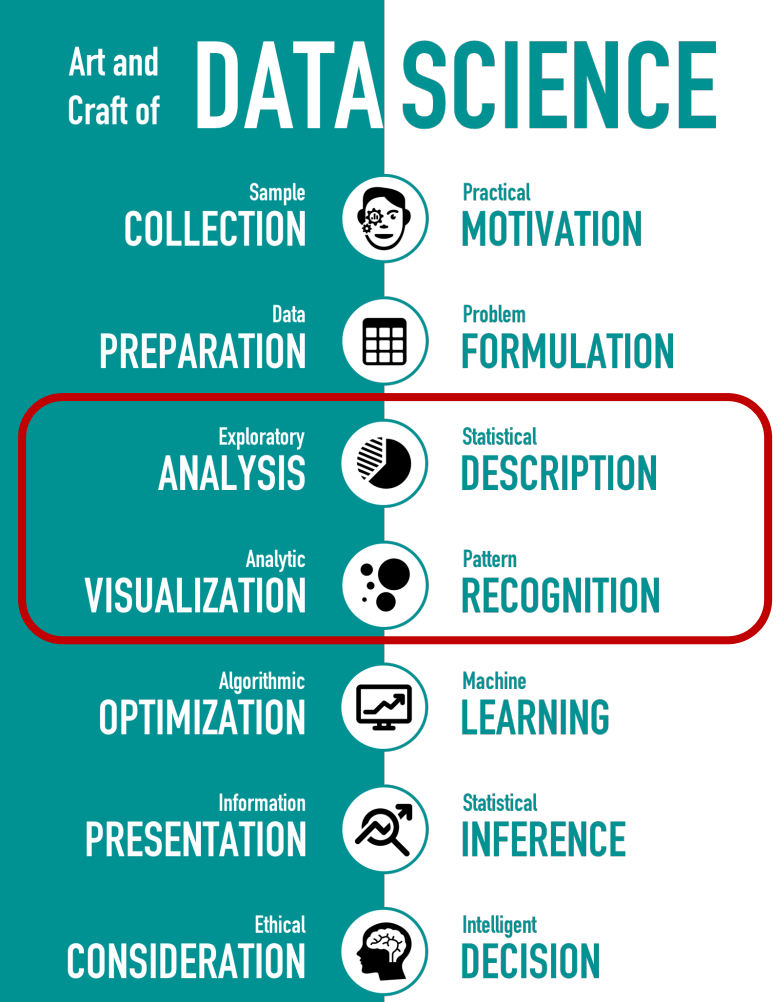
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Uni-Variate Statistics

Sourav SEN GUPTA
Lecturer, SCSE, NTU



Data Science

Uni-Variate Statistics

Exploratory Analysis

What are the Variables in the Data?

How to characterize the Variables?

How to find relation between them?

How to intelligently explore acquired Data?

Data Science

The Pokemon Dataset

#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
430	Honchkrow	Dark	Flying	505	100	125	52	105	52	71	4	False
338	Solrock	Rock	Psychic	440	70	95	85	55	65	70	3	False
32	Nidoran♂	Poison	NaN	273	46	57	40	40	40	50	1	False
442	Spiritomb	Ghost	Dark	485	50	92	108	92	108	35	4	False
480	Uxie	Psychic	NaN	580	75	75	130	75	130	95	4	True
536	Palpitoad	Water	Ground	384	75	65	55	65	55	69	5	False
360	Wynaut	Psychic	NaN	260	95	23	48	23	48	23	3	False
478	Froslass	Ice	Ghost	480	70	80	70	80	70	110	4	False
76	Golem	Rock	Ground	495	80	120	130	55	65	45	1	False
177	Natu	Psychic	Flying	320	40	50	45	70	45	70	2	False

Source : Kaggle Datasets | **Pokemon with stats** by Alberto Barradas | https://www.kaggle.com/abcsds/pokemon

