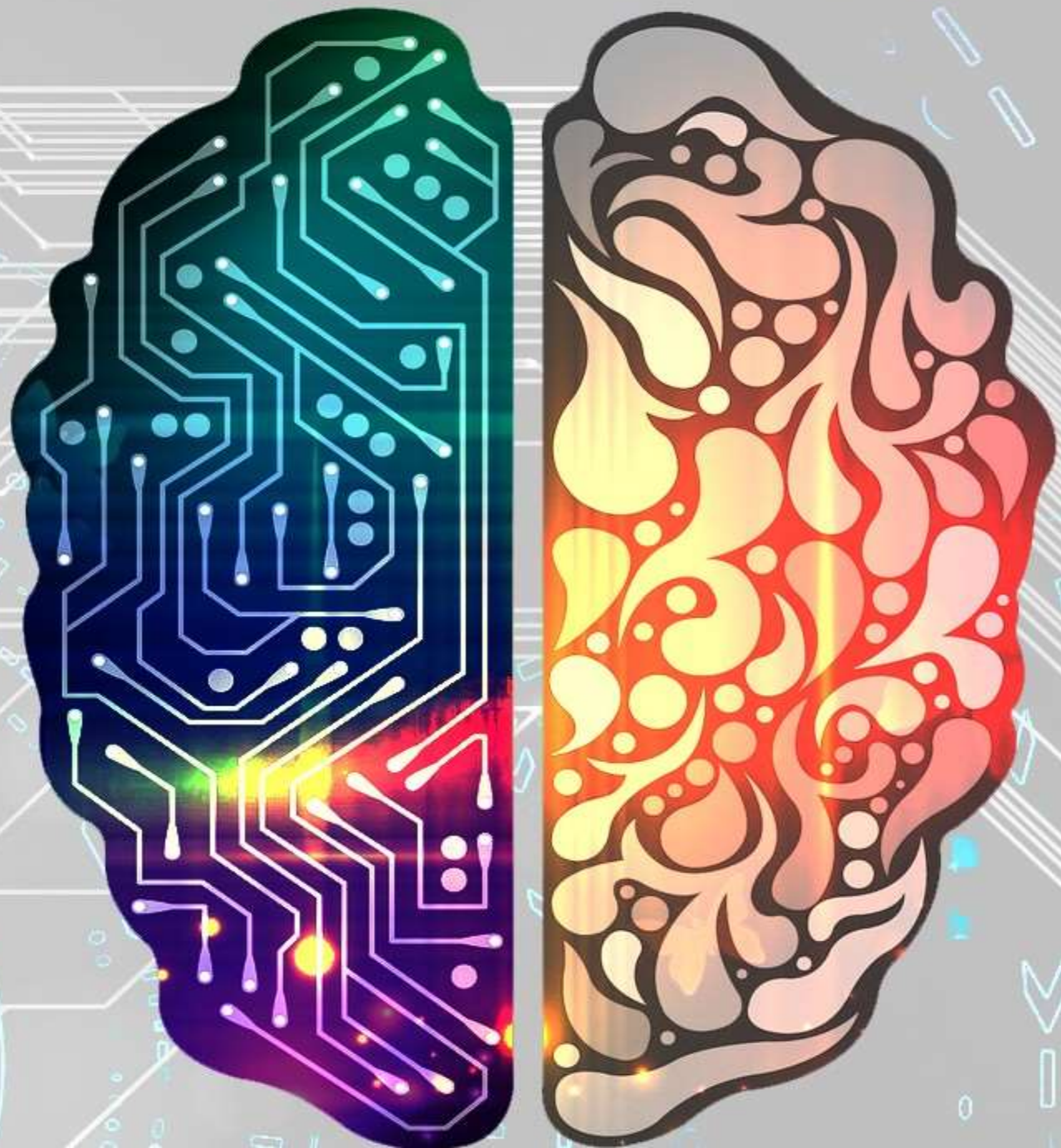
MACHINE LEARNING

Using

Python



By Eng. Mohammed Marwan Shahin

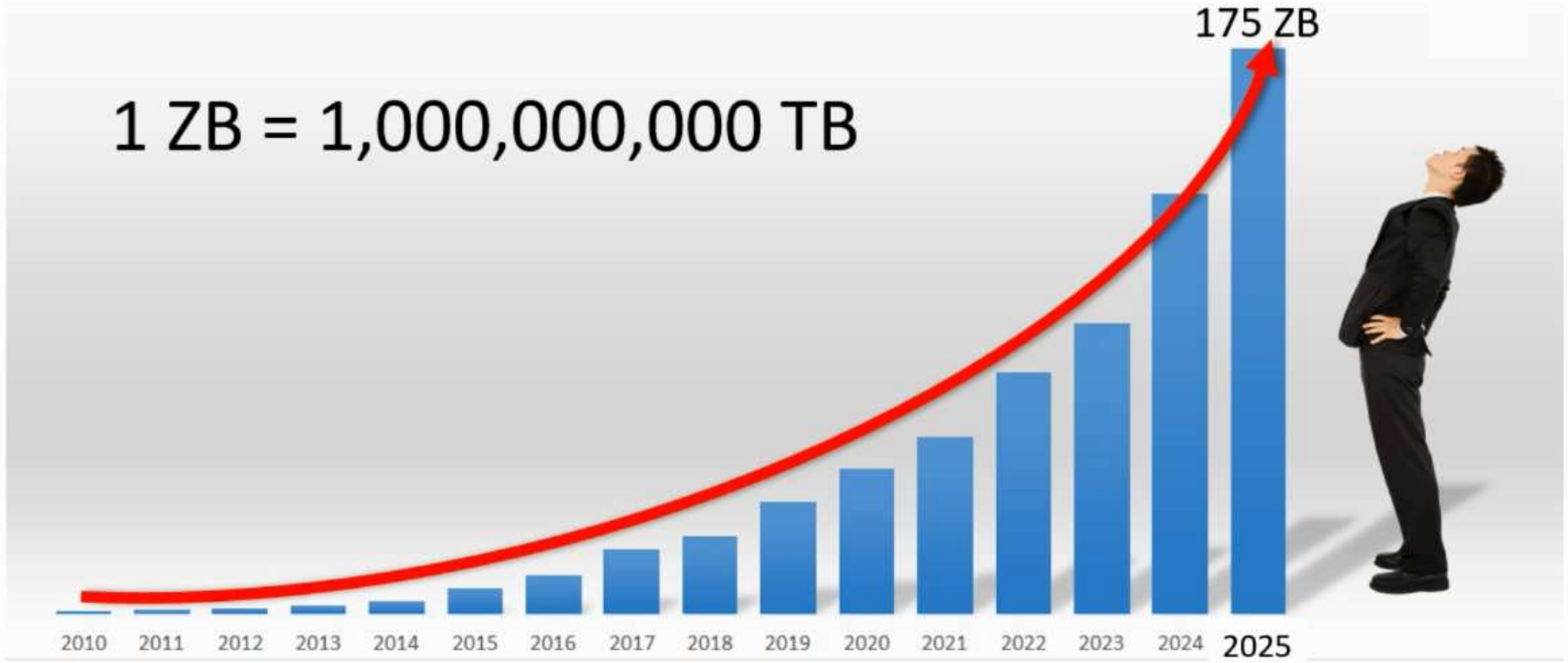
Outlines

- Intro to ML
- Pre-Processing & Visualization techniques
- Math and statistics
- **Supervised ML:**
  - Simple Linear Regression (Regression)
  - Multilinear Regression (Regression)
  - Polynomial Regression(Regression)
  - Decision Tree & Random Forest (Regression)
  - Logistic Regression (Classification)
  - Support Vector Machine (Classification)
  - Decision Tree & Random forest (Classification)
- Evaluate Classification & Regression models
- **Unsupervised ML:**
  - Clustering using (K-mean)
  - Recommendation Engine (Association using Apriori Algorithm)
  - Anomaly Detection

Training Plan

Day no.	Subject
Day 1	Pre-Processing & Visualization techniques
Day 2	Math and Statistic
Day 3 & 4	Regression • Simple Linear Regression • Multilinear Regression • Polynomial Regression • Decision Tree & Random Forest
Day 5	Logistic regression
Day 6	Support Vector Machine
Day 7	Decision Tree & Random forest
Day 8	Clustering Recommendation Engine (Association) Anomaly Detection

Data Growth – IDC-Seagate November, 2018



It’s not just about…..

NETFLIX

Customers who viewed this item also viewed these products

Dualit Food XL1500 Processor

$560

Add to cart

Kenwood kMix Manual Espresso Machine

★★★★☆

$250

Select options

Weber One Touch Gold Premium Charcoal Grill-57cm

$225

Add to cart

NoMU Salt Pepper and Spice Grinders

$3

View options

Application of Data Science and Machine Learning

Automobile

Preventive Maintenance

Banking

Fraud/Default Prevention

Healthcare

Predict Disease

Media

Content Personalisation

Telecom

Increase sale

Application of Data Science and Machine Learning

Application of Data Science and Machine Learning

Banking

Fraud/Default Prevention

Application of Data Science and Machine Learning

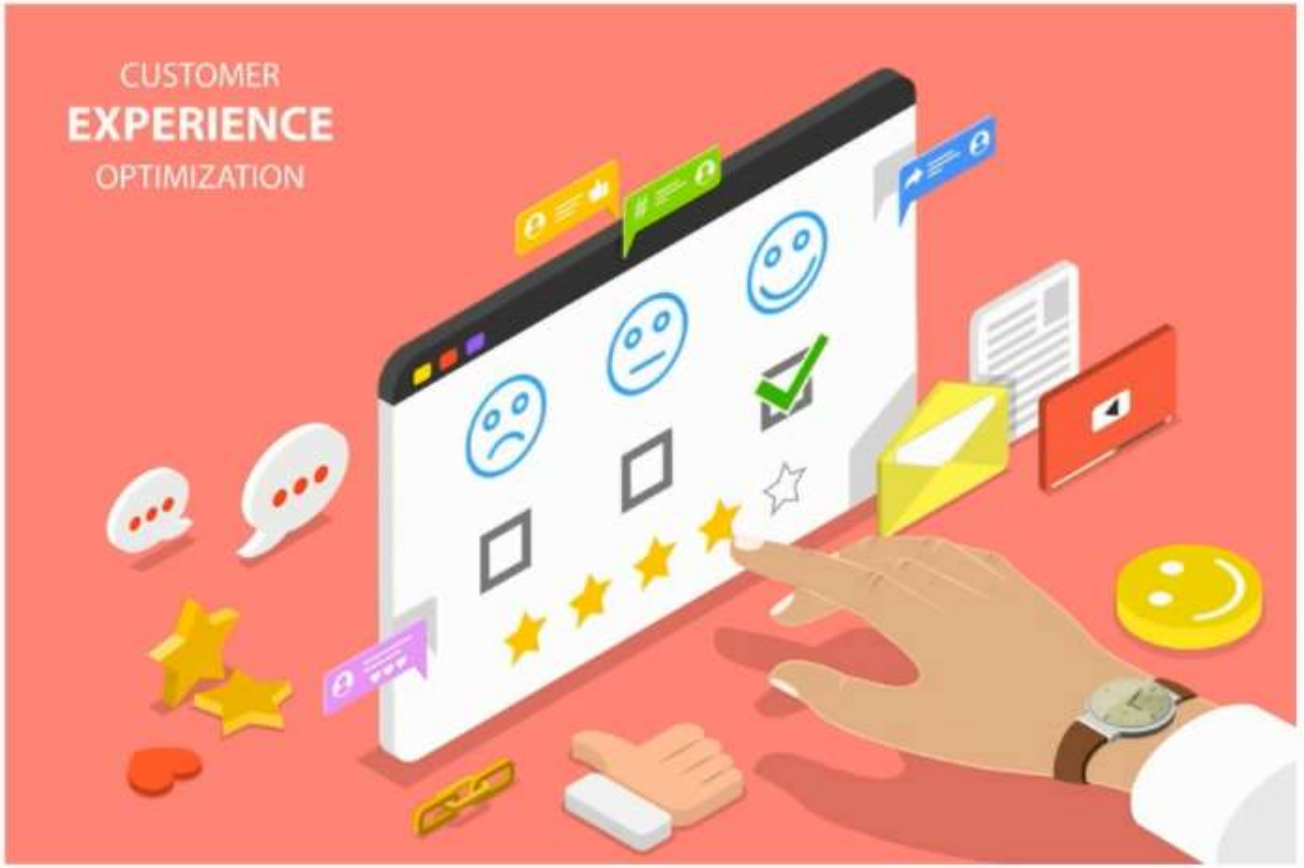
Healthcare

Predict Disease

Application of Data Science and Machine Learning

Media

Content Personalisation



Application of Data Science and Machine Learning

Telecom

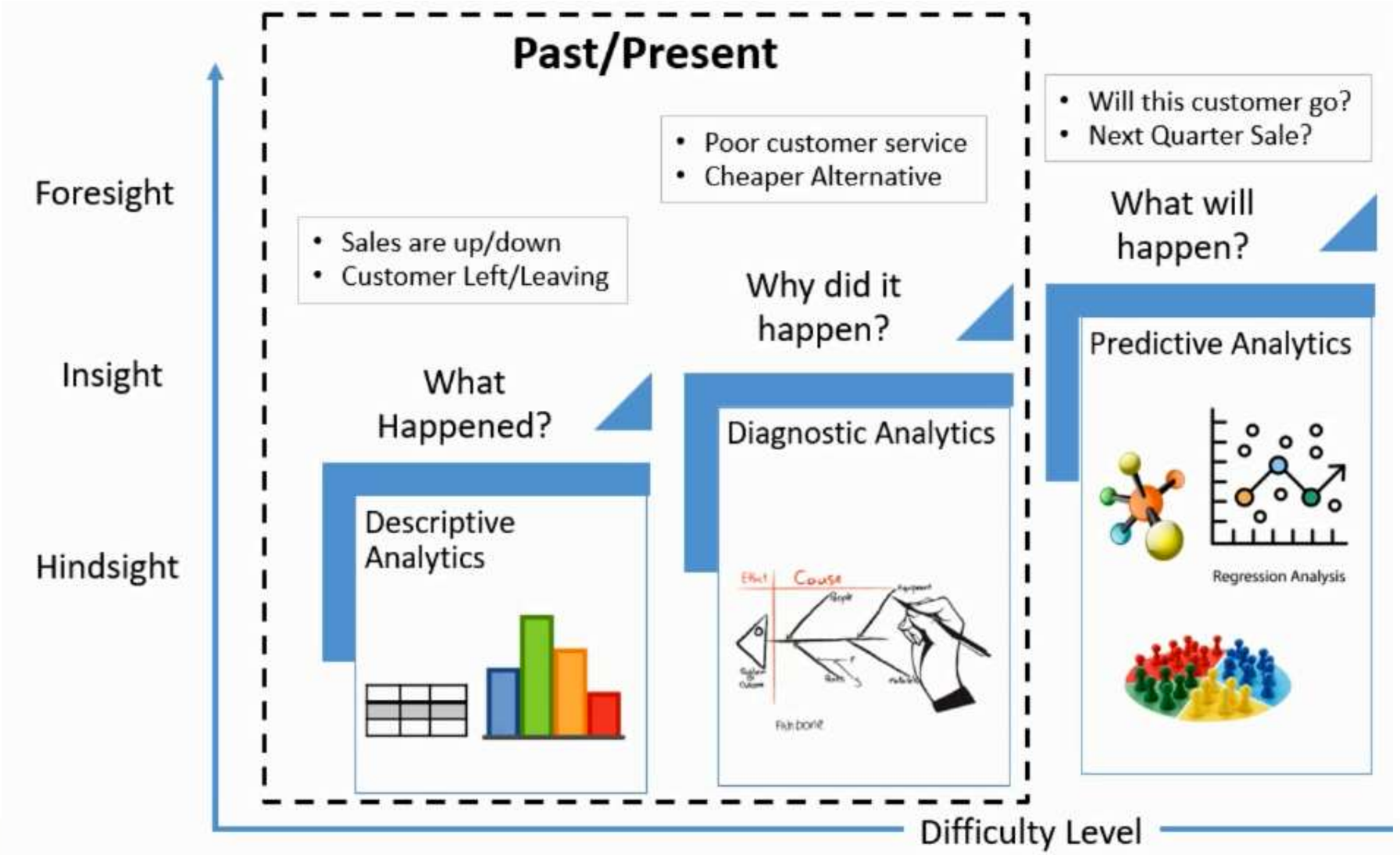
Increase sale

With the help of Data Science, we can identify the customer issues in advance to retain them.

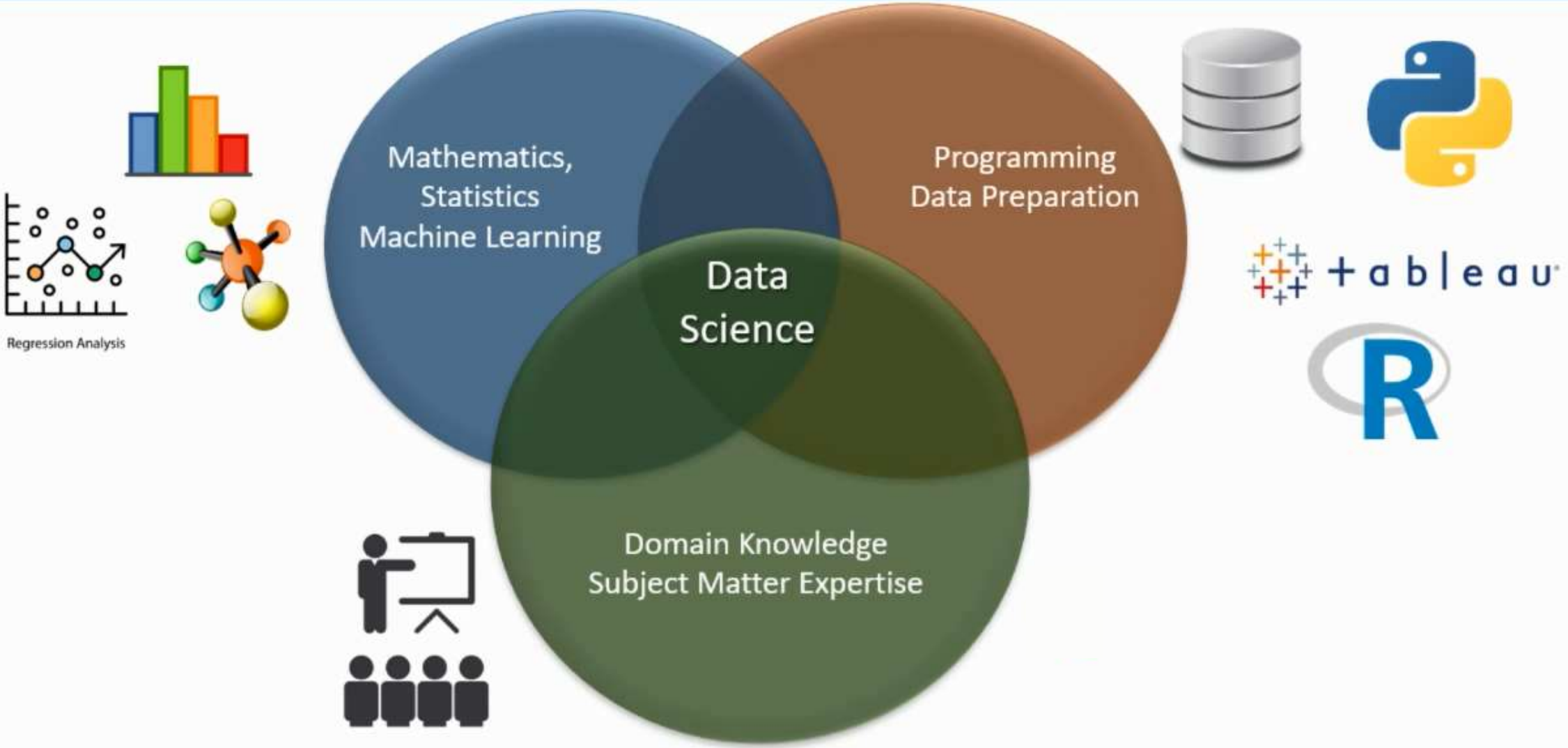
Mr. Data Scientist. Our customers are leaving us and new sale is hard to get. What do you suggest?

By predicting which customer will be more likely to buy our product, we can increase our sale.

Types of Analytics



What is Data Science?



What Is Machine Learning?

- **Machine learning** is the subfield of computer science that gives computers the ability to learn without being explicitly programmed.
  – Arthur Samuel, 1959

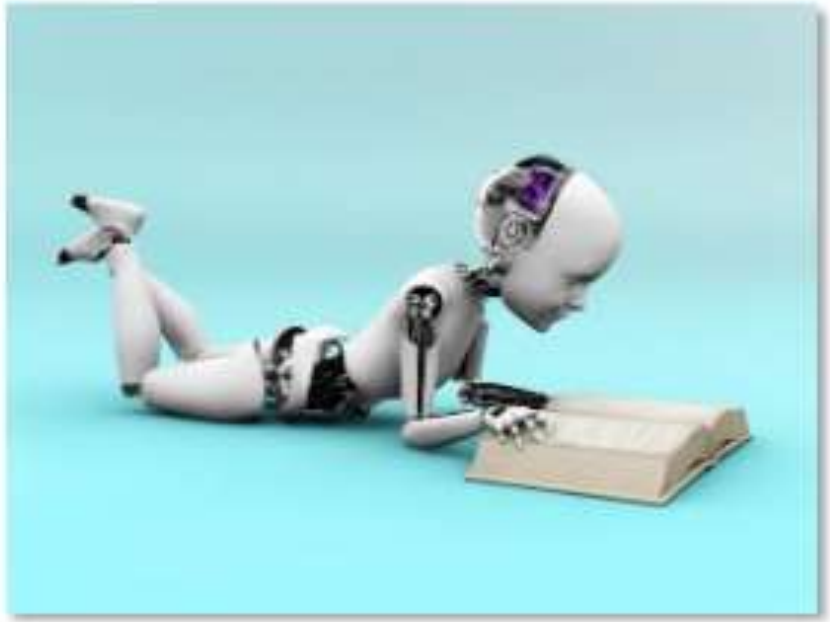
- Learns from past behaviour and make predictions or decisions

- Extraction of knowledge from past data

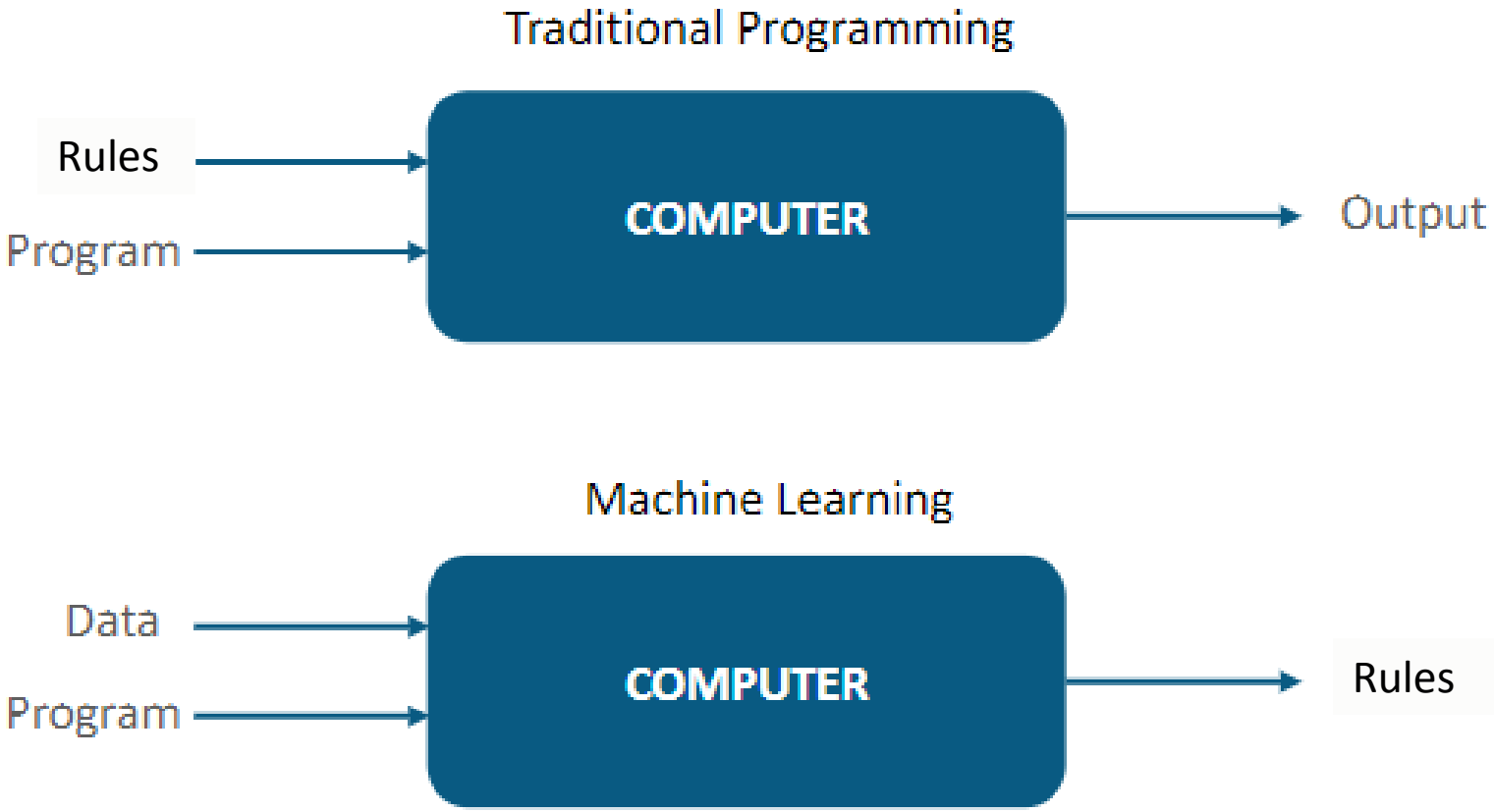
What is Machine Learning?

