Click here to add text

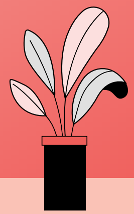
Introduction to Data Analytics

IBM Data Analyst Course: Course One, Week Three

AGENDA

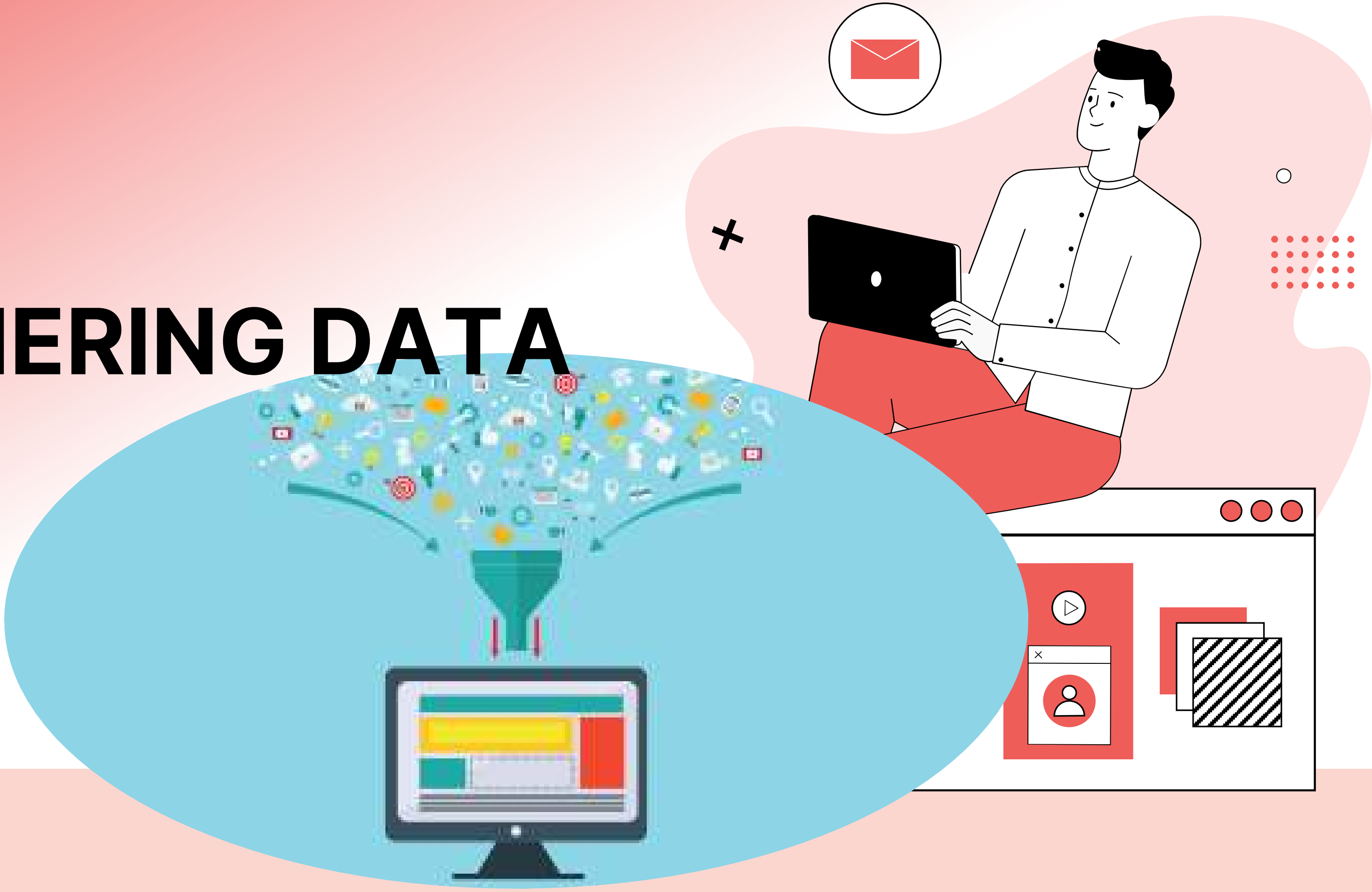
Gathering Data

Wrangling Data

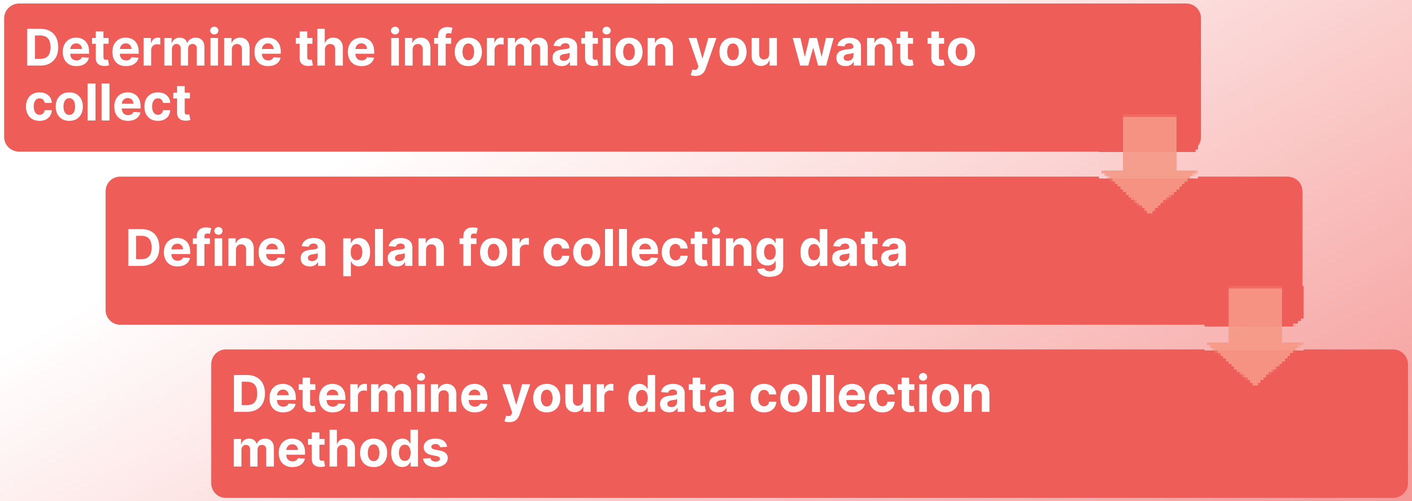


01

GATHERING DATA



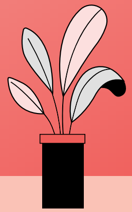
PROCESS FOR IDENTIFYING DATA



Determine the information you want to collect

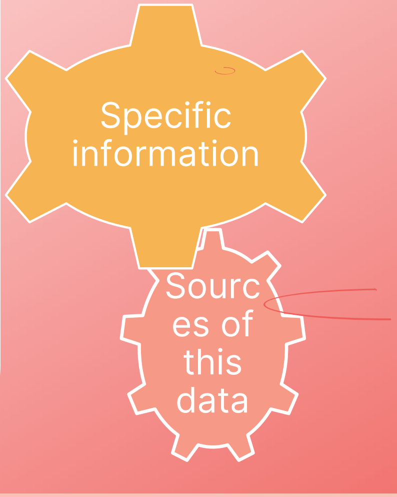
Define a plan for collecting data

Determine your data collection methods



Determine the information you want to collect

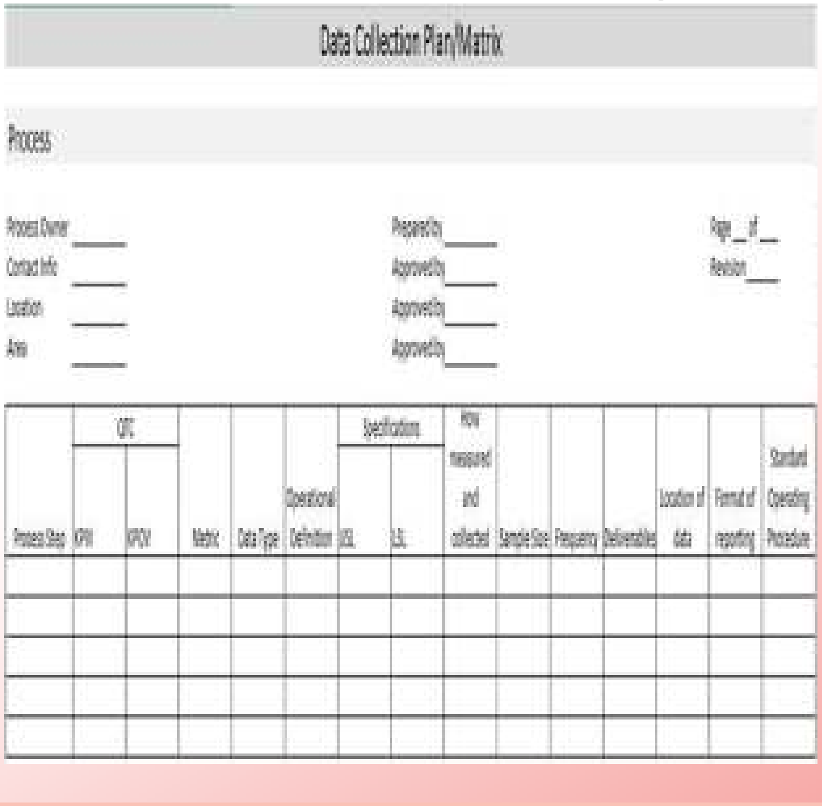
AUGUST
KEEP TRACK OF YOUR FINANCES
TAKE A BREAK
SET GOALS
TIPS ON IMPROVING YOUR SMALL BUSINESS
PERFECT YOUR BUSINESS PRESENTATION
MONITOR TRENDS
the balance



Specific information

Sources of this data

Define a plan for collecting data



Data Collection Plan/Matrix

Process

Process Owner ______ Prepared by ______ Page __ of __

Contact Info ______ Approved by ______ Revision ___

Location ______ Approved by ______

Area ______ Approved by ______

	CTC					Specifications								
Process Step	KPIV	KPOV	Metric	Data Type	Operational Definition	USL	LSL	How measured and collected	Sample Size	Frequency	Deliverables	Location of data	Format of reporting	Standard Operating Procedure

Establish a timeframe

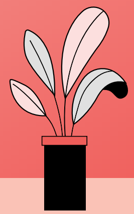
How much data is sufficient for credible analysis?

