EXPLOLATORY DATA ANALYSIS

Content :

- Introduction to EDA
- Importance of EDA
- Data types
- Python Packages for EDA
- Lists of Graphs
- Practical EDA

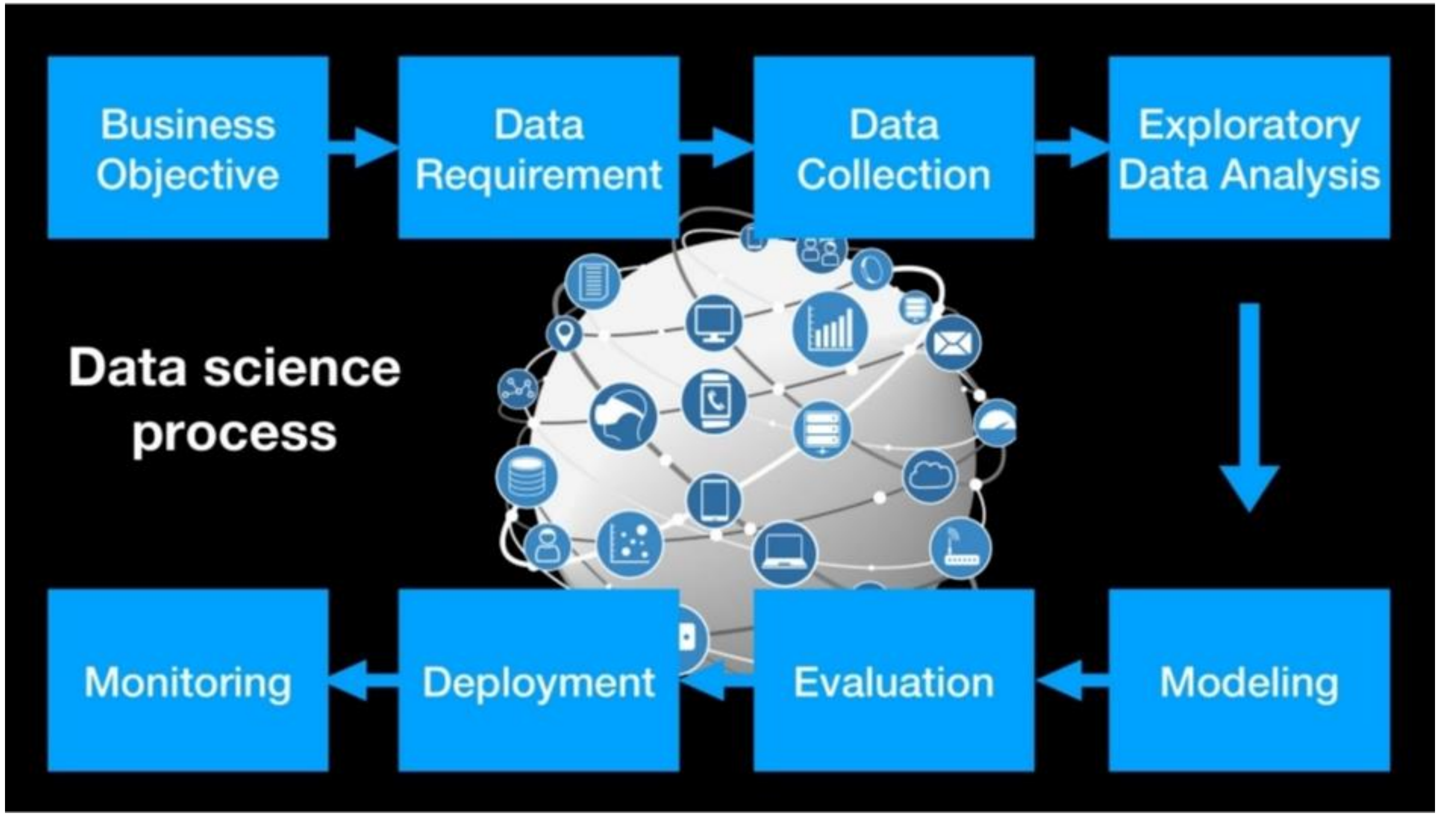
1. INTRODUCTION TO EDA

- Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations .
- It is a good practice to understand the data first and try to gather as many insights from it.

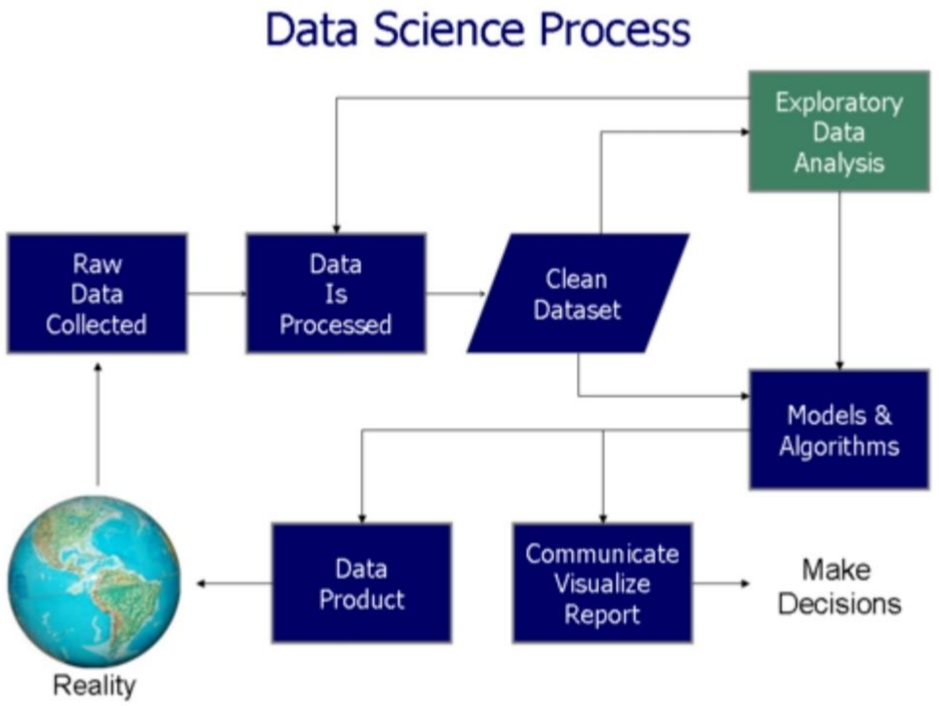
2. IMPORTANCE OF EDA

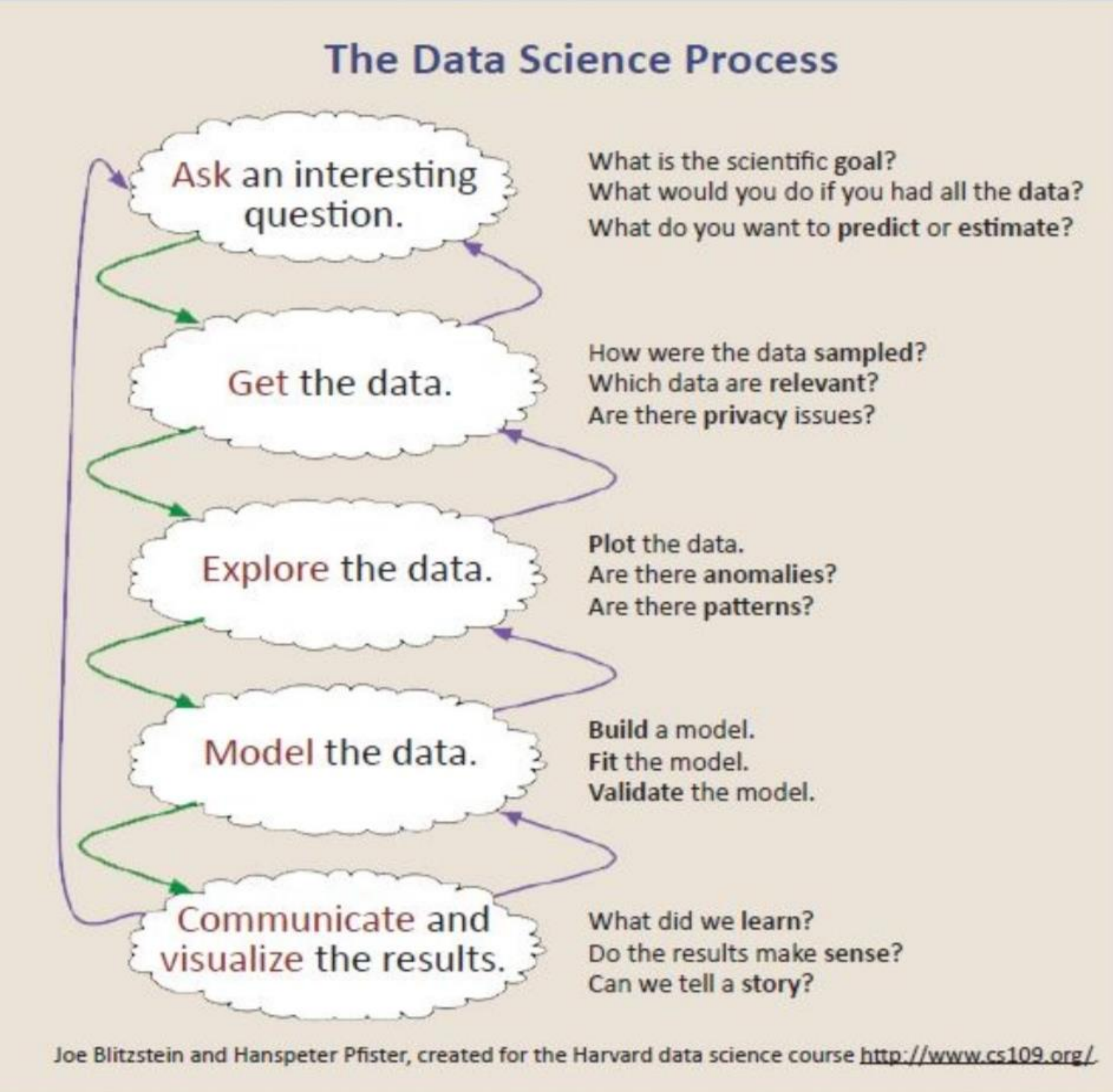
- Identifying the most important variables/features in your dataset.
- Testing a hypothesis or checking assumptions related to the dataset.
- To check the quality of data for further processing and cleaning.
- Deliver data driven insights to business stakeholders.
- Verify expected relationships actually exists in the data .
- To find unexpected structure or insights in the data.