- Machine Learning is a class of algorithms which is data-driven, i.e. unlike "normal" algorithms it is the data that "tells" what the "good answer" is
- Getting computers to program themselves and also teaching them to make decisions using data

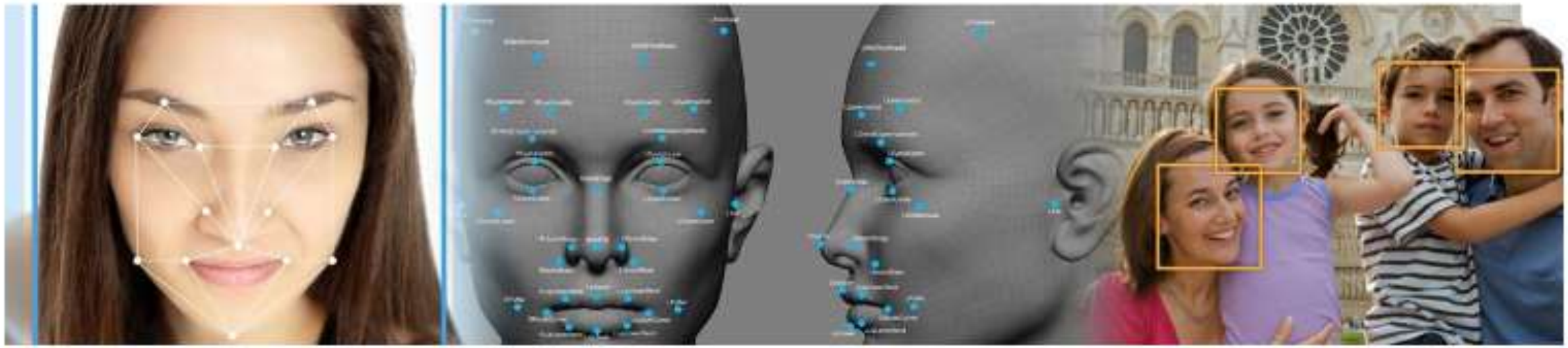
"Where writing software is the bottleneck, let the data do the work instead."



How it works?


Traditional Programming
Rules
Program
COMPUTER
Output
Machine Learning
Data
Program
COMPUTER
Rules

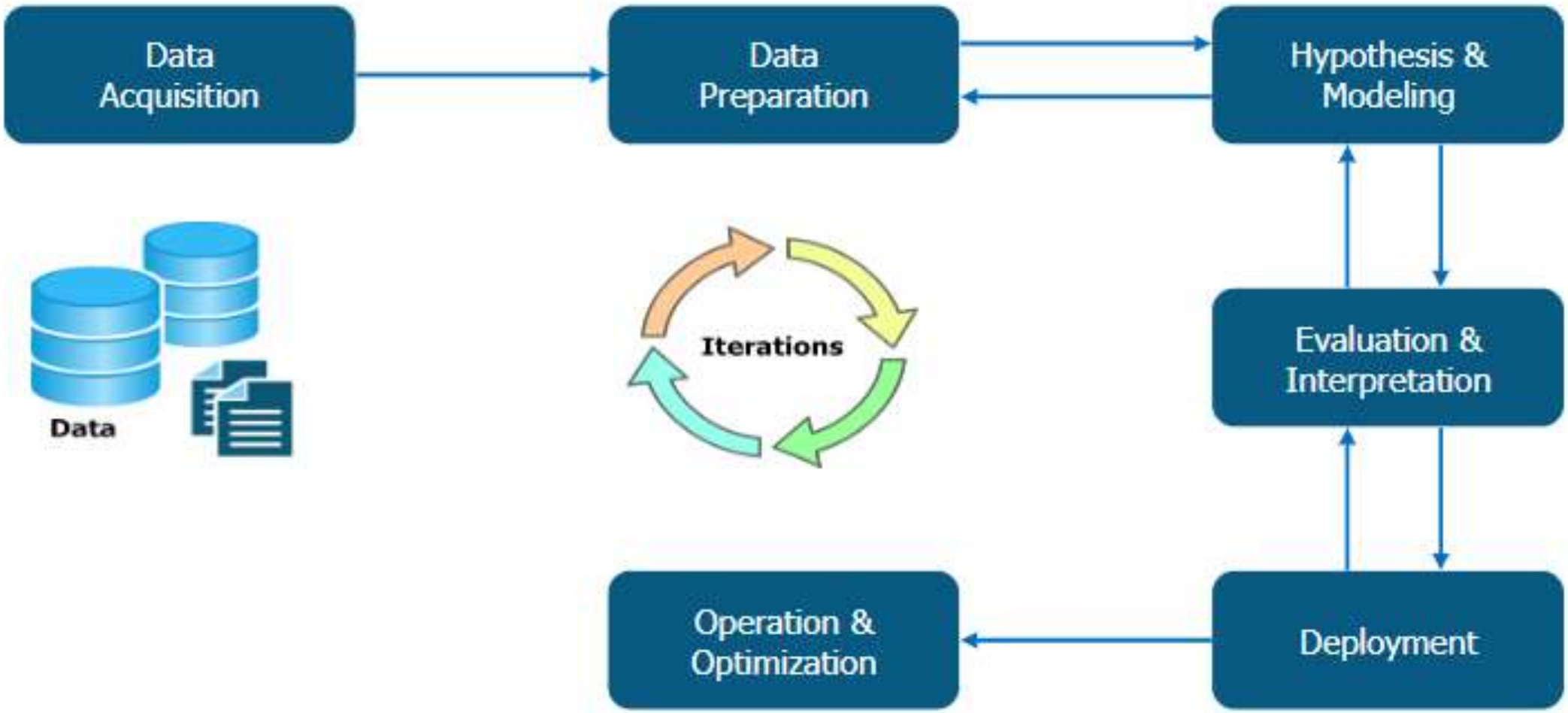
Machine Learning - Example



- In a hypothetical non-machine learning algorithm for face recognition in images, we would have to define what a face is
- But in machine learning algorithm there will be no such definition, instead it will “learn-by-examples“. We will show several images of faces and eventually a good algorithm will be made .It will be able to predict whether or not an unseen image is a face or not

Life Cycle of Data Science

Life Cycle of Data Science



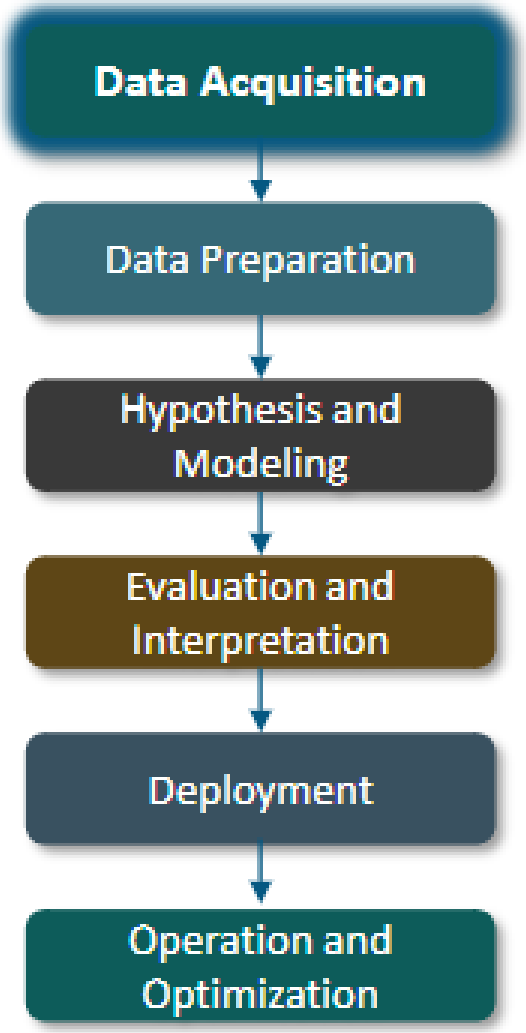
Consider a real estate company

Hi I am john, I need some baseline for pricing my apartments

I will help you with that.

Phases of Life Cycle of Data Science

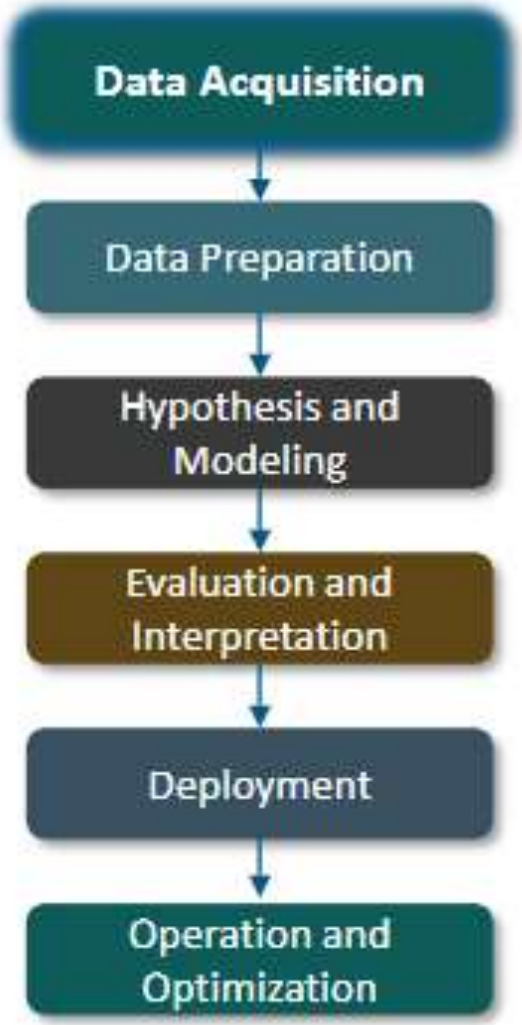
Data Acquisition/ Data Collection



Now to help John let's see what data we can collect from different locations and how it affects the pricing of an apartment

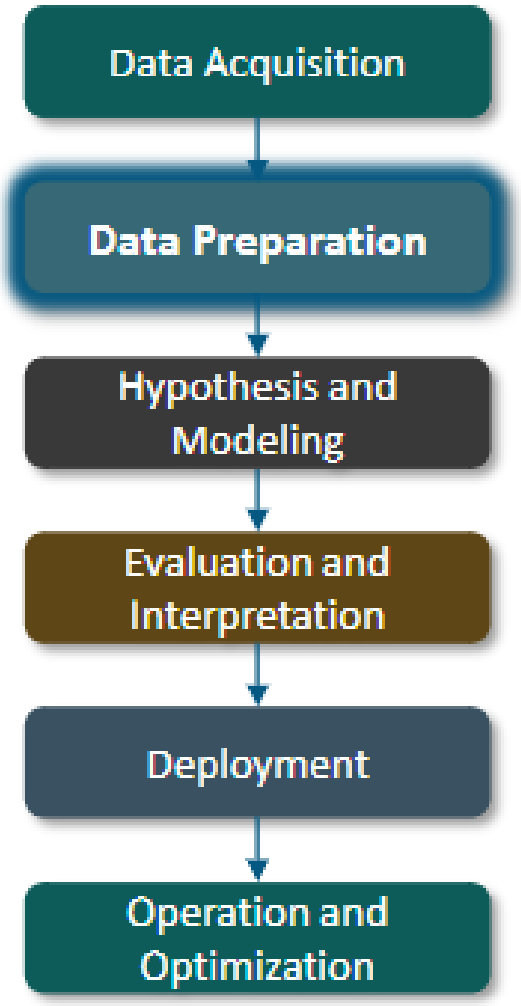
Price	Apartment name/no.	No of bedrooms	Floor number	Criminal rate per year	Pollution level	Distance to nearby Educational institution
30L	xv	3	2	3	15	900 m
20L	cs	2	4	2		2 km
28L	df	2	G	5	13	1.5 km
25L	re	1	3	1	12	1.7 m
30L	sd	2	0	3	13	700 m

Data Acquisition/ Data Collection



- Data acquisition involves acquiring data from all the identified internal and external sources that can help answer the business question
- This data could be
  - logs from webservers
  - social media data
  - census datasets
  - data streamed from online sources via APIs

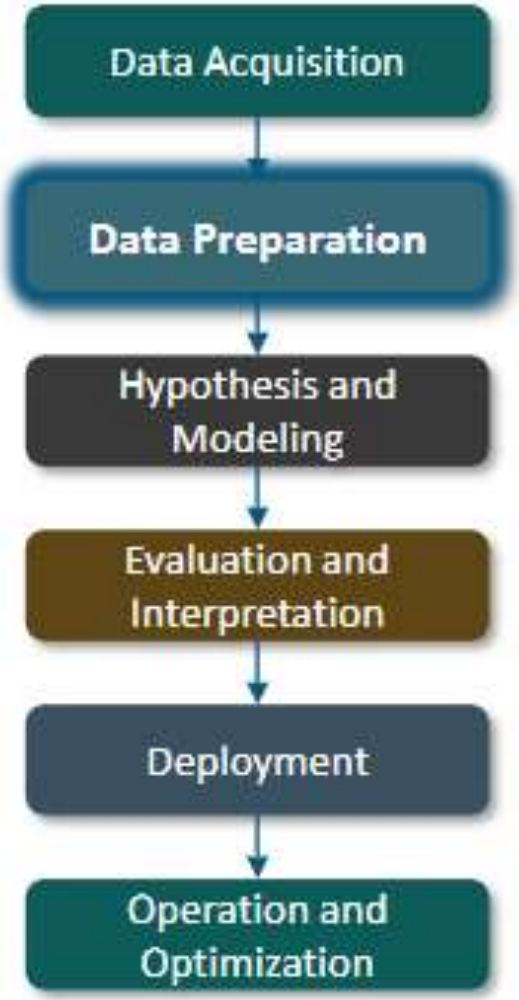
Data Preparation



- The data we have collected is not clean, there are some errors which need to cleansed
- Also we may need to change the values of columns as per requirements

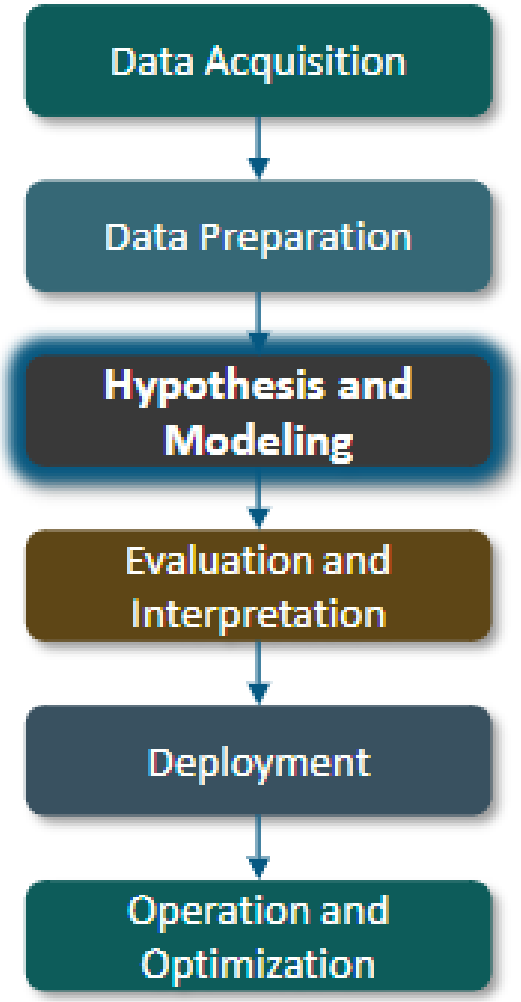
Price	Apartment name/no.	No of bedrooms	Floor number	Criminal rate per year	Pollution level	Educational institution within 1km radius
30L	xv	3	2	3	15	Yes
20L	cs	2	4	2		No
28L	df	2	G	5	13	No
25L	re	1	3	1	12	No
30L	sd	2	0	3	13	Yes

Data Preparation

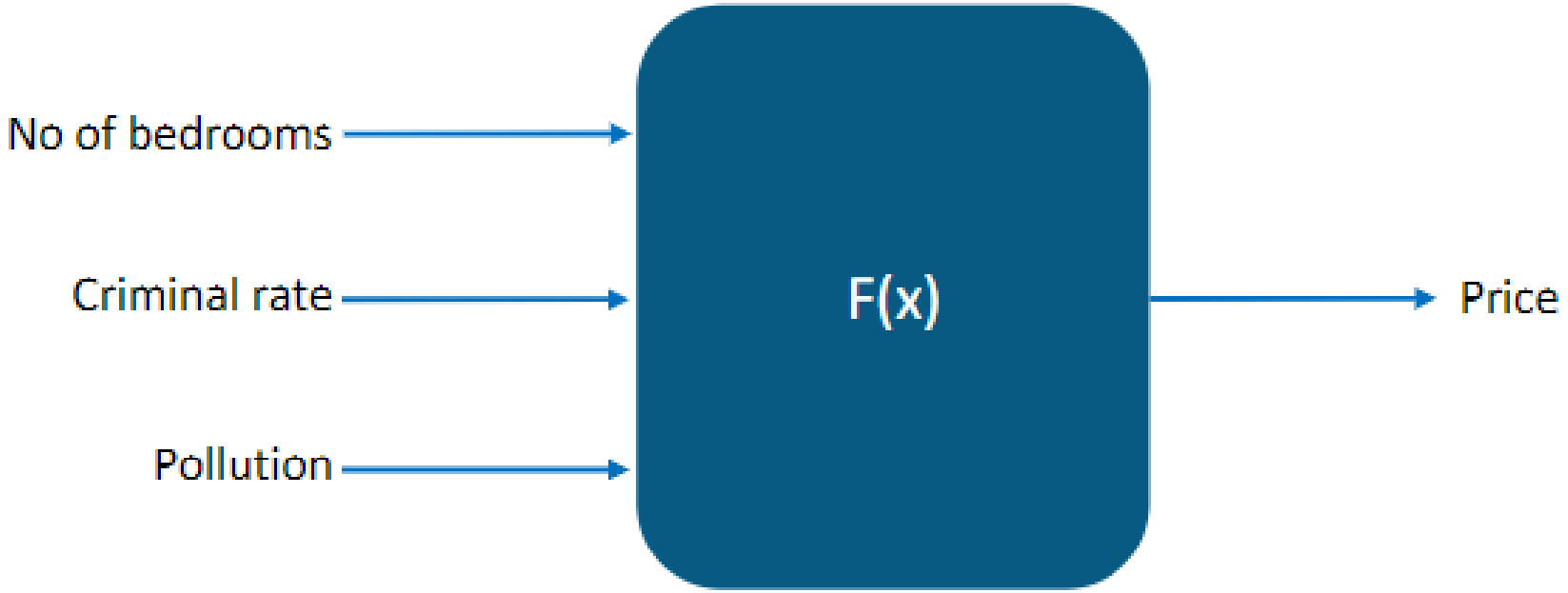


- Data Wrangling is the process of cleaning and unifying messy and complex data sets
- Data after reformatting can be converted to JSON, CSV or any other format that makes it easy to load into one of the data science tools

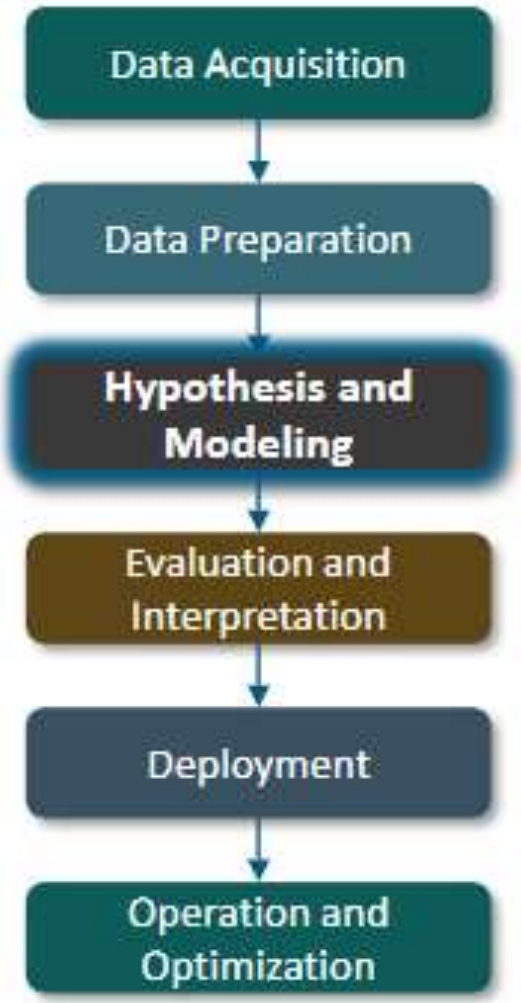
Hypothesis and Modeling



Based on the requirements, a model is created using the dataset

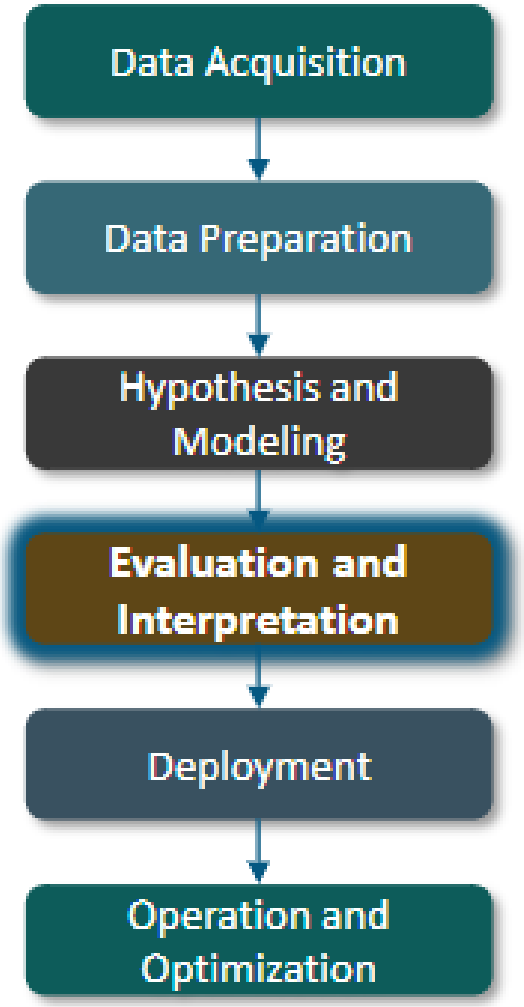


Hypothesis and Modeling

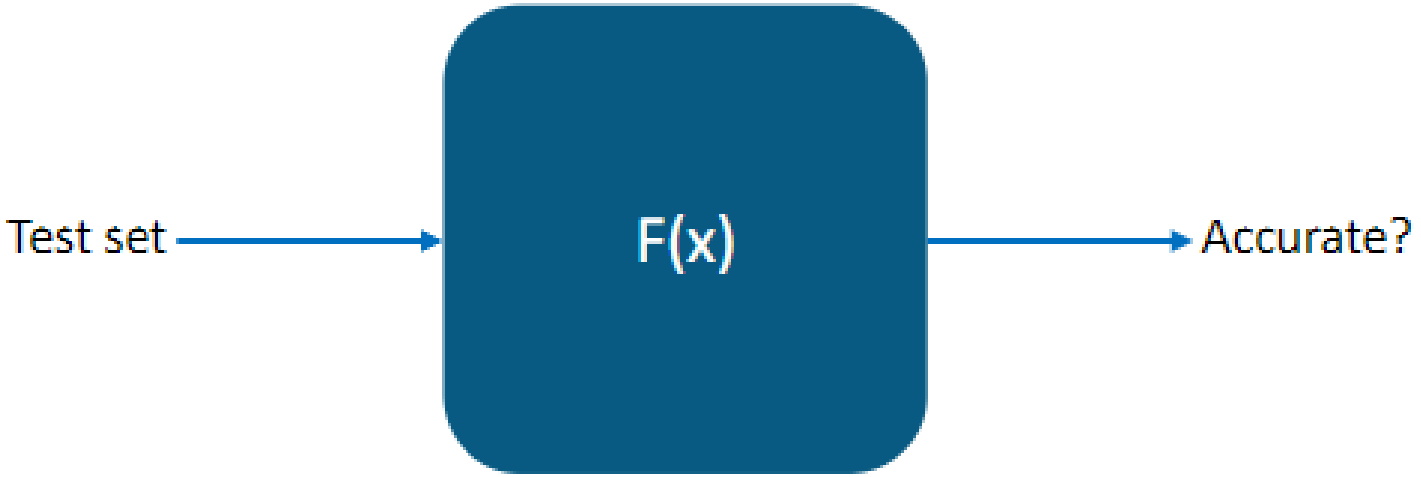


- Involves forming and testing hypotheses about the data and the processes that generate it
- Requires writing, running and refining the programs to analyze and derive meaningful business insights from data
- Mostly written in languages like Python, R, Spark