Define dependencies, risk and mitigation plan

Determine your data collection methods

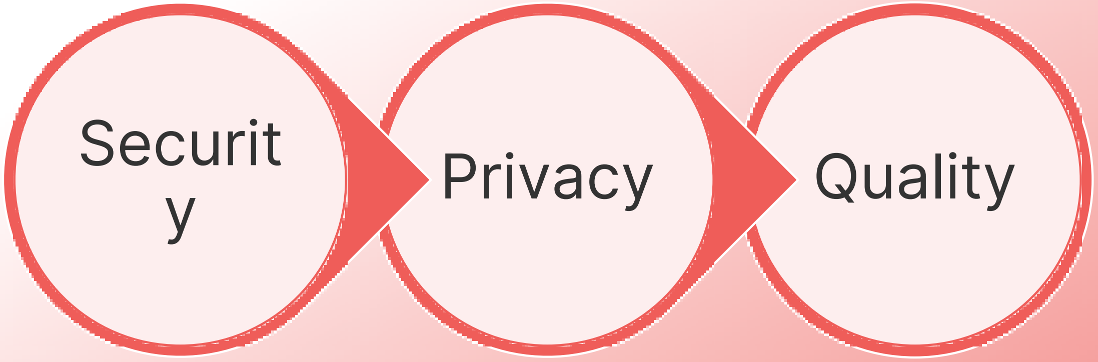
The methods of data collection depends upon a number of factors:

- Source of Data
- Type of Data
- Timeframe over which you need data
- Volume of Data
- Degree of accuracy required
- Funds

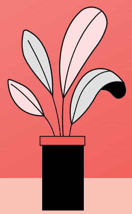


Key Consideration while collecting data

The data you identify , the source of that data and the practices you employ for gathering the data have implications for



In order for data to be reliable , it should be free of errors, complete, consistent, relevant and accessible.