Data Science Modeling Process



Data Science Process





The Data Science Process

Ask an interesting question.

What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?

Get the data.

How were the data **sampled**?
Which data are **relevant**?
Are there **privacy** issues?

Explore the data.

Plot the data.
Are there **anomalies**?
Are there **patterns**?

Model the data.

Build a model.
Fit the model.
Validate the model.

Communicate and visualize the results.

What did we **learn**?
Do the results **make sense**?
Can we tell a **story**?

Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course http://www.cs109.org/.

Two categories of Data

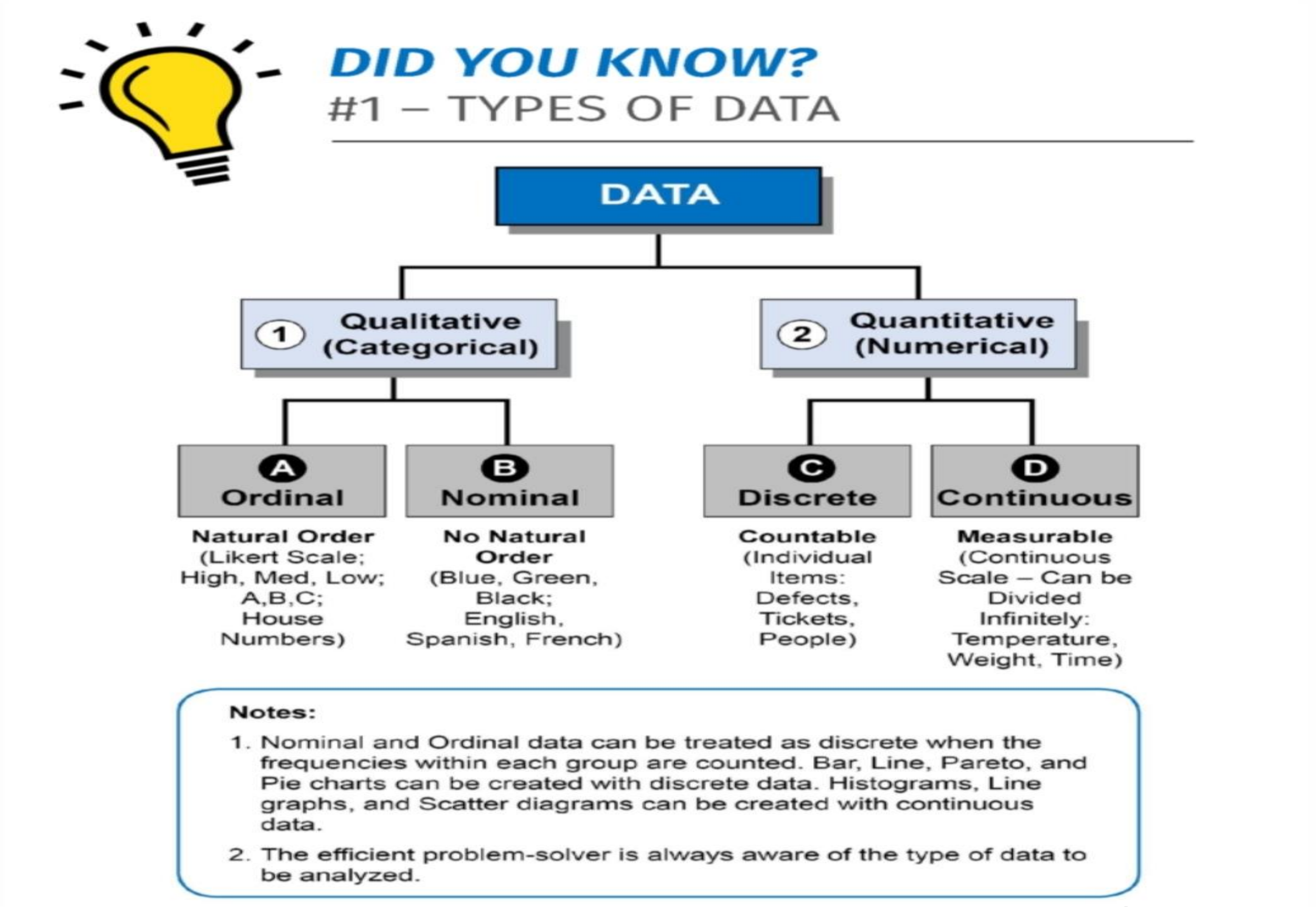
- Structured Data types

  example : CSV file , Excel file , Database file

- Unstructured Data types

  example : Images , videos , audio

Data Types



DID YOU KNOW?

#1 – TYPES OF DATA

DATA

1 Qualitative (Categorical)

2 Quantitative (Numerical)

A Ordinal

Natural Order (Likert Scale; High, Med, Low; A,B,C; House Numbers)

B Nominal

No Natural Order (Blue, Green, Black; English, Spanish, French)

C Discrete

Countable (Individual Items: Defects, Tickets, People)

D Continuous

Measurable (Continuous Scale – Can be Divided Infinitely: Temperature, Weight, Time)

Notes:

1. Nominal and Ordinal data can be treated as discrete when the frequencies within each group are counted. Bar, Line, Pareto, and Pie charts can be created with discrete data. Histograms, Line graphs, and Scatter diagrams can be created with continuous data.
2. The efficient problem-solver is always aware of the type of data to be analyzed.

Structured Data Types

Categorial - This is any data that is not a number .

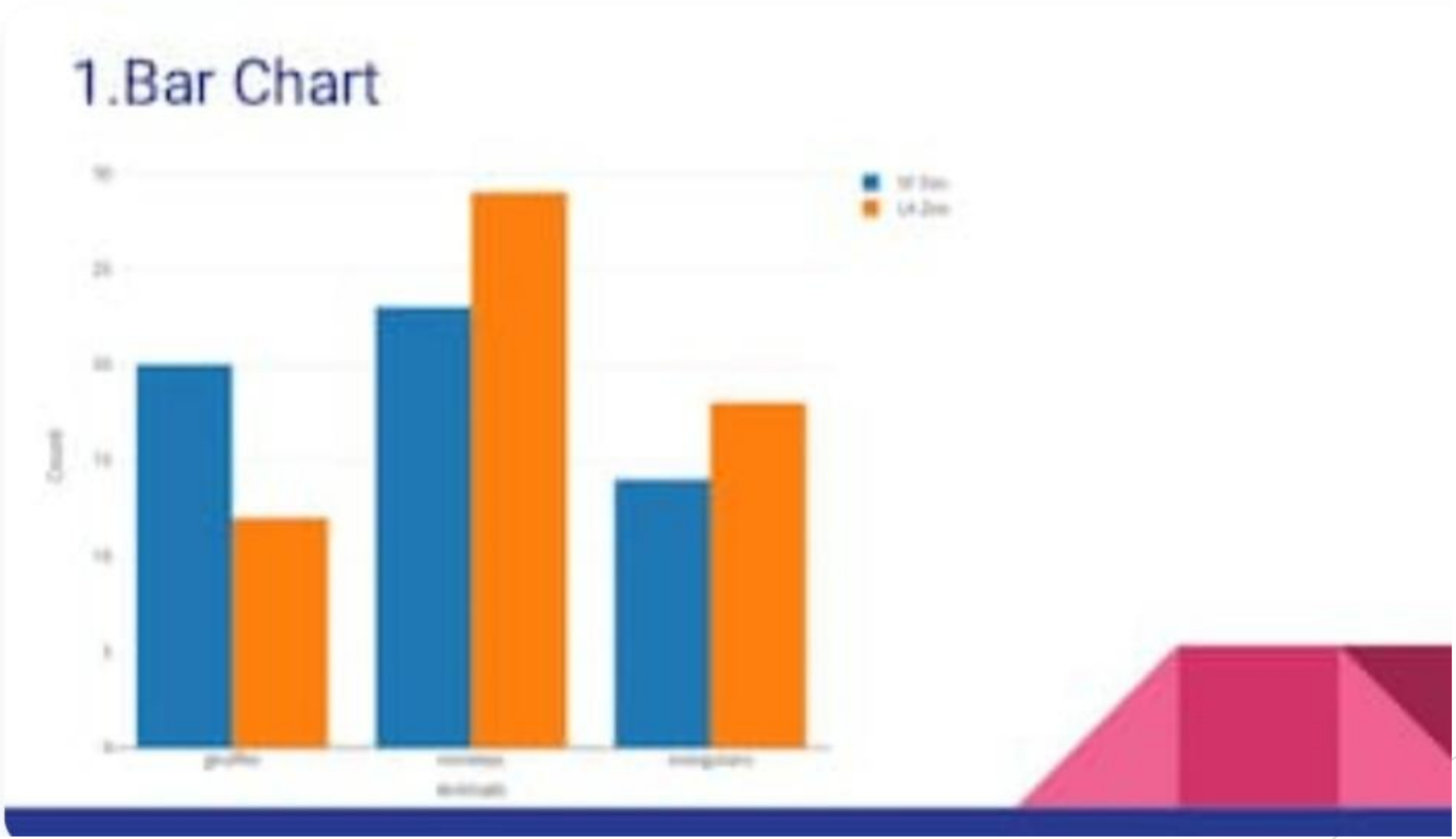
- Ordinal - have a set of order eg. Rating happiness on a scale of 1-10 .
- Binary – have only two values eg. Male or female
- Nominal – no set of order eg . Countries

