Understanding the Overall Data Trends

Mohammed Osman
SENIOR SOFTWARE DEVELOPER

@cognitiveosman www.cognitiveosman.com

Overview

Revisiting ML pipeline

Why data analysis?

Data Analysis techniques

- Numerical
- Graphical

Demos

Data Preparation

Problem definition

Data Sourcing

Data Preparation

Data Segregation

Model Training

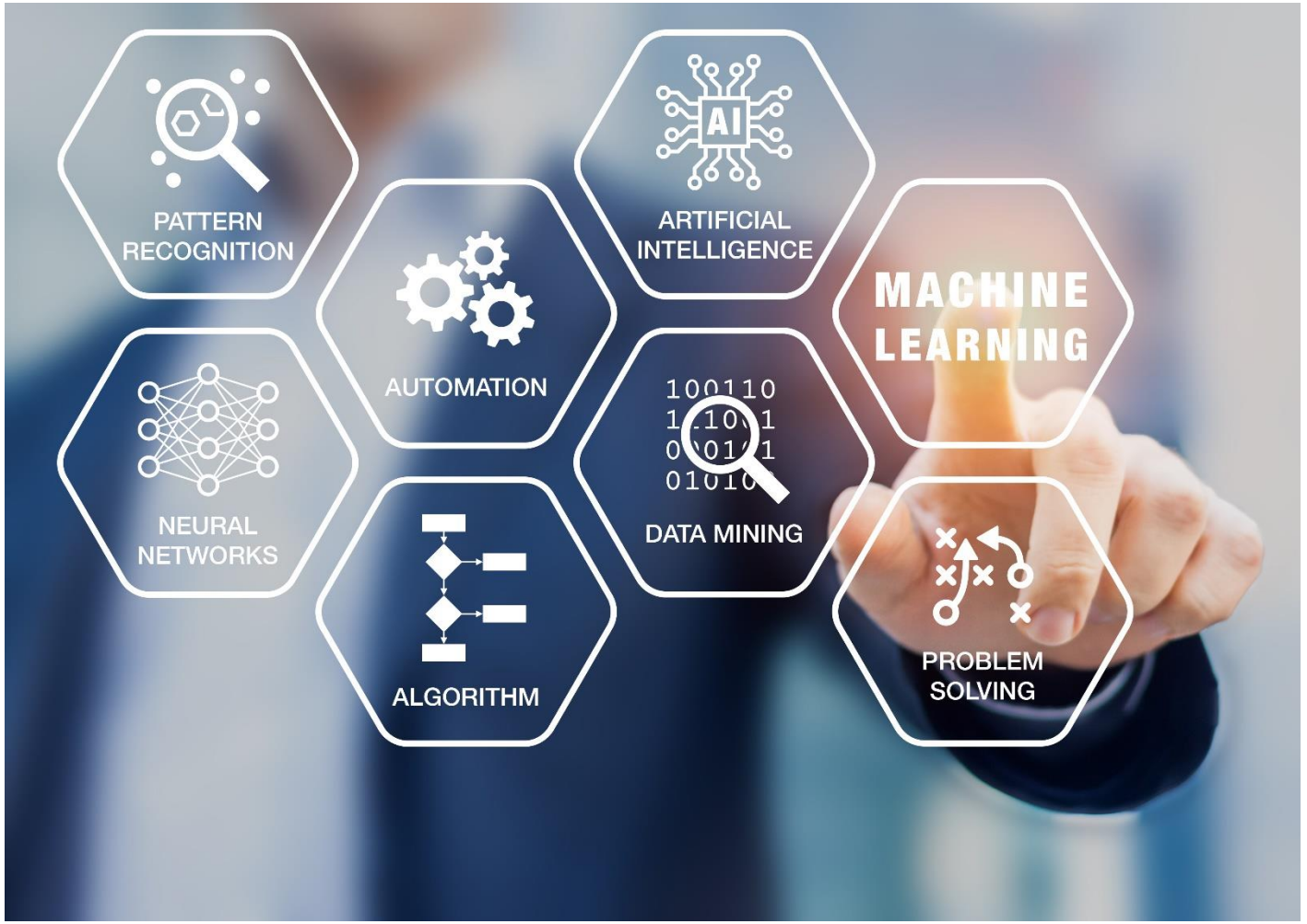
Model Evaluation

Model Deployment

Model Monitoring

Data Analysis

Interdisciplinary Field



Data Analysis

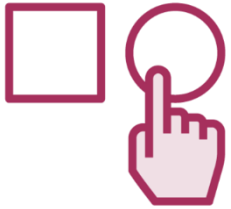
Data analysis is a process of **inspecting**, **cleansing**, **transforming** and **modeling** data with the goal of **discovering** useful information, **informing** conclusion and **supporting** decision-making.

Wikipedia

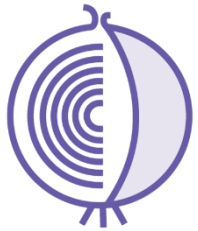
Why Data Preparation and Analysis?