45,	60,	80,	80,	39,	58,	78,	78,	78,	44,	59,	79,	79,
45,	50,	60,	40,	45,	65,	65,	40,	63,	83,	83,	30,	55,
40,	65,	35,	60,	35,	60,	50,	75,	55,	70,	90,	46,	61,
81,	70,	95,	38,	73,	115,	140,	40,	75,	45,	60,	75,	35,
60,	60,	70,	10,	35,	40,	65,	50,	80,	40,	65,	55,	90,
40,	65,	90,	25,	40,	55,	55,	70,	80,	90,	50,	65,	80,
40,	80,	40,	55,	80,	50,	65,	90,	95,	95,	25,	50,	52,
35,	60,	65,	90,	80,	105,	30,	50,	30,	45,	60,	60,	35,
60,	85,	30,	55,	40,	60,	60,	95,	50,	60,	50,	50,	90,
40,	65,	80,	105,	250,	65,	105,	105,	30,	55,	45,	80,	30,
60,	40,	70,	65,	65,	65,	65,	65,	75,	20,	95,	95,	130,
48,	55,	130,	65,	65,	65,	35,	70,	30,	60,	80,	80,	160,
90,	90,	90,	41,	61,	91,	106,	106,	106,	100,	45,	60,	80,
39,	58,	78,	50,	65,	85,	35,	85,	60,	100,	40,	55,	40,
70,	85,	75,	125,	20,	50,	90,	35,	55,	40,	65,	55,	70,
90,	90,	75,	70,	100,	70,	90,	35,	55,	75,	55,	30,	75,
65,	55,	95,	65,	95,	60,	95,	60,	48,	190,	70,	50,	75,
100,	65,	75,	75,	60,	90,	65,	70,	70,	20,	80,	80,	55,
60,	90,	40,	50,	50,	100,	55,	35,	75,	45,	65,	65,	45,
75,	75,	75,	90,	90,	85,	73,	55,	35,	50,	45,	45,	45,
95,	255,	90,	115,	100,	50,	70,	100,	100,	106,	106,	100,	40,
50,	70,	70,	45,	60,	80,	80,	50,	70,	100,	100,	35,	70,
38,	78,	45,	50,	60,	50,	60,	40,	60,	80,	40,	70,	90,
40,	60,	40,	60,	28,	38,	68,	68,	40,	70,	60,	60,	60,
80,	150,	31,	61,	1,	64,	84,	104,	72,	144,	50,	30,	50,
70,	50,	50,	50,	50,	50,	60,	70,	70,	30,	60,	60,	40,
70,	70,	60,	60,	65,	65,	50,	70,	100,	45,	70,	70,	130,
170,	60,	70,	70,	70,	60,	80,	60,	45,	50,	80,	50,	70,
45,	75,	75,	73,	73,	70,	70,	50,	110,	43,	63,	40,	60,
66,	86,	45,	75,	20,	95,	70,	60,	44,	64,	64,	20,	40,
99,	65,	65,	65,	95,	50,	80,	80,	70,	90,	110,	35,	55,
55,	100,	43,	45,	65,	95,	95,	40,	60,	80,	80,	80,	80,
80,	80,	80,	80,	80,	100,	100,	100,	100,	105,	105,	100,	50,

Data Science

Uni-Variate Statistics

Numeric Uni-Variate Data

HP Hit Point of Pokemon

Numeric Variable

800 Values in Total

Pertinent Questions

- How to describe the Data?
- How to analyze the Data?

Source : Kaggle Datasets | **Pokemon with stats** by Alberto Barradas | https://www.kaggle.com/abcsds/pokemon

45,	60,	80,	80,	39,	58,	78,	78,	78,	44,	59,	79,	79,
45,	50,	60,	40,	45,	65,	65,	40,	63,	83,	83,	30,	55,
40,	65,	35,	60,	35,	60,	50,	75,	55,	70,	90,	46,	61,
81,	70,	95,	38,	73,	115,	140,	40,	75,	45,	60,	75,	35,
60,	60,	70,	10,	35,	40,	65,	50,	80,	40,	65,	55,	90,
40,	65,	90,	25,	40,	55,	55,	70,	80,	90,	50,	65,	80,
40,	80,	40,	55,	80,	50,	65,	90,	95,	95,	25,	50,	52,
35,	60,	65,	90,	80,	105,	30,	50,	30,	45,	60,	60,	35,
60,	85,	30,	55,	40,	60,	60,	95,	50,	60,	50,	50,	90,
40,	65,	80,	105,	250,	65,	105,	105,	30,	55,	45,	80,	30,
60,	40,	70,	65,	65,	65,	65,	65,	75,	20,	95,	95,	130,
48,	55,	130,	65,	65,	65,	35,	70,	30,	60,	80,	80,	160,
90,	90,	90,	41,	61,	91,	106,	106,	106,	100,	45,	60,	80,
39,	58,	78,	50,	65,	85,	35,	85,	60,	100,	40,	55,	40,
70,	85,	75,	125,	20,	50,	90,	35,	55,	40,	65,	55,	70,
90,	90,	75,	70,	100,	70,	90,	35,	55,	75,	55,	30,	75,
65,	55,	95,	65,	95,	60,	95,	60,	48,	190,	70,	50,	75,
100,	65,	75,	75,	60,	90,	65,	70,	70,	20,	80,	80,	55,
60,	90,	40,	50,	50,	100,	55,	35,	75,	45,	65,	65,	45,
75,	75,	75,	90,	90,	85,	73,	55,	35,	50,	45,	45,	45,
95,	255,	90,	115,	100,	50,	70,	100,	100,	106,	106,	100,	40,
50,	70,	70,	45,	60,	80,	80,	50,	70,	100,	100,	35,	70,
38,	78,	45,	50,	60,	50,	60,	40,	60,	80,	40,	70,	90,
40,	60,	40,	60,	28,	38,	68,	68,	40,	70,	60,	60,	60,
80,	150,	31,	61,	1,	64,	84,	104,	72,	144,	50,	30,	50,
70,	50,	50,	50,	50,	50,	60,	70,	70,	30,	60,	60,	40,
70,	70,	60,	60,	65,	65,	50,	70,	100,	45,	70,	70,	130,
170,	60,	70,	70,	70,	60,	80,	60,	45,	50,	80,	50,	70,
45,	75,	75,	73,	73,	70,	70,	50,	110,	43,	63,	40,	60,
66,	86,	45,	75,	20,	95,	70,	60,	44,	64,	64,	20,	40,
99,	65,	65,	65,	95,	50,	80,	80,	70,	90,	110,	35,	55,
55,	100,	43,	45,	65,	95,	95,	40,	60,	80,	80,	80,	80,
80,	80,	80,	80,	80,	100,	100,	100,	100,	105,	105,	100,	50,