Numerical - Data inform of numbers

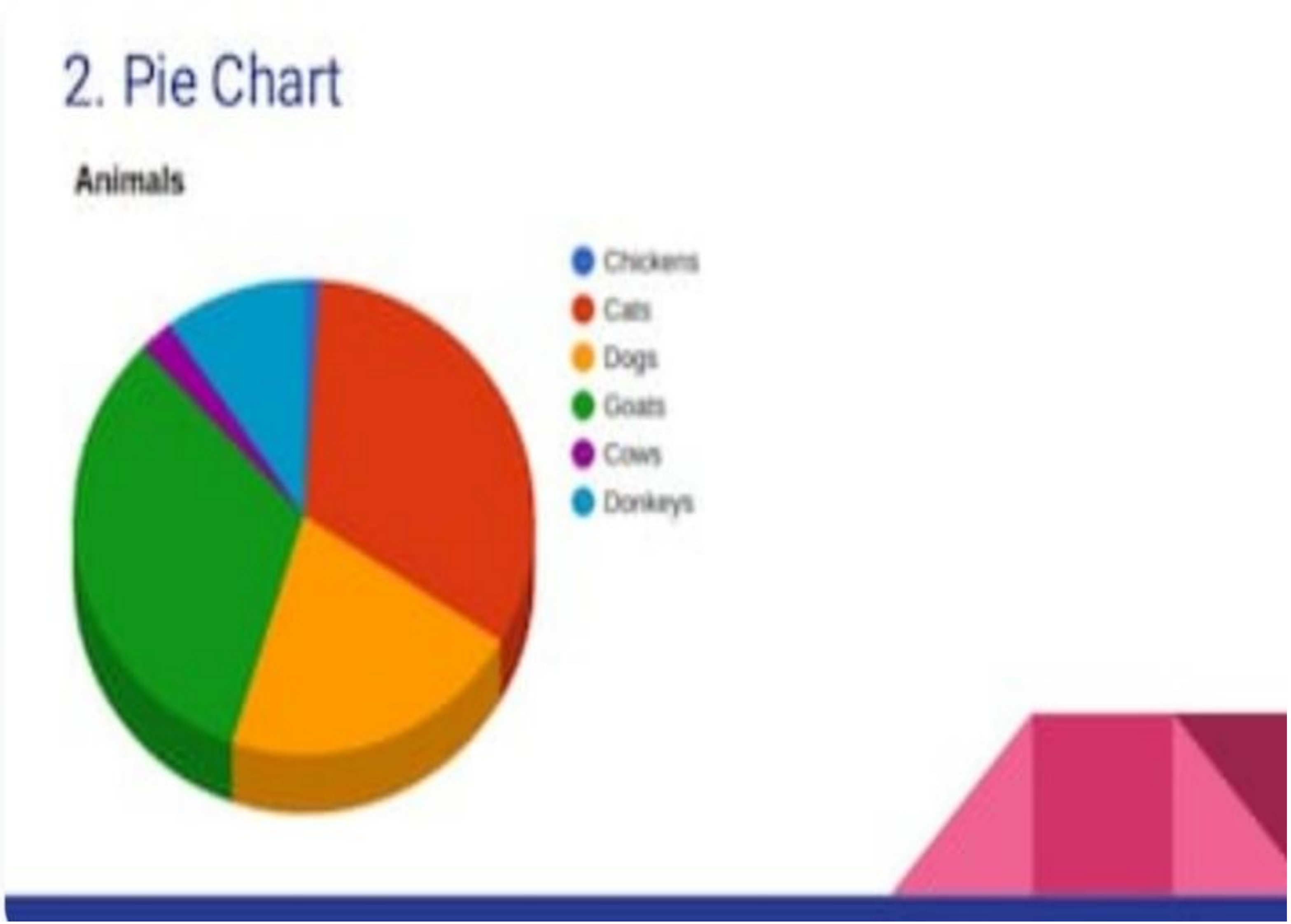
- Continious - numbers that don’t have a logical end to them eg. Heights
- Discrete - have a logical end to them eg. Days in the Month