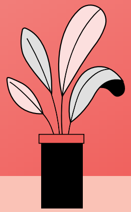


Data Sources

Primary

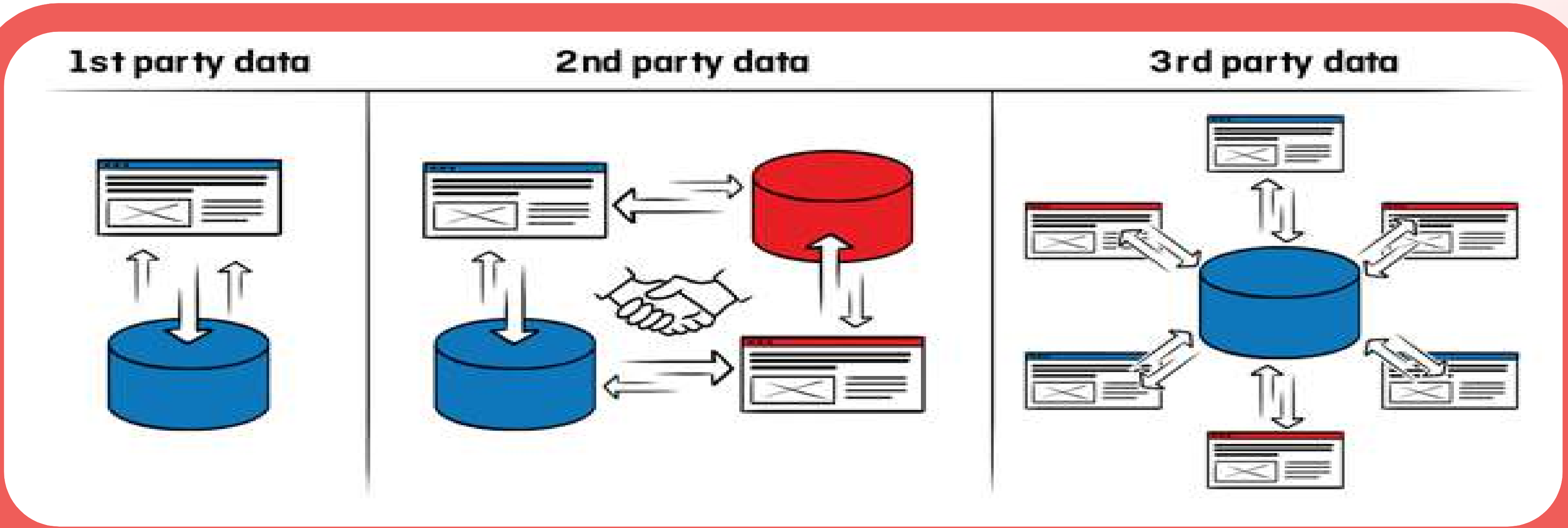
Secondar
y

Third
Party



Slide 9

1 geetu sodhi, 5/27/2021



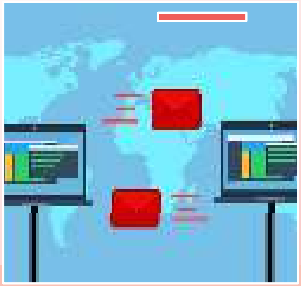
Primary data refers to information obtained directly from the source. Data from organization’s CRM, HR or workflow applications. This can be obtained directly from surveys, focus groups , interviews or questionnaires

Secondary data refers to the information obtained from existing sources such as external databases, research articles, training material . This can also be obtained directly from surveys, focus groups , interviews or questionnaires .

Third party refers to data which is purchased from aggregators who collect data from various sources and combine it into comprehensive datasets.

Sources for gathering data

Some of the data sources from which you could be gathering data include databases, the web, social media, interactive platforms, sensor devices, data exchanges, surveys and observation studies.



ACTIVITY

Suggest a correct source of gathering data for below specified scenarios:

- **Scenario 1:**

A marketing company is interested in the proportion of people who will buy a particular product.

- **Scenario 2:**

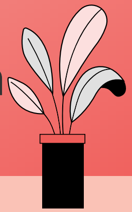
A Community College instructor is interested in the mean number of days math students are absent from class during a quarter.

- Scenario 3:

John requires an information about mechanical, orderly tasks, like checking the number of manual interventions required in a day to keep an assembly line functioning smoothly.

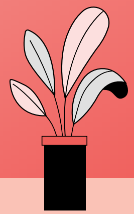
- Scenario 4:

Sneha wants to launch her online store. She wants to run a quick analysis through which she can decide on the type of online purchases made by users frequently.

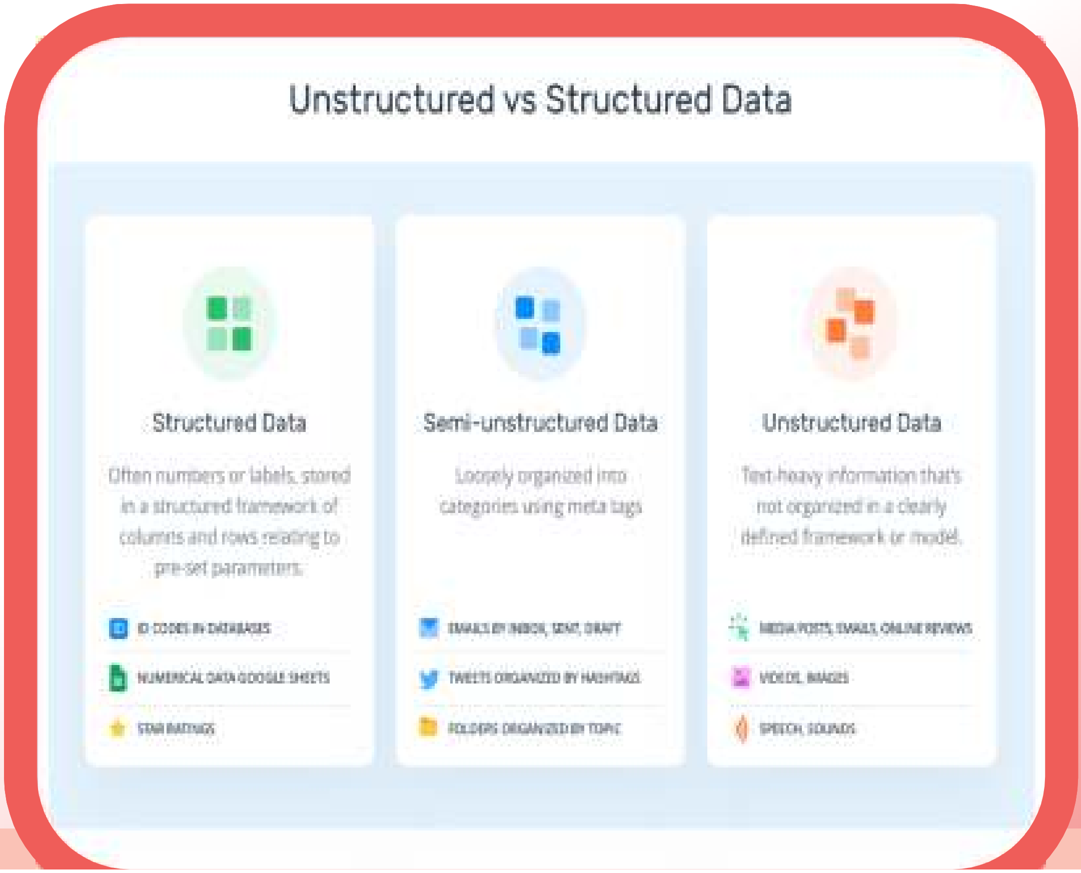


How to gather and import data?