Evaluation and Interpretation

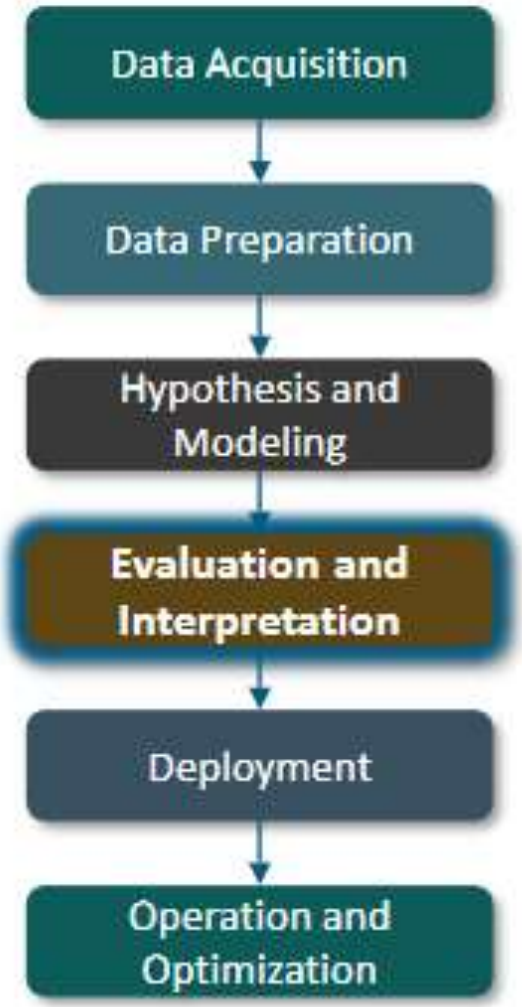


This model is evaluated using test data set



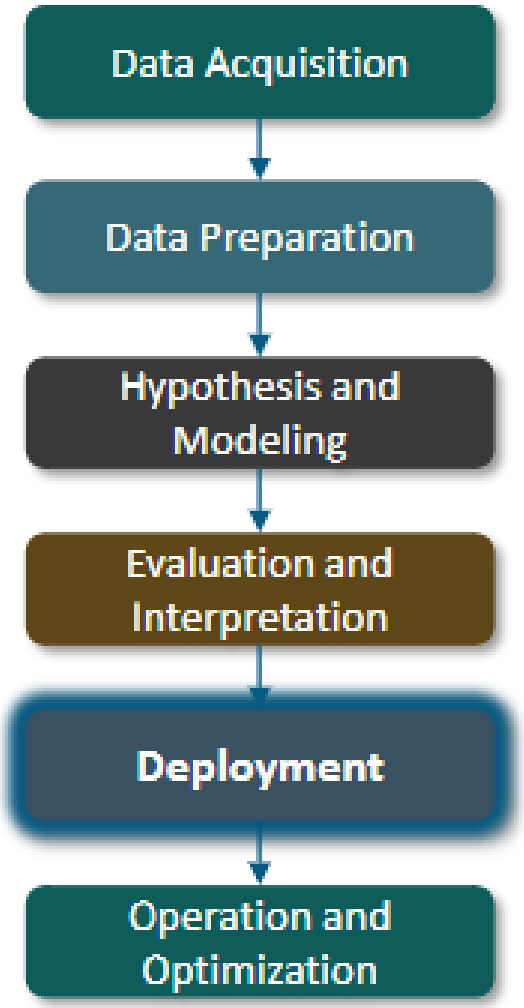
If accuracy is low, the above steps are repeated until a good model is found

Evaluation and Interpretation

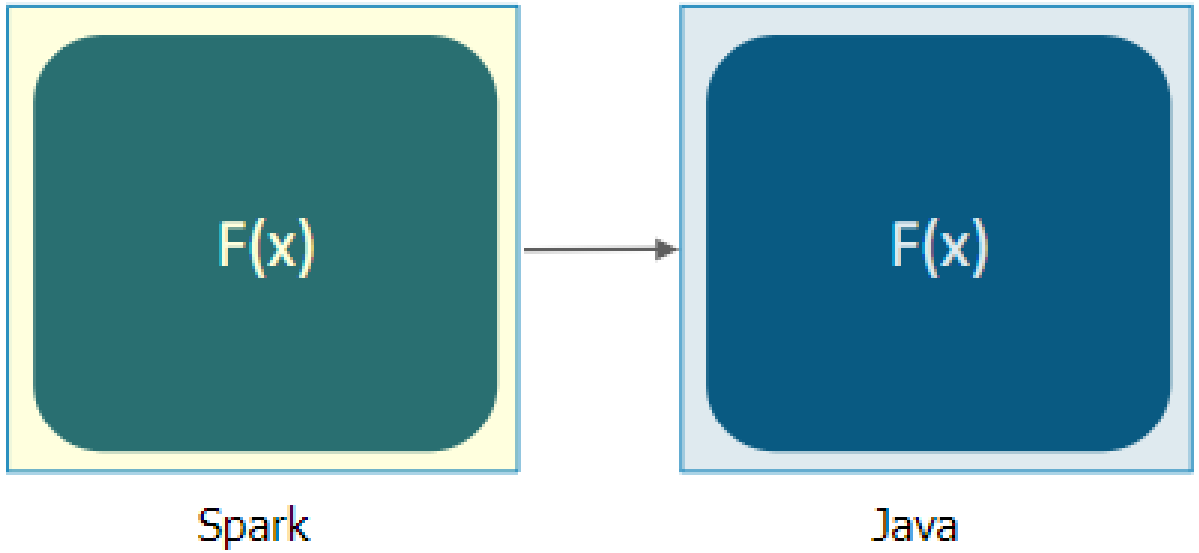


- Model performances should be measured and compared using validation and test datasets
- Models should have a high accuracy for implementation

Deployment



Data scientist might have done this in python or spark, but if the production environment supports only Java then he needs to recode it

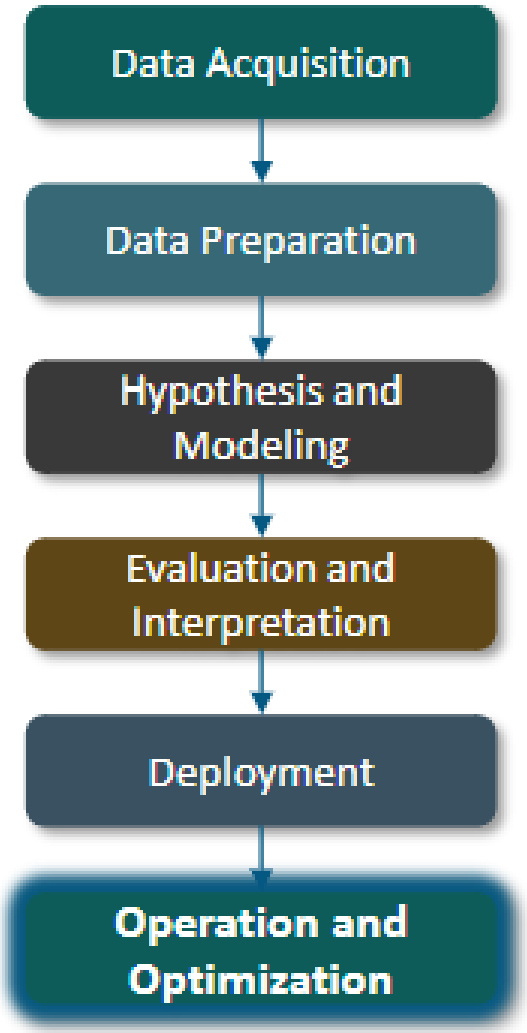


Deployment

- Data Acquisition
- Data Preparation
- Hypothesis and Modeling
- Evaluation and Interpretation
- **Deployment**
- Operation and Optimization

- In this step the model we created is deployed in to the market
- Models generally have to be recoded before deployment (e.g., data scientists may favor Python, but production environments may require Java)

Operation and Optimization



Retraining the model using new factor from data set

F(x)

Y(x)

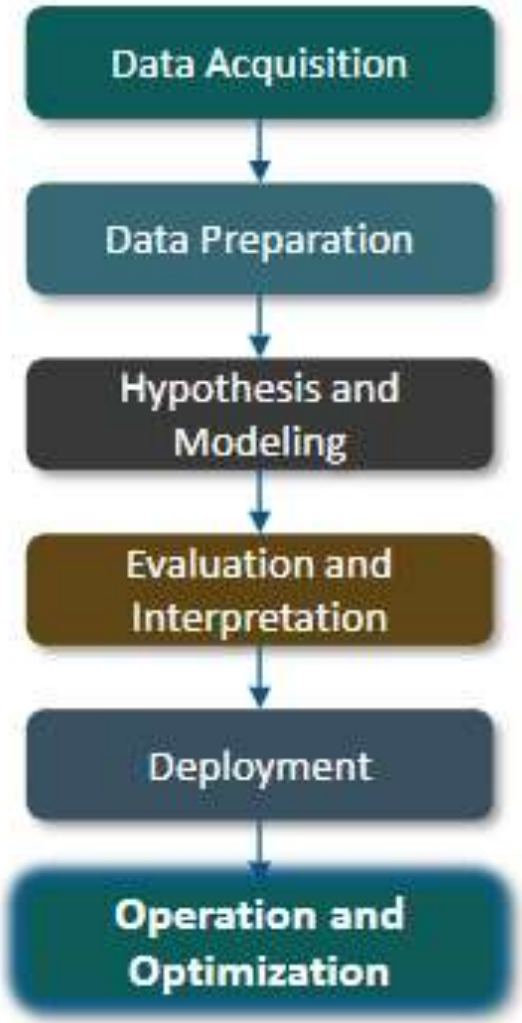
Accurate?

New factor from data set

(E.g. Educational institutions)

After the model is retrained we evaluate the model and deploy it

Operation and Optimization



- Involves developing a plan for monitoring and maintaining the data science project in the long run
- Performance downgrade is monitored in this phase
- Model is retrained whenever a new dataset is added or there is a downgrade in performance

Important aspects of Data Science

Statistics

Statistics is a crucial part of data science. If you think about the 3 phases of a typical data science project, Data Collection, Data Analysis and Results Communications, statistics is critical in the first two. You need to apply appropriate sampling techniques so data collected are not biased.

Machine Learning

Machine learning as a technology helps analyze large chunks of data, easing the tasks of data scientists in an automated process and is gaining a lot of prominence and recognition. It has changed the way data extraction and interpretation works by involving automatic sets of generic methods.

Software Engineering

Software is the generalization of a specific aspect of a data analysis. If specific parts of a data analysis require implementing or applying a number of procedures or tools together, software is the encompassing of all these tools into a specific module or procedure that can be repeatedly applied in a variety of settings.

Why Python?

Python's popularity for Data Science is largely due to the strength of it's core libraries (Numpy, Scipy, pandas, matplotlib, Ipython), high productivity for prototyping, and building small and reusable systems

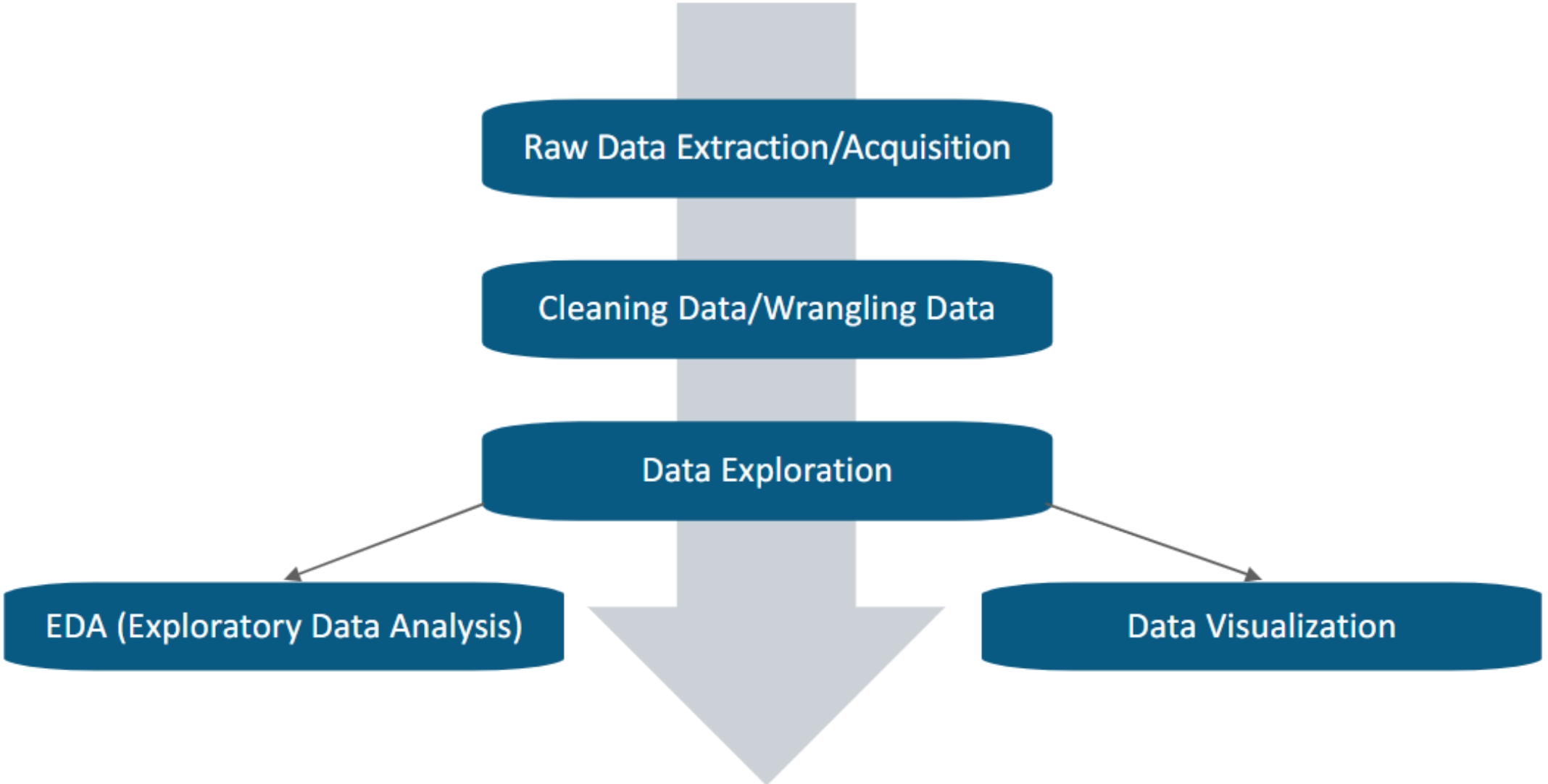
Python is easy to learn, scalable, awesome visualisation, packages and excellent Python community where you find Data Science libraries

Data Science

Python

Data Extraction, Wrangling and Visualization

Steps Involved in the Process of Handling Data



Data Acquisition

- Data Acquisition is the process of retrieving data out of (usually unstructured or poorly structured) data sources for further data processing or data storage

- We can classify data into,
  - Static Data and Dynamic data
  - Hard and Soft Data
  - Structured, Unstructured and Semi-Structured data

Structured and Unstructured Data

Structured Data

- Structured data refers to any data that resides in a fixed field within a record or file
- Example:
  - Data present in relational databases and spreadsheets

Date	Time	Number	Direction
1/22/2008	16:38:53	605-996-1003	Incoming
1/22/2008	16:40:44	605-996-1003	Incoming
1/22/2008	16:42:40	605-996-1003	Incoming
1/23/2008	08:13:55	605-996-1003	Outgoing
1/23/2008	08:14:31	605-996-1003	Incoming
1/23/2008	08:14:31	605-996-6244	Outgoing
1/23/2008	08:36:02	605-996-1003	Incoming

Unstructured Data

- Data that does not have a recognizable structure.
- Examples:
  - Audio
  - Video
  - Images
  - email

Semi-Structured Data

- It is the data that is not organized in a rational model
- It contains tags or other markers to separate semantic elements

  Example:

  - JSON files
  - XML files

```
<?xml version="1.0"?>
- <players>
  - <user>
      <username> sam </username>
      <password> testing</password>
    </user>
  - <user>
      <username> bob </username>
      <password> yop </password>
    </user>
  - <user>
      <username>Noscoper</username>
      <password>palp</password>
    </user>
  - <user>
      <username>test</username>
      <password>test</password>
    </user>
</players>
```

Raw Data

- Raw data cannot be used in analysis directly
- Structural and semantic errors will be present in raw data
- So for analysis it is important to clean the data

Some examples of raw data are,

- Binary files from sensors
- Web scrapped data
- Log files
- Videos/ Images etc

Data collected and extracted from different sources are "Raw data".

Reading CSV File

- For reading data from CSV file, we have to import pandas library

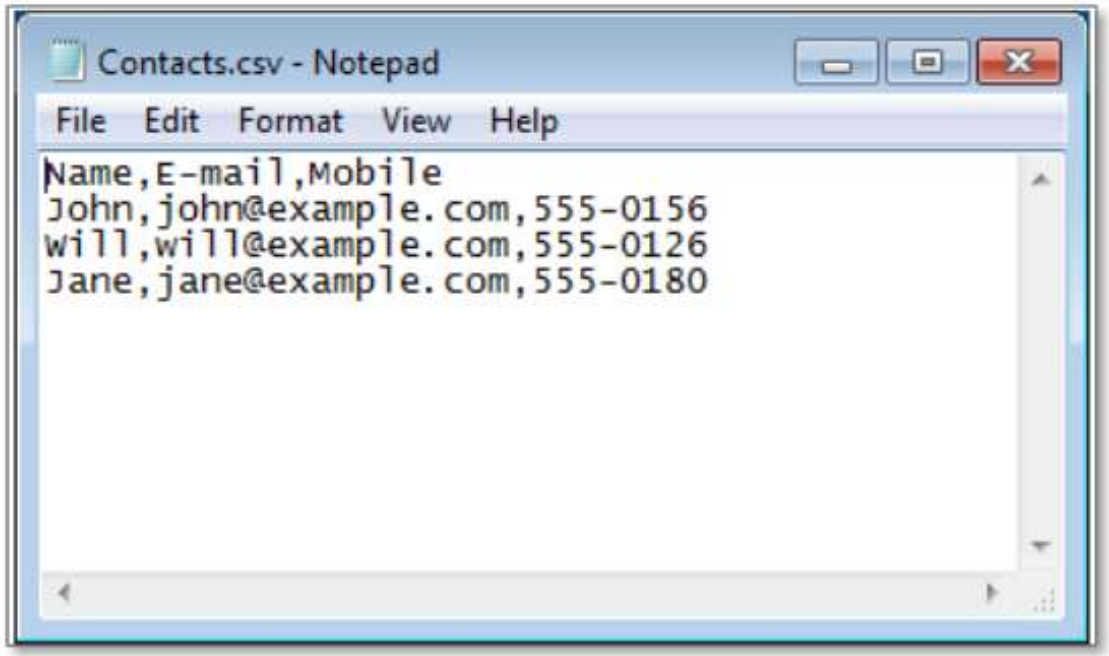
```
import pandas as pd

df=pd.read_csv("iris.csv")
```

- If the file is present in another directory then you can write file's path also

```
import pandas as pd

df=pd.read_csv("C:\Users\adhyapakss\PycharmProjects\
                ML_m3\iris.csv")
```



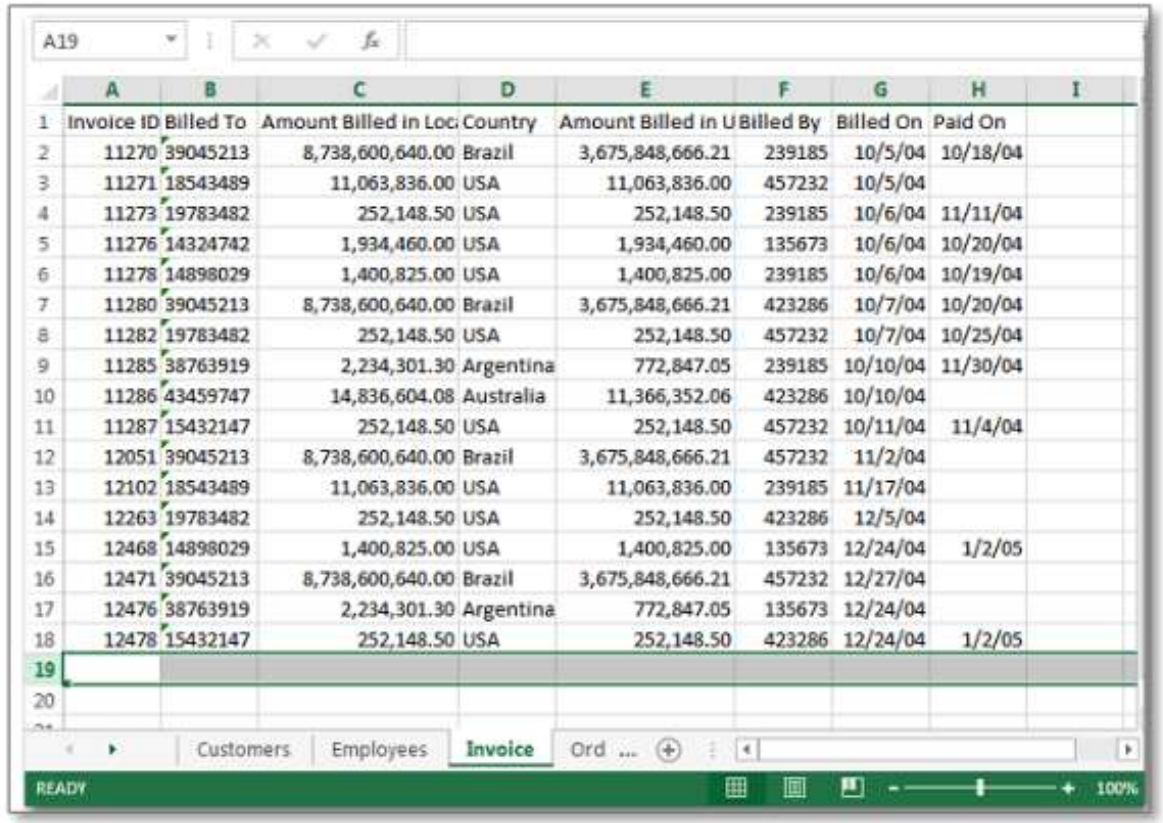
Reading Excel File

- For reading data from excel file, we have to import pandas library

```
import pandas as pd
df=pd.read_excel("Employee.xlsx")
```

- If the file is present in another directory then you can write file's path also

```
import pandas as pd
df=pd.read_excel("C:\Users\adhyapakss\Desktop\DS master's Program\
                  Python\Certification Project\Employee.xlsx")
```