Python Packages for EDA

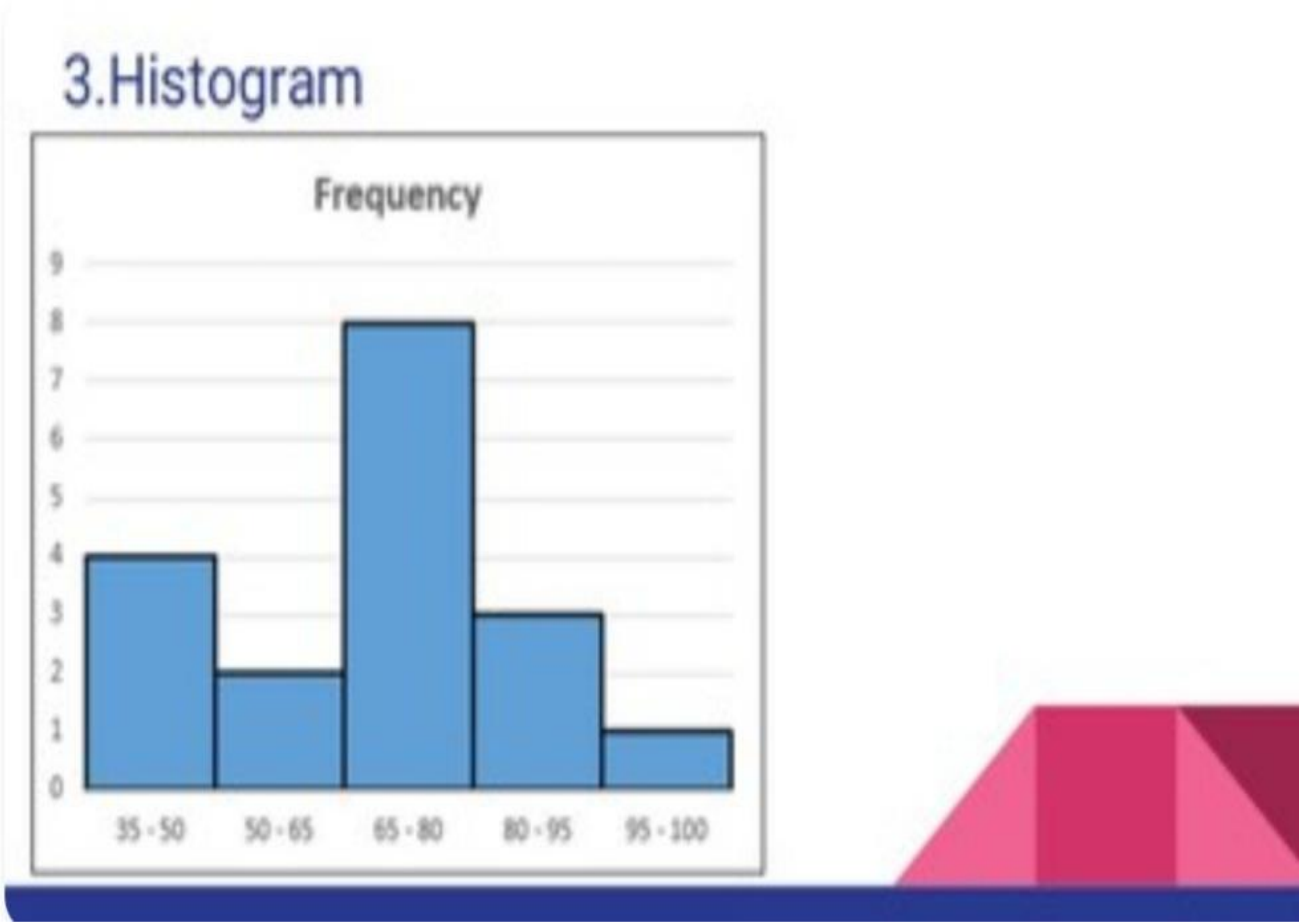
1. Bar Chart



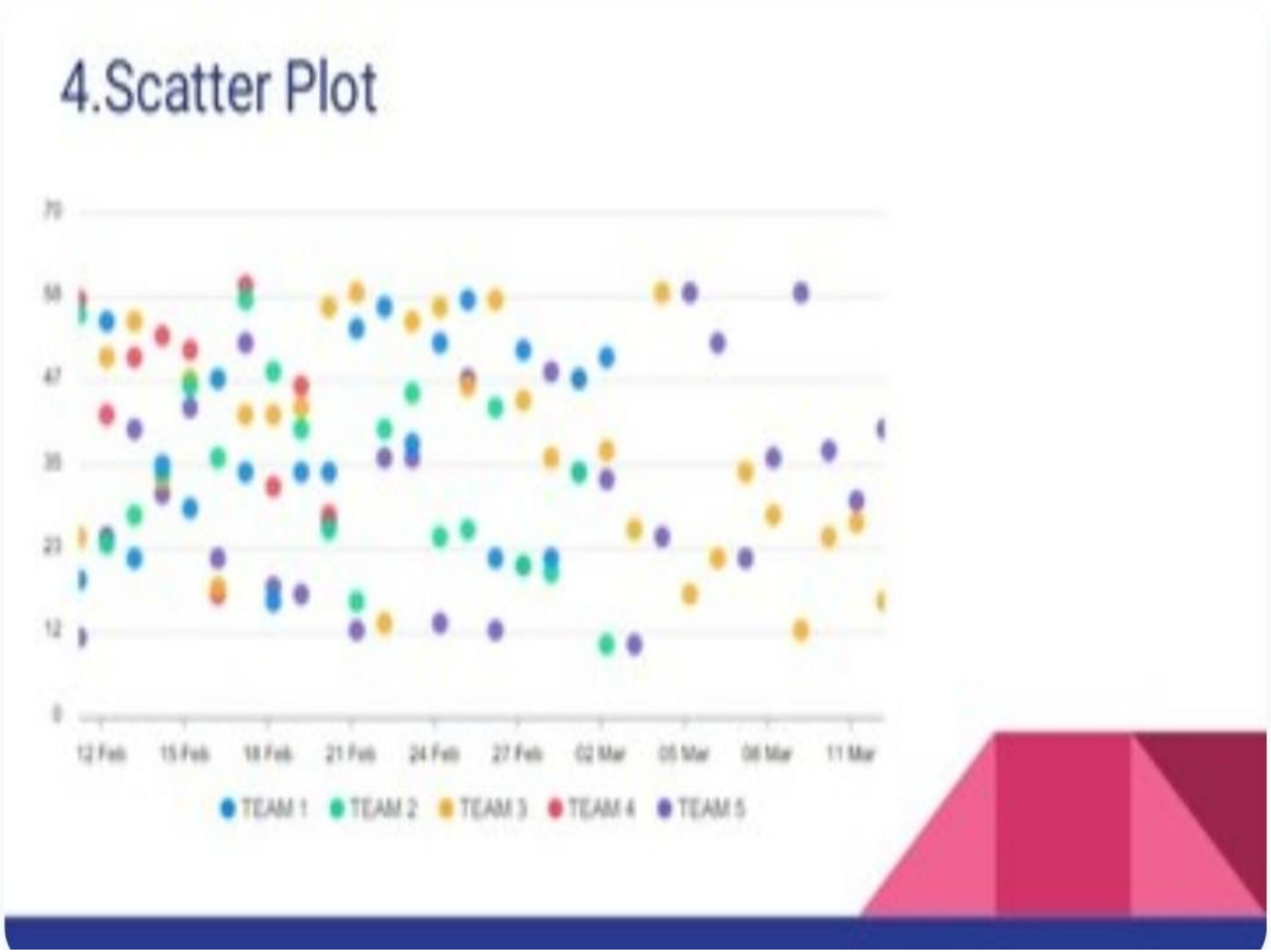
1.Bar Chart



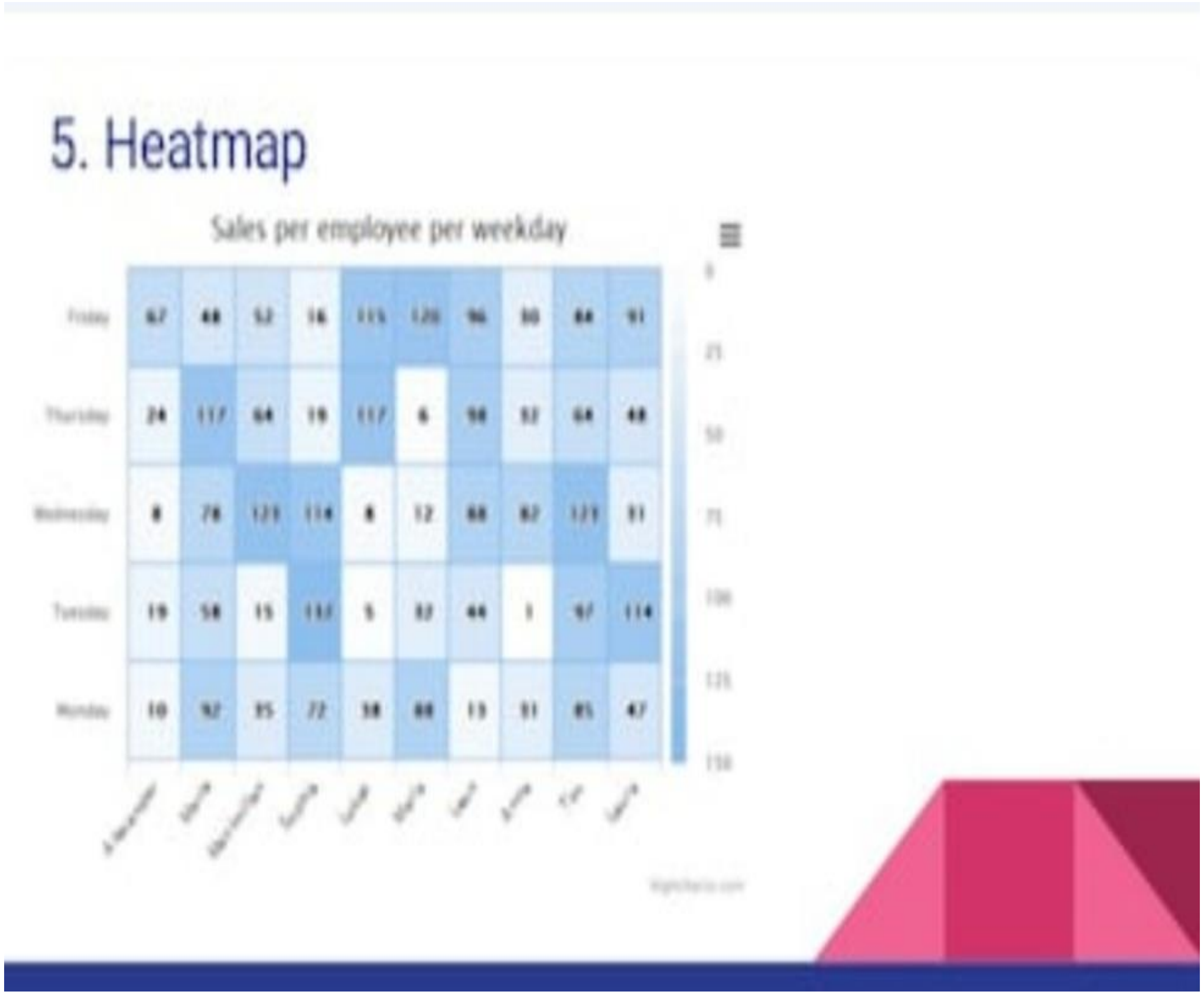
2. Pie Chart



3.Histogram



4.Scatter Plot

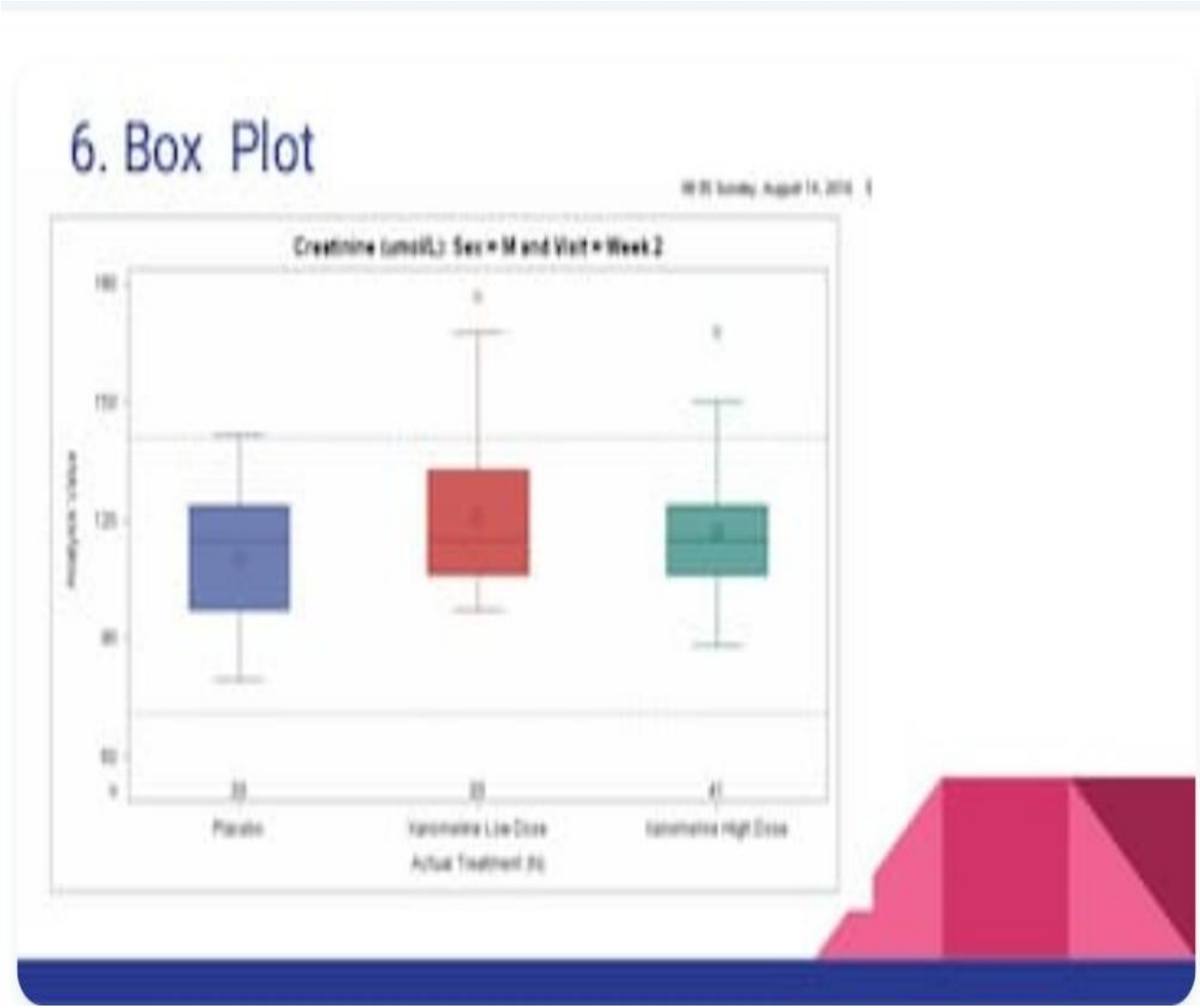


5. Heatmap

Sales per employee per weekday

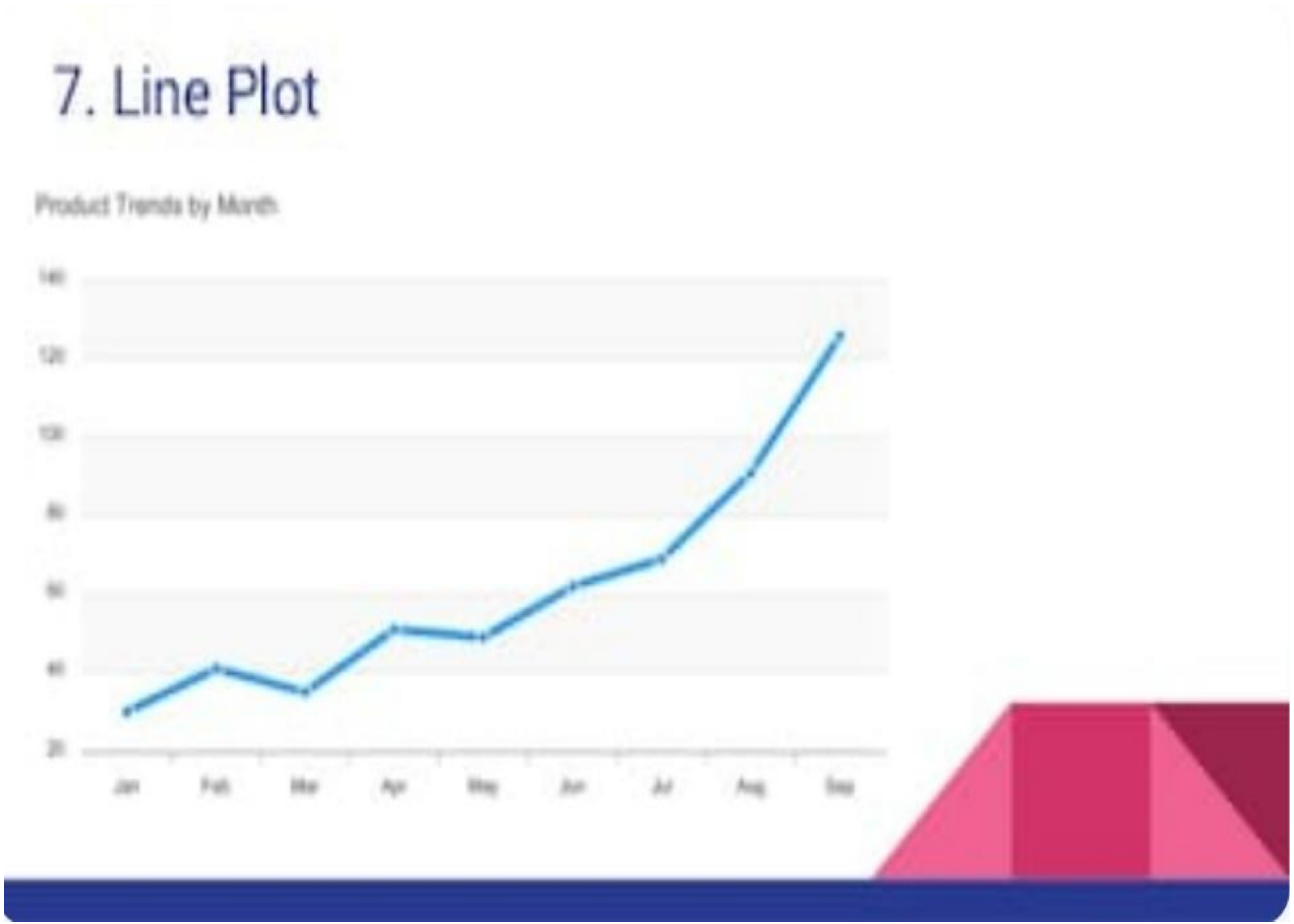
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Friday	67	48	52	16	115	120	96	30	84	91
Thursday	24	117	64	19	117	6	98	32	64	48
Wednesday	8	78	123	114	8	12	88	82	123	31
Tuesday	19	58	15	132	5	32	44	1	97	114
Monday	10	92	35	72	38	88	13	31	85	47

