Data Exploration
- EDA(Exploratory Data Analysis)

Dept. of Mechanical System Design Engineering,
Seoul National University of Science and Technology

Prof. Ju Yeon Lee
(jylee@seoultech.ac.kr)

Review

File Handling (Pandas)

Format Type	Data Description	Reader	Writer
text	CSV	read_csv	to_csv
text	Fixed-Width Text File	read_fwf	
text	JSON	read_json	to_json
text	HTML	read_html	to_html
text	LaTeX		Styler.to_latex
text	XML	read_xml	to_xml
text	Local clipboard	read_clipboard	to_clipboard
binary	MS Excel	read_excel	to_excel
binary	OpenDocument	read_excel	

File Handling (Pandas)

binary	HDF5 Format	read_hdf	to_hdf
binary	Feather Format	read_feather	to_feather
binary	Parquet Format	read_parquet	to_parquet
binary	ORC Format	read_orc	to_orc
binary	Stata	read_stata	to_stata
binary	SAS	read_sas	
binary	SPSS	read_spss	
binary	Python Pickle Format	read_pickle	to_pickle
SQL	SQL	read_sql	to_sql
SQL	Google BigQuery	read_gbq	to_gbq

EDA(Exploratory Data Analysis)

Descriptive Statistics

Index	mpg	cylinders	displacement	weight	acceleration	model_year	origin
count	398	398	398	398	398	398	398
mean	23.5146	5.45477	193.426	2970.42	15.5681	76.0101	1.57286
std	7.81598	1.701	104.27	846.842	2.75769	3.69763	0.802055
min	9	3	68	1613	8	70	1
25%	17.5	4	104.25	2223.75	13.825	73	1
50%	23	4	148.5	2803.5	15.5	76	1
75%	29	8	262	3608	17.175	79	2
max	46.6	8	455	5140	24.8	82	3

- **Count : the number of available data**
- **Mean : arithmetic mean value**
- **Min : minimum value**
- **Max : maximum value**
- **Q1 : ~25%**
- **Q2 : ~50% (median)**
- **Q3 : ~75%**
- **Q4 : ~max**
- **Mode: most frequent value**
- **Std : standard deviation**
- **Min – Max : a range of values**