A19

	A	B	C	D	E	F	G	H	I
1	Invoice ID	Billed To	Amount Billed in Loc	Country	Amount Billed in U	Billed By	Billed On	Paid On	
2	11270	39045213	8,738,600,640.00	Brazil	3,675,848,666.21	239185	10/5/04	10/18/04	
3	11271	18543489	11,063,836.00	USA	11,063,836.00	457232	10/5/04		
4	11273	19783482	252,148.50	USA	252,148.50	239185	10/6/04	11/11/04	
5	11276	14324742	1,934,460.00	USA	1,934,460.00	135673	10/6/04	10/20/04	
6	11278	14898029	1,400,825.00	USA	1,400,825.00	239185	10/6/04	10/19/04	
7	11280	39045213	8,738,600,640.00	Brazil	3,675,848,666.21	423286	10/7/04	10/20/04	
8	11282	19783482	252,148.50	USA	252,148.50	457232	10/7/04	10/25/04	
9	11285	38763919	2,234,301.30	Argentina	772,847.05	239185	10/10/04	11/30/04	
10	11286	43459747	14,836,604.08	Australia	11,366,352.06	423286	10/10/04		
11	11287	15432147	252,148.50	USA	252,148.50	457232	10/11/04	11/4/04	
12	12051	39045213	8,738,600,640.00	Brazil	3,675,848,666.21	457232	11/2/04		
13	12102	18543489	11,063,836.00	USA	11,063,836.00	239185	11/17/04		
14	12263	19783482	252,148.50	USA	252,148.50	423286	12/5/04		
15	12468	14898029	1,400,825.00	USA	1,400,825.00	135673	12/24/04	1/2/05	
16	12471	39045213	8,738,600,640.00	Brazil	3,675,848,666.21	457232	12/27/04		
17	12476	38763919	2,234,301.30	Argentina	772,847.05	135673	12/24/04		
18	12478	15432147	252,148.50	USA	252,148.50	423286	12/24/04	1/2/05	
19									
20									

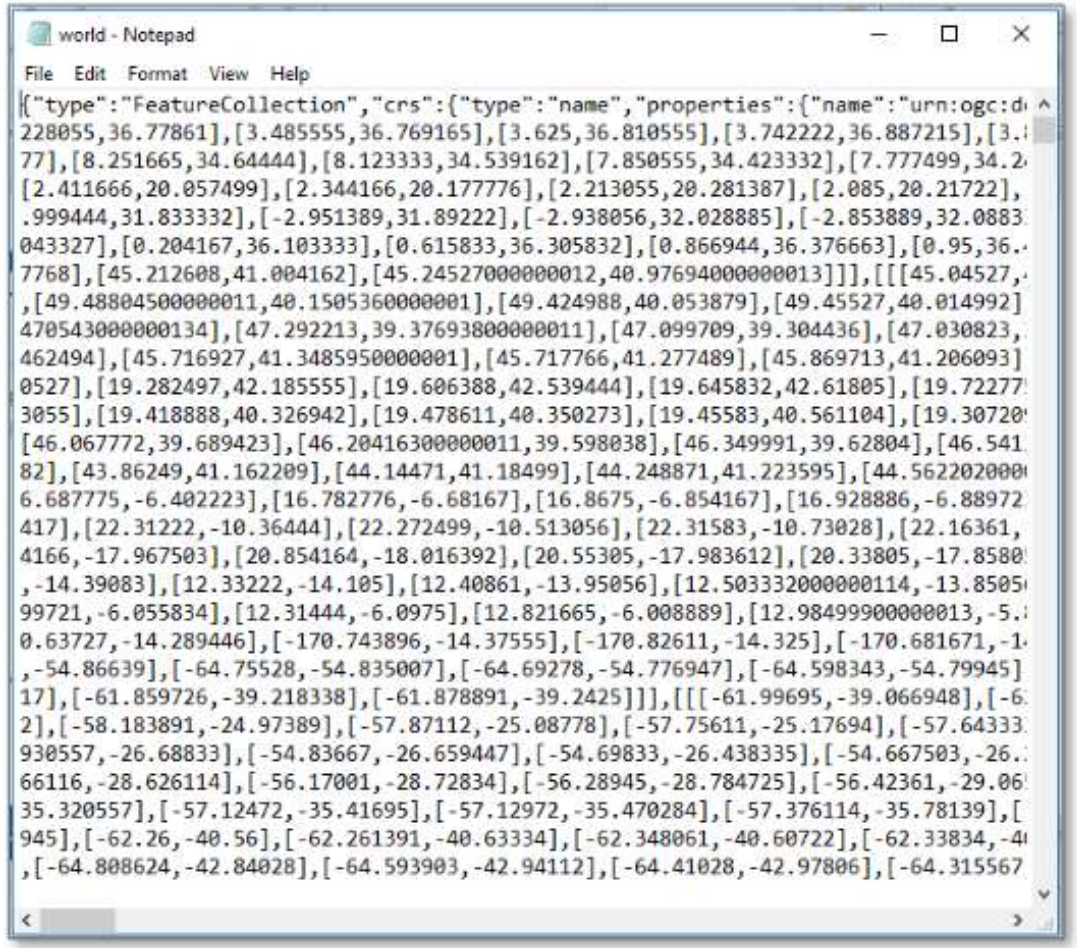
Customers | Employees | Invoice | Ord ...

READY 100%

Reading JSON File

- For reading data from json file, we have to import pandas library

```
import pandas as pd
jsonl=pd.read_json("world.json")
```



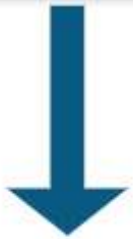
Data Wrangling

- Data Wrangling is the process of cleaning and unifying messy and complex data sets
- Data Scientists spend 50-80% of their time stuck in the mundane labor of collecting and organizing data before it can be utilized

```
John 19/01/1994 New York Paris Michelle 29/11/1993 21/09/1995 stephen London oliver bangkook 11/04/93
```

Raw data

Its unstructured and contains errors



	A	B	C	D
1		Name	DOB	Hometown
2	1	John	19/01/1994	New York
3	2	Michelle	29/11/1993	Paris
4	3	Stephen	21/09/1995	London
5	4	Oliver	11/4/1993	Bangkok

Processed Data

Techniques used in Data Wrangling

1. Sampling data
2. Filtering data
3. Removing Null values
4. Text manipulation

Data Wrangling – Sampling Data

- Sampling data allows data scientists to work with a small, manageable amount of data in order to build and run analytical models more quickly, while still producing accurate findings
- Sampling can be particularly useful with data sets that are too large to efficiently analyze in full

```
import pandas as pd
df=pd.read_csv("Movie_metadata.csv")
result=df[(df['color']=='color')]
result.to_csv("Movie_metadata.csv")
```

	color	director_name	num_critic_for_reviews	duration	director_facebook_likes
1	Color	James Cameron	723	178	0
2	Color	Gore Verbinski	302	169	563
3	Color	Sam Mendes	602	148	0
4	Color	Christopher Nolan	813	164	22000
6	Color	Andrew Stanton	462	132	475
7	Color	Sam Raimi	392	156	0
8	Color	Nathan Greno	324	100	15
9	Color	Joss Whedon	635	141	0
10	Color	David Yates	375	153	282
11	Color	Zack Snyder	673	183	0
12	Color	Bryan Singer	434	169	0
13	Color	Marc Forster	403	106	395
14	Color	Gore Verbinski	313	151	563
15	Color	Gore Verbinski	450	150	563
16	Color	Zack Snyder	733	143	0
17	Color	Andrew Adamson	258	150	80
18	Color	Joss Whedon	703	173	0
19	Color	Rob Marshall	448	136	252
20	Color	Barry Sonnenfeld	451	106	188

Data Wrangling – Filtering Data

- Filtering data refers to a wide range of strategies or solutions for refining data sets
- This means the data sets are refined into simply what a user (or set of users) needs, without including other data that can be repetitive, irrelevant or even sensitive

	color	director_name	num_critic_for_reviews	duration	director_facebook_likes
1	Color	James Cameron	723	178	0
2	Color	Gore Verbinski	302	169	563
3	Color	Sam Mendes	602	148	0
4	Color	Christopher Nolan	813	164	22000
6	Color	Andrew Stanton	462	132	475
7	Color	Sam Raimi	392	156	0
8	Color	Nathan Greno	324	100	15
9	Color	Joss Whedon	635	141	0
10	Color	David Yates	375	153	282
11	Color	Zack Snyder	673	183	0
12	Color	Bryan Singer	434	169	0
13	Color	Marc Forster	403	106	395
14	Color	Gore Verbinski	313	151	563
15	Color	Gore Verbinski	450	150	563
16	Color	Zack Snyder	733	143	0
17	Color	Andrew Adamson	258	150	80
18	Color	Joss Whedon	703	173	0
19	Color	Rob Marshall	448	136	252
20	Color	Barry Sonnenfeld	451	106	188

Data Wrangling – Removing NULL Values

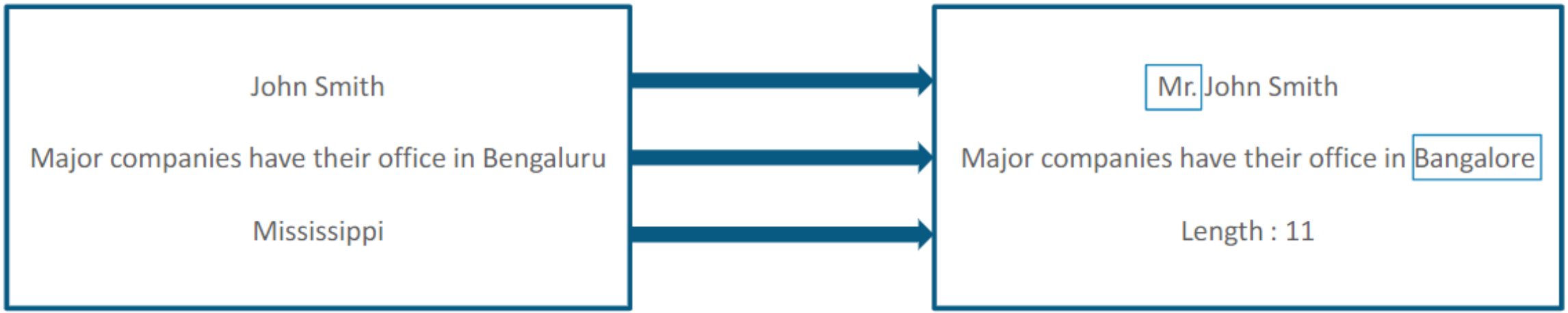
- Our dataset contains some Null values represented by 'NA' as shown in the screen shot

```
import pandas as pd
df=pd.read_csv("Movie_metadata.csv")
pd.isnull(df).any()
```

color	director_name	num_critic_for_reviews	duration	director_facebook_likes	actor_3_facebook_likes	actor
Color	James Cameron	723	178	0	855	Joel Da
Color	Gore Verbinski	302	169	563	1000	Orland
Color	Sam Mendes	602	148	0	161	Rory K
Color	Christopher Nolan	813	164	22000	23000	Christi
	Doug Walker	NA	NA	131	NA	Rob W
Color	Andrew Stanton	462	132	475	530	Saman
Color	Sam Raimi	392	156	0	4000	James
Color	Nathan Greno	324	100	15	284	Donna
Color	Joss Whedon	635	141	0	19000	Robert
Color	David Yates	375	153	282	10000	Daniel
Color	Zack Snyder	673	183	0	2000	Lauren
Color	Bryan Singer	434	169	0	903	Marlon
Color	Marc Forster	403	106	395	393	Mathie
Color	Gore Verbinski	313	151	563	1000	Orland
Color	Gore Verbinski	450	150	563	1000	Ruth W
Color	Zack Snyder	733	143	0	748	Christ
Color	Andrew Adamson	258	150	80	201	Pierfra
Color	Joss Whedon	703	173	0	19000	Robert
Color	Rob Marshall	448	136	252	1000	Sam C
Color	Barry Sonnenfeld	451	106	188	718	Micha

Data Wrangling – Text Manipulation

- Text manipulation can be defined as a range of activities to change a previously written text
- Some of the fundamental operations of text manipulation include: adding prefix/suffix into lines, extraction of string segments, string matching, their comparison, discovering their length and replacing substrings by other strings



Demo: Import Pandas Library to Read and Slice the data.

Demo: Understand the data you are dealing with.

Demo: Handling Missing Values

So far in Python….

- Import Library
- Read CSV and TSV files
- Understanding the Data
- Replace Missing Values

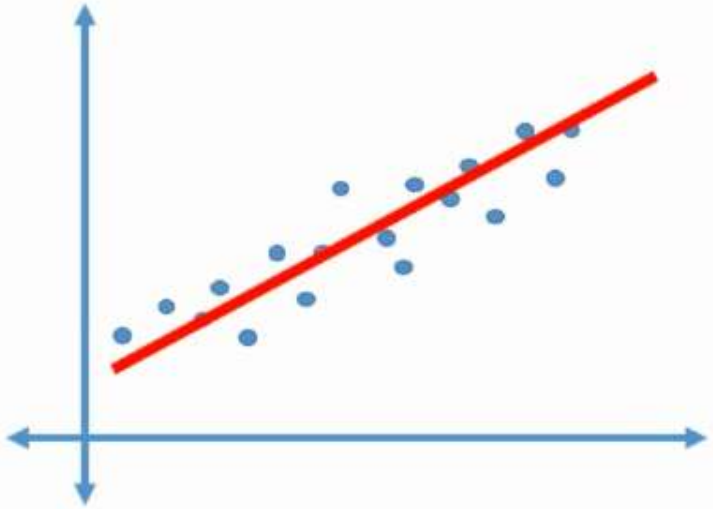
```
# Drop the rows with missing values
cleandata = dataset.dropna()
cleandata = dataset.dropna(subset=['Loan_Stat[redacted])

# Replace with Mode for categorical
dt = dataset.copy()
cols = ['Gender', 'Area', 'Loan_Status']
dt[cols] = dt[cols].fillna(dt.mode().iloc[0])
dt.isnull().sum()

# Replace with mean for numerical
cols2 = ['ApplicantIncome', 'CoapplicantIncome', 'LoanAmount']
dt[cols2] = dt[cols2].fillna(dt.mean())
dt.isnull().sum()
```

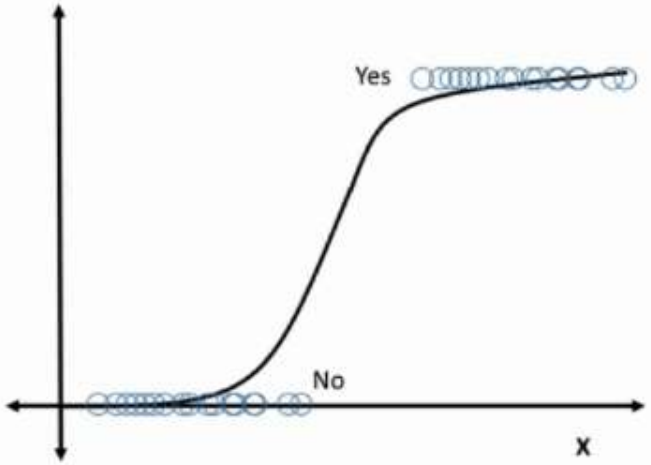
Mathematics as basis of Machine Learning

Regression



$$y = a + b * x$$

Classification

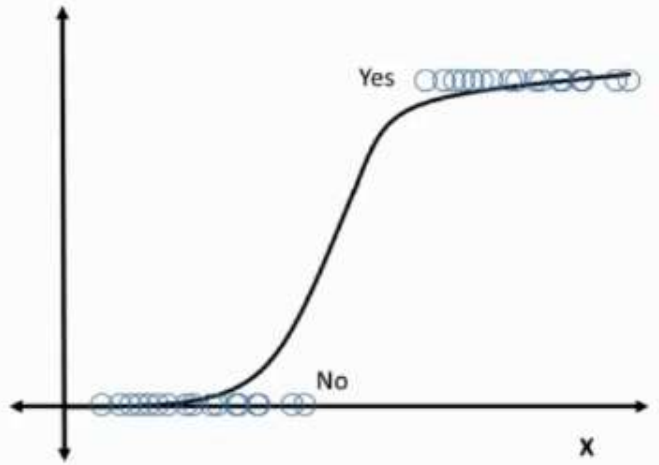


$$\log\left(\frac{P}{1-P}\right) = b_0 + b_1 x$$

Categorical Variables

Loan_ID	Gender	ApplicantIncome	oapplicantIncom	LoanAmount	Area	Loan_Status
LP001002	Male	5849.00	0.00	nan	urban	Y
LP001003	Male	4583.00	nan	128.00	semi	N
LP001005	Male	3000.00	0.00	66.00	semi	Y
LP001006	Female	2583.00	2358.00	120.00	semi	Y
LP001008	Male	nan	0.00	141.00	urban	Y
LP001011	Male	5417.00	4196.00	267.00	semi	Y
LP001013	Male	2333.00	1516.00	nan	semi	Y
LP001014	Female	3036.00	2504.00	158.00	semi	N
LP001018	Male	4006.00	1526.00	168.00	semi	Y
LP001020	Male	12841.00	10968.00	349.00	semi	N
LP001024	Female	3200.00	700.00	70.00	urban	Y
LP001027	Male	2500.00	1840.00	109.00	urban	Y
LP001028	Female	nan	8106.00	nan	urban	Y
LP001029	Male	1853.00	2840.00	114.00	urban	N
LP001030	Male	1299.00	1086.00	17.00	semi	Y
LP001032	Male	4950.00	0.00	125.00	semi	Y

Classification



$$\log\left(\frac{P}{1-P}\right) = b_0 + b_1 x$$

Categorical Variables

Loan_ID	Gender	ApplicantIncome	oapplicantIncom	LoanAmount	Area	Loan_Status
LP001002	Male	5849.00	0.00	nan	urban	Y
LP001003	Male	4583.00	nan	128.00	semi	N
LP001005	Male	00.00	0.00	66.00	semi	Y
LP001006	Female	2 00	2358.00	120.00	semi	Y
LP001008	Male	nan	0.00	141.00	urban	Y
LP001011	Male	5417	4196.00	267.00	semi	Y
LP001013	Male	2		n	semi	Y
LP001014	Female	3		8.00	semi	N
LP001018	Male	4		8.00	semi	Y
LP001020	Male	12841.00	10968.00	349.00	semi	N
LP001024	Female	3200.00	700.00	70.00	urban	Y
LP001027	Male	2500.00	1840.00	109.00	urban	Y
LP001028	Female	nan	8106.00	nan	urban	Y
LP001029	Male	1853.00	2840.00	114.00	urban	N
LP001030	Male	1299.00	1086.00	17.00	semi	Y
LP001032	Male	4950.00	0.00	125.00	semi	Y

Male → 0
Female → 1

Y → 0
N → 1

Urban → 0
Semi → 1
Rural → 2

Demo: Label Encoding

LabelEncoder

```
22
23 # Categorical to Numeric label encoding using Pandas
24 dt.dtypes
25
26 dt[cols] = dt[cols].astype('category')
27 dt.dtypes
28
29 for columns in cols:
30     dt[columns] = dt[columns].cat.codes
```

LP001002	1	5849.00	0.00	140.92	1	1
LP001003	1	4583.00	2509.33	128.00	0	0
LP001005	1	3000.00	0.00	66.00	0	1
LP001006	0	2583.00	2358.00	120.00	0	1
LP001008	1	4103.57	0.00	141.00	1	1
LP001011	1	5417.00	4196.00	267.00	0	1
LP001013	1	2333.00	1516.00	140.92	0	1
LP001014	0	3036.00	2504.00	158.00	0	0
LP001018	1	4006.00	1526.00	168.00	0	1
LP001020	1	12841.00	10968.00	349.00	0	0
LP001024	0	3200.00	700.00	70.00	1	1
LP001027	1	2500.00	1840.00	109.00	1	1
LP001028	0	4103.57	8106.00	140.92	1	1
LP001029	1	1853.00	2840.00	114.00	1	0
LP001030	1	1299.00	1086.00	17.00	0	1
LP001032	1	4950.00	0.00	125.00	0	1

Problem with LabelEncoder ONLY

Area	LabelEncoder
Urban	1
Semi-Urban	2
Rural	3

3 > 2 > 1 → Rural > Semi-Urban > Urban

City	LabelEncoder
London	1
New York	2
Delhi	3

1 + 2 = 3 → London + New York = Delhi

(1 + 3)/2 = 2 → (London + Delhi)/2 = New York

One-Hot Encoder

City	LabelEncoder	London	New York	Delhi
London	1	1	0	0
New York	2	0	1	0
Delhi	3	0	0	1

Demo: Hot-Encoding for Categorical Data

What is Normalization?

"In the simplest cases, normalization of ratings means adjusting values measured on different scales to a notionally common scale, often prior to averaging"

--Wikipedia

What is Normalization?

City	Temperature
New York	92 °F
Chicago	87 °F
Boston	94 °F
Detroit	91 °F

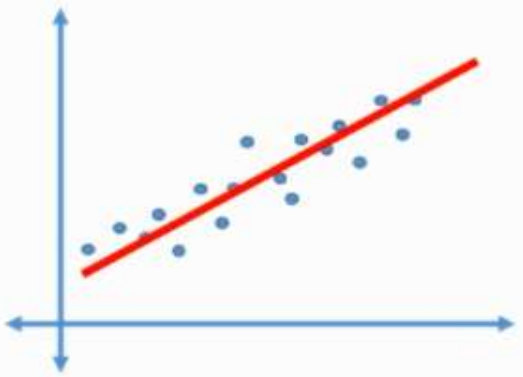
City	Temperature
London	28 °C
Paris	24 °C
Delhi	34 °C
Tokyo	31 °C

City	Temperature
New York	92 °F
Chicago	87 °F
Boston	94 °F
Detroit	91 °F
London	82.4 °F
Paris	75.2 °F
Delhi	93.2 °F
Tokyo	87.8 °F

City	Temperature
New York	33.3 °C
Chicago	30.5 °C
Boston	34.4 °C
Detroit	32.8 °C
London	28 °C
Paris	24 °C
Delhi	34 °C
Tokyo	31 °C

Why should we normalize the data?