Why Data Analysis: Understanding Our Data

Identifying dataset distribution



Choosing right Machine Learning algorithm



Extracting the right features

Why Data Analysis: Evaluating Our ML Models

Why Data Analysis: Presenting Our Results

Exploratory Data Analysis

Numerical Summaries

Graphical Summaries

Numerical Summaries

Univariant Numerical Measures

Mean

Median

Percentiles

Other

Story: Fair Pay Assessment

Weber PLC

Disclaimer: Company name is fictitious

How Much Weber PLC Pays?

Employee	Salary
Adam	500 $
Sara	300 $
Dina	10000 $
Ali	2000 $
Hans	80000 $
Carl	300 $
John	6000 $
Lisa	1000 $
Maya	12000 $
Khalid	3000 $

Mean

$$\text{Mean} = \frac{\text{Sum of Values}}{\text{Number of Values}}$$

Weber PLC Pay Mean:

$$\frac{500 + 300 + \cdots + 3000}{10}$$

= 11510 $

Adam , Sara and Carl: 500, 300$

Hans: 80000$

+ Mean considers all the values

- Mean is sensitive for extreme values (Carl and Hans salaries)

Median

Median is the value separating lower half from the upper half of the data

Weber PLC Pay Median:

300,300,500,1000,2000,3000,6000,10000,12000,80000

= 2500 $

Hans, Maya, Dina: 80000, 12000, 10000$

\+ Insensitive to extreme values

\- Does not consider dataset distribution

Percentiles

Percentile is a measure used indicating certain percentage of the dataset is below that value.

25% , 50% (Median) and 75%

300
300
500 } 25% = 500
1000
2000
3000 } 50% = 2500
6000
10000 } 75% = 10000
12000
80000

+ More expressive

- Multiple measures

Standard Deviation

Standard deviation is measure that tells the typical difference between the a data value and the mean

$$\sigma = \sqrt{\frac{\sum(\bar{x} - u)^2}{N}}$$

Weber PLC Standard deviation = 23172 $

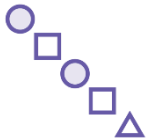
+ Considers all items

+ Considers data distribution

- Harder to calculate

Other Measures

Maximum and Minimum = 80,000 & 300



Count = 10

Mode = 300

Range = 80000-300 = 79700

Outliers = 80000 (larger than mean + 2*standard deviation)

Bivariate Measures

Looking from more than one angle!

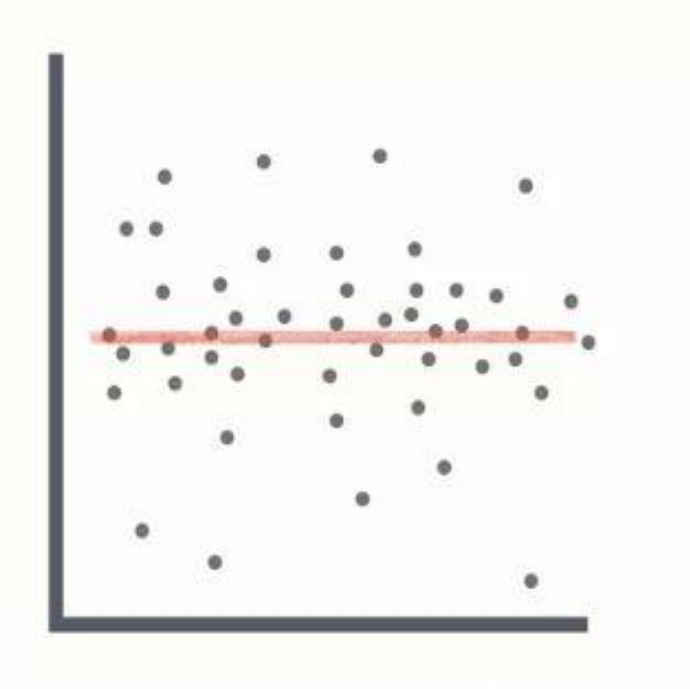
Correlation

Is a measure defining to what extent two or more variables are linearly related

Correlation Cases

Positive Correlation

Negative Correlation



No Correlation

Source: http://bit.ly/2MxmFmT

Correlation

Is a measure defining to what extent two or more variables fluctuate together

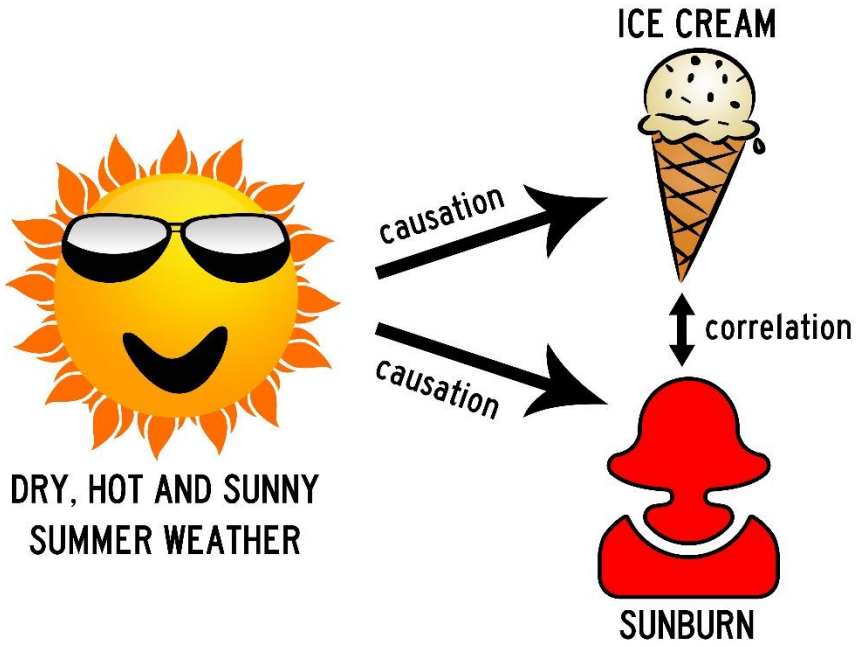
It can be (strong) positive or (strong) negative correlation or no correlation

