Course Outcomes (COs) & CO-PO Mapping (3-Strong; 2-Medium; 1-Weak Correlation)

COs	Upon completion of course the students will be able to	PO3	PO4	PO5	PO9	PO12	PSO2
CO1	create python shell script for data validation	3	3	3	3	3	3
CO2	demonstrate how to import data into tableau	3	3	3	3	3	3
CO3	apply the tableau concepts of dimensions and measures	3	3	3	3	3	3
CO4	develop programs, map visual layouts and graphical properties	3	3	3	3	3	3
CO5	create a dashboard that links multiple visualizations	3	3	3	3	3	3

List of Experiments

Week	Title/Experiment
	Data Exploration

DEV Lab – 24CSPC37

1	Data: understand, find, explore, cleanup (format, outliers, duplicates, normalize and standardize data).
2	Develop the python script to parse the pdf files using pdfminer.
3	Develop the python script for data cleanup on child labour and marriage data.xlsx a) check duplicates and missing data b) cleans line breaks, spaces and special characters.
4	Draw the chart between perceived corruption scores compared to the child labour percentages using matplotlib.
5	Write a python script to download & display content of robot.txt for en.wikipedia.org.
	Data Visualization
6	a) Create a first visualization with tableau software for data file formats b) Create basic charts (line, bar charts, tree maps) using the show me panel
7	a) Tableau calculations: sum, avg, aggregate, create custom calculations and fields. b) Visualizations: formatting, tools and menus, specific parts of the view for data calculations
8	Editing, formatting axes, manipulating data in tableau data and pivoting tableau data.
9	Tableau data: Structuring, sorting, filtering and pivoting.
10	Advanced visualization tools: a) using filters, detail panel, size panels b) customizing filters using tooltips, formatting data with colors.
11	Dashboards and storytelling: Create and design different interactive displays, distribute and publish the data.
12	Create custom charts, cyclical data and circular area charts, dual axis charts.

References

1. Data Exploration and Visualization Lab Manual, Dept. of CSE, CMRIT.

Micro-Projects: Student should submit a report on one of the following/any other micro-project(s) approved by the lab faculty before commencement of lab internal examination.

DATA	INFO'N
- Data is unorganized & unrefined facts.	- Infon comprises organised data presented in a meaningful context.
- Data is an individual unit that contains raw materials which do not carry any specific meaning.	- Infon is a group of data that collectively carries a logical meaning.
- Data doesn't depend on infon	- Infon depends on data.
- Raw data alone is insufficient for decision making.	- Infon is sufficient for decision making.
- Ex: stud test score.	- Ex:- Avg score of a class is the infon derived from given data

Data Exploration :-

- It is the 1st step in the journey of extracting insights from raw datasets.
- Sometimes also referred as "exploratory data analysis".

- Data expl'n is the 1st step in data analysis involving the tools of DV tools & statistical techniques to uncover dataset characteristics & initial patterns.
- D.E is the initial phase of data analysis where raw data is examined to understand its characteristics, identify patterns, & detect anamolies.
- It's a crucial step before dividing into more complex analysis or modeling.
- DE involves techniques like dv, statistical summaries, data profiling to gain insights into the Data's structure, quality & potential relationships.

Data Exploration & Visualization :-

Exploration Data Analysis EDA or DE is a process of examining the dataset to discover patterns, spot anomalies (odd / abnormal), test hypotheses, & check assumptions using statistical data.

- Why EDA ?

To examine what data can tell us before actually going through formal modeling or hypothesis formulation.

- Data analysis phases :-

- Data req's
- Data coll'n → organising & managing
- Data preprocessing → process of pre curating before analysis.
- Data cleaning → incomplete, duplicate, missing, error.
- EDA
- Modeling & alg'n

Significance of EDA

- Science, economics, engineering, marketing, db's.
- Data scientists use this EDA process to understand what type of modeling & hypotheses can be created.
- EDA is the 1st step in data mining to make further decisions from collected data.

steps in EDA :-

- prblm defn – obj of analysis, deliverables, roles, responsibilities, status of data, cost/benefit analysis
  - based on prblm defn, exec plan is created.
- Data preparation – we find sources of data, schemas, tables, understand data, clean, delete non relevant, transform, divide for analysis.
- Data analysis – deals with descriptive statistics & analysis of the data. Main tasks are summarizing data, finding hidden correlation & r/ships among data, developing predictive models, evaluating models & calculating accuracies.
- Development & representation of the results :-
  presenting dataset to target audience in form of graphs, summary tables, maps & diagrams

scatter plots, character " , histograms, box plot, residual plot, mean plot.