Regression



$$y = a + b * x_1 + c * x_2$$

LoanTerm	Salary
12	$ 68,000
60	$ 85,000
120	$ 115,000

Normalization Defined Statistically

- A method to standardise the range of independent variables or features of data
- Variables are fitted within a certain range (Generally between 0 and 1)
- Applied on numeric columns

Normalize data – Transformation Methods

ZScore

$$Z = \frac{X - mean(x)}{stdev\ (x)}$$

MinMax

$$Z = \frac{X - min(x)}{Max\ (x) - min(x)}$$

Logistic

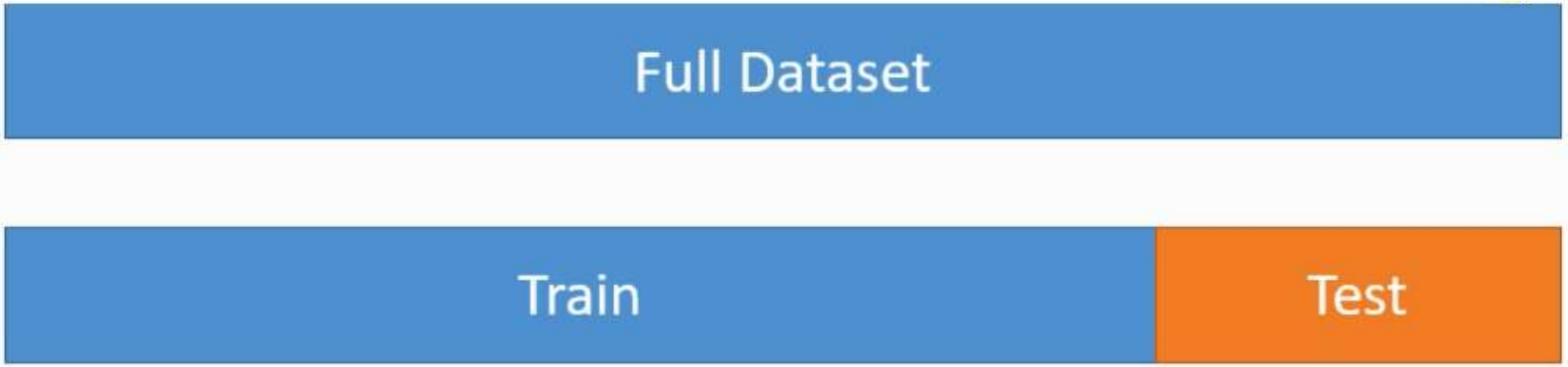
$$Z = \frac{1}{1 + exp(-x)}$$

Most commonly used transformation methods

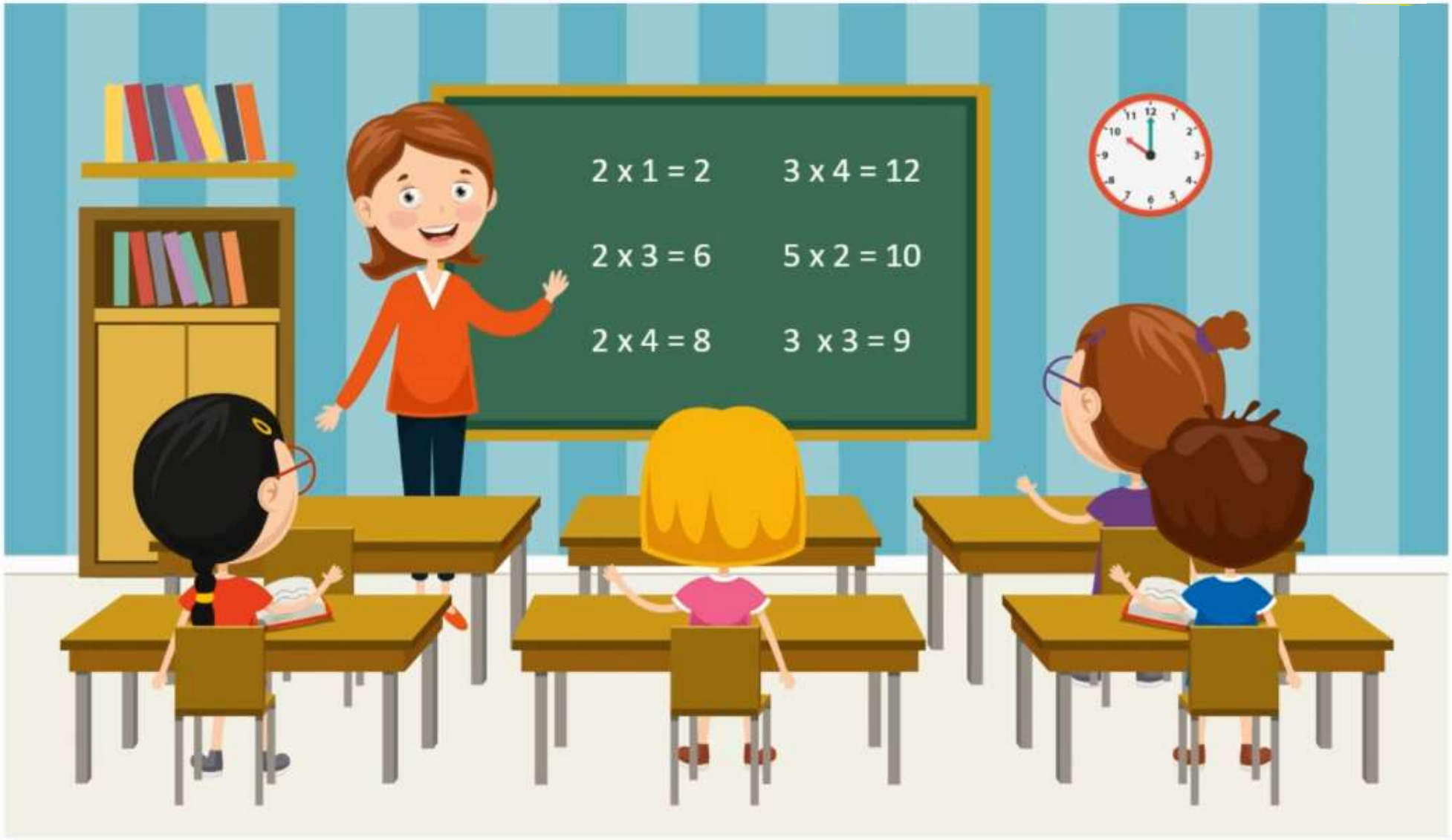
Demo: Data Normalization using Standard Scaler z-Transformation

Demo: Data Normalization using minmax

Train and Test Data



Going back to school



Going back to school

2 x 1 = ? 3 x 4 = ?

2 x 3 = ? 5 x 2 = ?

2 x 4 = ? 3 x 3 = ?

Going back to school

4 x 1 = ? 2 x 6 = ?

4 x 5 = ? 4 x 5 = ?

3 x 7 = ? 3 x 8 = ?

High Test Accuracy

Going back to school

4 x 1 = ? 2 x 6 = ?

4 x 5 = ? 4 x 5 = ?

3 x 7 = ? 3 x 8 = ?

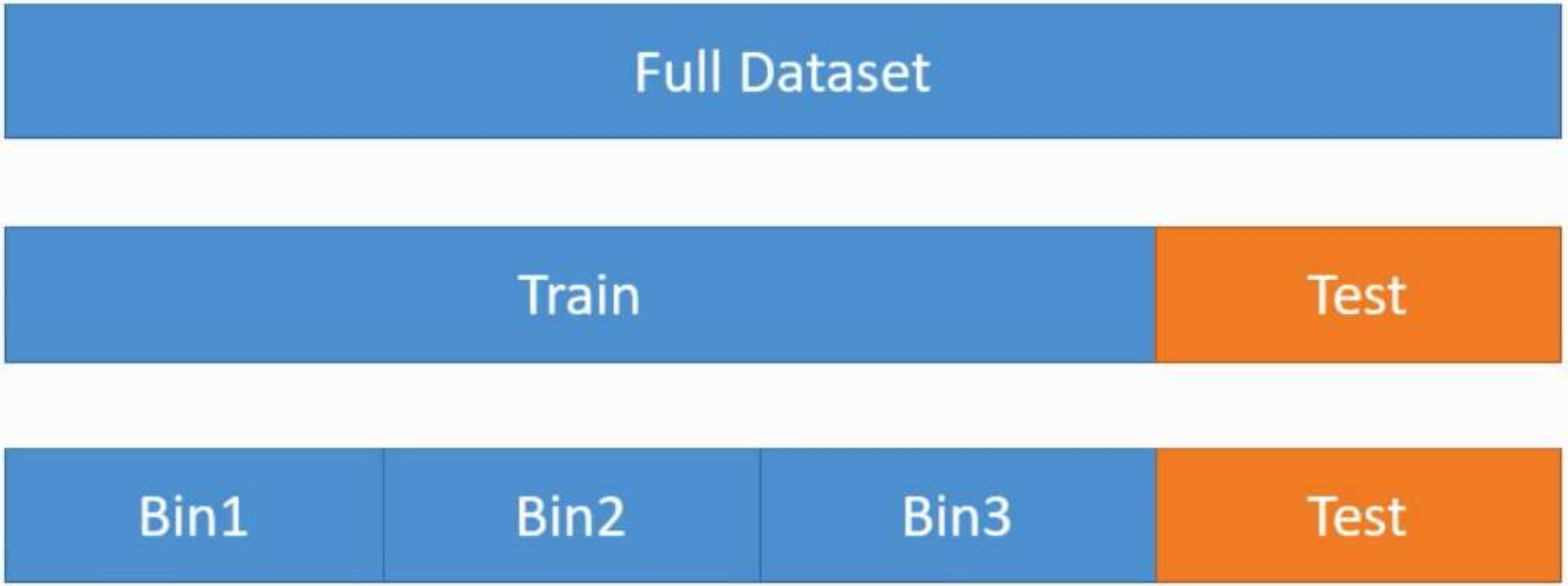
Low Test Accuracy

Going back to school

High Test Accuracy

Low Test Accuracy

Train and Test Data



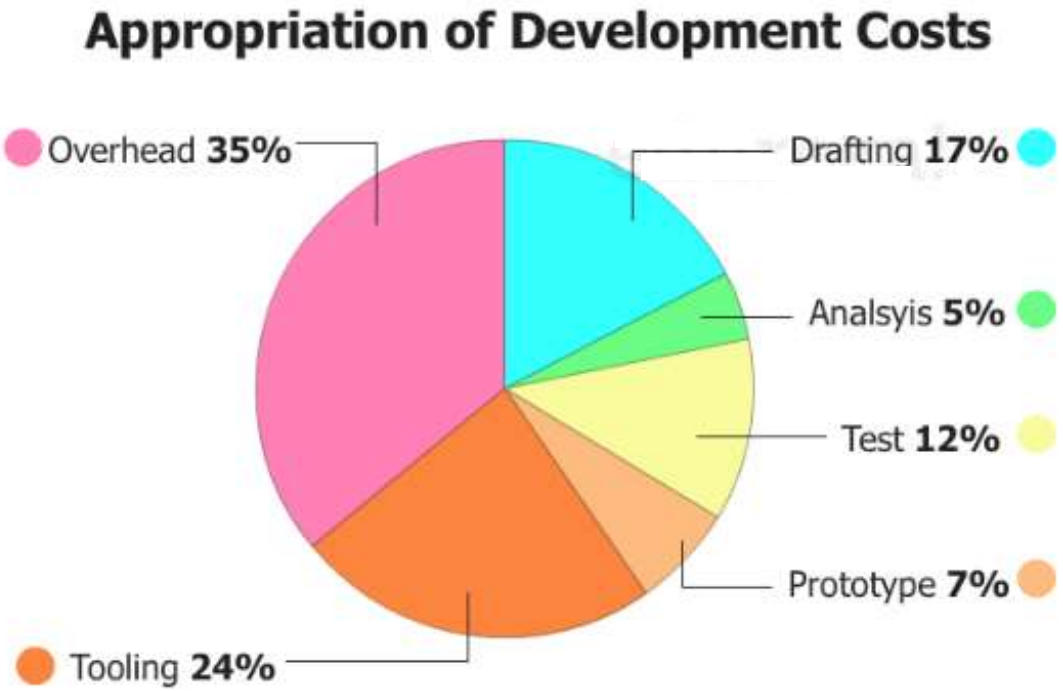
Demo: Train and Test Data Split

DATA VISUALIZATION

Data Visualisation

- Data visualization is the presentation of data in a pictorial or graphical format
- It enables decision makers to see patterns, trends and correlations that might go undetected in text-based data

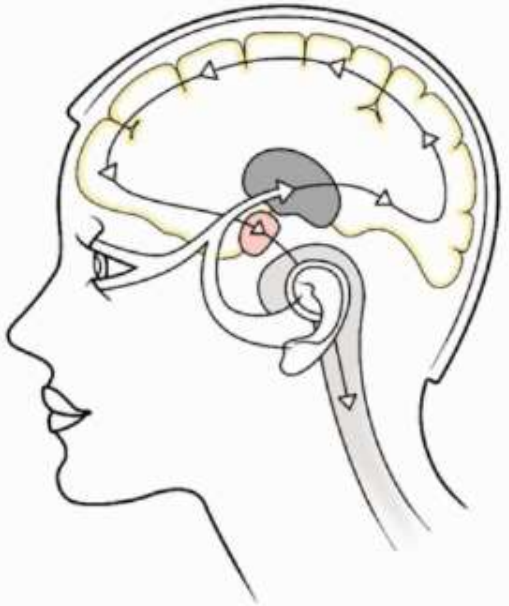
Costs	Total Development Costs ($M)
Drafting	58.2
Analsyis	15.2
Test	38.8
Prototyping	22.6
Tooling	79.5
Overhead	120.5



Why to visualise the data?

Why to visualise the data?

MIT neuroscientists find the brain can identify images seen for as little **as 13 milliseconds**.



Picture is worth thousand words

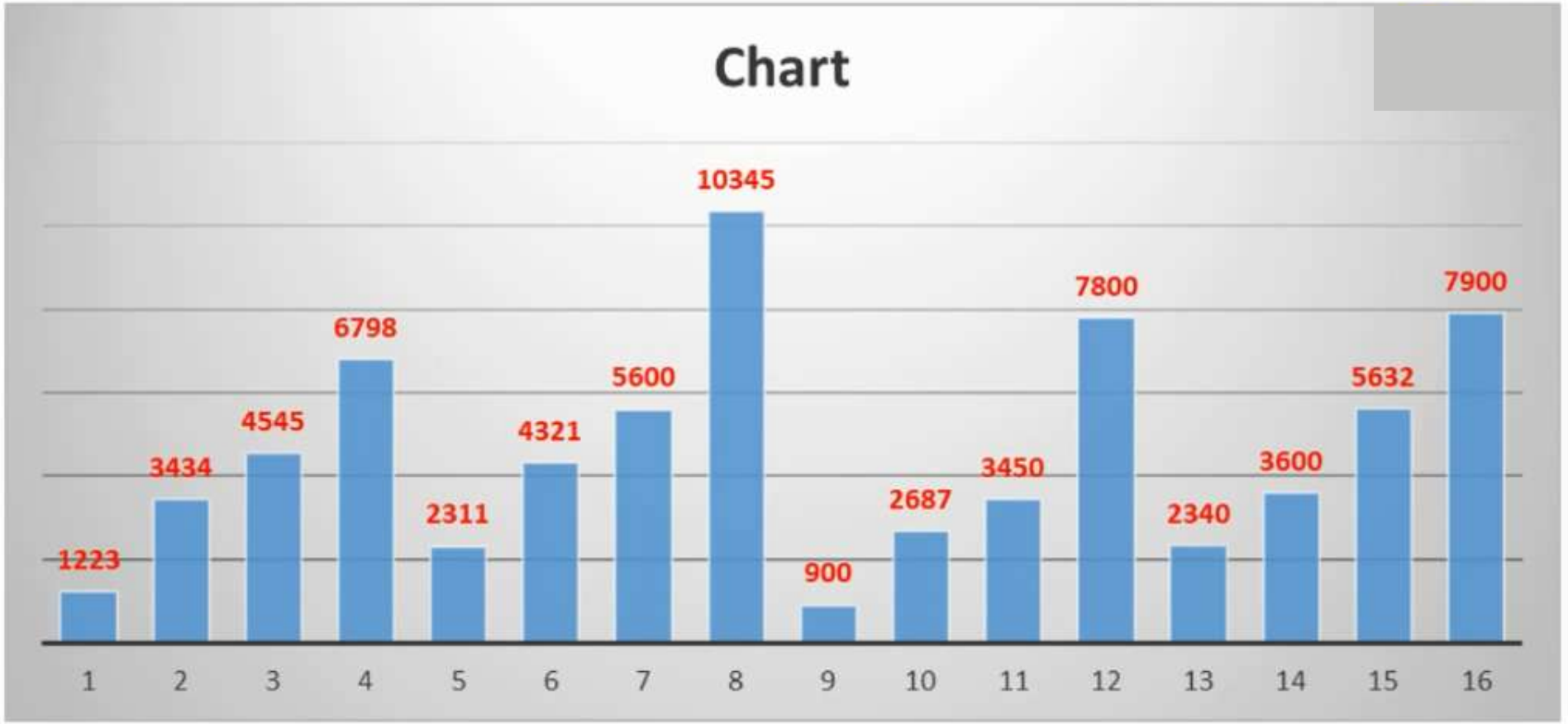
1223
3434
4545
6798
2311
4321
5600
10345
900
2687
3450
6700
2340
3600
5632
7900

Tell me three things about this data,

1. What's the maximum value?
2. Does this data follow a trend?
3. If there is a trend, describe it?

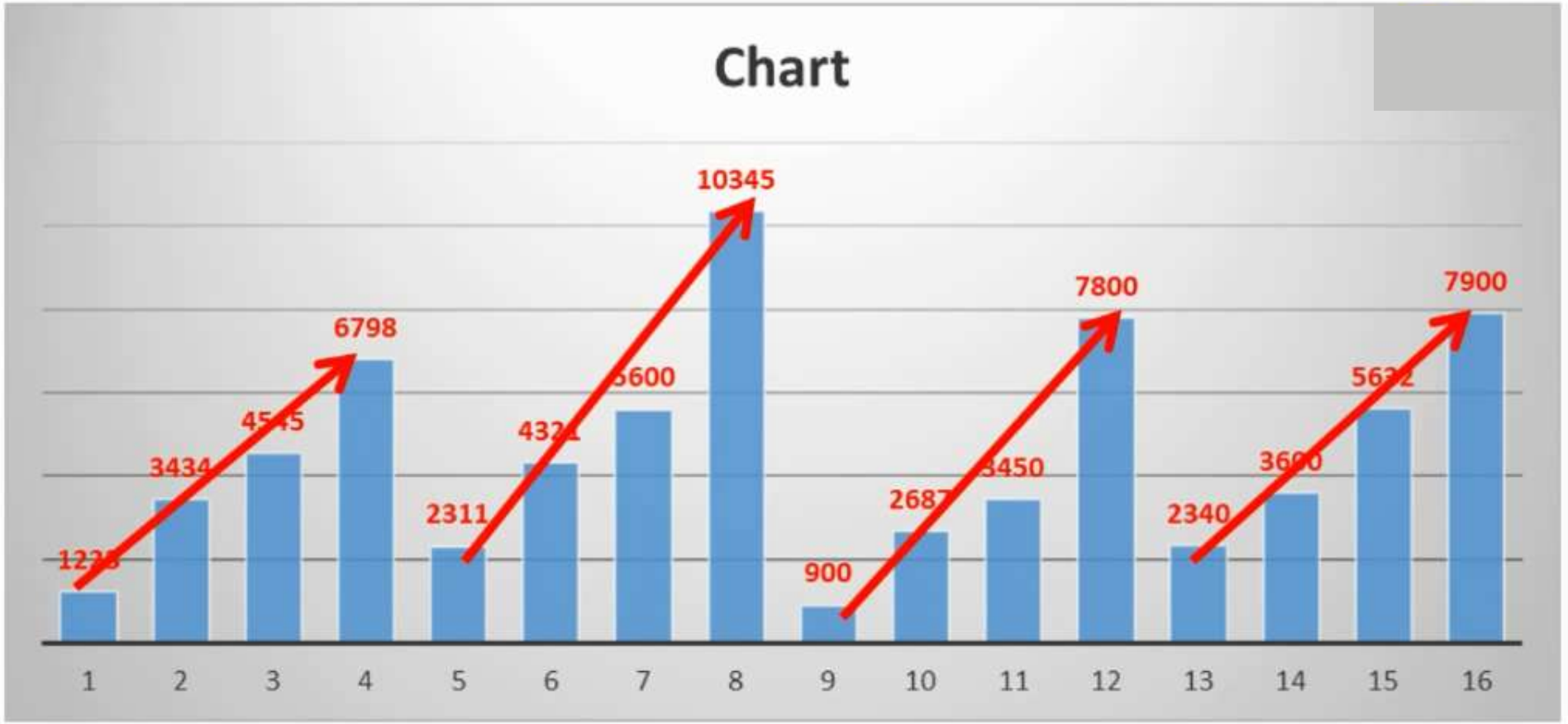
Picture is worth thousand words

1223
3434
4545
6798
2311
4321
5600
10345
900
2687
3450
6700
2340
3600
5632
7900



Picture is worth thousand words

1223
3434
4545
6798
2311
4321
5600
10345
900
2687
3450
6700
2340
3600
5632
7900



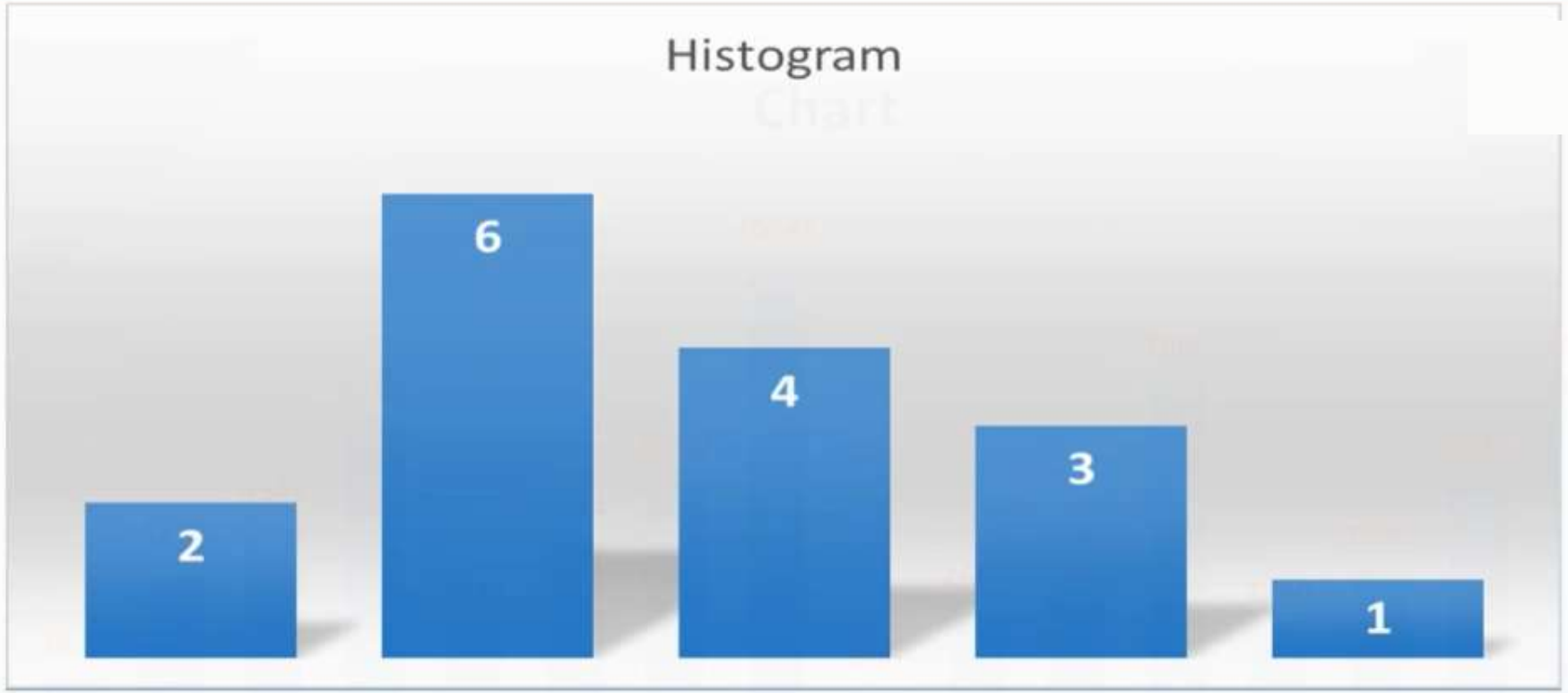
Frequency Table

1223
3434
4545
6798
2311
4321
5600
10345
900
2687
3450
6700
2340
3600
5632
7900

0-2000	2-4000	4-6000	6-8000	> 8000
1223	3434	4545	6798	10345
900	2311	4321	6700	
	2687	5600	7900	
	3450	5632		
	2340			
	3600			
2	**6**	**4**	**3**	**1**

Histogram

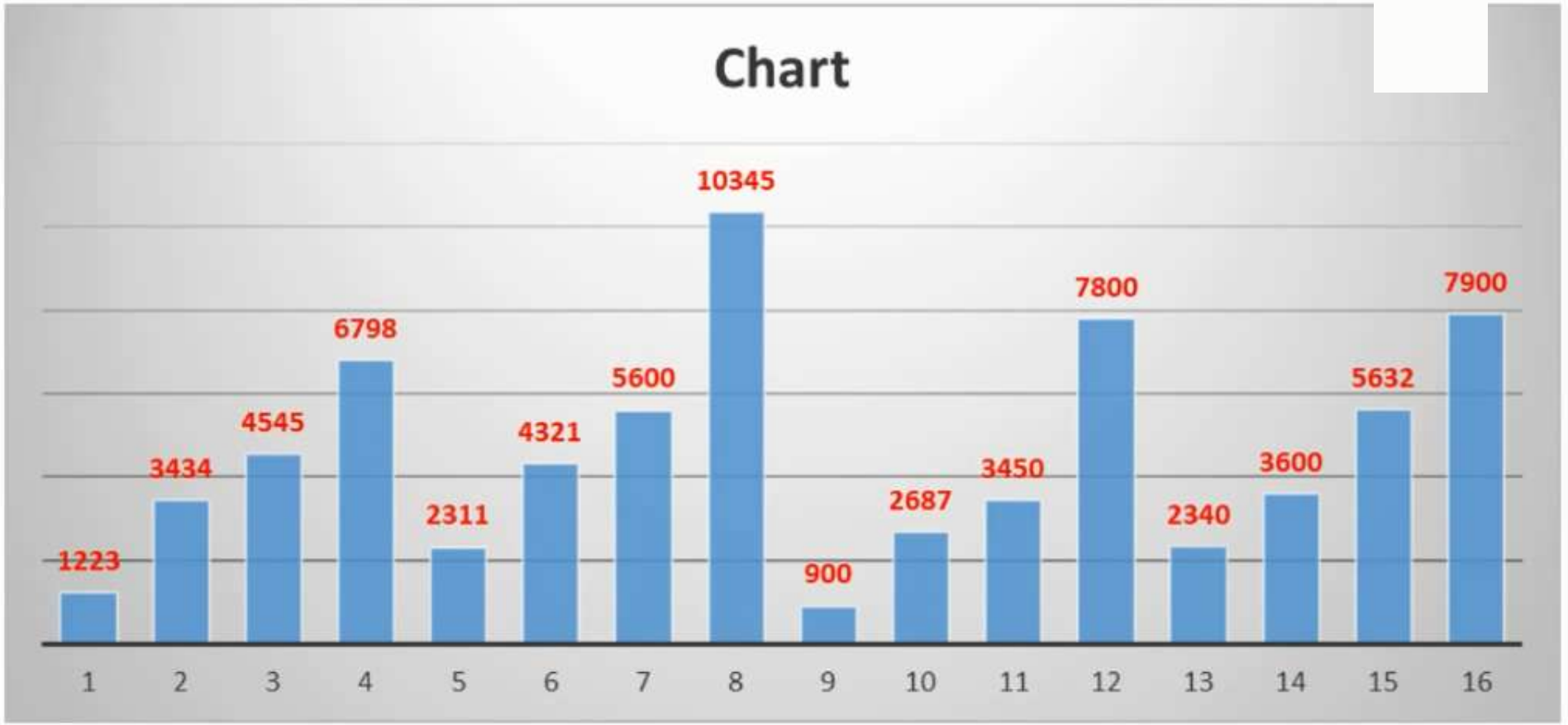
1223
3434
4545
6798
2311
4321
5600
10345
900
2687
3450
6700
2340
3600
5632
7900