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

σ = population standard deviation

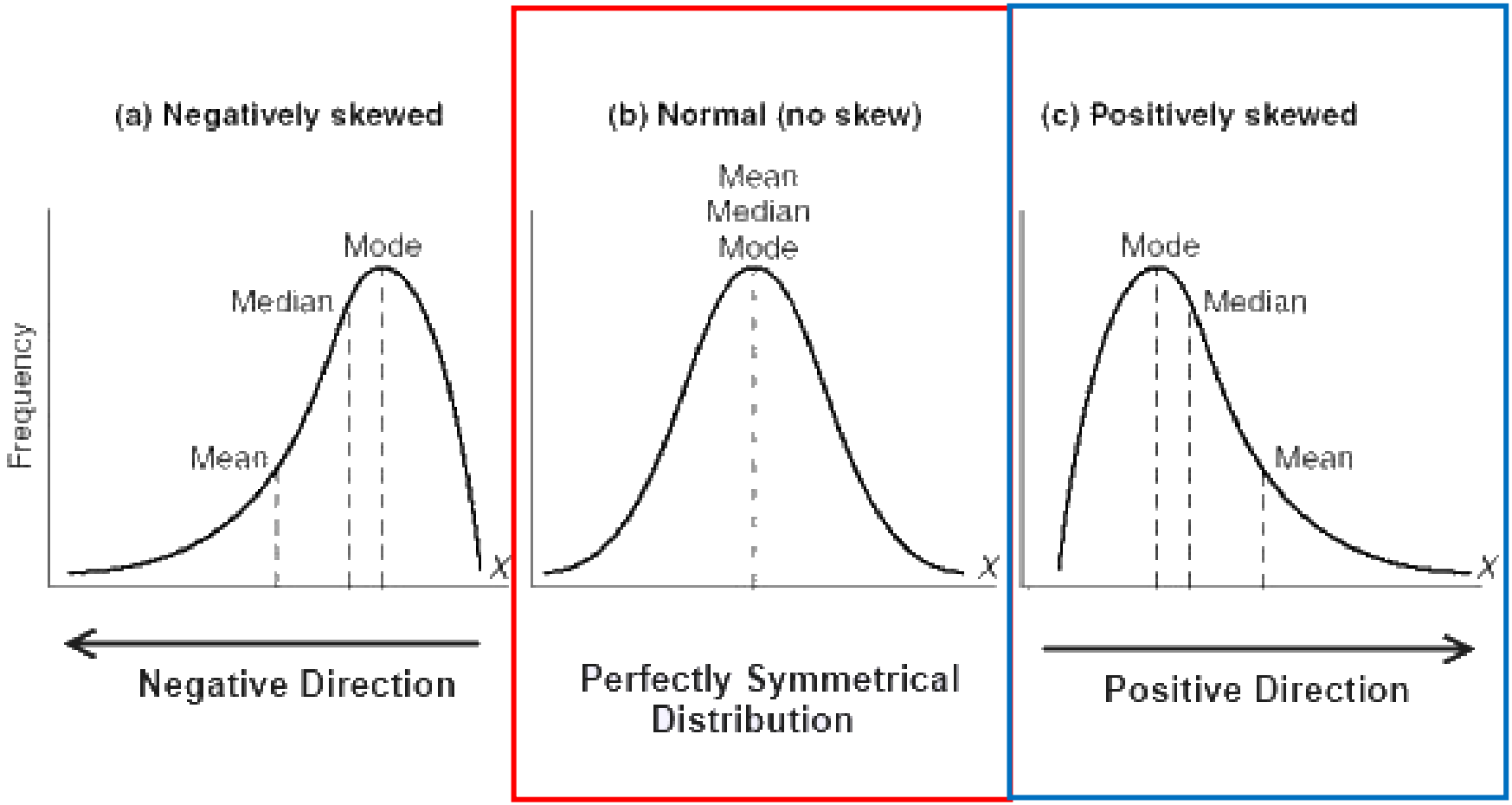
N = the size of the population

x_i = each value from the population

μ = the population mean

Skewness

Mean = Median = Mode



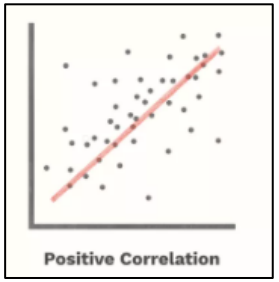
Auto MPG Dataset

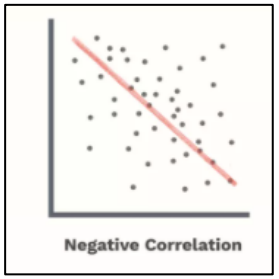
Correlation Analysis

Index	mpg	cylinders	displacement	weight	acceleration	model_year	origin
mpg	1	-0.775396	-0.804203	-0.831741	0.420289	0.579267	0.56345
cylinders	-0.775396	1	0.950721	0.896017	-0.505419	-0.348746	-0.562543
displacement	-0.804203	0.950721	1	0.932824	-0.543684	-0.370164	-0.609409
weight	-0.831741	0.896017	0.932824	1	-0.417457	-0.306564	-0.581024
acceleration	0.420289	-0.505419	-0.543684	-0.417457	1	0.288137	0.205873
model_year	0.579267	-0.348746	-0.370164	-0.306564	0.288137	1	0.180662
origin	0.56345	-0.562543	-0.609409	-0.581024	0.205873	0.180662	1