- SQL Queries
- API
- Web Scraping / Web Harvesting
- Data Streams
- Data exchanges
- Other



Import Data



Unstructured vs Structured Data

Structured Data

Often numbers or labels, stored in a structured framework of columns and rows relating to pre-set parameters.

- ID CODES IN DATABASES
- NUMERICAL DATA GOOGLE SHEETS
- STAR RATINGS

Semi-unstructured Data

Loosely organized into categories using meta tags

- EMAILS BY INBOX, SENT, DRAFT
- TWEETS ORGANIZED BY HASHTAGS
- FOLDERS ORGANIZED BY TOPIC

Unstructured Data

Text-heavy information that's not organized in a clearly defined framework or model.

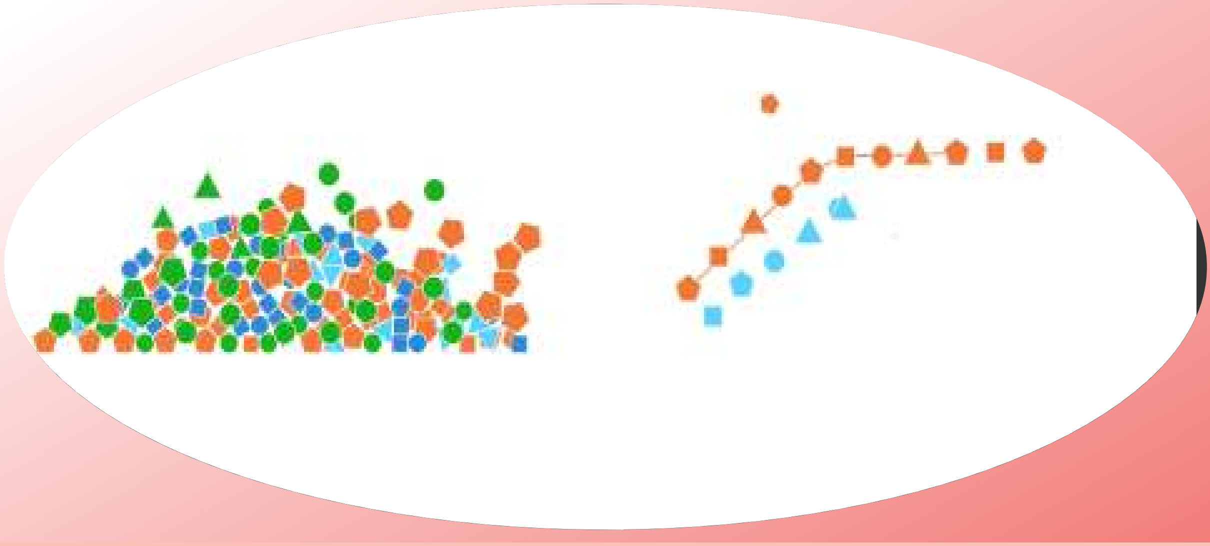
- MEDIA POSTS, EMAILS, ONLINE REVIEWS
- VIDEOS, IMAGES
- SPEECH, SOUNDS

Structured data can be stored in relational databases with well defines schemas

Semi Structured data can be stored in noSQL clusters

Unstructured data can be stored in noSQL databases or datalakes.