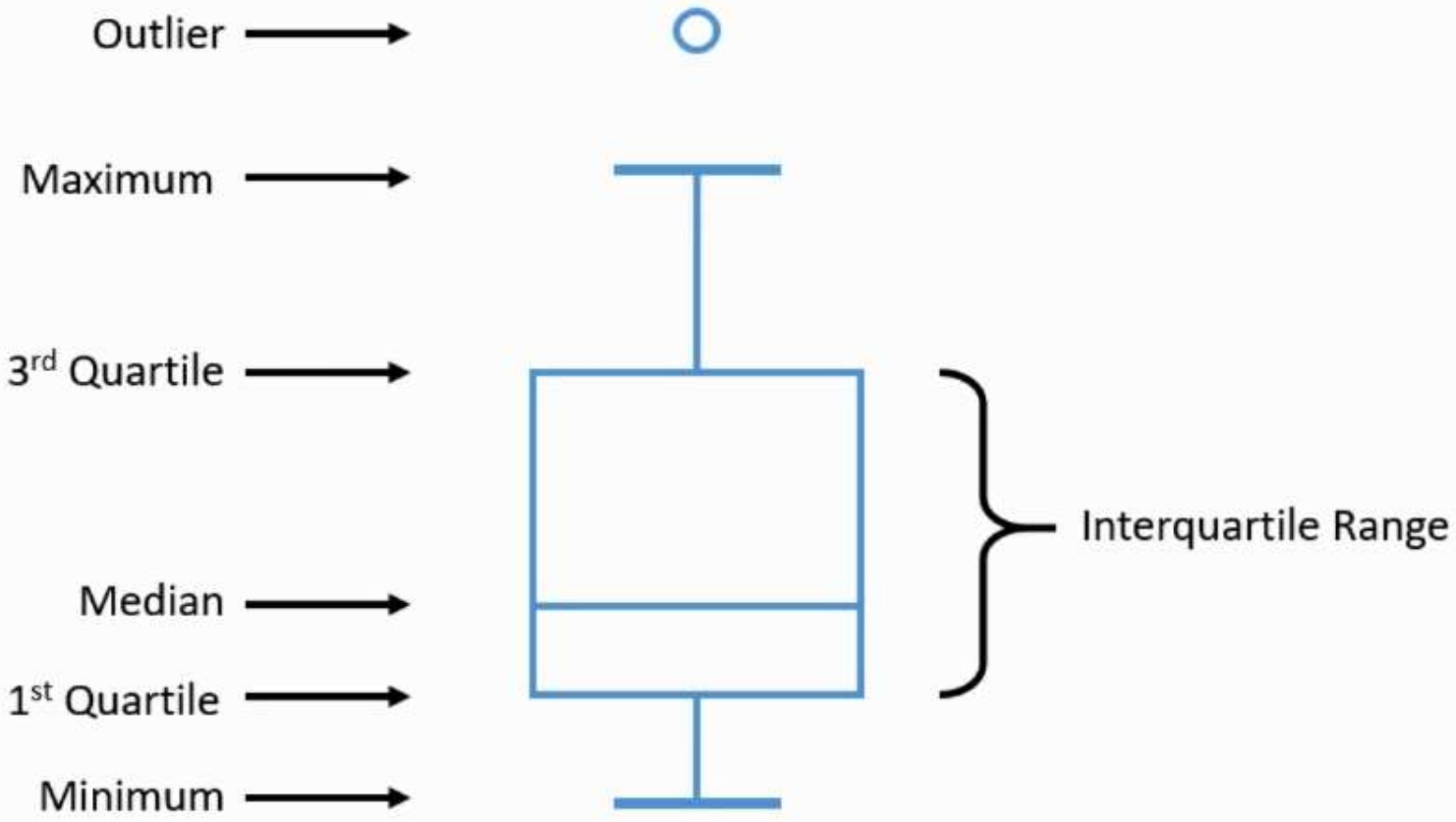
0-2000	2-4000	4-6000	6-8000	> 8000
2	6	4	3	1

Bar Chart

1223
3434
4545
6798
2311
4321
5600
10345
900
2687
3450
6700
2340
3600
5632
7900



Box Plot


Outlier
Maximum
3rd Quartile
Median
1st Quartile
Minimum
Interquartile Range

Quartiles

What are the quartiles?

The **first quartile**, or **25th percentile x_L** (also written as **Q_1**), is the number for which 25% of values in the data set are smaller than x_L.

The **second quartile** or **50th percentile**, **x_m** (also written as **Q_2**) is also known as the median. It represents the value for which 50% of observations are lower and 50% are higher.

The **third quartile** or **75th percentile**, **x_H** (**Q_3**) is the value such that 75% of the observations are less than x_H

Inter Quartile Range (IQR)

1st Quartile

Median

3rd Quartile

Row Number	Salary
1	$ 3,725
2	$ 4,155
3	$ 4,627
4	$ 5,147
5	$ 5,718
6	$ 6,347
7	$ 7,039
8	$ 7,210
9	$ 7,423
10	$ 7,556
11	$ 8,369
12	$ 8,810
13	$ 8,940
14	$ 9,200
15	$ 9,458

Q3 – Q1
Inter Quartile Range
IQR

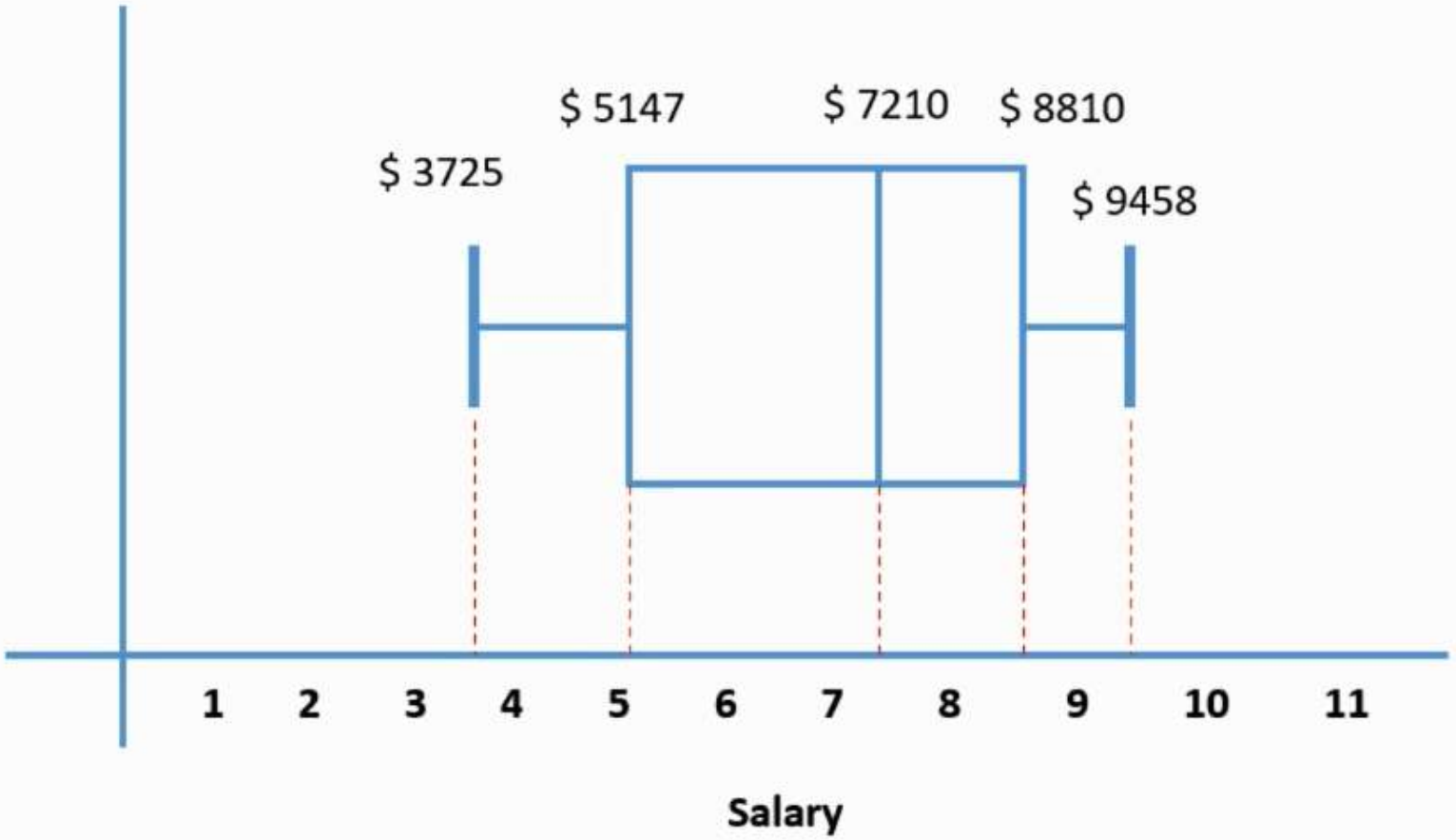
$8,810 – $5,147 = $3,663

Box Plot

Salary
$ 3,725
$ 4,155
$ 4,627
$ 5,147
$ 5,718
$ 6,347
$ 7,039
$ 7,210
$ 7,423
$ 7,556
$ 8,369
$ 8,810
$ 8,940
$ 9,200
$ 9,458

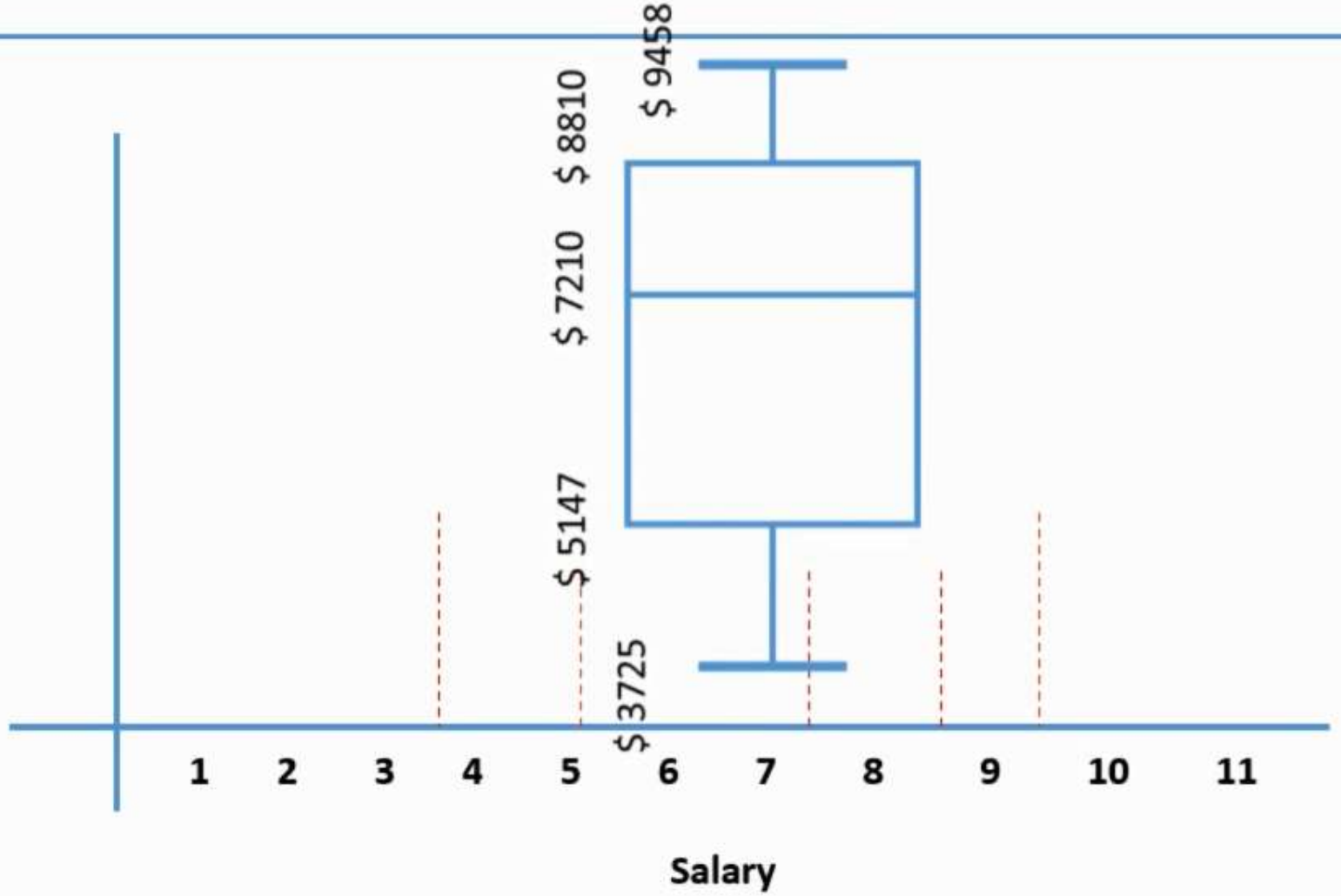
Box Plot

Salary
$ 3,725
$ 4,155
$ 4,627
$ 5,147
$ 5,718
$ 6,347
$ 7,039
$ 7,210
$ 7,423
$ 7,556
$ 8,369
$ 8,810
$ 8,940
$ 9,200
$ 9,458

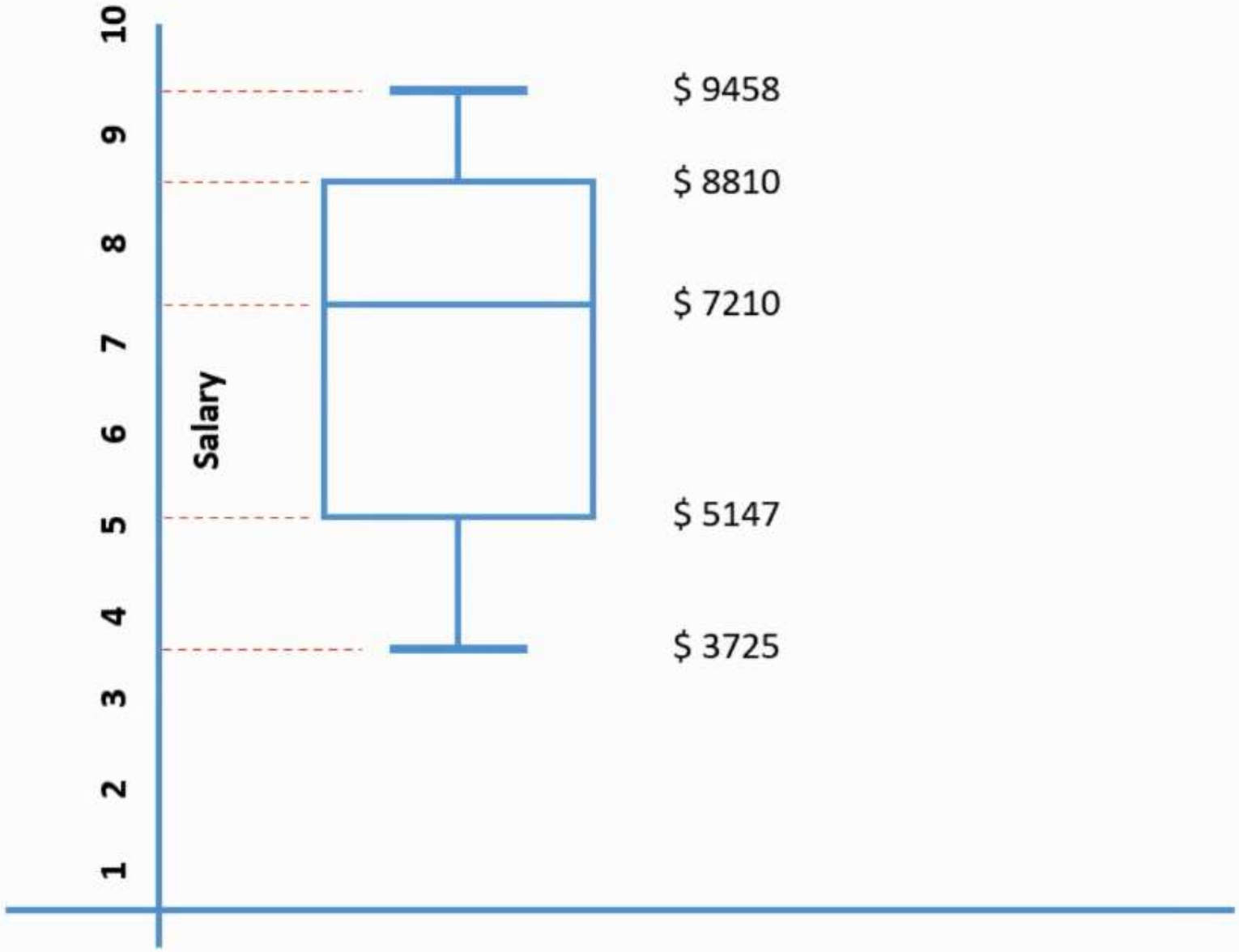


Box Plot

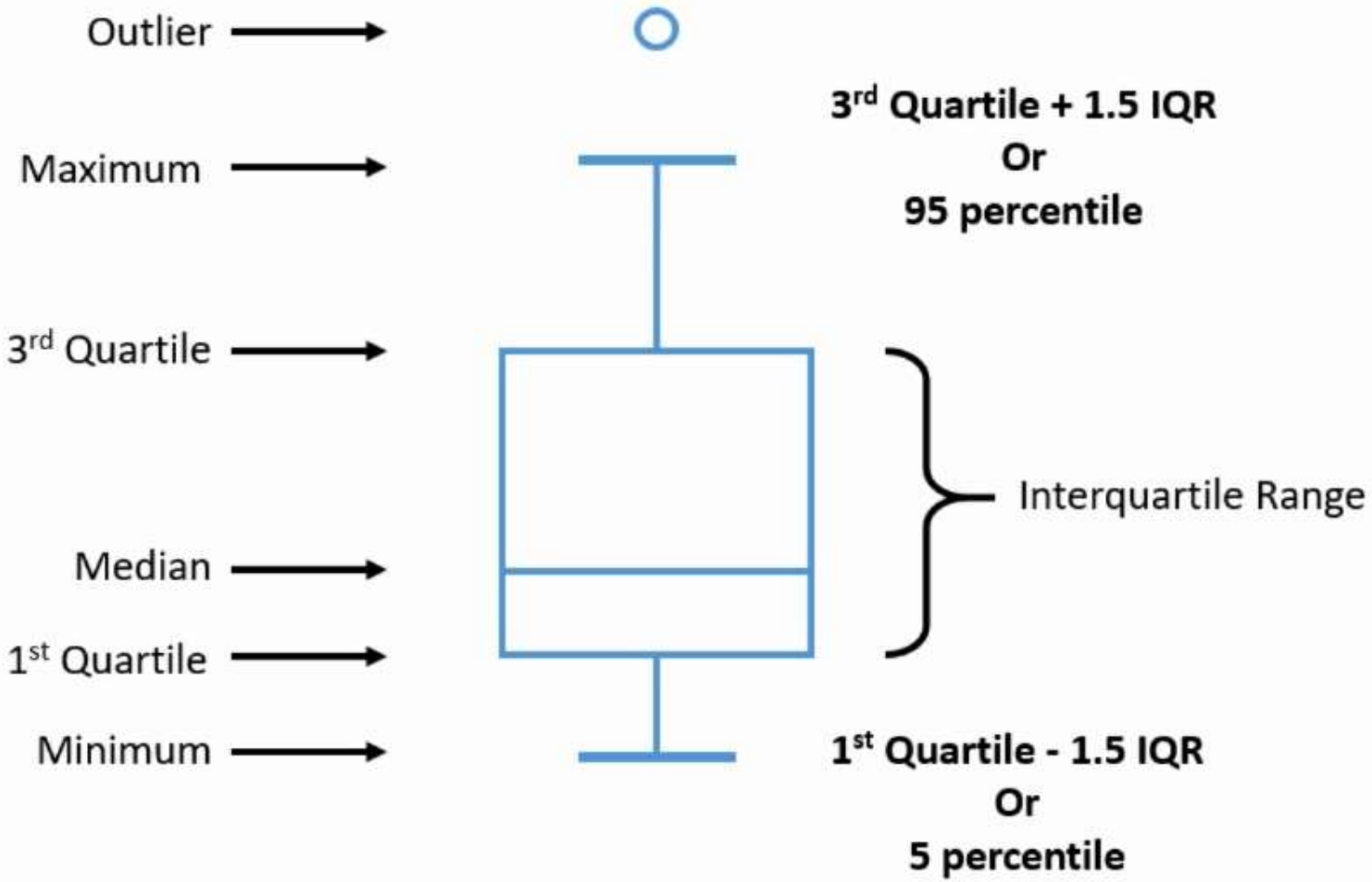
Salary
$ 3,725
$ 4,155
$ 4,627
$ 5,147
$ 5,718
$ 6,347
$ 7,039
$ 7,210
$ 7,423
$ 7,556
$ 8,369
$ 8,810
$ 8,940
$ 9,200
$ 9,458



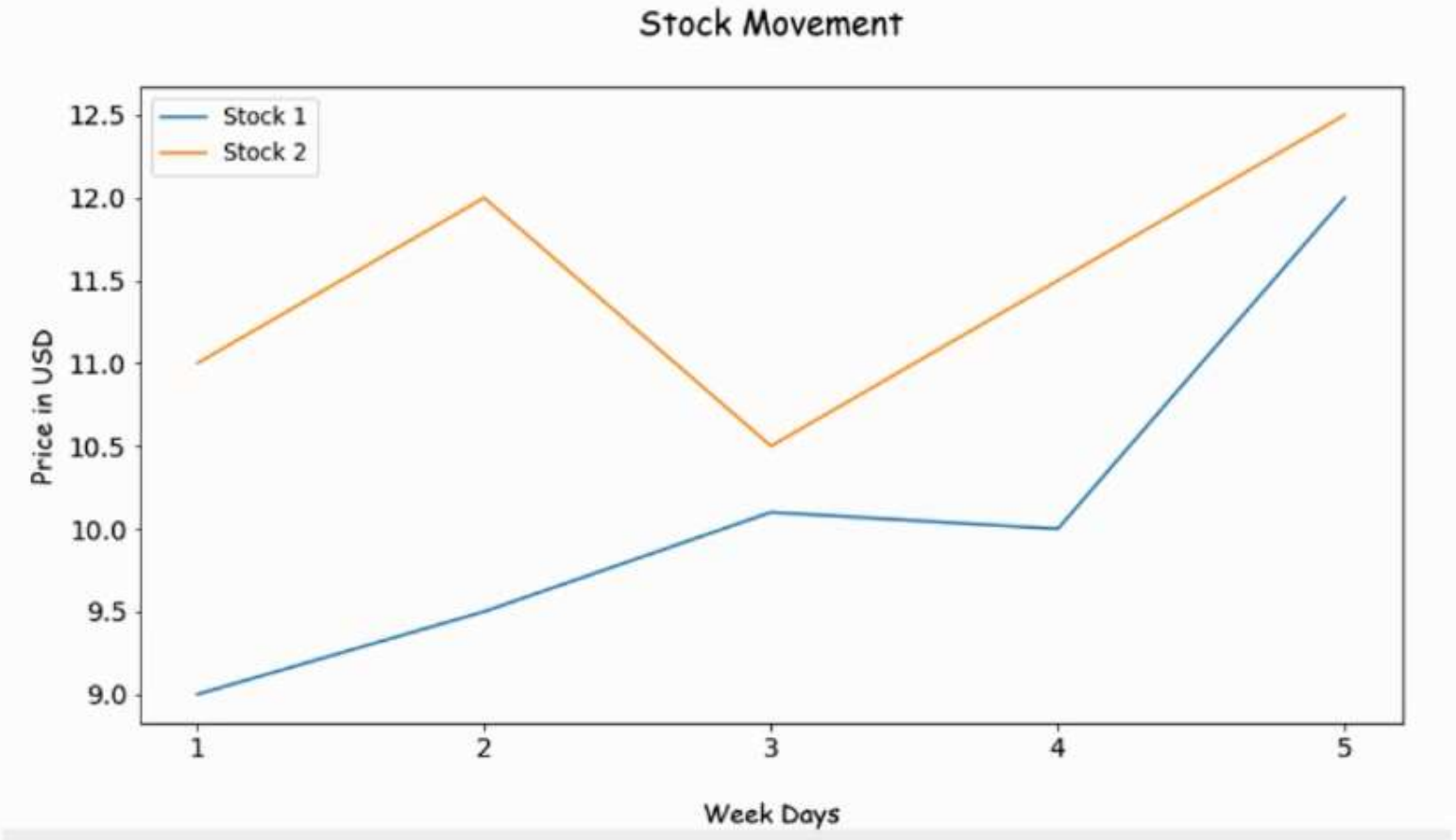
Salary
$ 3,725
$ 4,155
$ 4,627
$ 5,147
$ 5,718
$ 6,347
$ 7,039
$ 7,210
$ 7,423
$ 7,556
$ 8,369
$ 8,810
$ 8,940
$ 9,200
$ 9,458