- Understand / types of data → numerical / categorical

1. **Numerical data** — "QUANTITATIVE DATA"

Ex:- heartrate, temp, no. of teeth, weight, BP, etc

- Numerical data can be :

discrete	continuous
• values can be listed out. 1, 2, 3, 4, 5 & so on.	• infinite no. of numerical values with specific range. Ex: temp of our city today

2. **categorical data** represents characteristics of an obj.

Ex: gender, marital status, type of addr, movie categories, types. etc. This data is often referred to as "QUALITATIVE dataset"

- Dichotomous variable / binary categorical variable :- can take exactly 2 values.
- Polytomous variables :- are categorical var's that can take more than 2 possible values.

Most of the categorical dataset follows either nominal or ordinal measurement scales.

- Measurement scales :- 4 types of measurement scales described in statistics.

1. nominal
2. ordinal
3. interval
4. ratio

1. nominal = generally referred to as 'labels'
- These are practiced for labelling variables without any quantitative value.
- values are mutually exclusive & do not carry any numerical importance.
- To visualise the nominal dataset, we can either use a pie chart or a bar chart.

Ex:- Gender - male, female, prefer not to ans

In case of nominal dataset, we can know the folling:-

frequency: is the rate at which a label occurs over a period of time within the dataset.

Proportion: can be calculated by dividing the frequency to tal no. of events

%: % of each proportion can be computed.:

2. Ordinal = main difference in nominal & order is the ORDER. the order of values is a significant factor. We can consider ordinal scales, as an order of ranking (1st, 2nd, —).

Median item is allowed as the measure of central tendency. But, avg is not permitted.

Ex:- very likely, unhappy, ok, happy.

3. Interval = In interval scales, both the order & exact differences b/n values are significant. Ex:- time, temp. [Mean, Median, Mode are allowed on interval data]

4. Ratio:

* Ratio scales contain order, exact values, & absolute zero, which makes it possible to be used in descriptive & inferential statistics.

Example of nominal & ordinal data, discrete & continuous data

cust name	item	size	quantity	prize
King	shirt	S	2	499.99
Mahi	Pant	M	3	1599.45
Pavan	Jeans	XL	4	3999.99
			↓ Discrete data	↓ continuous data.
nominal data		ordinal data		