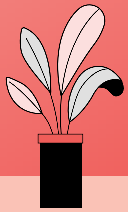
DATA WRANGLING



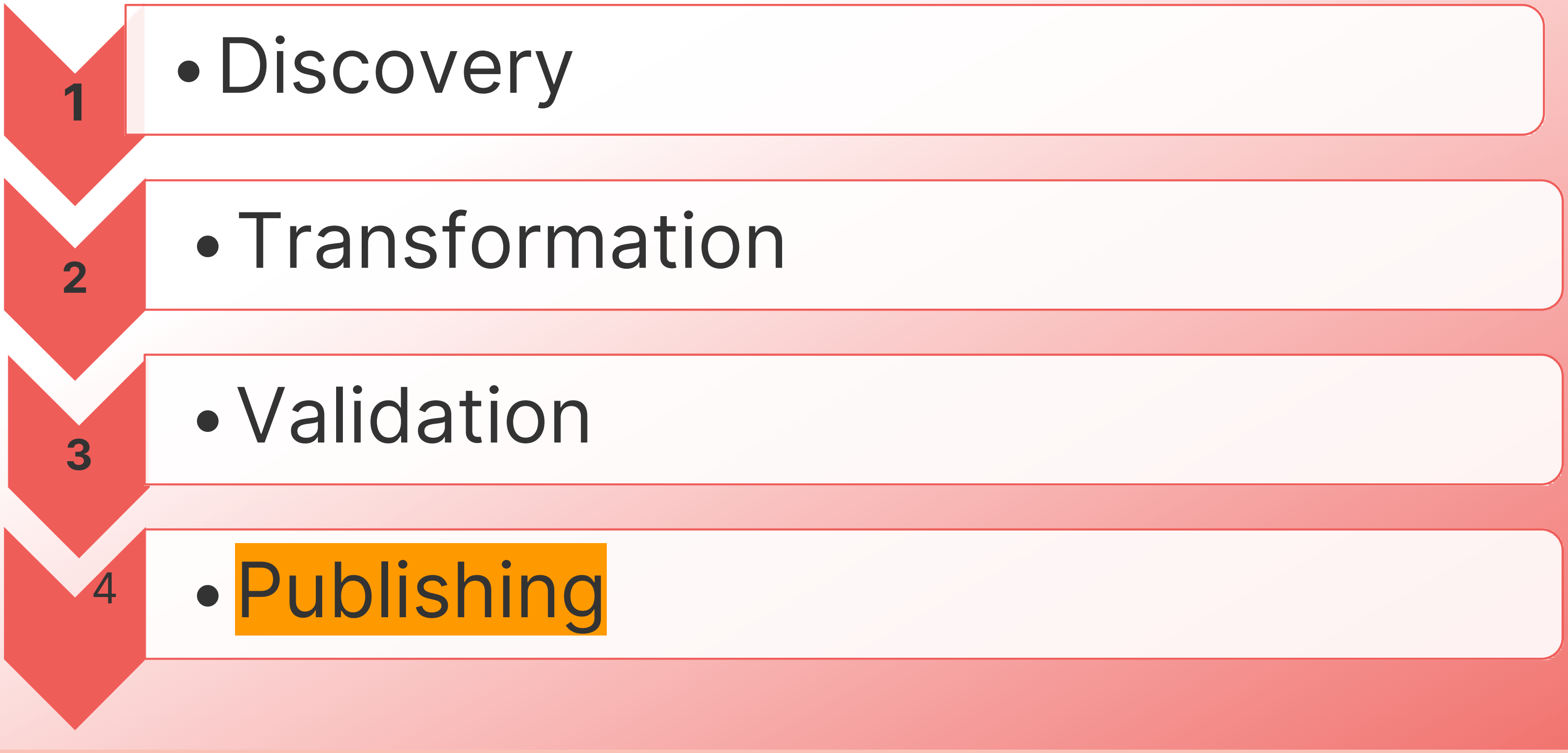
What is Data Wrangling?

Data wrangling, sometimes referred to as **data munging**, is the process of transforming and mapping data from one "raw” data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics. The goal of data wrangling is to assure quality and useful data.

It is a 4-step process that involves—Discovery, Transformation, Validation, and Publishing.



The Data Wrangling Process



1

- Discovery

2

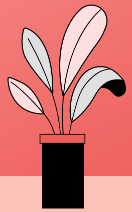
- Transformation

3

- Validation

4

- Publishing



Discover
y

The **Discovery phase**, also known as the **Exploration phase**, is about understanding your **data** better with respect to your use case.

Creating a plan for cleaning, structuring, organizing and mapping your data.

Transformation

Transformation phase, forms the bulk of the **data wrangling** process. It involves the task to **transform** the data, such as

structuring,

normalizing, de normalizing,

cleaning, and

enriching data.

Structuring the data

Structuring refers to actions that change the form or schema of your data. Splitting columns, pivoting rows and deleting fields are all forms of structuring.

Join and Union are most common structural transformation to combine data from one or more tables.