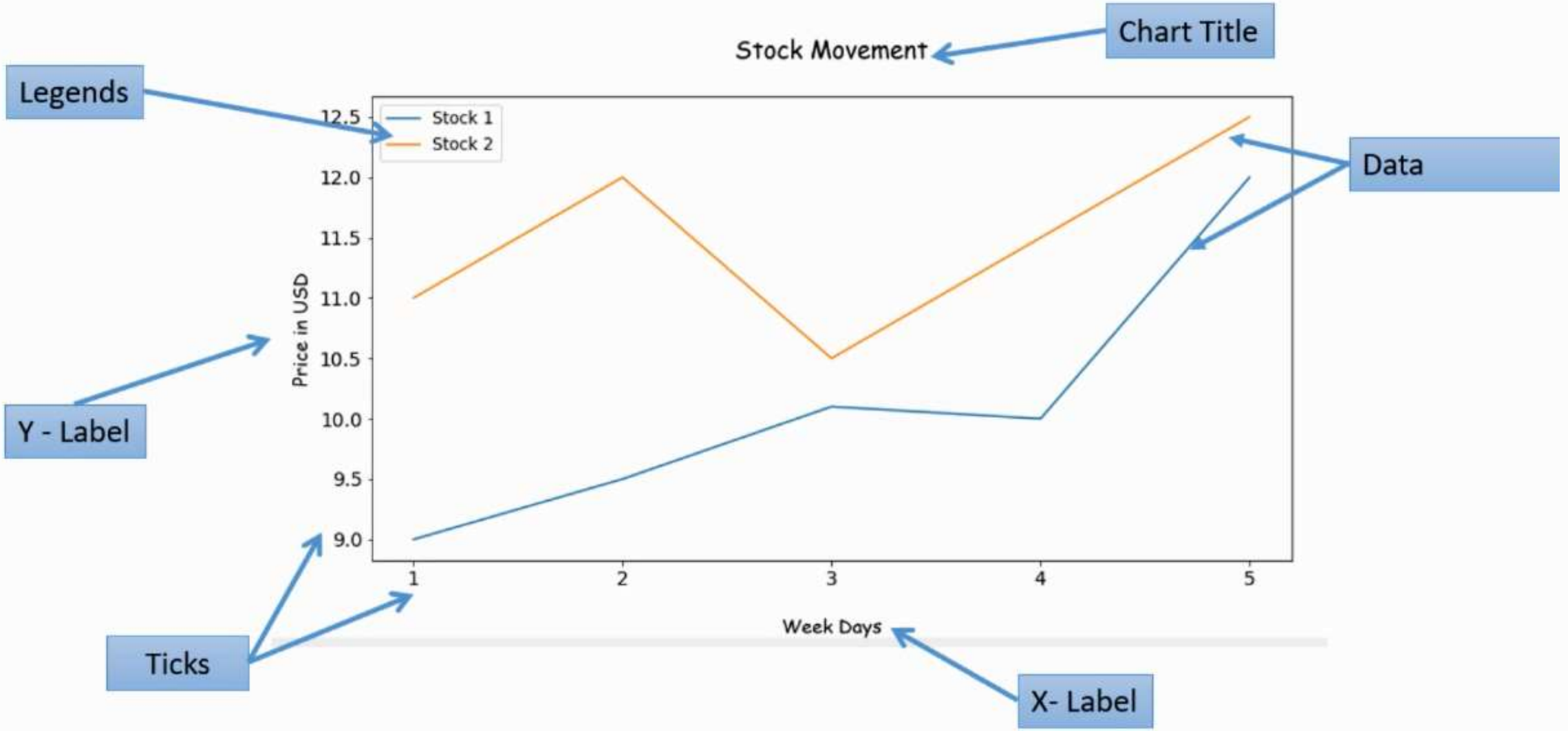
Box Plot


Outlier
Maximum
3rd Quartile
Median
1st Quartile
Minimum
3rd Quartile + 1.5 IQR
Or
95 percentile
Interquartile Range
1st Quartile - 1.5 IQR
Or
5 percentile

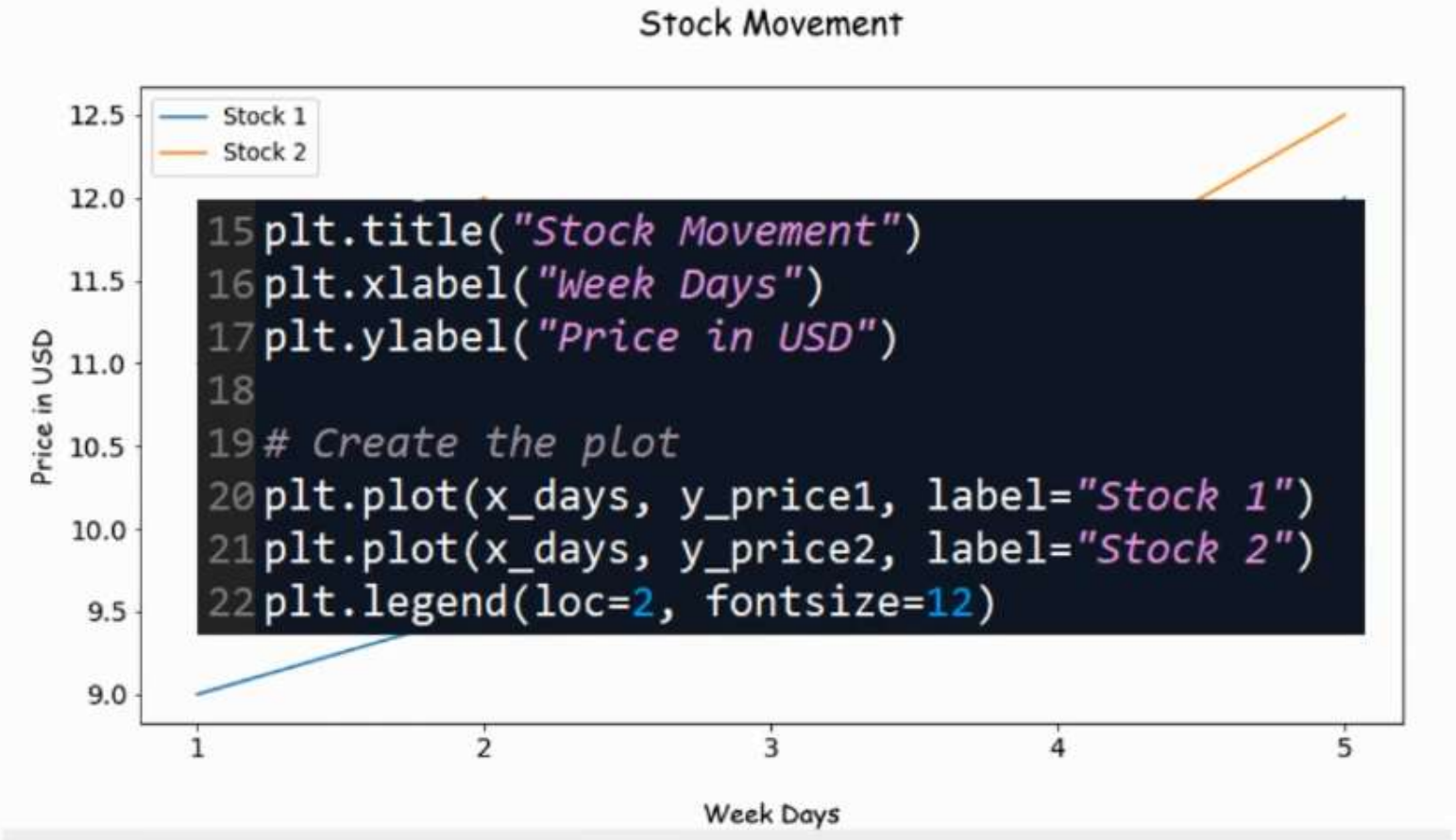
Basic Elements of the plot



Basic Elements of the plot

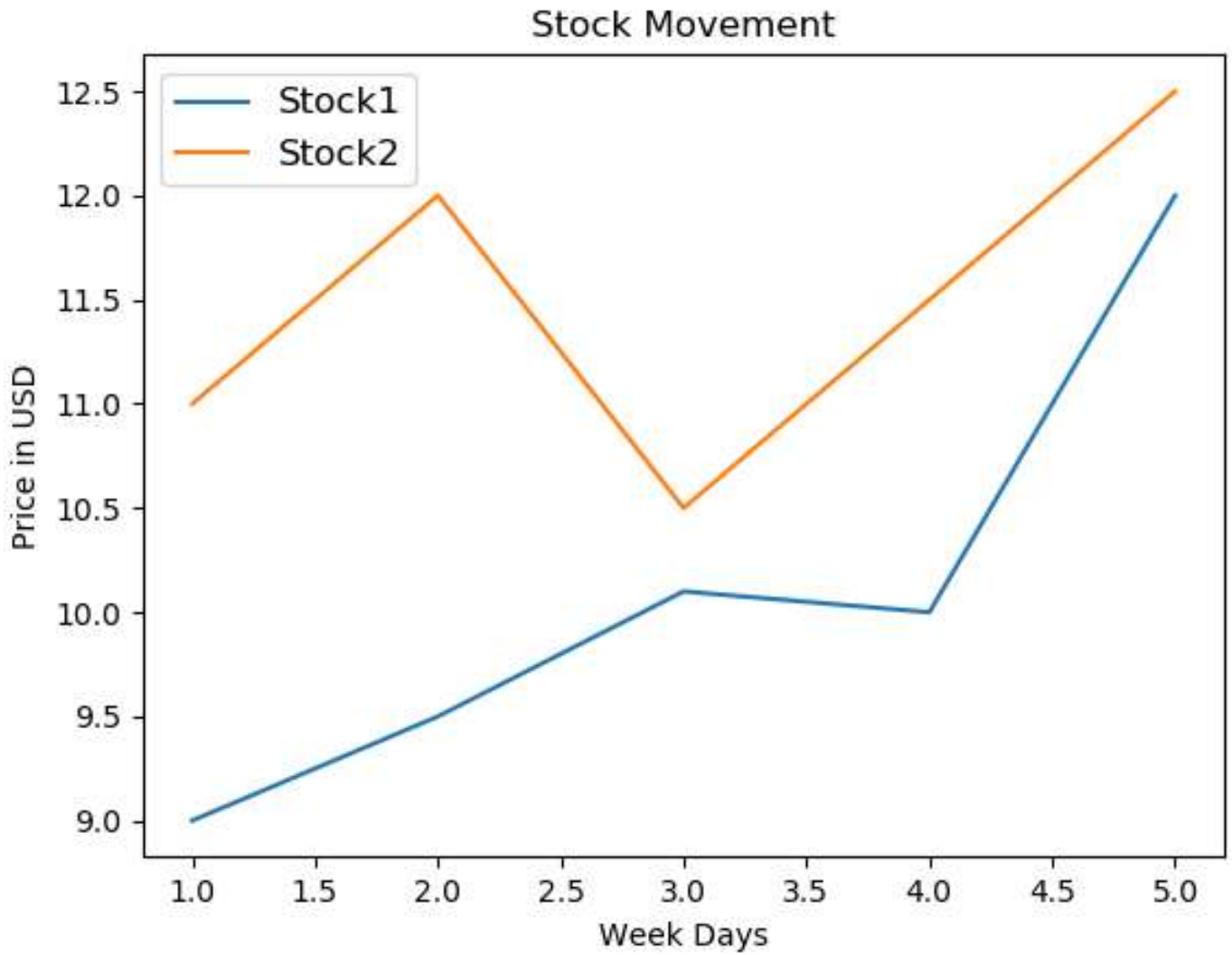


Basic Elements of the plot

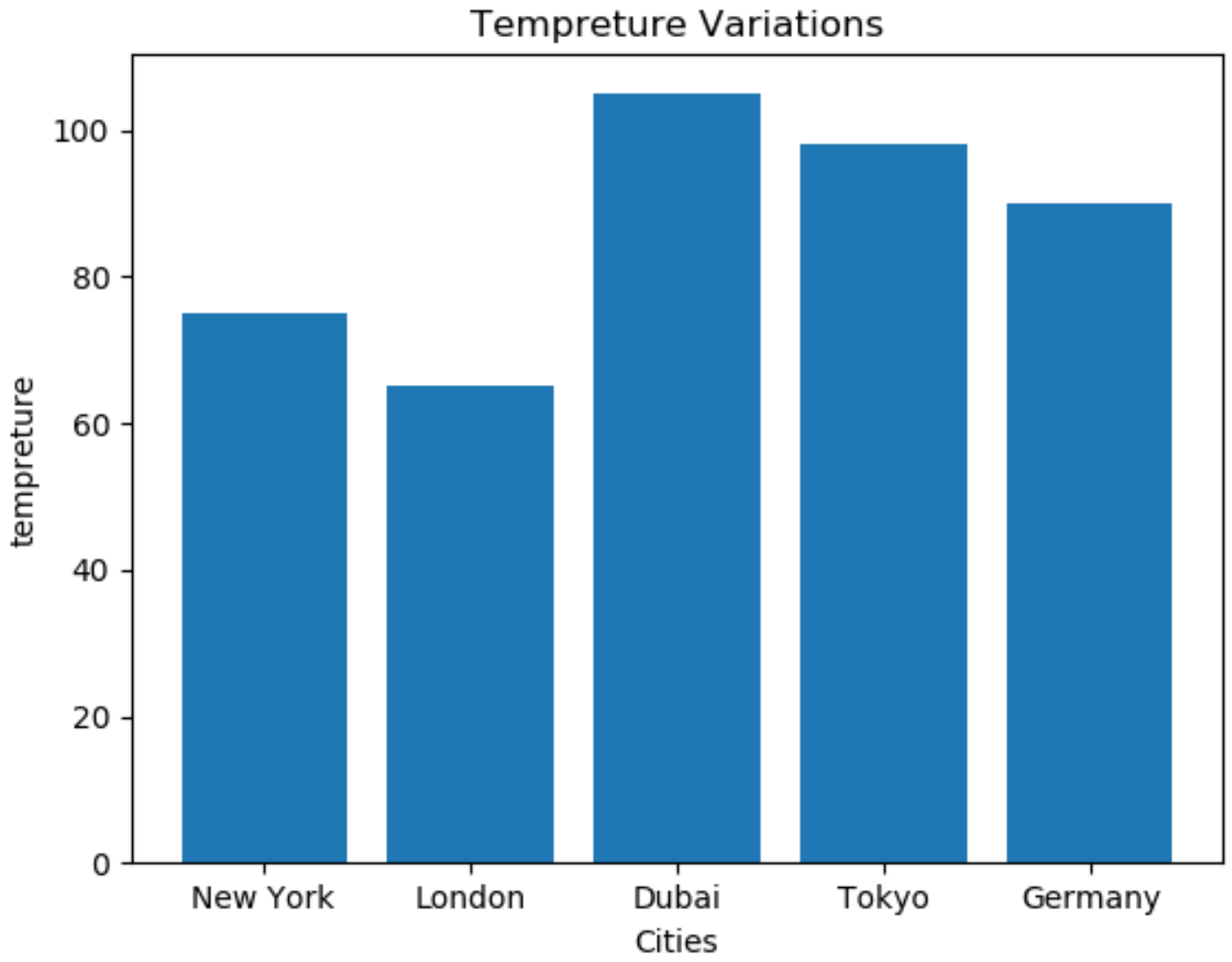



```
15 plt.title("Stock Movement")
16 plt.xlabel("Week Days")
17 plt.ylabel("Price in USD")
18
19 # Create the plot
20 plt.plot(x_days, y_price1, label="Stock 1")
21 plt.plot(x_days, y_price2, label="Stock 2")
22 plt.legend(loc=2, fontsize=12)
```

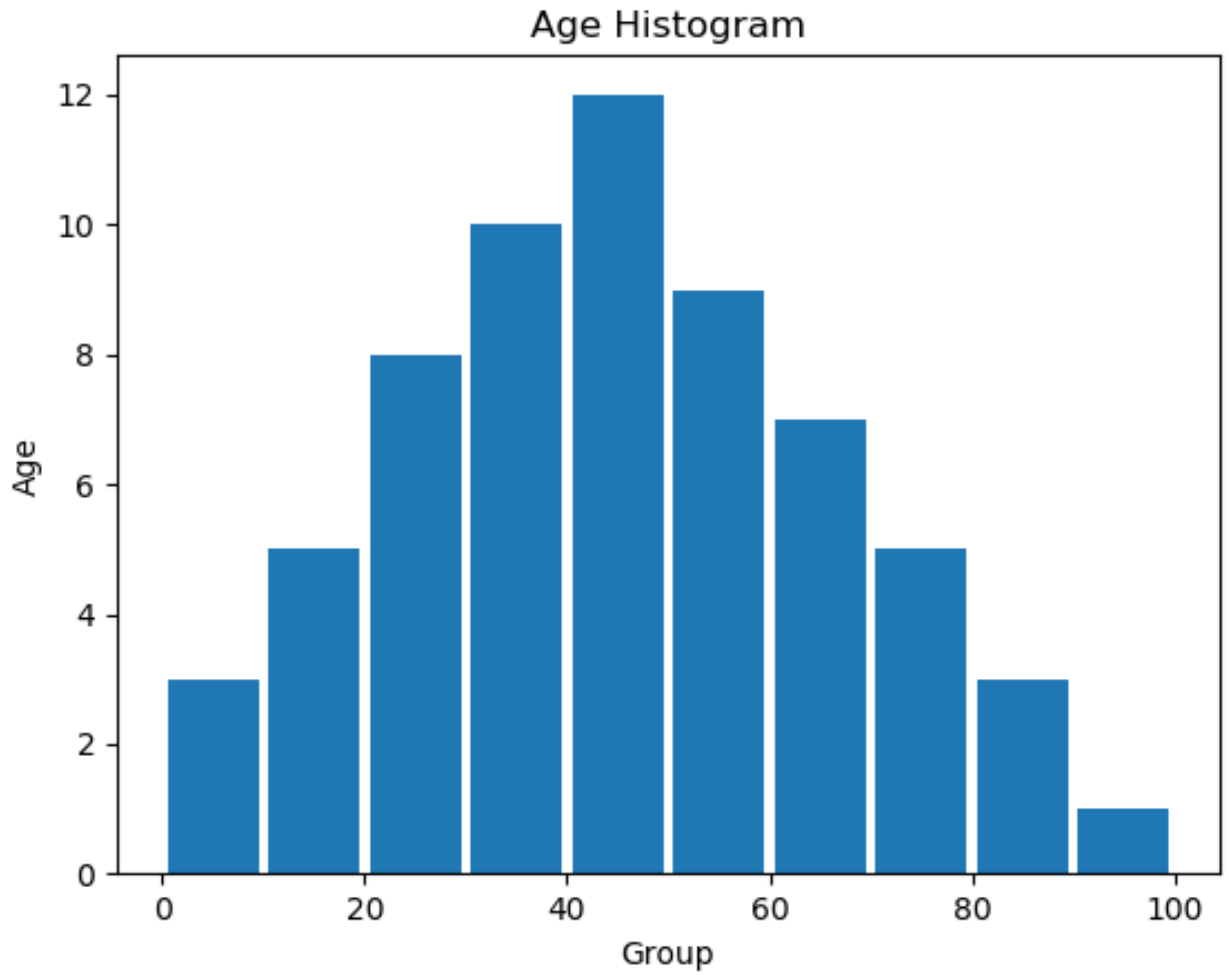
Hands on – Create Line Plots



Hands On - Create your first Bar Chart



Hands On - Create Histogram of Data



Hands On - Plotting Boxplot

DATA VISUALIZATION

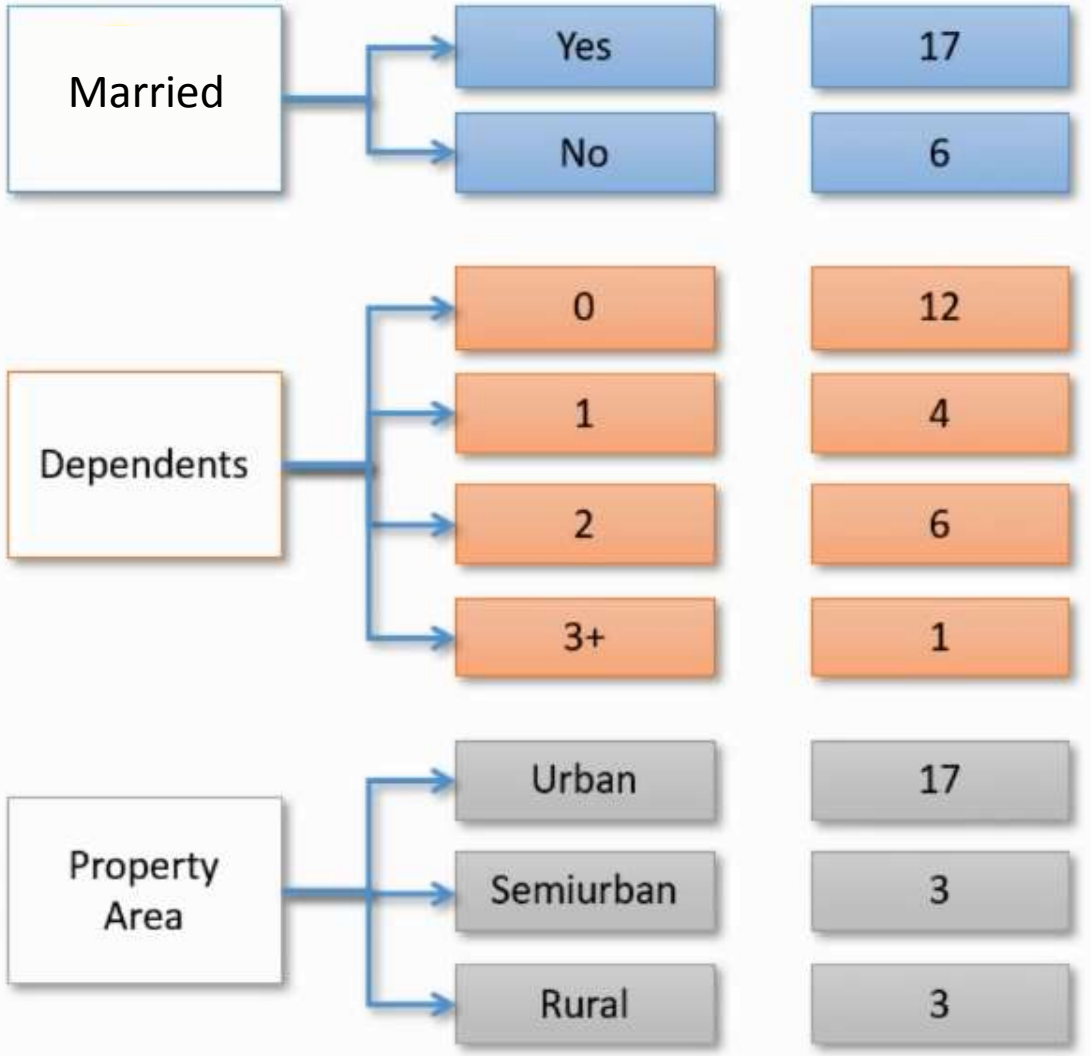
For Categorical Data

Sample dataset

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	Credit_History	Property_Area	Loan_Status
LP001002	Male	No	0	Graduate	No	1	Urban	Y
LP001003	Male	Yes	1	Graduate	No	1	Rural	N
LP001005	Female	Yes	0	Graduate	Yes	1	Urban	Y
LP001006	Male	Yes	0	Not Graduate	No	1	Urban	Y
LP001008	Male	No	0	Graduate	No	1	Urban	Y
LP001011	Female	Yes	2	Graduate	Yes	1	Urban	Y
LP001013	Male	Yes	0	Not Graduate	No	1	Urban	Y
LP001014	Male	Yes	3+	Graduate	No	0	Semiurban	N
LP001018	Female	Yes	2	Graduate	No	1	Urban	Y
LP001020	Male	Yes	1	Graduate	No	1	Semiurban	N
LP001024	Male	Yes	2	Graduate	No	1	Urban	Y
LP001027	Male	Yes	2	Graduate	No	1	Urban	Y
LP001028	Male	Yes	2	Graduate	No	1	Urban	Y
LP001029	Female	No	0	Graduate	No	1	Rural	N
LP001030	Male	Yes	2	Graduate	No	1	Urban	Y
LP001032	Male	No	0	Graduate	No	1	Urban	Y
LP001034	Male	No	1	Not Graduate	No	0	Urban	Y
LP001036	Female	No	0	Graduate	No	0	Urban	N
LP001038	Male	Yes	0	Not Graduate	No	1	Rural	N
LP001041	Male	Yes	0	Graduate	No	1	Urban	Y
LP001043	Male	Yes	0	Not Graduate	No	0	Urban	N
LP001046	Female	Yes	1	Graduate	No	1	Urban	Y
LP001047	Male	Yes	0	Not Graduate	No	0	Semiurban	N