Highcharts.com



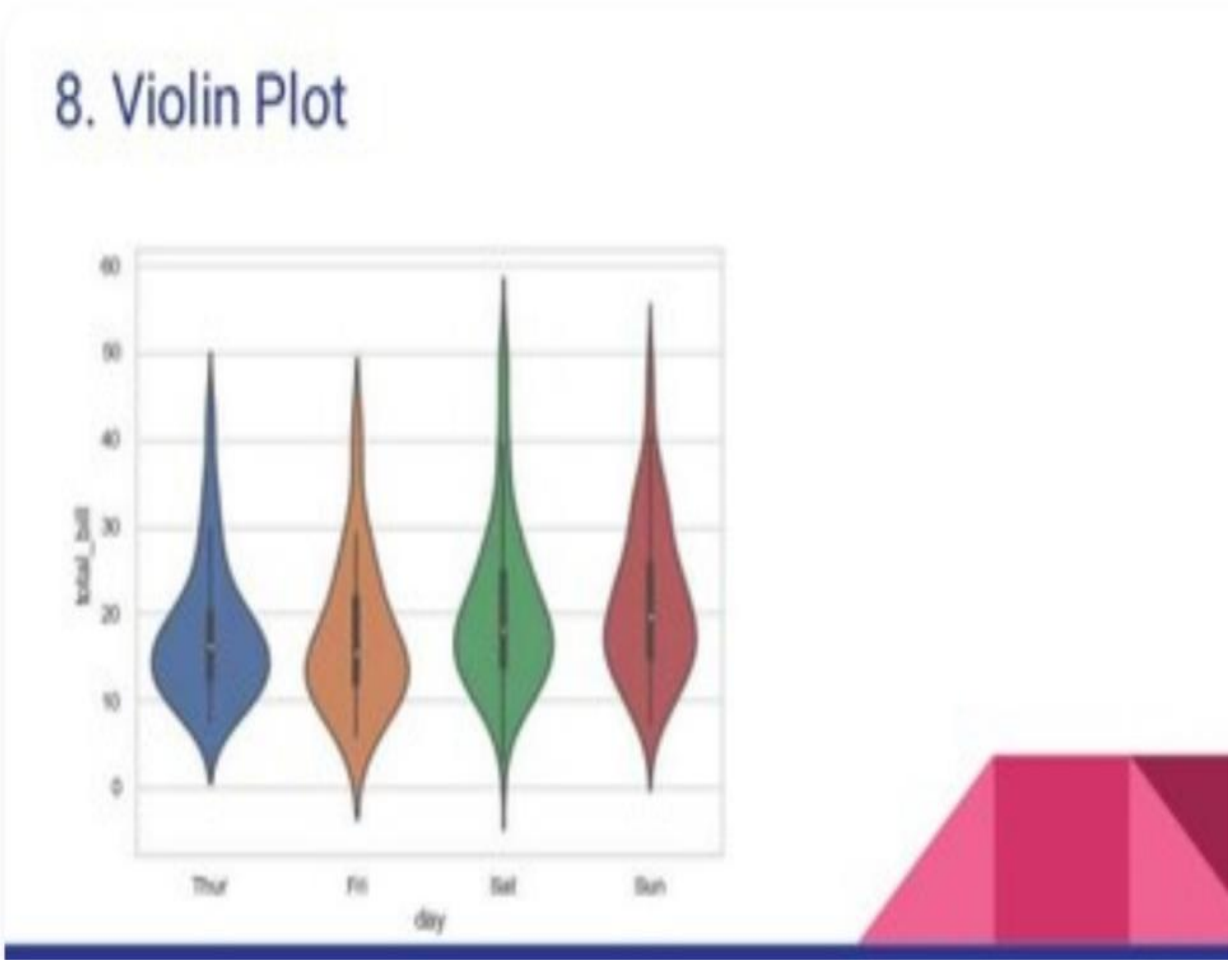
6. Box Plot

[illegible]
Creatinine (umol/L): Sex = M and Visit = Week 2
Analysis Value
180
150
120
90
60
N
33
33
41
Placebo
Xanomeline Low Dose
Xanomeline High Dose
Actual Treatment (N)