JOIN

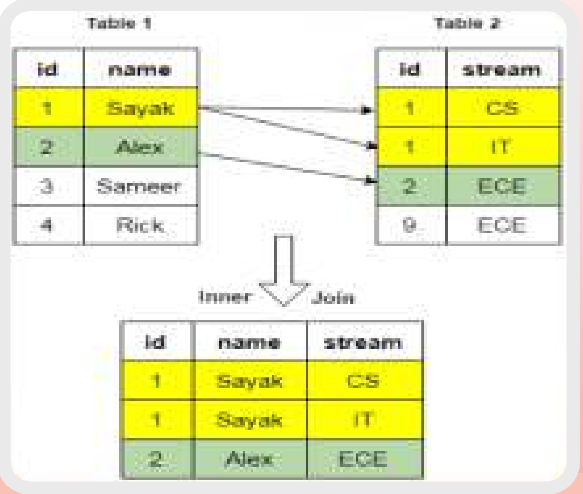


Table 1

id	name
1	Sayak
2	Alex
3	Sameer
4	Rick

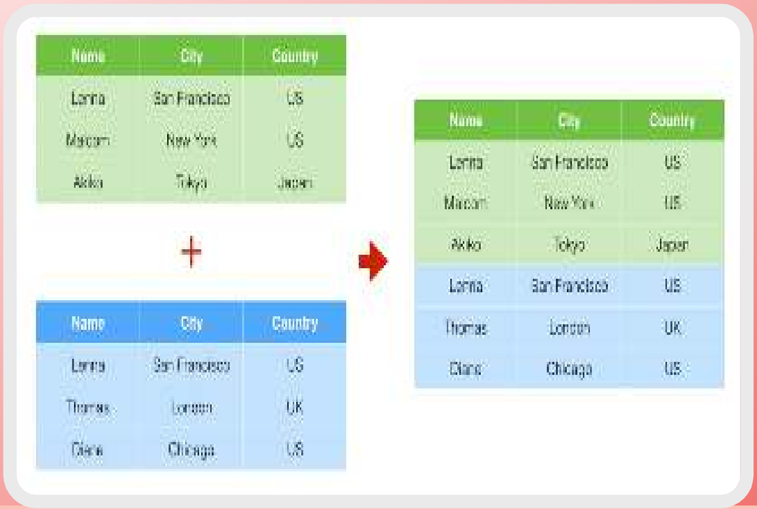
Table 2

id	stream
1	CS
1	IT
2	ECE
9	ECE

Inner Join

id	name	stream
1	Sayak	CS
1	Sayak	IT
2	Alex	ECE

UNION



Name	City	Country
Lorna	San Francisco	US
Malcom	New York	US
Akiko	Tokyo	Japan

+

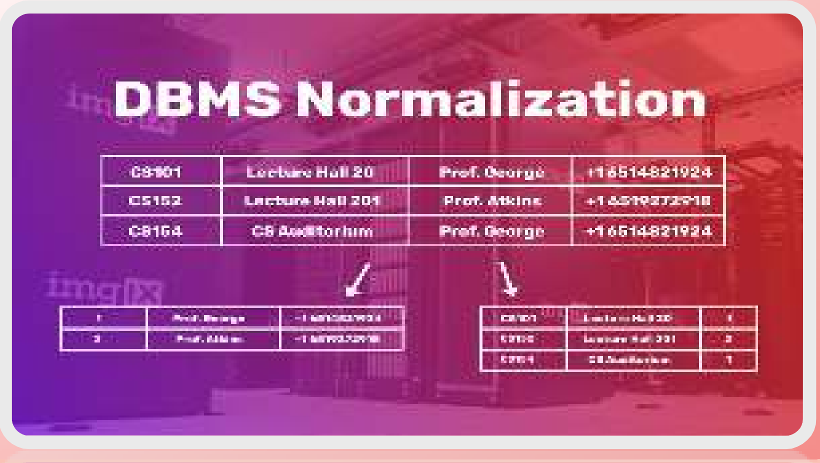
Name	City	Country
Lorna	San Francisco	US
Thomas	London	UK
Diana	Chicago	US

Name	City	Country
Lorna	San Francisco	US
Malcom	New York	US
Akiko	Tokyo	Japan
Lorna	San Francisco	US
Thomas	London	UK
Diana	Chicago	US

Normalizing and De normalizing Data

Normalization, from a statistical view, often has to do with calculating new values from a dataset to standardize the **data** on a particular scale.

It can also imply how well the transaction data is handled for reducing redundancy and inconsistency.



DBMS Normalization

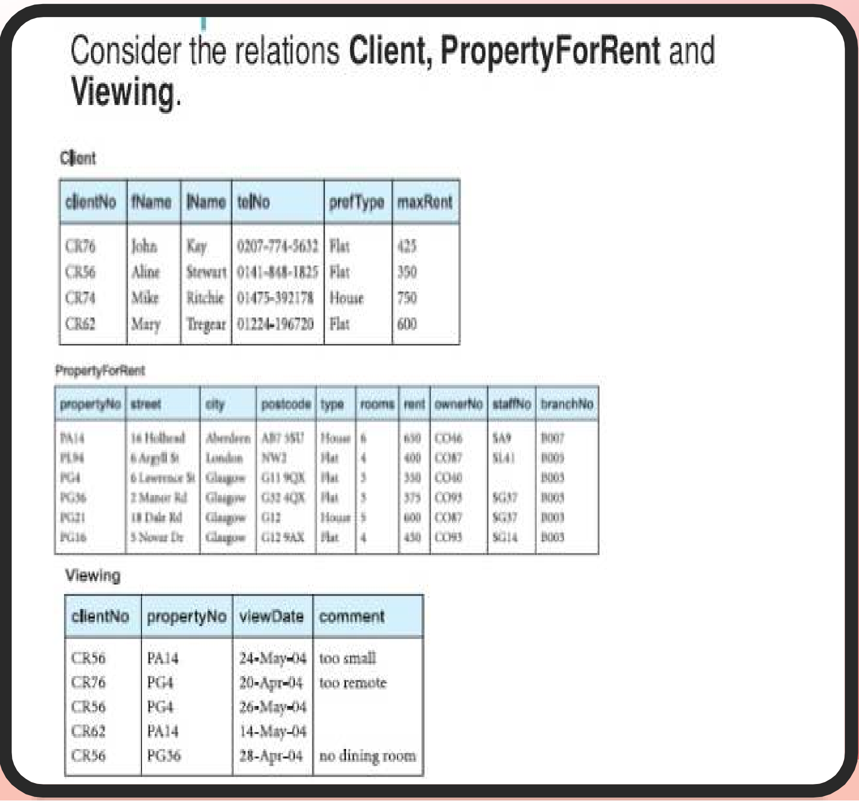
CS101	Lecture Hall 20	Prof. George	+16514821924
CS152	Lecture Hall 201	Prof. Atkins	+16519272918
CS154	CS Auditorium	Prof. George	+16514821924