$$Cor(x,y) = \frac{\sum(\bar{x} - u_x)(\bar{y} - u_y)}{\sqrt{\sum(\bar{x} - u_x)^2(\bar{y} - u_y)^2}}$$

Weber PLC correlation between Salary and Years of Experience= 0.94

Salary	Years of Experience
300	1
300	2
500	2
1000	3
2000	4
3000	4
6000	4
10000	7
12000	10
80000	22

The Correlation Fallacy



Correlation does not imply casuation!
("with this, therefore because of this“ fallacy)

Think of Weber PLC case

Demo

To be updated

Graphical Summaries

Mountains

Trees

Lakes

Lighthouse

Greenness

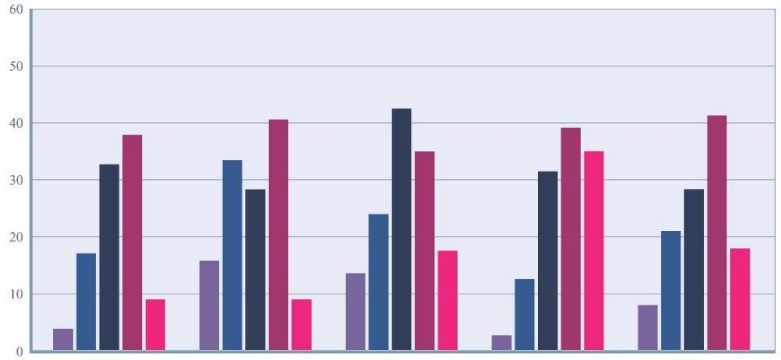
Clear Sky

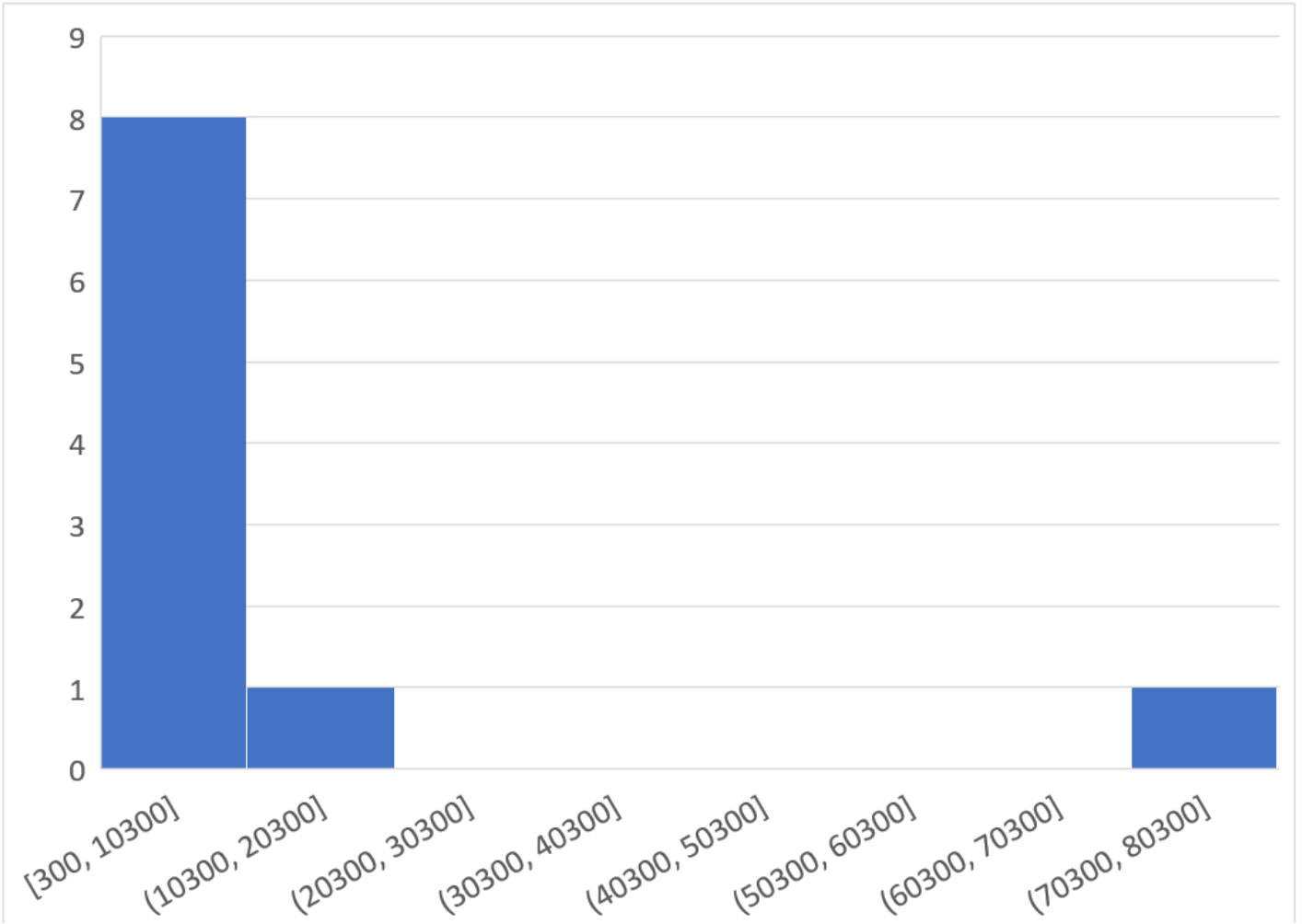
Cottages

Picture superiority effect

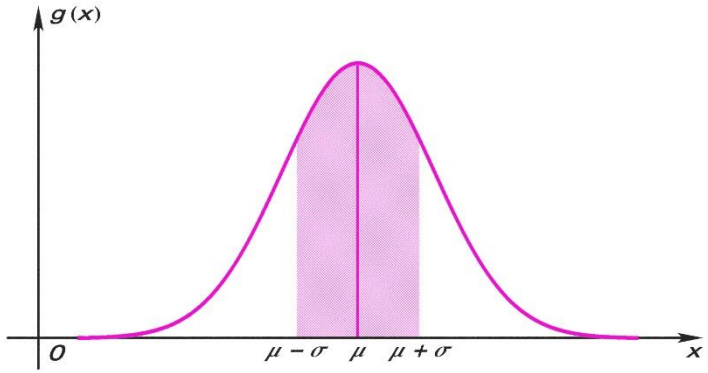
Mike Utilizies Research

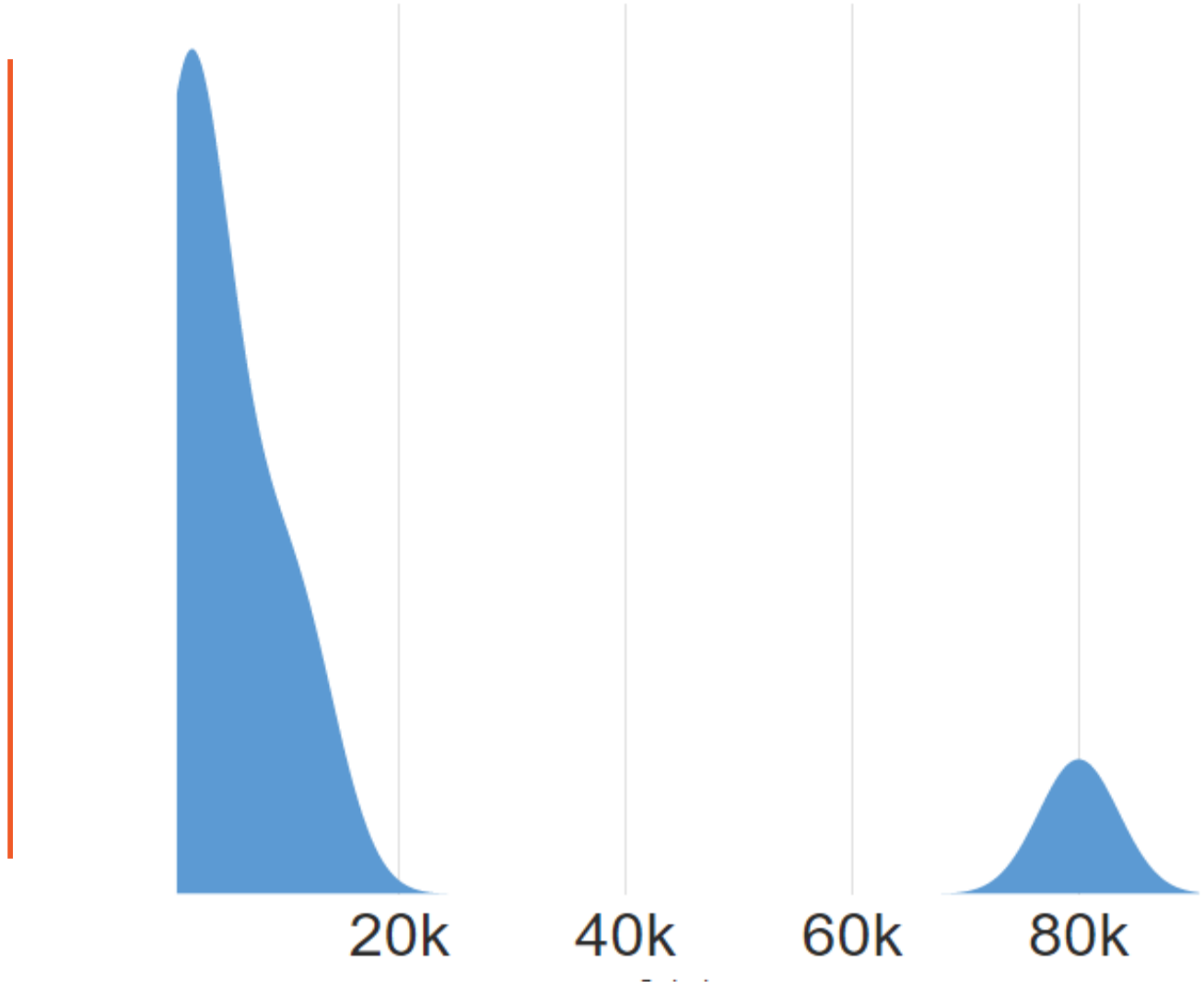
Histograms


60
50
40
30
20
10
0

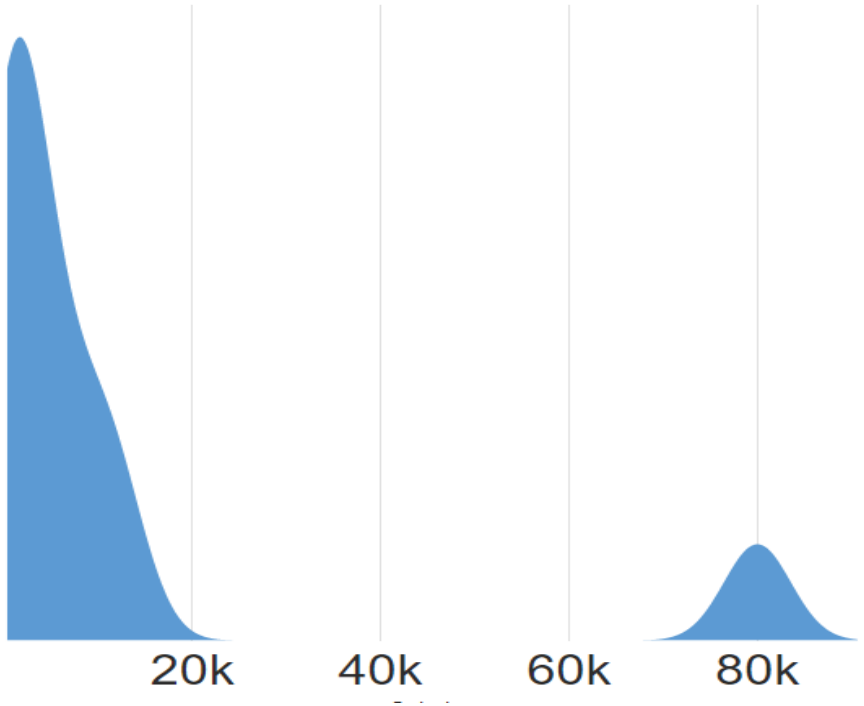

9
8
7
6
5
4
3
2
1
0
[300, 10300]
(10300, 20300]
(20300, 30300]
(30300, 40300]
(40300, 50300]
(50300, 60300]
(60300, 70300]
(70300, 80300]