Data Science

Uni-Variate Statistics

Basic Summary of the Data

HP The Average Hit Point
Deviation from Average
Maximum and Minimum

Statistical Questions

- What is the Central Tendency?
- What is the Spread of the Data?

Source : Kaggle Datasets | **Pokemon with stats** by Alberto Barradas | https://www.kaggle.com/abcsds/pokemon

Data Science

Uni-Variate Statistics

Central Tendency : Mean

Natural Intuition

Average Hit Point of a Pokemon

Statistical Definition

Sum of Data / Count of Data

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Source : Kaggle Datasets | **Pokemon with stats** by Alberto Barradas | https://www.kaggle.com/abcsds/pokemon

Data Science

Uni-Variate Statistics

Dispersion : Standard Deviation

Natural Intuition

Average Deviation from the Mean

Statistical Definition

Sum of Deviation / Count of Data

$$\sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n}}$$

Source : Kaggle Datasets | **Pokemon with stats** by Alberto Barradas | https://www.kaggle.com/abcsds/pokemon

Data Science

Uni-Variate Statistics

Central Tendency : Median

Natural Intuition

Mid-Value of Pokemon Hit Points

Statistical Definition

Marker to Divide the Data 50:50

$$P(x \leq x_M) = P(x \geq x_M) = 0.5$$

Source : Kaggle Datasets | **Pokemon with stats** by Alberto Barradas | https://www.kaggle.com/abcsds/pokemon

Data Science

Uni-Variate Statistics

Dispersion : Quantiles

Natural Intuition

Distribution of Pokemon Hit Points

Statistical Definition

Markers to Divide the Data 25:50:25

$$P(x \leq x_{Q1}) = 0.25, P(x \geq x_{Q3}) = 0.25$$
$$P(x_{Q1} \leq x \leq x_{Q3}) = 0.5$$

Source : Kaggle Datasets | **Pokemon with stats** by Alberto Barradas | https://www.kaggle.com/abcsds/pokemon

count	800.000000
mean	69.258750
std	25.534669
min	1.000000
25%	50.000000
50%	65.000000
75%	80.000000
max	255.000000