Sample dataset

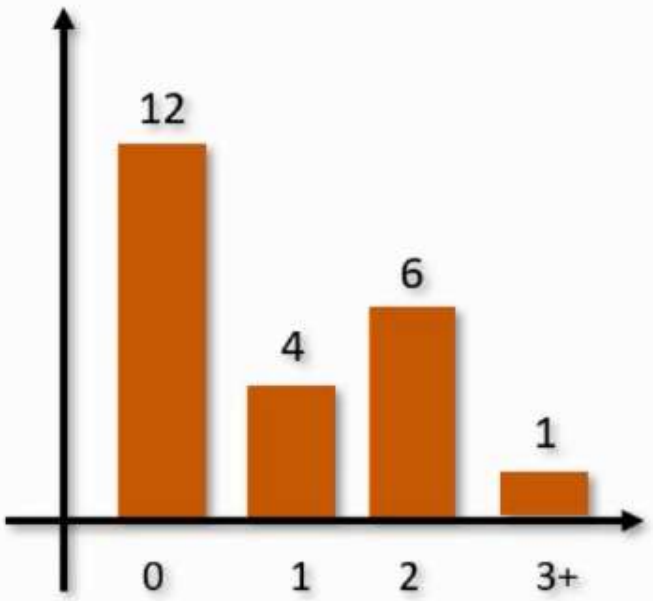
Gender	Married	Dependents	Education	Self_Employed	Credit_History	Property_Area	Loan_Status
Male	No	0	Graduate	No	1	Urban	Y
Male	Yes	1	Graduate	No	1	Rural	N
Female	Yes	0	Graduate	Yes	1	Urban	Y
Male	Yes	0	Not Graduate	No	1	Urban	Y
Male	No	0	Graduate	No	1	Urban	Y
Female	Yes	2	Graduate	Yes	1	Urban	Y
Male	Yes	0	Not Graduate	No	1	Urban	Y
Male	Yes	3+	Graduate	No	0	Semiurban	N
Female	Yes	2	Graduate	No	1	Urban	Y
Male	Yes	1	Graduate	No	1	Semiurban	N
Male	Yes	2	Graduate	No	1	Urban	Y
Male	Yes	2	Graduate	No	1	Urban	Y
Male	Yes	2	Graduate	No	1	Urban	Y
Female	No	0	Graduate	No	1	Rural	N
Male	Yes	2	Graduate	No	1	Urban	Y
Male	No	0	Graduate	No	1	Urban	Y
Male	No	1	Not Graduate	No	0	Urban	Y
Female	No	0	Graduate	No	0	Urban	N
Male	Yes	0	Not Graduate	No	1	Rural	N
Male	Yes	0	Graduate	No	1	Urban	Y
Male	Yes	0	Not Graduate	No	0	Urban	N
Female	Yes	1	Graduate	No	1	Urban	Y
Male	Yes	0	Not Graduate	No	0	Semiurban	N



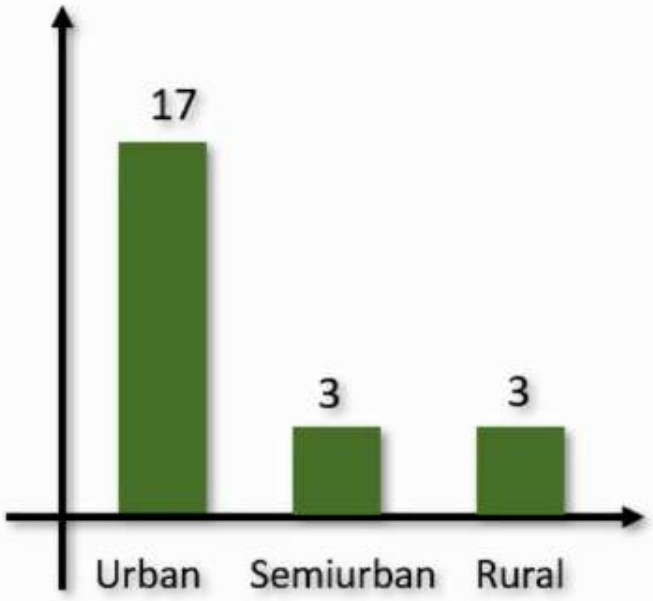
Bar Chart

Marital Status

Number of Dependents

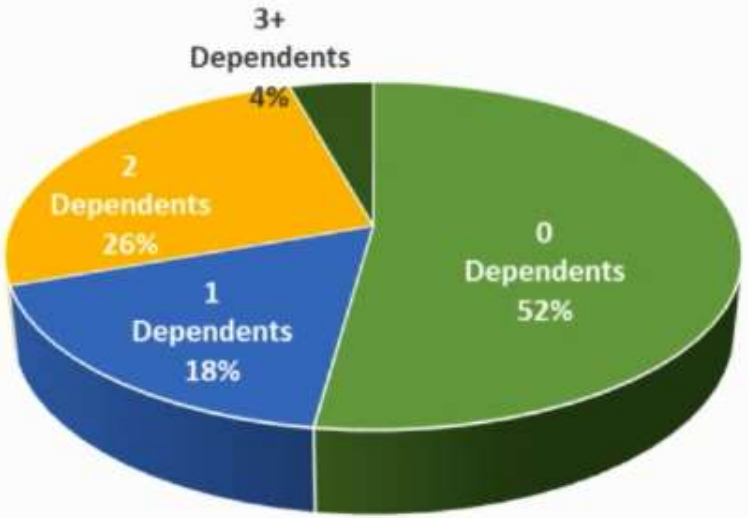


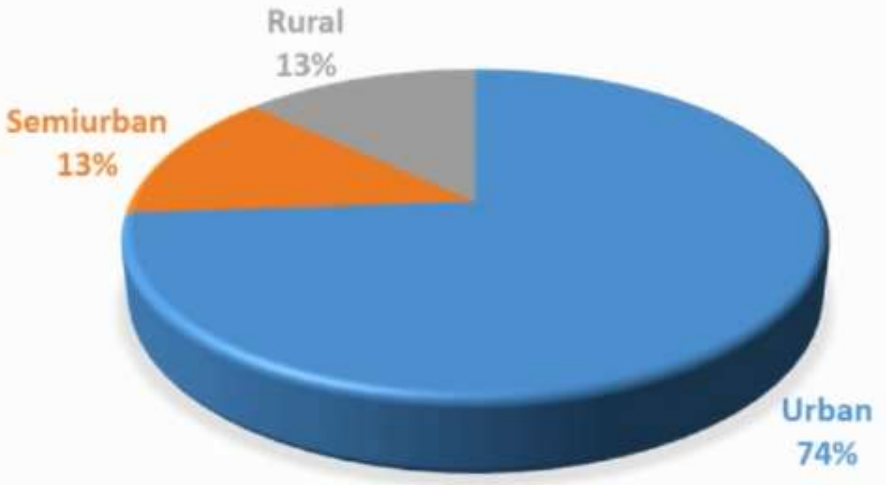
Property Area



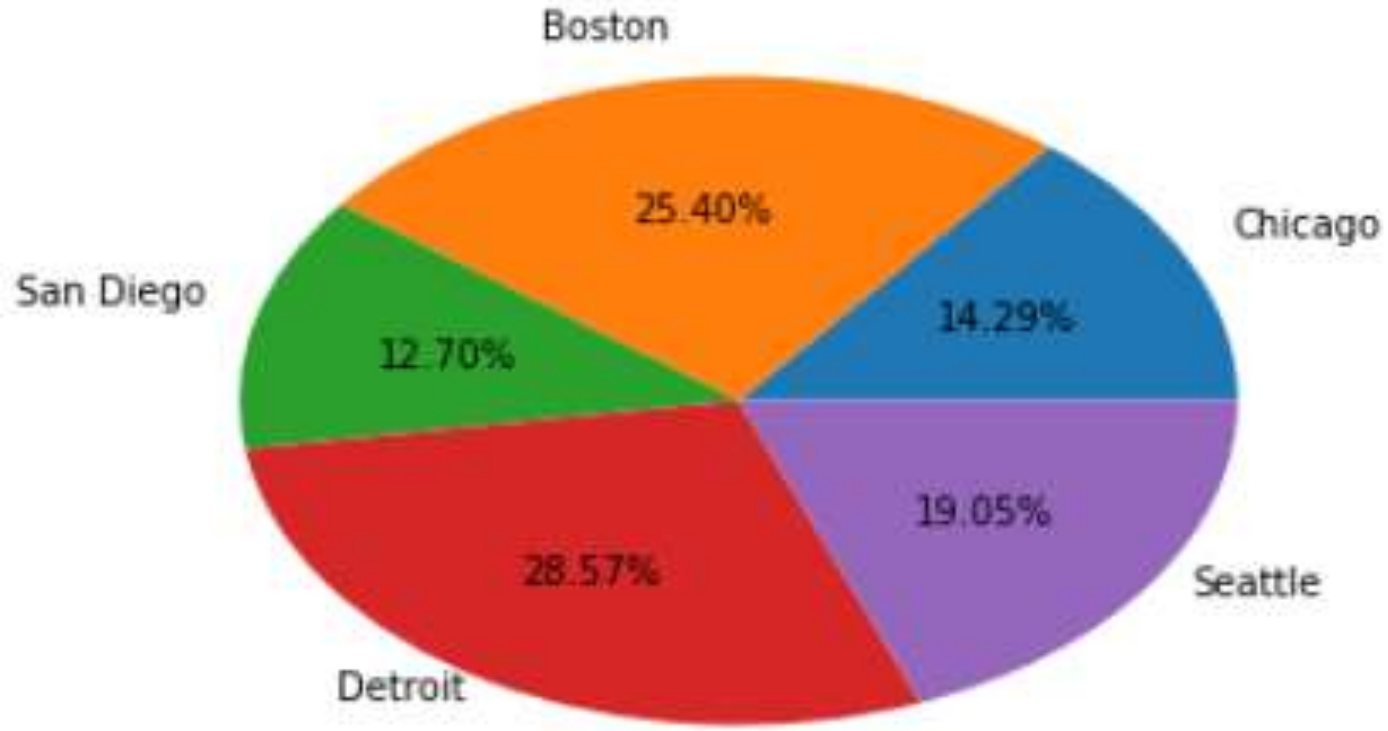
Pie Chart


Unmarried
26%
Married
74%
Married
Unmarried

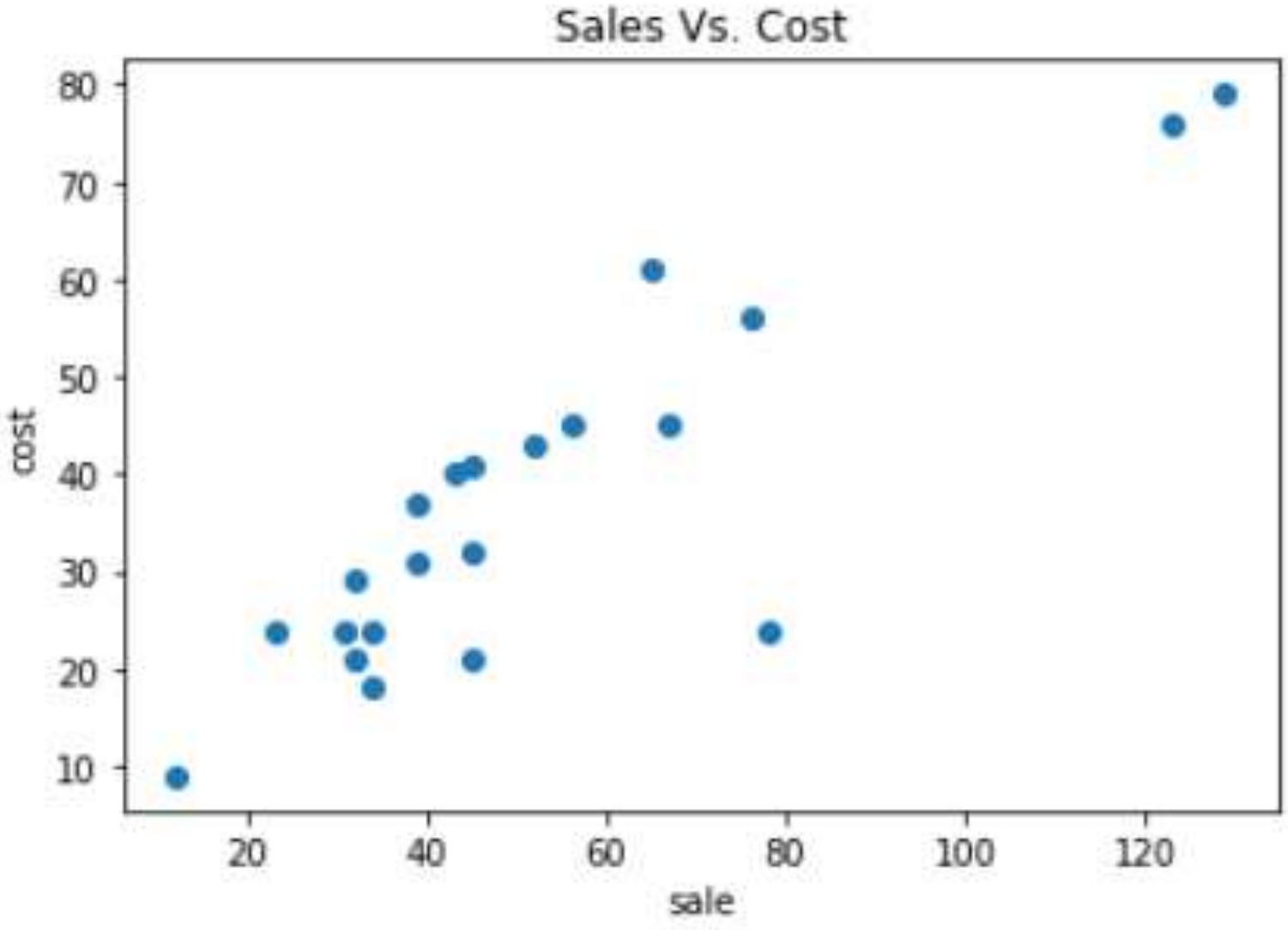

3+
Dependents
4%
2
Dependents
26%
1
Dependents
18%
0
Dependents
52%


Rural
13%
Semiurban
13%
Urban
74%

Hands On - Pie Charts Part 1

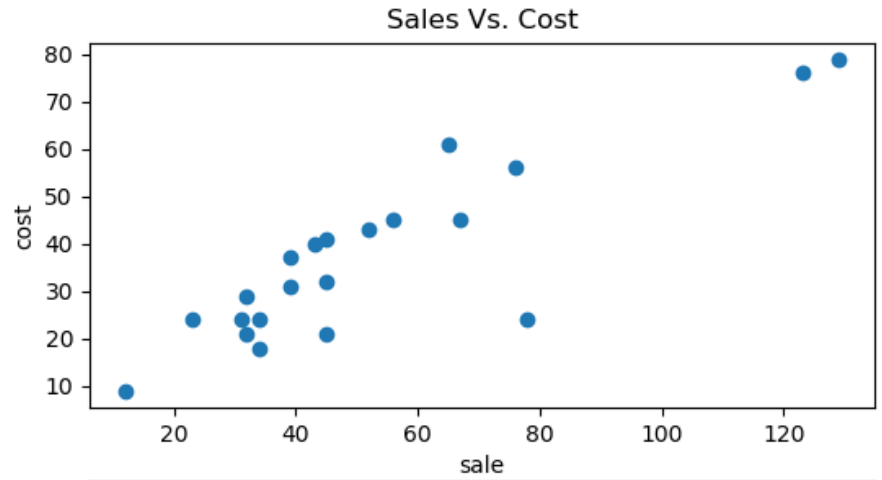


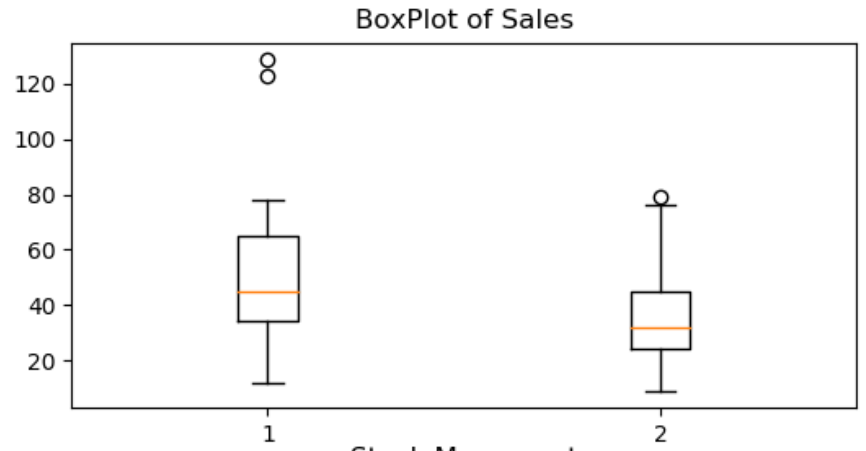
Hands On - Scatter Plots

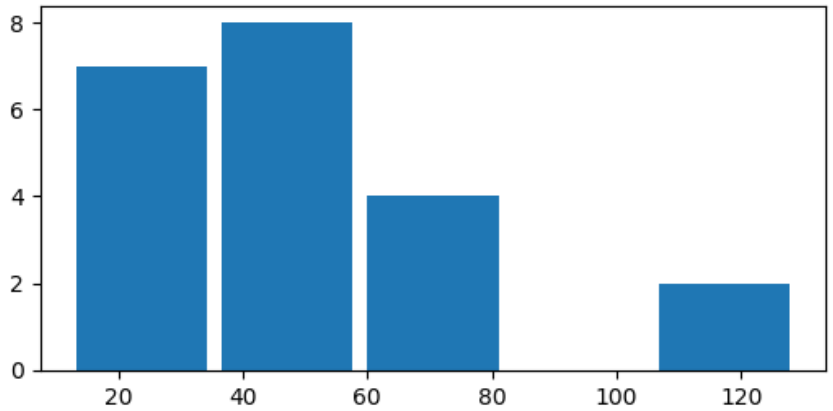


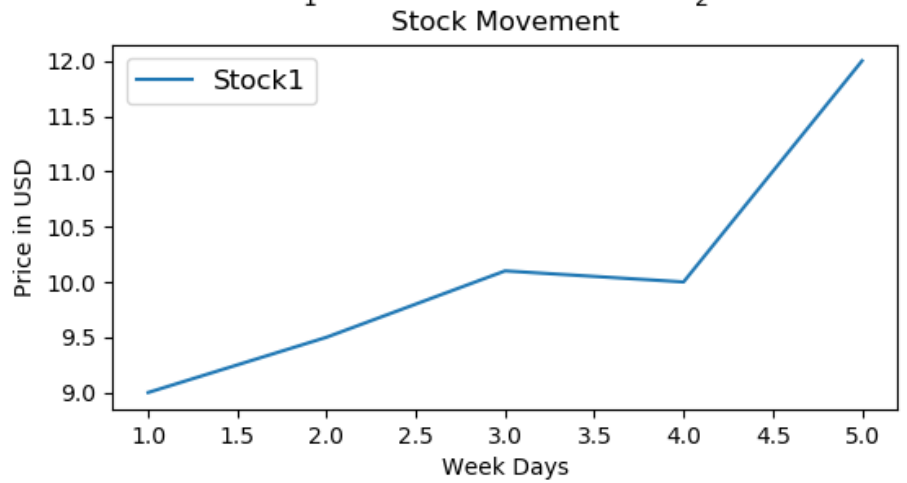
Hands On - MatplotLib Figures for creating multiple plots

Hands On - Subplots for plotting multiple plots in one figure



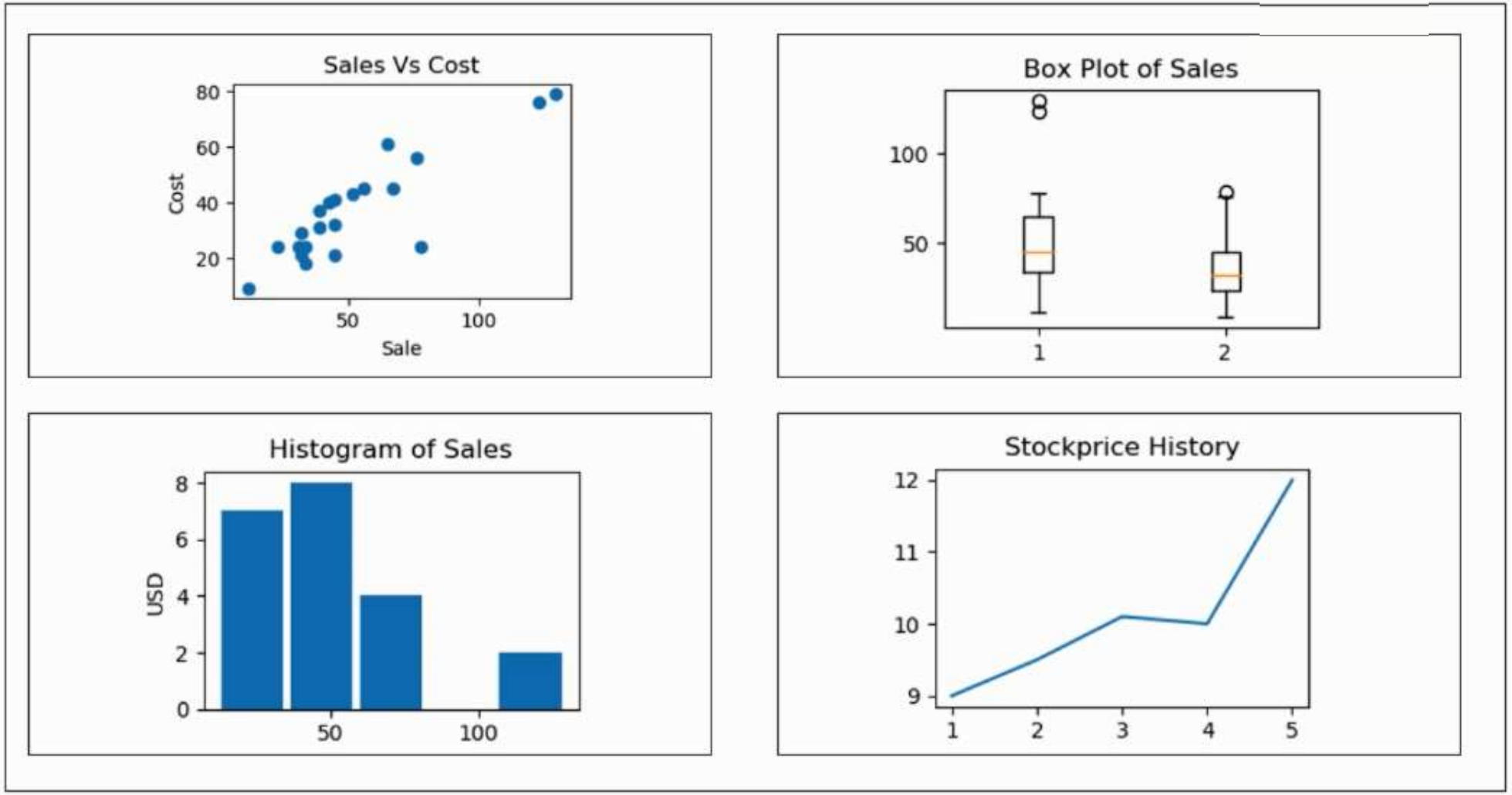






Plots in a grid

Plots in a grid



Plots in a grid