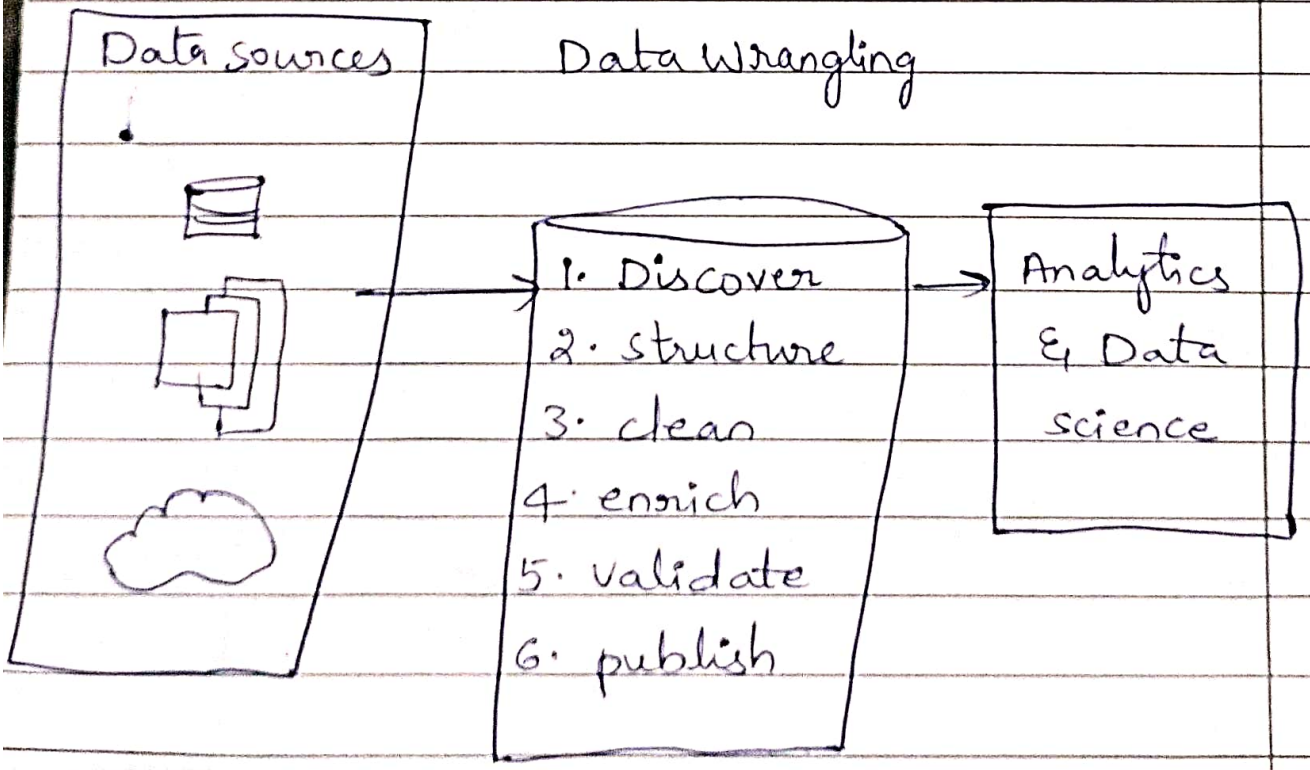
↓

found in public datasets, API's, web scraping, govt's & pvt data, research institutions

Data Wrangling focuses on cleaning, transforming, & structuring the data into a usable format for analysis. ie., a preprocessing technique.

- they are not sequential steps but rather an iterative process, where insights from exploration guide wrangling efforts, & wrangling refines data for further exploration.
- not a 1 time task, its an iterative process that often involves revisiting & refining the data based on insights gained during data exploration.
- Also known as "Data Munging".

Steps / process of Data Wrangling :-



1. Discover :-

- understand & explore data, this involves identifying data sources, assessing data quality & gaining insights into the structure & format of your data.
- Goal is to establish a foundation for the sub sequent data preparation steps by recognizing potential challenges & opportunities in the data.

2. Structure :-

- In data structuring, we organise & format the raw data in a way that facilitates efficient analysis.
- structuring involves reshaping data, handling missing values, & converting data types.
- ensures data is presented in a coherant & standarized manner, laying the groundwork for further manipulation & exploration.

3. clean :-

- Data cleansing is a crucial step to address inconsistencies, errors, & outliers within dataset.
- focus is on enhancing data accuracy & reliability for downstream processes.

4. enrich :- includes merging datasets, extracting relevant features.
goal : to augment the original dataset, making it more comprehensive & valuable for analysis.
(If you add data, structure & clean & then add)

5. validate : ensures quality & reliability of your processed data. We'll check for inconsistencies, verify data integrity & confirm that data adheres to pre defined standards.
- validation helps in meeting the req's for meaningful analysis.

6. Publish :
- By this step, curated & validated dataset is prepared for analysis.
- involves documenting data lineage.
- publishing facilitates coll'n & allows others to use the data for their analysis or decision making processes.

+ves :- improved data quality & consistency, increased analysis efficiency & support for EDA, facilitates data integration & adaptability, preparation for ML, supports decision making & reduces error.

Finding Outlier Example :-

Q) 58, 89, 56, 58, 67, 70, 124, 4, 58, 2

Sol:-

step 1 : sort

2, 4, 56, 58, 58, 58, 67, 70, 89, 124.

step 2 : find Quartiles

Q1 (first Quartile) = Median of lower half

Q2 (Median) = Middle value

Q3 (Third Quartile) = Median of upper half

which means

Q1 = 2, 4, 56, 58, (58), 58, 67, 70, 89, 124

low half Mid upper half

∴ Q1 = 2, 4, (56), 58, 58, 58

$\frac{6}{2}$ (3)rd position

∴ Q1 = lower half Median

Q1 = 56

Q2 = Median (or) mid value

= 2, 4, 56, 58, (58), 58, 67, 70, 89, 124

10/2 = 5th position

5/2 = 2.5

∴ Median = total no's / 2

= 10/2 = 5th place ie, 58 BUT

- now, since total count is even (10), there is no single middle numbers. So, we take avg of 2 middle no's ie, 5th, 6th values.
- If its odd, directly middle numbers.
- here, median = $\frac{58+58}{2} = \frac{116}{2} = 58.$

Q3 = upperhalf's mid value

= 58, 67, 70, 89, 124

Q3 = 70

step3 :- calculate IQR (Inter Quartile Range):

$\therefore IQR = Q3 - Q1$

$= 70 - 56$

$IQR = 14$

step4: calculate lower & upper limits.