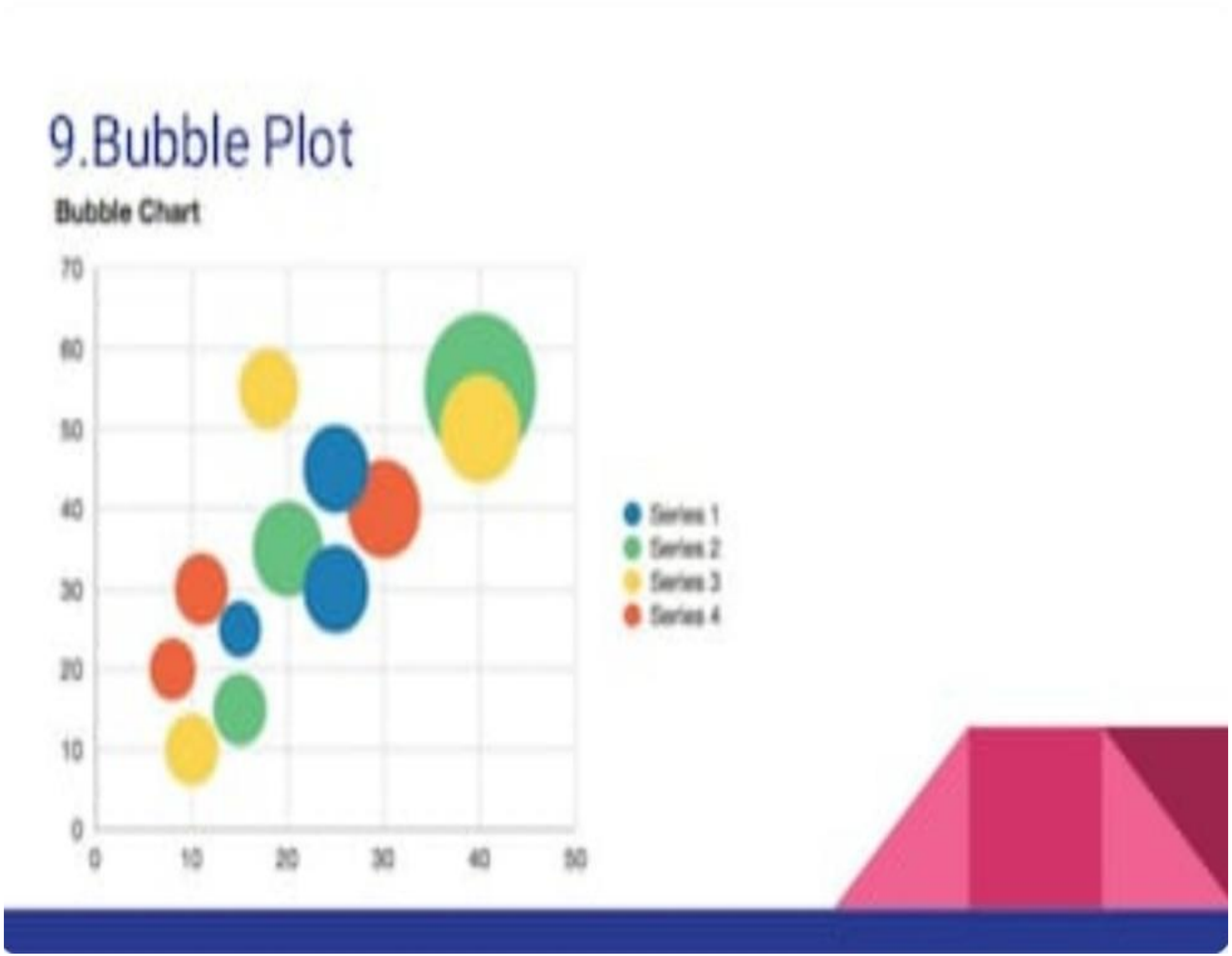


7. Line Plot

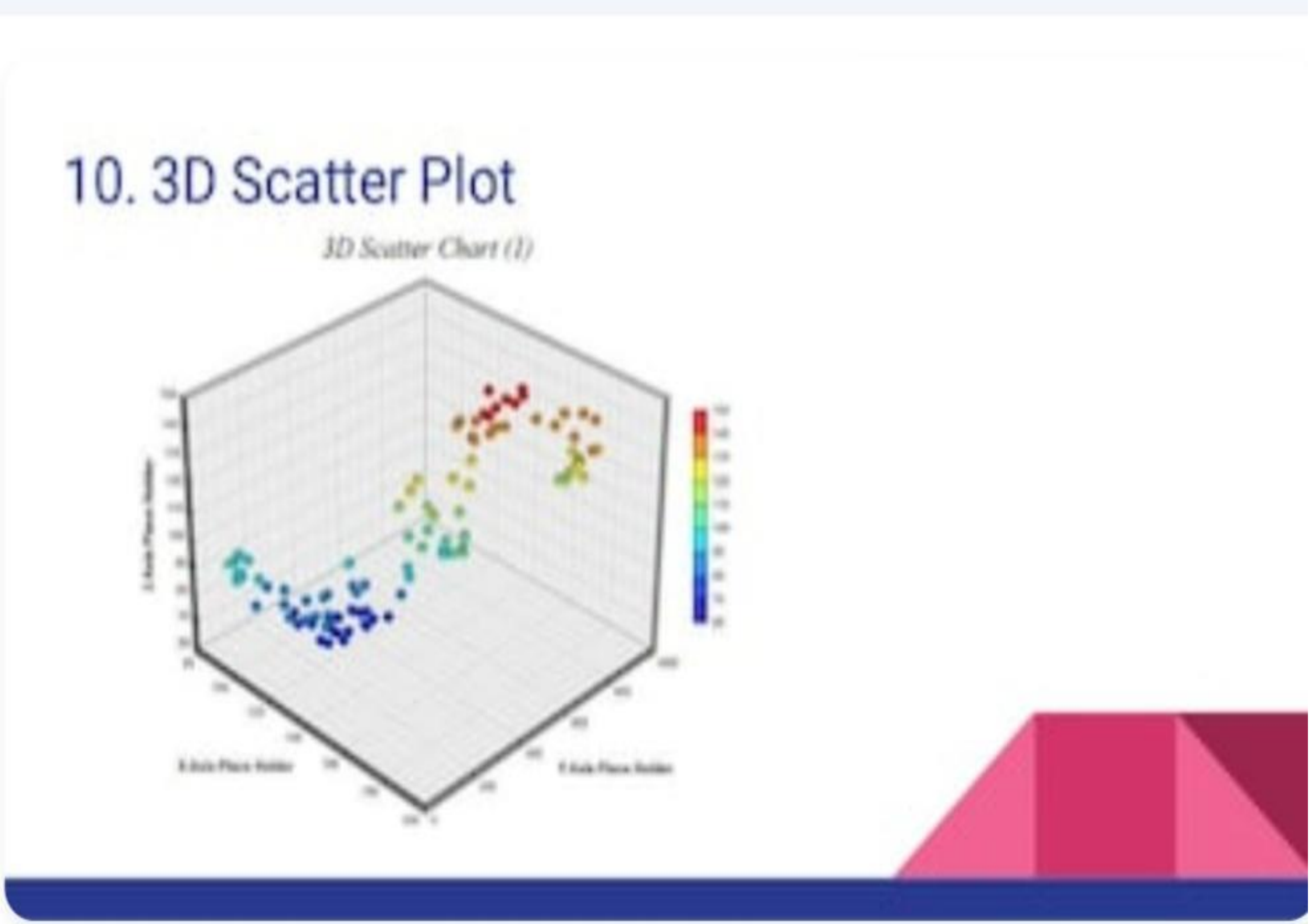
Product Trends by Month
140
120
100
80
60
40
20
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep



8. Violin Plot



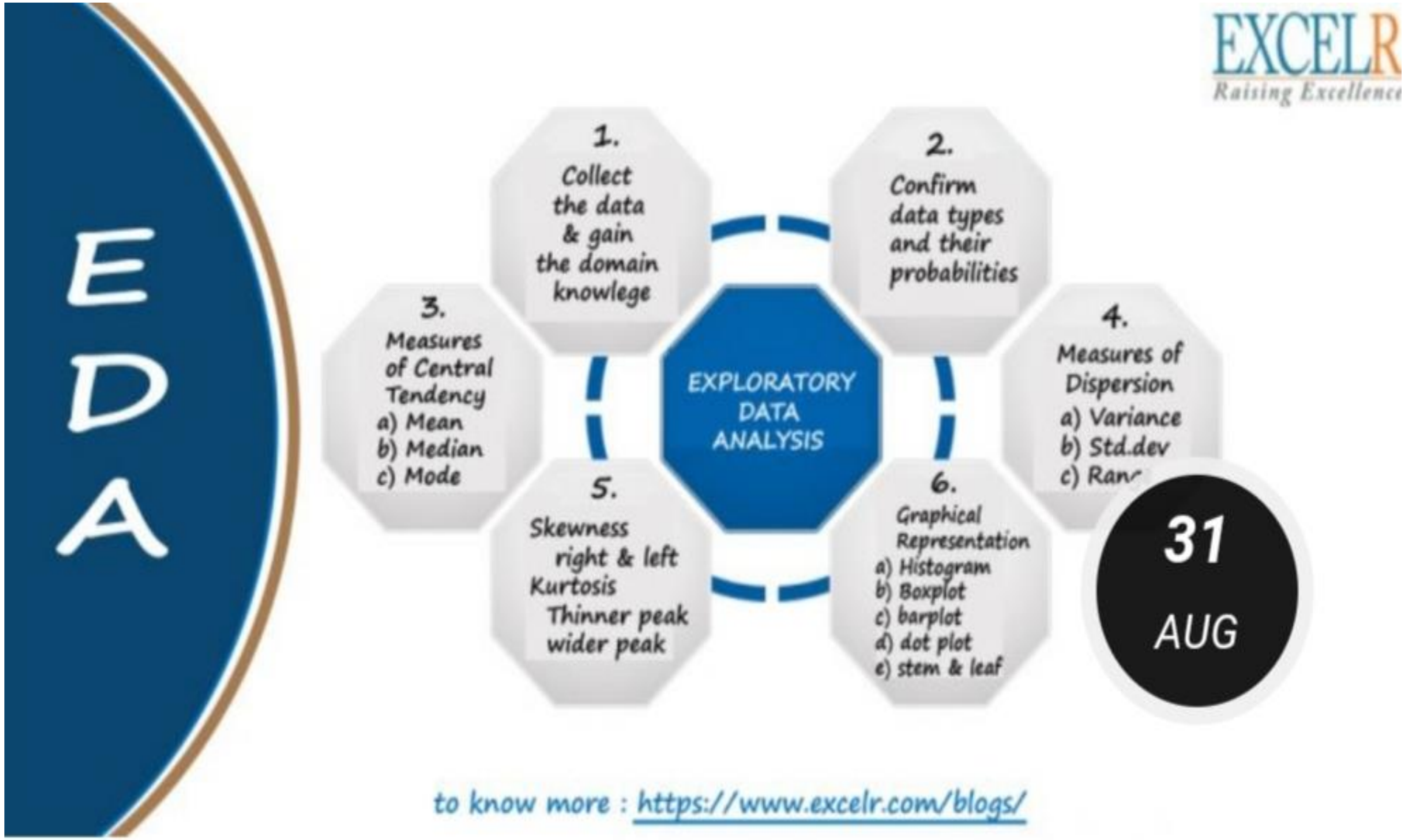
9.Bubble Plot



10. 3D Scatter Plot

3D Scatter Chart (1)