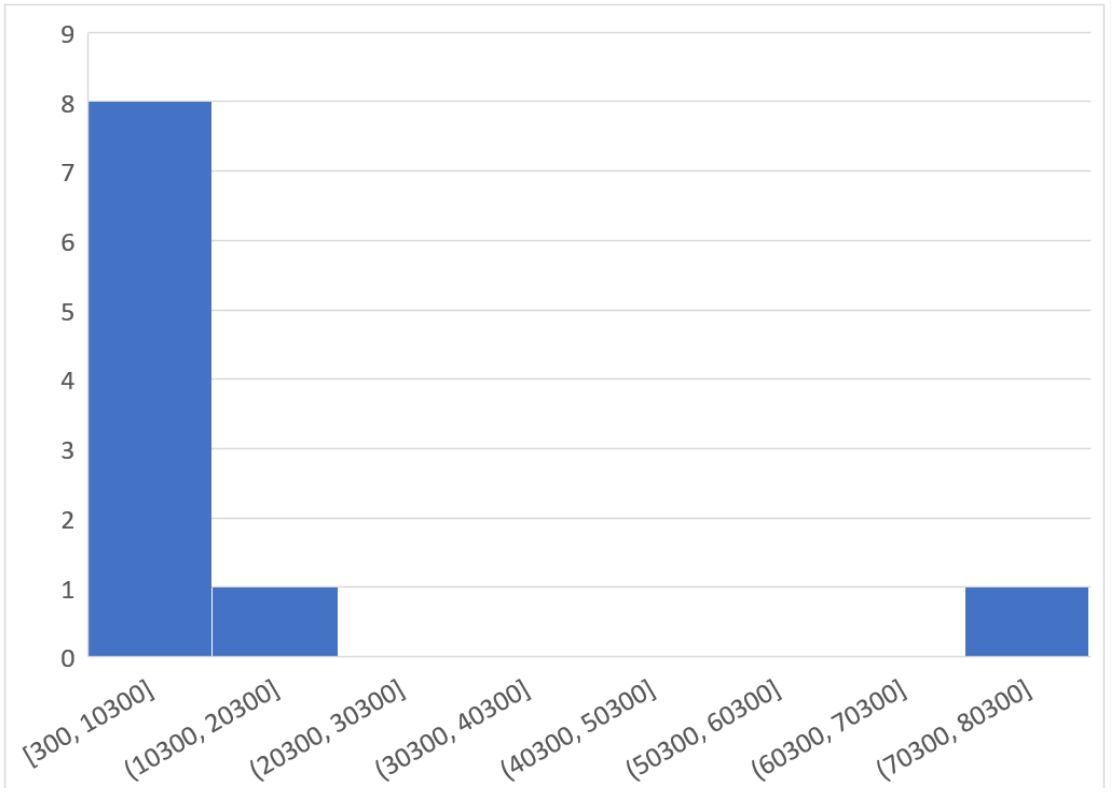
Density (Distribution) Graphs



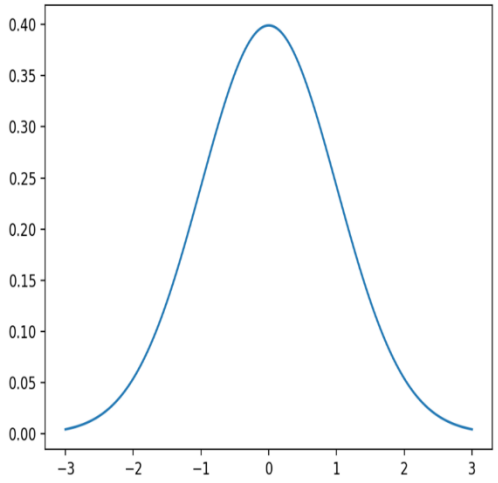


Comparison

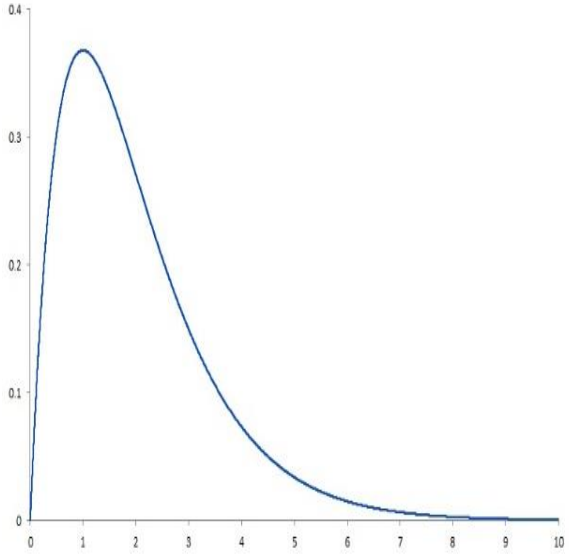

20k
40k
60k
80k


9
8
7
6
5
4
3
2
1
0
[300, 10300]
(10300, 20300]
(20300, 30300]
(30300, 40300]