- **Positive correlation :**
  the variables move in the same direction

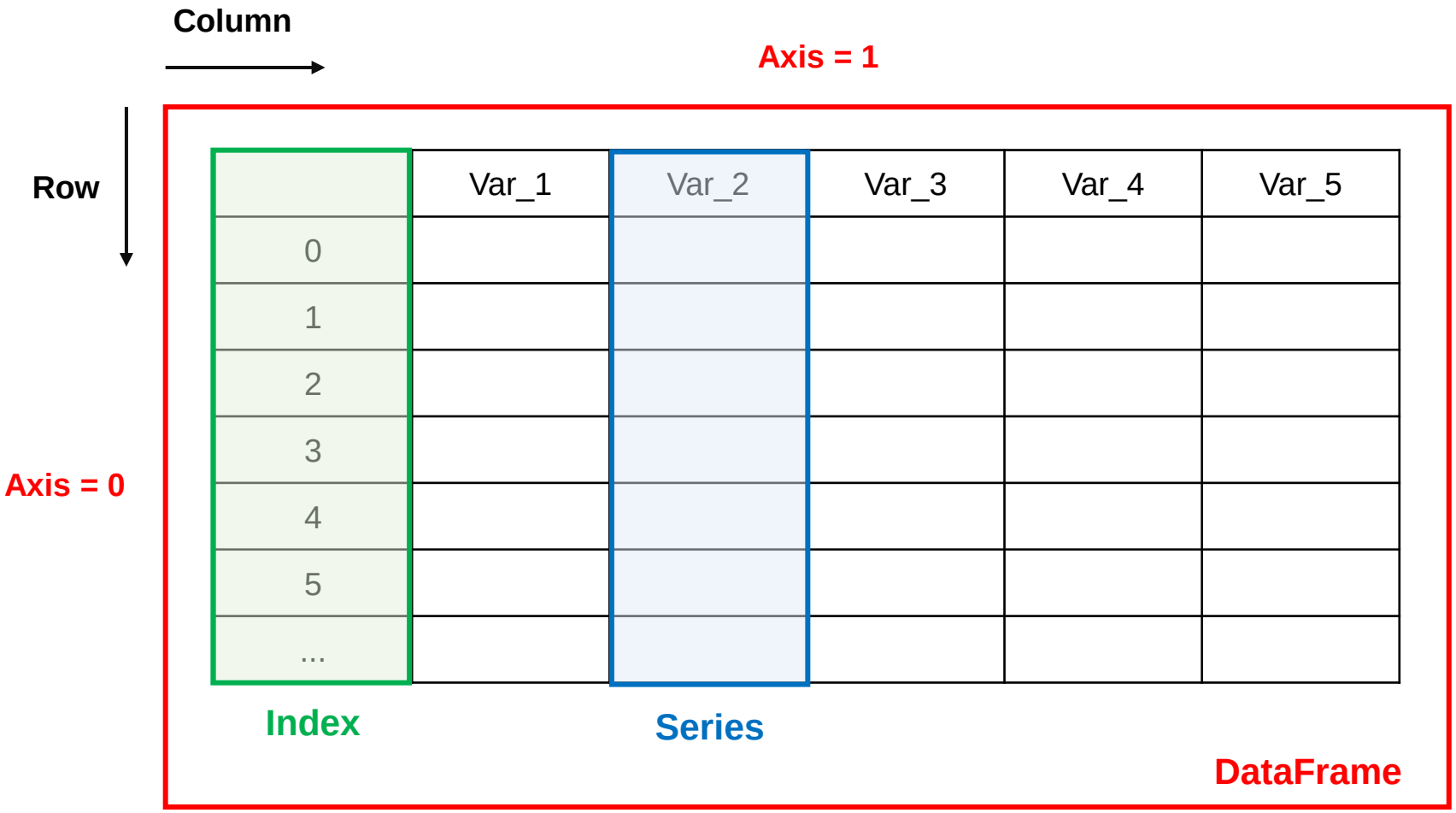
- **Negative correlation :**
  the variables move in opposite directions