EDA steps and Visualization



EXCELR
Raising Excellence

EDA

EXPLORATORY DATA ANALYSIS

1. Collect the data & gain the domain knowlege

2. Confirm data types and their probabilities

3. Measures of Central Tendency
a) Mean
b) Median
c) Mode

4. Measures of Dispersion
a) Variance
b) Std.dev
c) Ran

5. Skewness right & left
Kurtosis Thinner peak wider peak

6. Graphical Representation
a) Histogram
b) Boxplot
c) barplot
d) dot plot
e) stem & leaf

31 AUG

to know more : https://www.excelr.com/blogs/

EDA steps and Assumptions

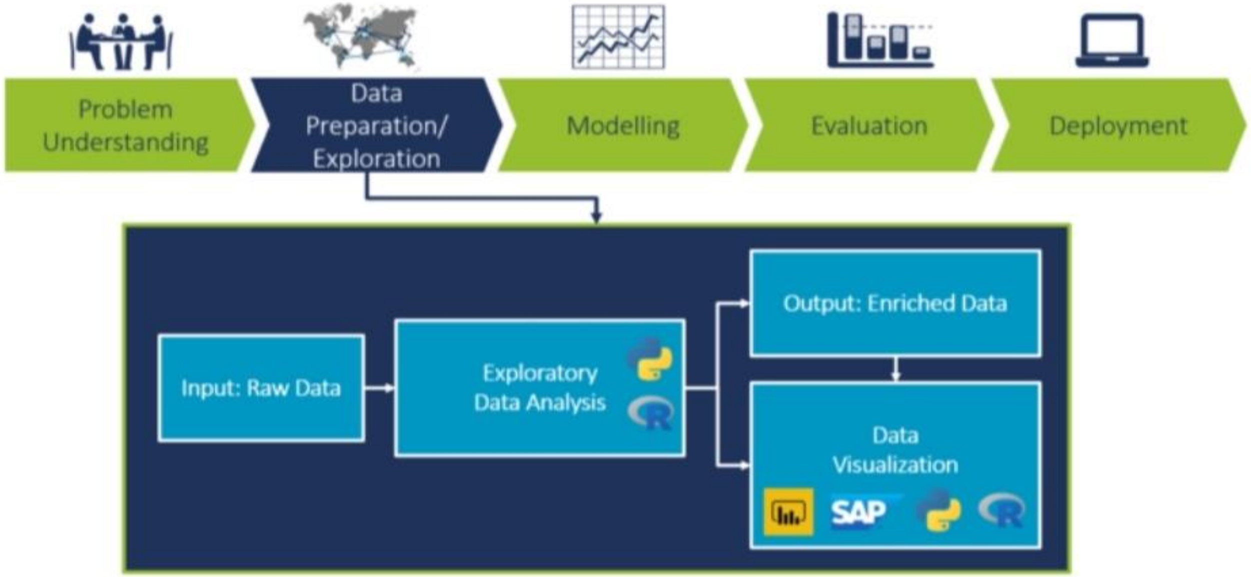
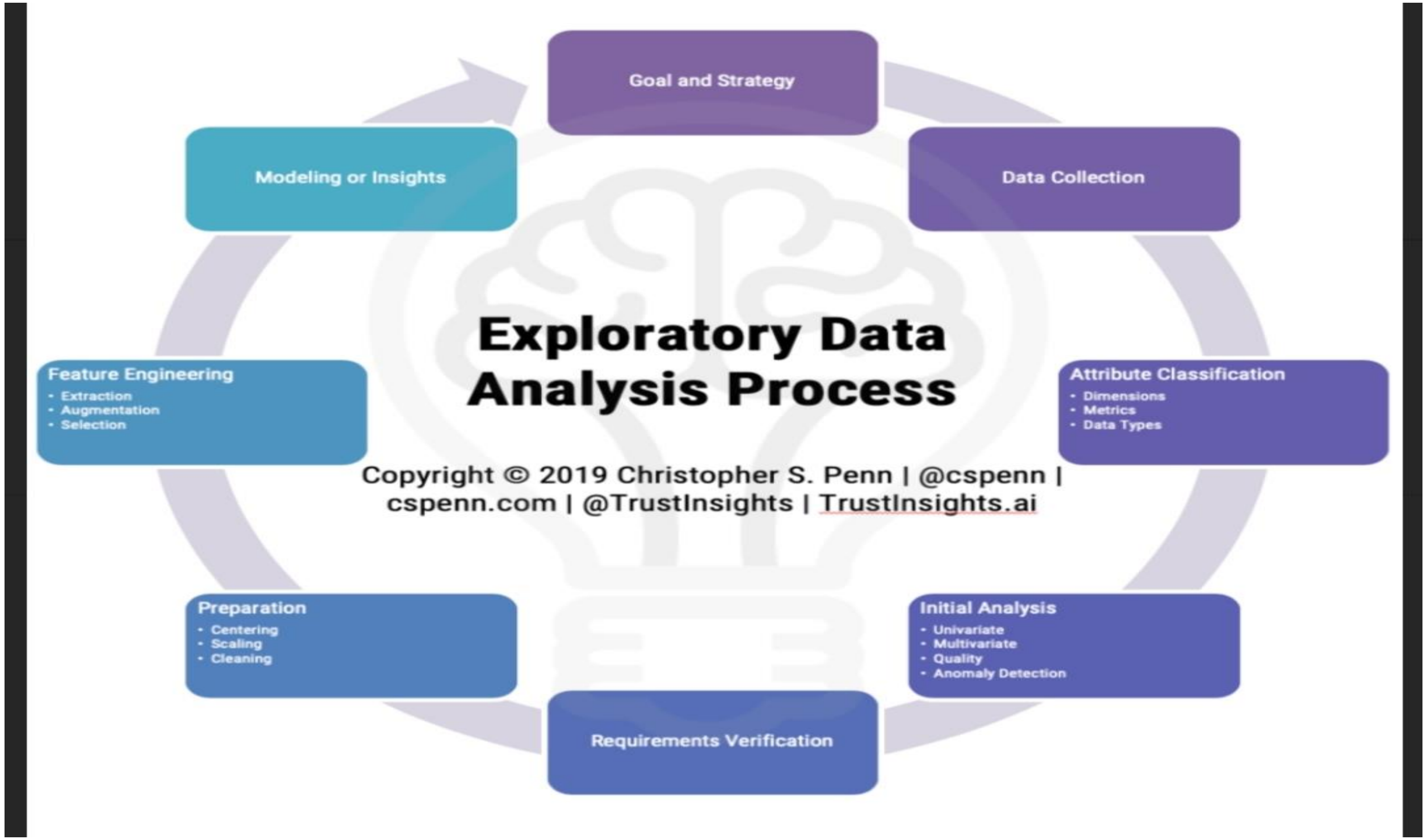


Fig. 1: Data Science Project Flow

EDA Analysis Process



Goal and Strategy

Data Collection

Attribute Classification
- Dimensions
- Metrics
- Data Types

Initial Analysis
- Univariate
- Multivariate
- Quality
- Anomaly Detection

Requirements Verification

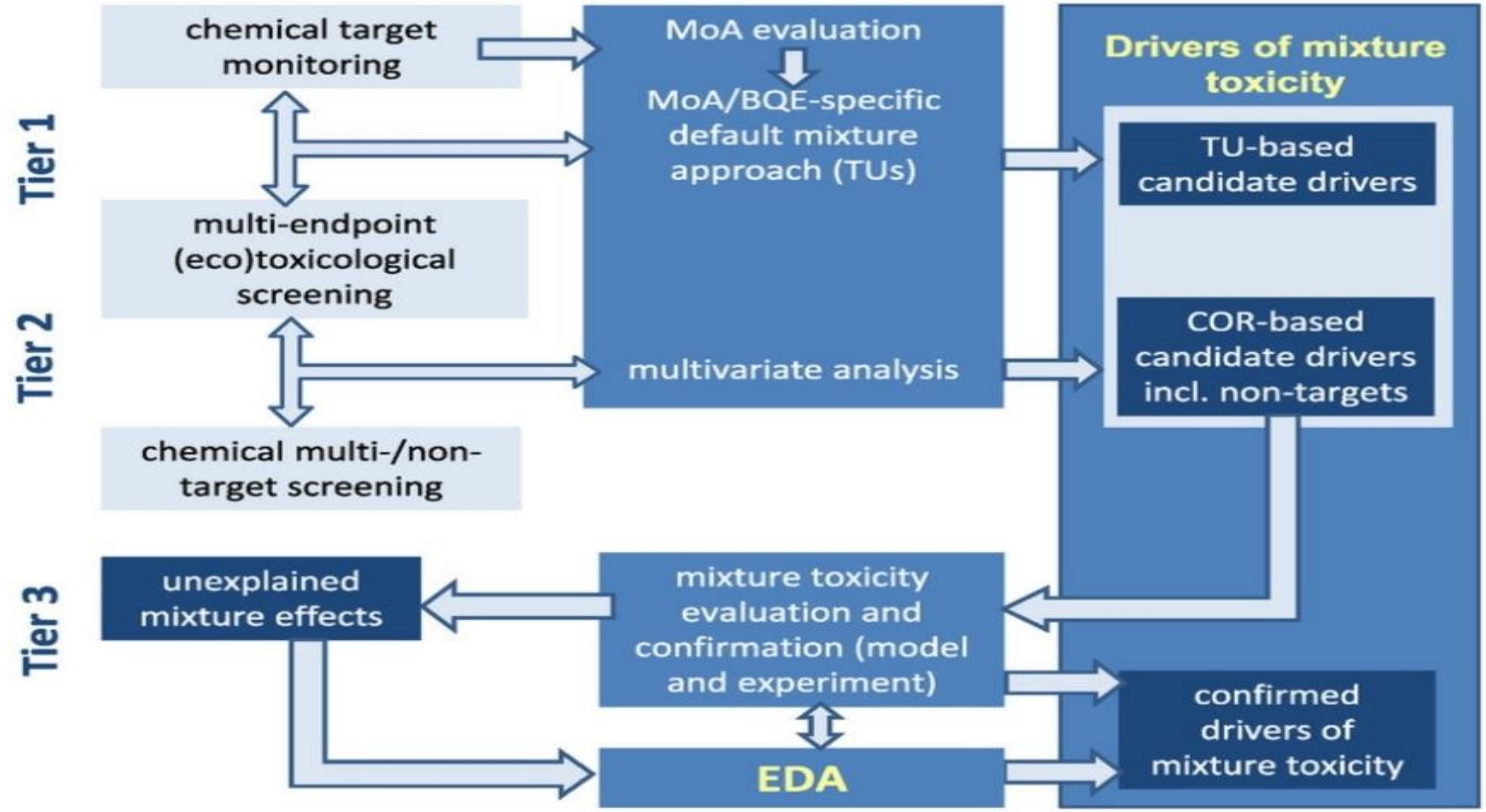
Preparation
- Centering
- Scaling
- Cleaning

Feature Engineering
- Extraction
- Augmentation
- Selection

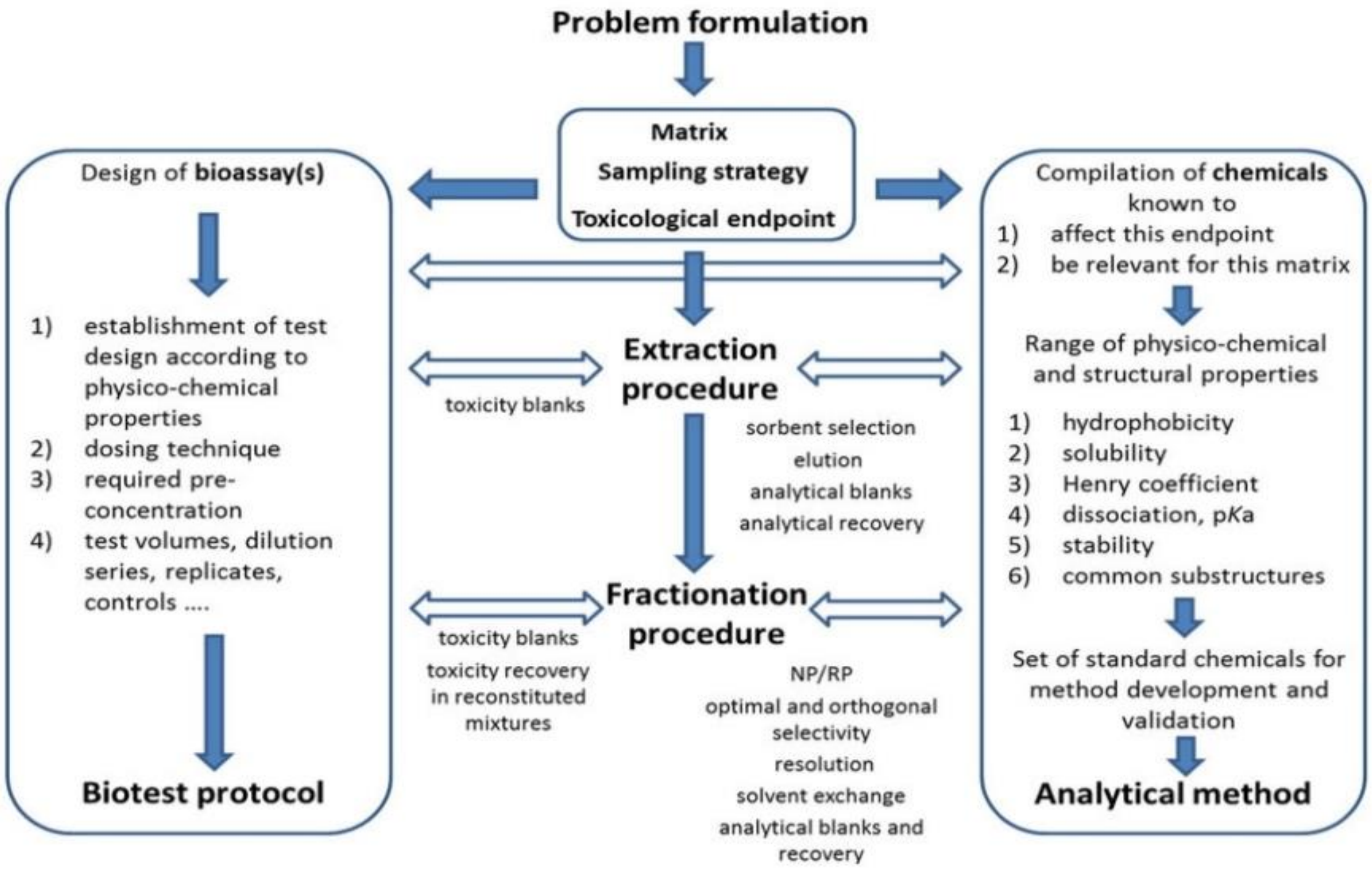
Modeling or Insights

Exploratory Data Analysis Process

EDA Concept


Tier 1
Tier 2
Tier 3
chemical target monitoring
multi-endpoint (eco)toxicological screening
chemical multi-/non-target screening
MoA evaluation
MoA/BQE-specific default mixture approach (TUs)
multivariate analysis
Drivers of mixture toxicity
TU-based candidate drivers
COR-based candidate drivers incl. non-targets
unexplained mixture effects
mixture toxicity evaluation and confirmation (model and experiment)
EDA
confirmed drivers of mixture toxicity