(40300, 50300]
(50300, 60300]
(60300, 70300]
(70300, 80300]

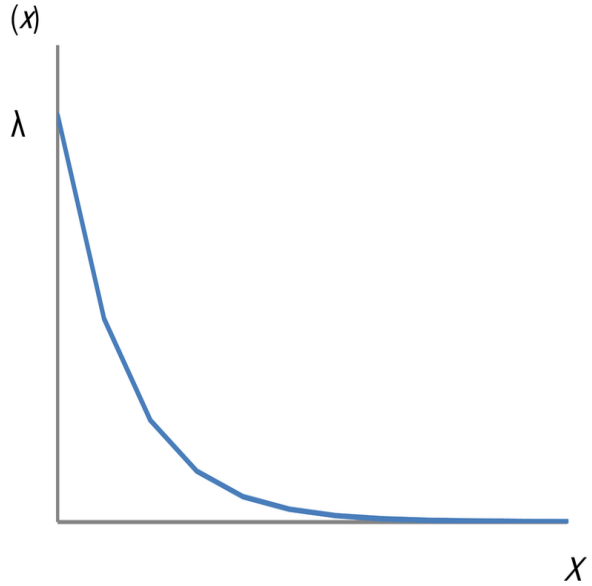
Common Distribution Types



Normal Distribution



Skewed Distribution



Exponential Distribution

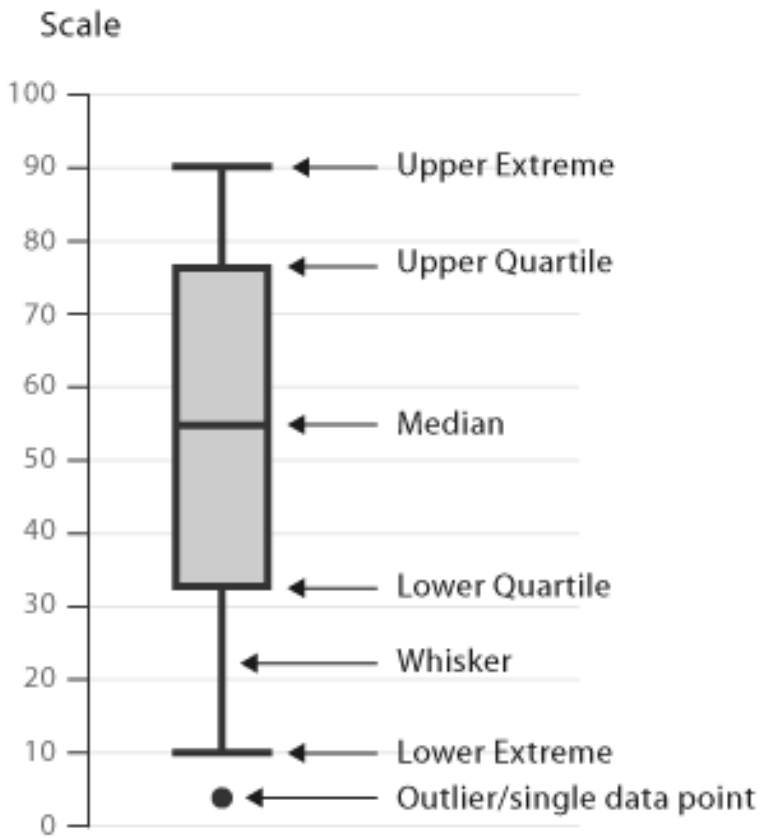
Why Histograms and Density Graphs?

