variables into numerical variables.
  - Dummy variables are also called Indicator Variables. It is useful to take categorical variable as a predictor in statistical models. Categorical variables can take values 0 and 1.
  - We can also create dummy variables for more than two classes of categorical variables with n or n-1 dummy variables.

	Candy Variety	Day	Type of Day	Weekend	Weekday
0	Chocolate Hearts	Sunday	Weekend	1	0
1	Sour Jelly	Saturday	Weekend	1	0
2	Candy Canes	Friday	Weekday	0	1
3	Sour Jelly	Sunday	Weekend	1	0
4	Fruit Drops	Sunday	Weekend	1	0
5	Sour Jelly	Thursday	Weekday	0	1

Emp_Code	Gender	Var_Male	Var_Female
A001	Male	1	0
A002	Female	0	1
A003	Female	0	1
A004	Male	1	0
A005	Female	0	1
A006	Male	1	0
A007	Male	1	0

Step 6: Data Transformation and Feature Engineering

- **Discretization:** Taking a set of values of data and grouping sets of them together in some logical fashion into bins (or buckets). Binning can apply to numerical values as well as to categorical values..

Name	Birthday		Birthday
John	31/12/1990	------>	90's
Mery	15/10/1978	------>	70's
Alice	19/04/2000	------>	00's
Mark	01/11/1997	------>	90's
Alex	15/03/2000	------>	00's
Peter	01/12/1983	------>	80's
Calvin	05/05/1995	------>	90's
Roxane	03/08/1948	------>	40's
Anne	05/09/1992	------>	90's
Paul	14/11/1992	------>	90's

Student_id	Age	Grade	Employed	marks	bucket
1	19	1st Class	yes	29	Poor
2	20	2nd Class	no	41	Below_average
3	18	1st Class	no	57	Average
4	21	2nd Class	no	29	Poor
5	19	1st Class	no	57	Average
6	20	2nd Class	yes	53	Average
7	19	3rd Class	yes	78	Above_Average
8	21	3rd Class	yes	70	Above_Average
9	22	3rd Class	yes	97	Excellent
10	21	1st Class	no	58	Average

Step 6: Data Transformation and Feature Engineering

Fraud detection Example

Features associated with the **transaction**:

- Date and time
- Transaction amount
- Merchant
- Product

Features associated with the **credit card holder**:

- Card type
- Credit card number
- Expiration date
- Billing address (street address, state, zip code),
- Phone number
- Email address

https://developer.nvidia.com/blog/leveraging-machine-learning-to-detect-fraud-tips-to-developing-a-winning-kaggle-solution/

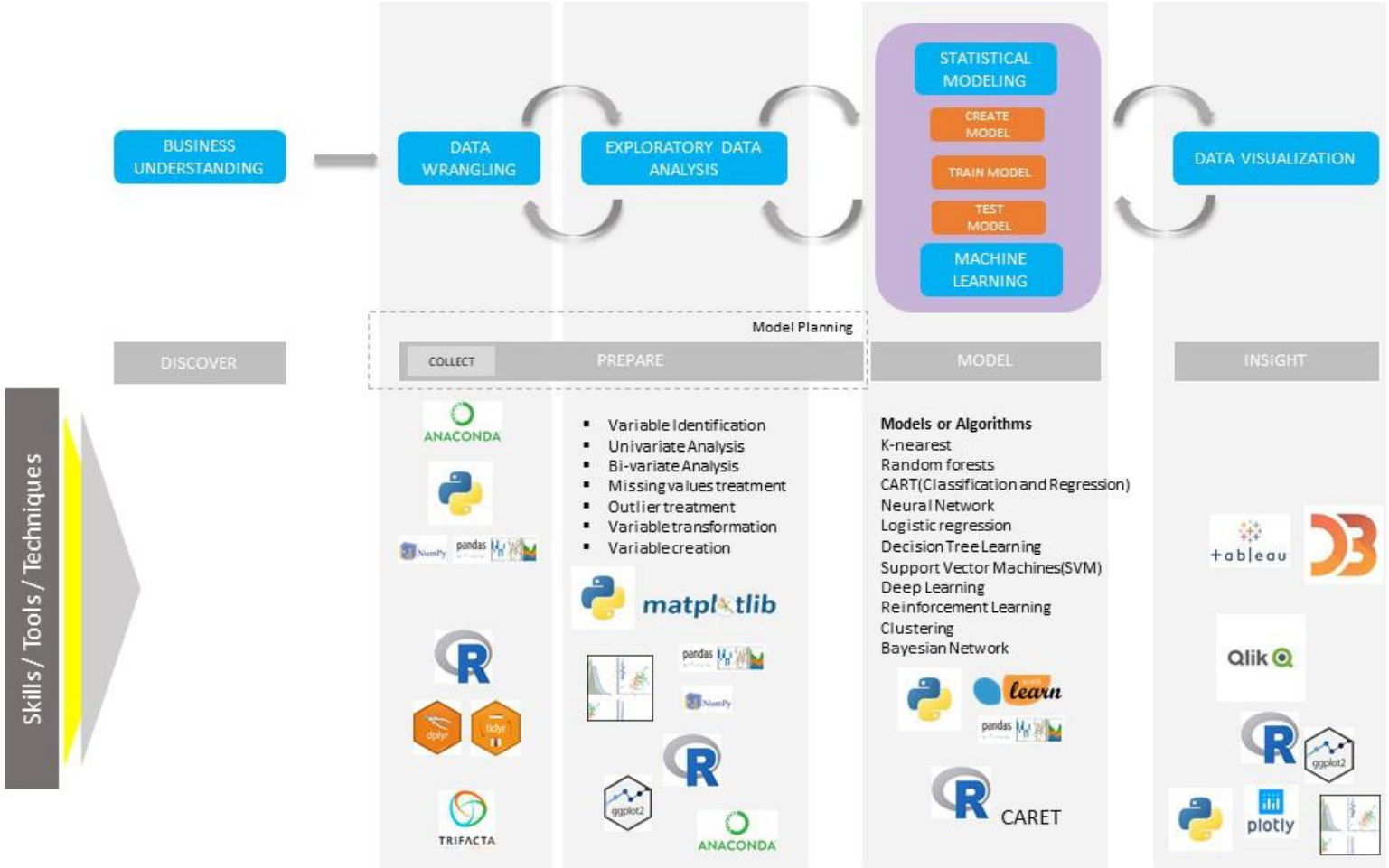
Step 6: Data Transformation and Feature Engineering

Fraud detection Example

Possible features generated from **transaction history**:

- Number of transactions a credit card holder has made in the last 24 hours (holder features plus device ID and IP address)
- Same or different credit card numbers?
- Same or different shipping addresses?
- The total amount a credit cardholder has spent in the last 24 hours
- Average amount last 24 hours compared to the historical daily average amount
- Number of transactions made from the same device or IP address in the last 24 hours
- Multiple online charges in close temporal proximity?
- Frequent fraud at a specific merchant?
- Group of fraud charges in certain areas?
- Multiple charges from a merchant within a short time span?

https://developer.nvidia.com/blog/leveraging-machine-learning-to-detect-fraud-tips-to-developing-a-winning-kaggle-solution/



https://blog.camelot-group.com/2019/03/exploratory-data-analysis-an-important-step-in-data-science/