EDA Method



Problem formulation

Matrix

Sampling strategy

Toxicological endpoint

Design of **bioassay(s)**

1) establishment of test design according to physico-chemical properties
2) dosing technique
3) required pre-concentration
4) test volumes, dilution series, replicates, controls ....

Biotest protocol

toxicity blanks

Extraction procedure

sorbent selection

elution

analytical blanks

analytical recovery

toxicity blanks

toxicity recovery in reconstituted mixtures

Fractionation procedure

NP/RP

optimal and orthogonal selectivity

resolution

solvent exchange

analytical blanks and recovery

Compilation of **chemicals** known to

1) affect this endpoint
2) be relevant for this matrix

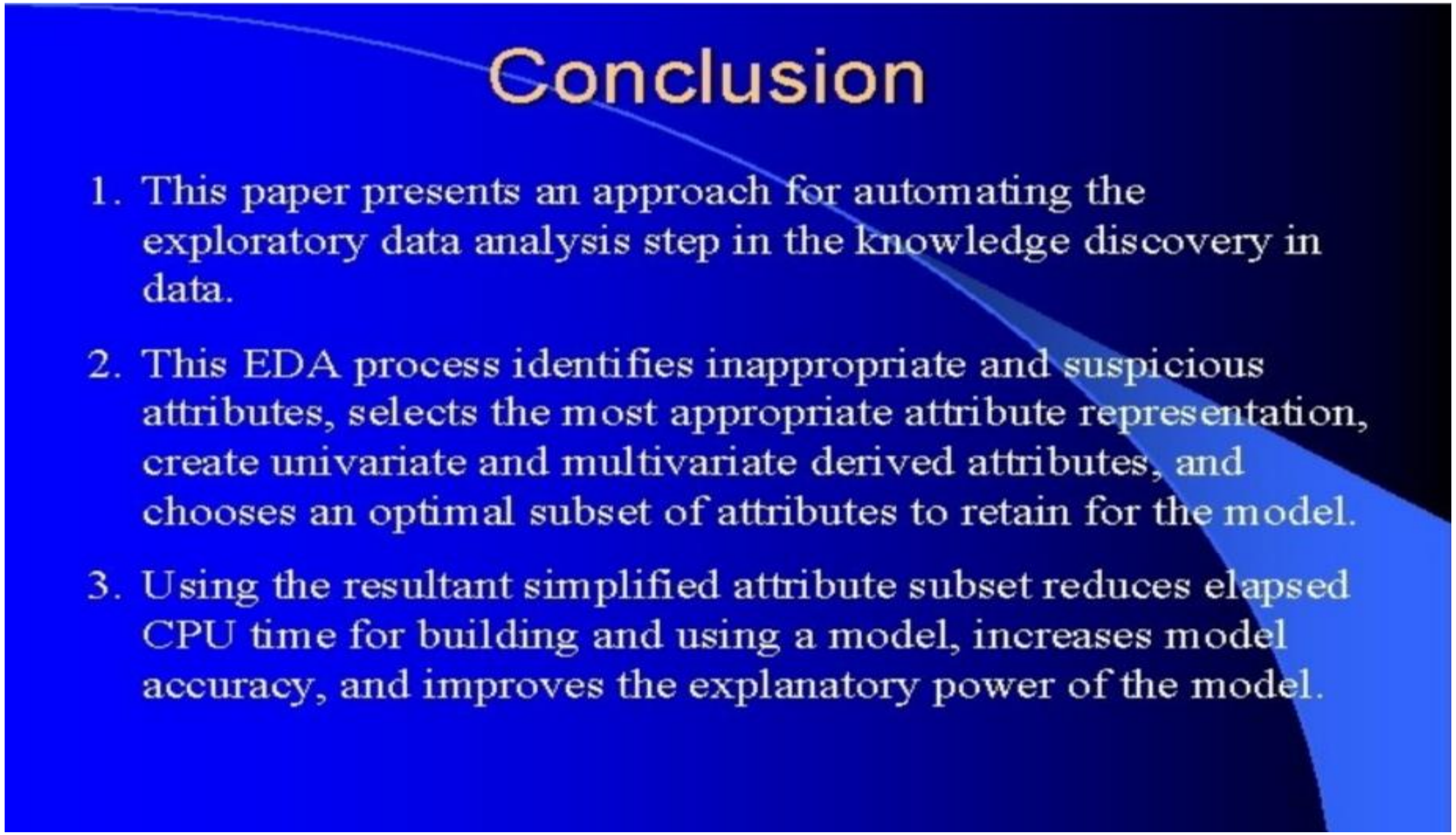
Range of physico-chemical and structural properties

1) hydrophobicity
2) solubility
3) Henry coefficient
4) dissociation, p*K*a
5) stability
6) common substructures

Set of standard chemicals for method development and validation

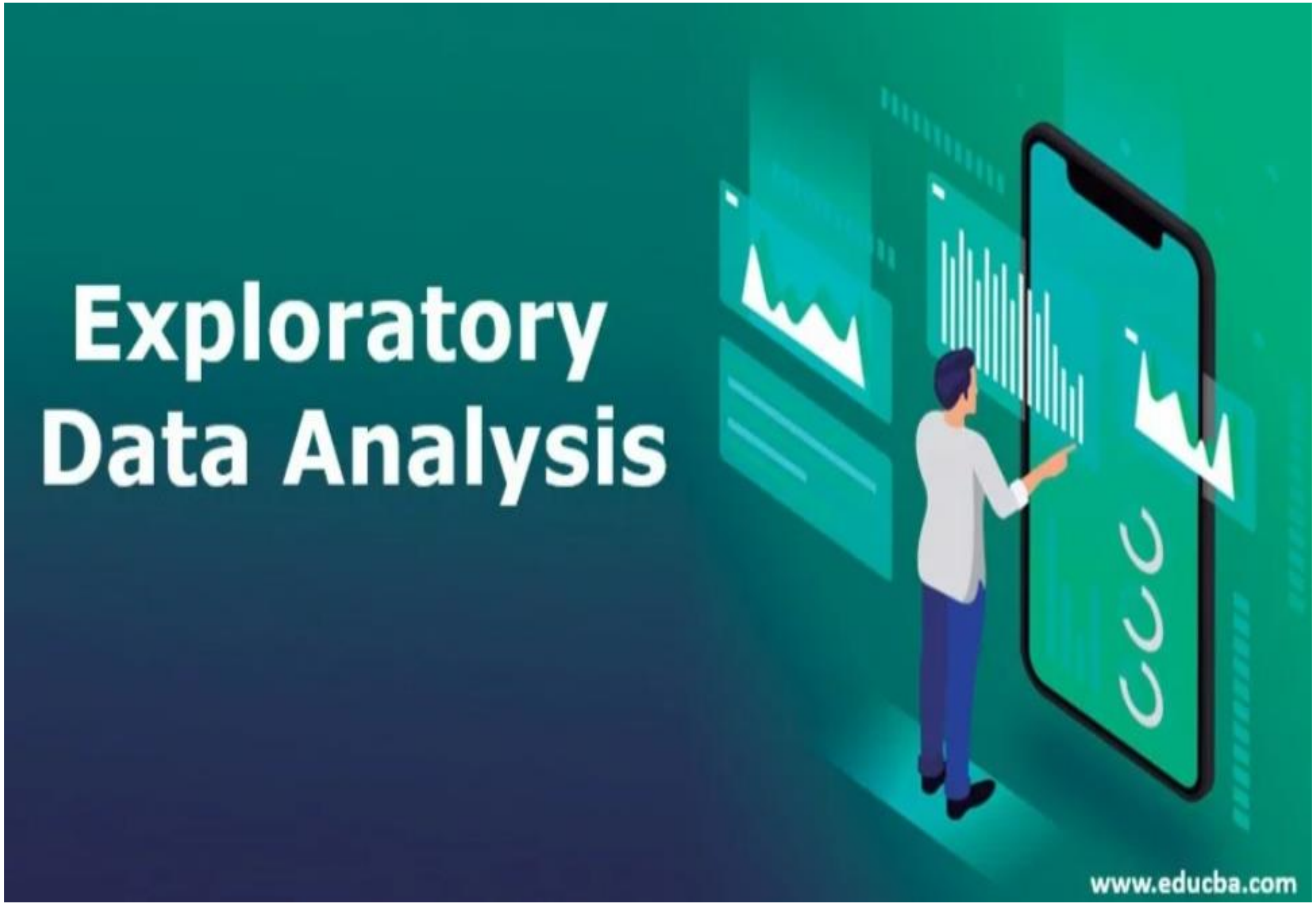
Analytical method

EDA Conclusion



Conclusion

1. This paper presents an approach for automating the exploratory data analysis step in the knowledge discovery in data.
2. This EDA process identifies inappropriate and suspicious attributes, selects the most appropriate attribute representation, create univariate and multivariate derived attributes, and chooses an optimal subset of attributes to retain for the model.
3. Using the resultant simplified attribute subset reduces elapsed CPU time for building and using a model, increases model accuracy, and improves the explanatory power of the model.


Exploratory
Data Analysis
www.educba.com