Data Science

Uni-Variate Statistics

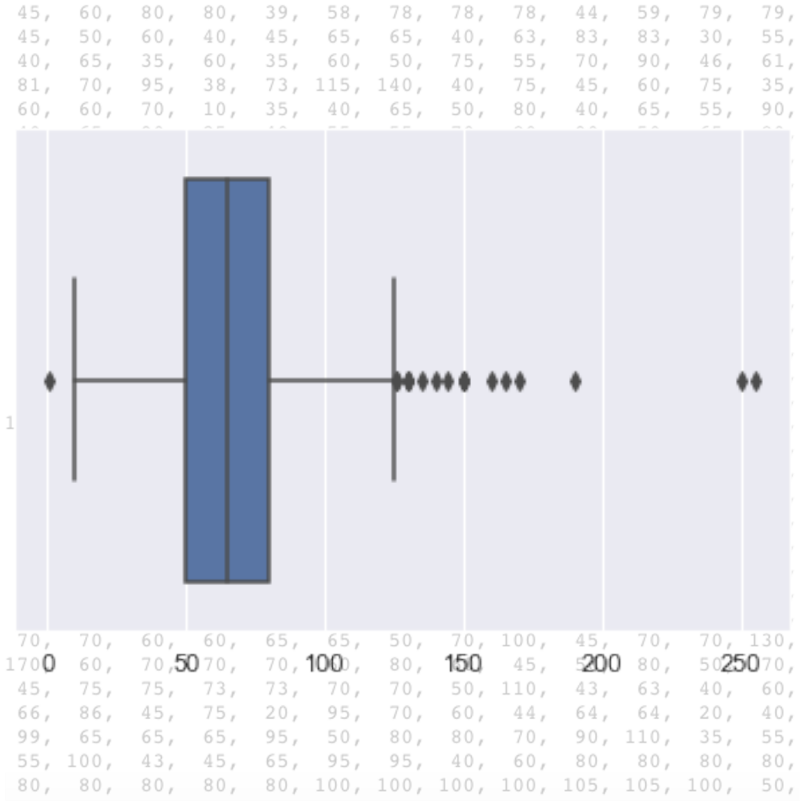
Statistical Summary of the Data

HP The Average Hit Point
Deviation from Average
Median and Quantiles

Statistical Questions

- What is the Central Tendency?
- What is the Spread of the Data?

Source : Kaggle Datasets | **Pokemon with stats** by Alberto Barradas | https://www.kaggle.com/abcsds/pokemon



Data Science

Uni-Variate Statistics

Statistical Summary of the Data

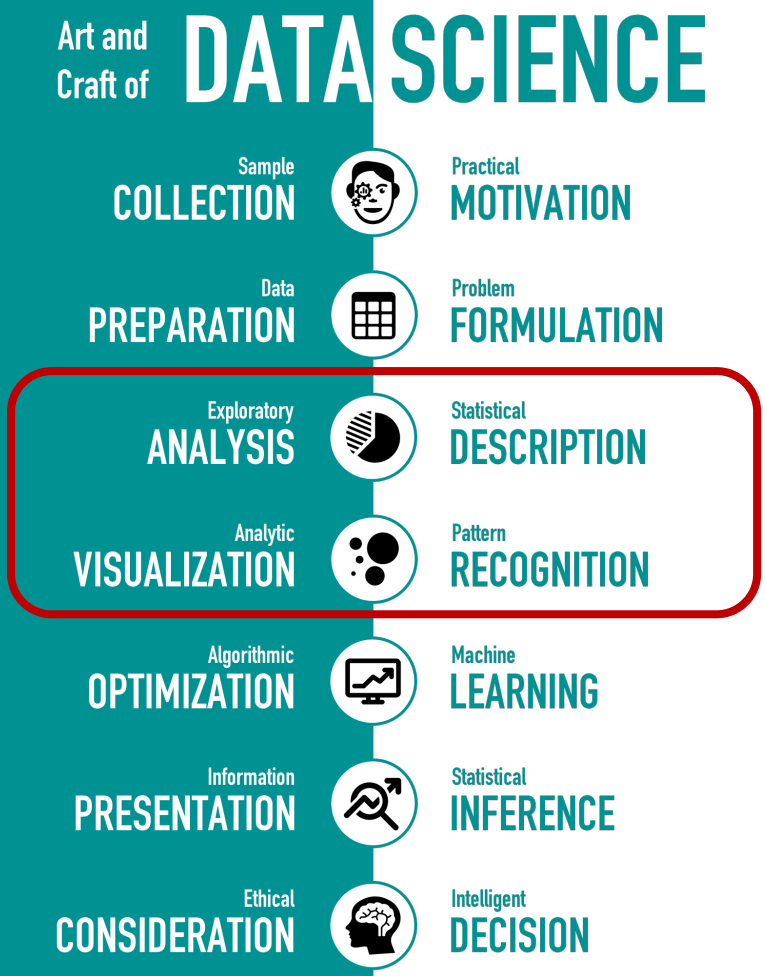
HP

The Average Hit Point
Deviation from Average
Median and Quantiles

Statistical Questions

- What is the Central Tendency?
- What is the Spread of the Data?

Source : Kaggle Datasets | **Pokemon with stats** by Alberto Barradas | https://www.kaggle.com/abcsds/pokemon



Data Science Pipeline

Exploratory Analysis

How to summarize the acquired Data?

How to visualize the acquired Data?

How to analyze the acquired Data?

How to intelligently explore acquired Data?