$\therefore$ Lower Limit $= Q1 - 1.5 \times IQR$

$= 56 - 1.5 \times 14$

lower Limit $= 56 - 21 \Rightarrow 35$

$\therefore$ Upper limit $= Q3 + 1.5 \times IQR$

$= 70 + 1.5 \times 14$

$= 70 + 21 \Rightarrow 91$

step 5 :- Identify outliers

- Any value < 35 or > 91 is considered OUTLIER.

from dataset; 2 & 4 are < 35, so outlier

124 > 91, so outlier.

$\therefore$ 2, 4, 124 = outliers.

↑ process of scaling no's to common range.

Normalization:- Common method of normalization is Min Max scaling.

It scales the data to a fixed range, usually [0,1]: and (-1,1).

∴ Min-Max normalization $x' = \frac{x - X_{min}}{X_{max} - X_{min}}$

x = original value

X_{min} = min value in dataset

X_{max} = max value in dataset

Example :- 200, 300, 400, 600, 1000.

$X_{min} = 200$, $X_{max} = 1000$

$x' = \frac{200 - 200}{1000 - 200} \Rightarrow \frac{0}{800} \Rightarrow 0$

$x' = \frac{300 - 200}{1000 - 200} = \frac{100}{800} = 0.125$

$x' = \frac{400 - 200}{1000 - 200} = \frac{200}{800} = 0.25$

$x' = \frac{600 - 200}{1000 - 200} = \frac{400}{800} = \frac{1}{2} = 0.5$

$x' = \frac{1000 - 200}{1000 - 200} = \frac{800}{800} = 1$

8) 100 (0.125
8
20
16
40
40
0

4) 10 (0.25
8
20
20
0

2) 10 (0.5
10
0

Standardization :- A data preprocessing technique that transforms features in a dataset to have a mean of 0 & SD 1.

- used to handle outliers, useful when data follows normal distribution.
- Also called 'Z-score normalization'; $Z = \frac{X - \mu}{\sigma}$

X = actual value
μ = mean of data
σ = standard deviation of data.

Example :- 200, 300, 400, 600, 1000.

$$\text{Mean } \mu = \frac{\Sigma x_i}{N}$$

$$\mu = \frac{200+300+400+600+1000}{5} = \frac{2500}{5} = 500$$

$$\text{S.D } \sigma = \sqrt{\frac{\Sigma (X_i - \mu)^2}{N}}$$

$(200 - 500)^2 \Rightarrow (-300)^2 \Rightarrow 90,000$

$(300 - 500)^2 \Rightarrow (-200)^2 \Rightarrow 40,000$

$(400 - 500)^2 \Rightarrow (-100)^2 \Rightarrow 10,000$

$(600 - 500)^2 \Rightarrow (100)^2 \Rightarrow 10,000$

$(1000 - 500)^2 \Rightarrow (500)^2 \Rightarrow 2,50,000$

$$= \sqrt{\frac{90,000 + 40,000 + 10,000 + 10,000 + 2,50,000}{5}}$$

$$= \sqrt{\frac{4,00,000}{5}} \Rightarrow \sqrt{80,000} \Rightarrow 282.5$$

$$Z = \frac{X - \mu}{\sigma}$$

$X = 200$; $Z = \frac{200 - 500}{282.5} \Rightarrow \frac{-380}{282.5} \Rightarrow -1.06$

$X = 300$; $Z = \frac{300 - 500}{282.5} \Rightarrow \frac{-200}{282.5} \Rightarrow -0.707$

$X = 400$; $Z = \frac{400 - 500}{282.5} \Rightarrow \frac{-100}{282.5} \Rightarrow -0.354$

$X = 600$; $Z = \frac{600 - 500}{282.5} \Rightarrow \frac{100}{282.5} \Rightarrow 0.354$

$X = 1000$; $Z = \frac{1000 - 500}{282.5} \Rightarrow \frac{500}{282.5} \Rightarrow 1.77$

X	Z-score
200	-1.06
300	-0.707
400	-0.354
600	0.354
1000	1.77

here, values less than 500 have -ve Z-score. values above mean have tve Z-scores

outlier = 1000, because its SD is 1.77 which means above the mean, indicating outlier.