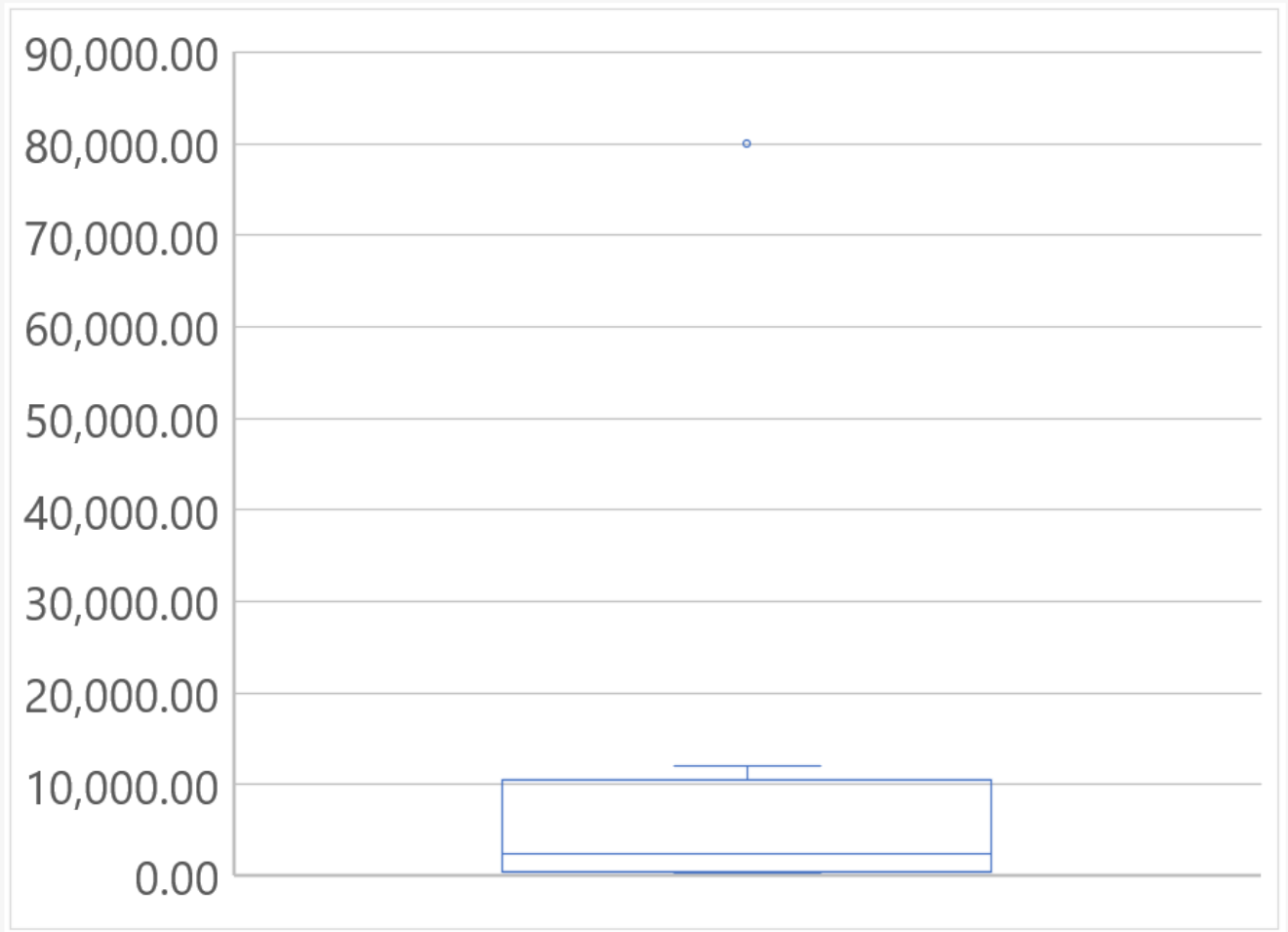
Detecting impossible values

Identifying the shape of the data

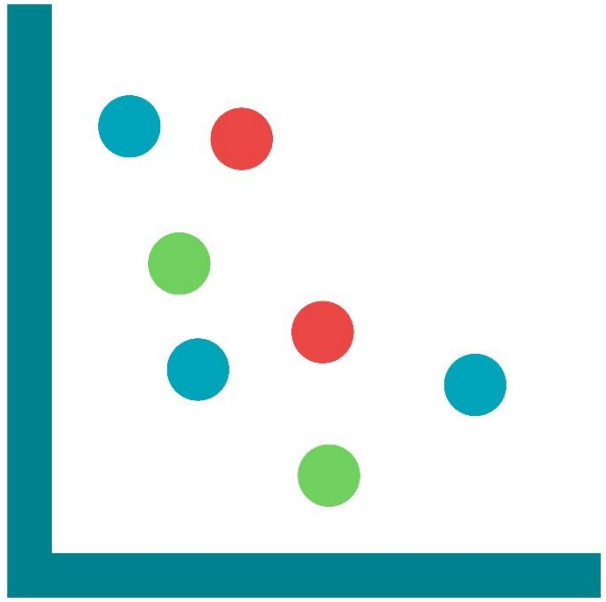
Detecting errors and mistakes in the data

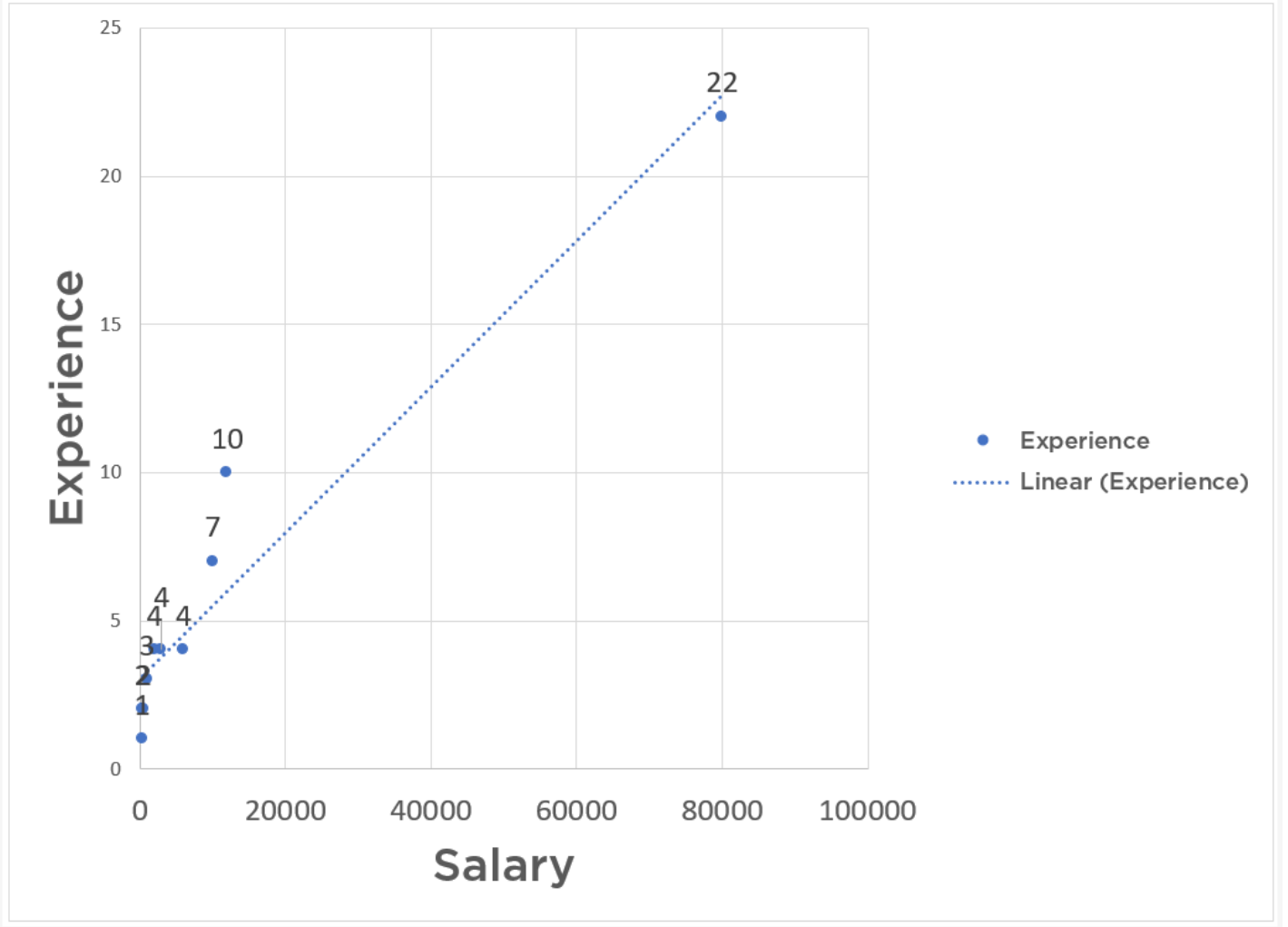
Box and Whisker Plot





Scatter Plot





Demo

To be updated

Summary

Refreshed our minds

Morale of data preparation

Exploratory data analysis

- Numerical (Univariate, Bivariate)
- Graphical

Demos