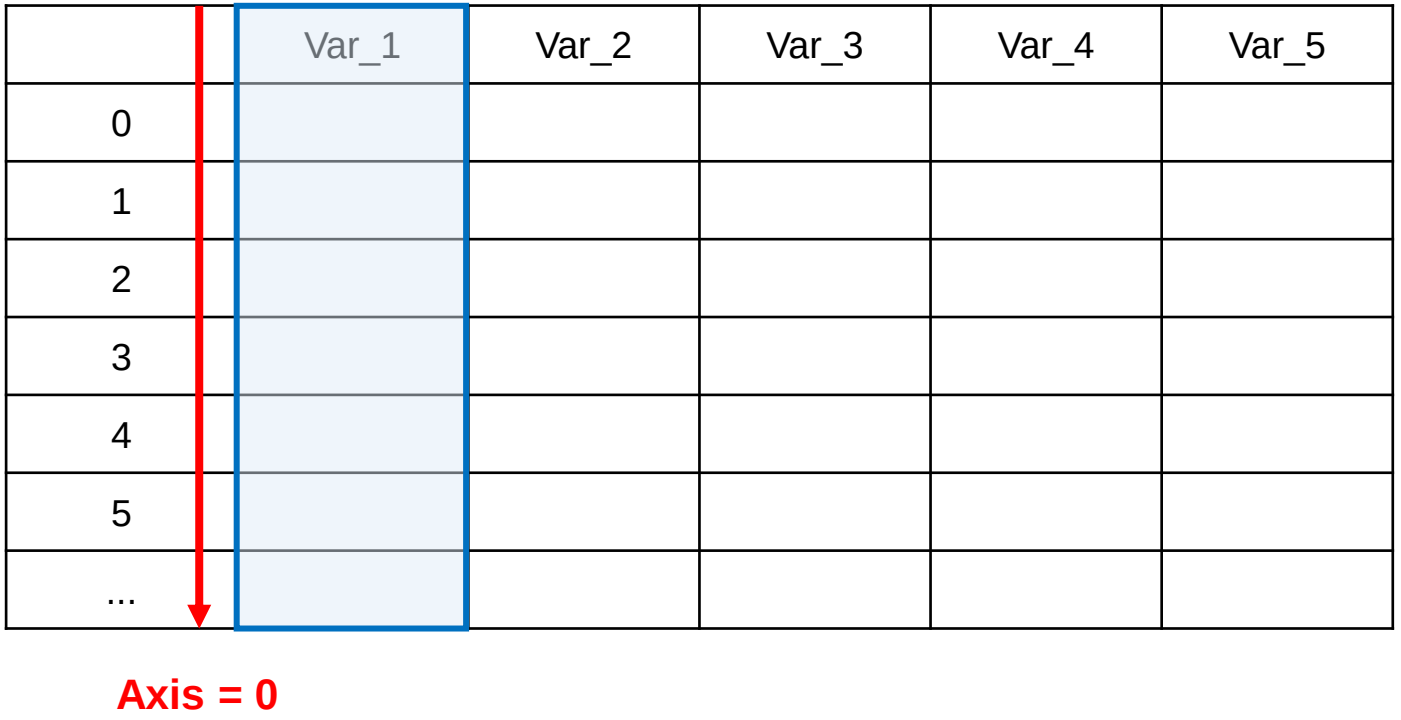
- **No correlation**

DataFrame Axis



DataFrame Axis

(Code) Function Parameter is Axis = 0,
(Mean) Direction is Axis = 0



	Var_1	Var_2	Var_3	Var_4	Var_5
0					
1					
2					
3					
4					
5					
...					

Axis = 0

DataFrame Axis

(Code) Function Parameter is Axis = 1,
(Mean) Direction is Axis = 1

	Var_1	Var_2	Var_3	Var_4	Var_5
0					
1					
2					
3					
4					
5					
...					


Thank you
Q & A