1	Prof. George	+16514821924
2	Prof. Atkins	+16519272918

CS101	Lecture Hall 20	1
CS152	Lecture Hall 201	2
CS154	CS Auditorium	1

DENORMALI ATION

Combining data from multiple tables for faster querying of data for reports and analysis



Consider the relations **Client, PropertyForRent** and **Viewing**.

Client

clientNo	fName	lName	telNo	prefType	maxRent
CR76	John	Kay	0207-774-5632	Flat	425
CR56	Aline	Stewart	0141-848-1825	Flat	350
CR74	Mike	Ritchie	01475-392178	House	750
CR62	Mary	Tregear	01224-196720	Flat	600

PropertyForRent

propertyNo	street	city	postcode	type	rooms	rent	ownerNo	staffNo	branchNo
PA14	16 Holhead	Aberdeen	AB7 5SU	House	6	650	CO46	SA9	B007
PL94	6 Argyll St	London	NW2	Flat	4	400	CO87	SL41	B005
PG4	6 Lawrence St	Glasgow	G11 9QX	Flat	3	350	CO40		B003
PG36	2 Manor Rd	Glasgow	G32 4QX	Flat	3	375	CO93	SG37	B003
PG21	18 Dale Rd	Glasgow	G12	House	5	600	CO87	SG37	B003
PG16	5 Novar Dr	Glasgow	G12 9AX	Flat	4	450	CO93	SG14	B003

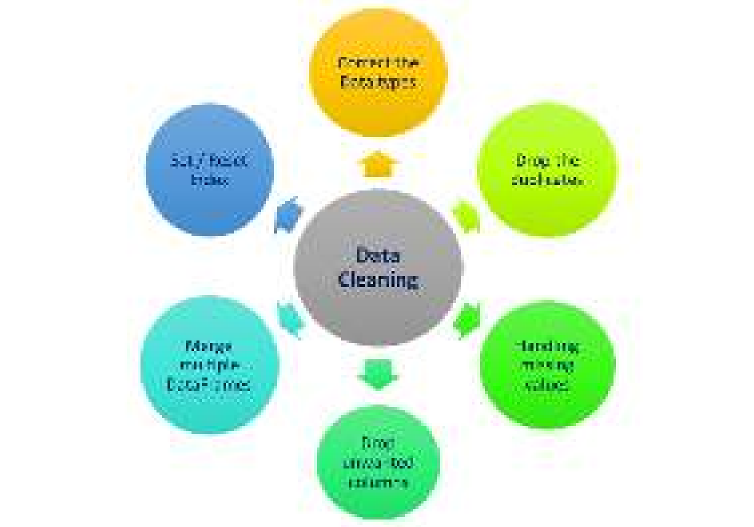
Viewing

clientNo	propertyNo	viewDate	comment
CR56	PA14	24-May-04	too small
CR76	PG4	20-Apr-04	too remote
CR56	PG4	26-May-04	
CR62	PA14	14-May-04	
CR56	PG36	28-Apr-04	no dining room

Cleaning data

During the cleaning stage, users identify data quality issues, such as missing or mismatched values, and apply the appropriate transformation to correct or delete these values from the dataset.

Example: In this dataset, if the data has also some negative values in time column, Given that it’s impossible for a time to be negative, we will simply remove those rows since they seem to represent errors in the dataset and might throw off our analysis.



Data Cleaning

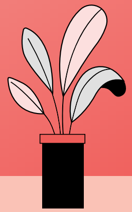
Data cleaning is the process of ensuring **data** is correct, consistent and usable

Missing data can reduce statistical power

Data inconsistency creates unreliable data

Removing duplicate and inaccurate data from databases can help business save valuable resources

Incorrect data may lead to bad decisions



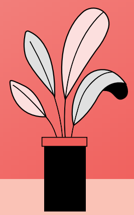
Data cleaning Workflow

The workflow is a sequence of three steps aiming at producing high-quality data and taking into account all the criteria we've talked about.

Inspection: Detect unexpected, incorrect, and inconsistent data. Fix or remove the anomalies discovered. Verifying: After **cleaning**, the results are inspected to verify correctness

2. **Data profiling** is the process of reviewing source data, understanding structure, content and interrelationships, and identifying potential for data projects..

3. **Verification:** After cleaning, the results are inspected to verify correctness.

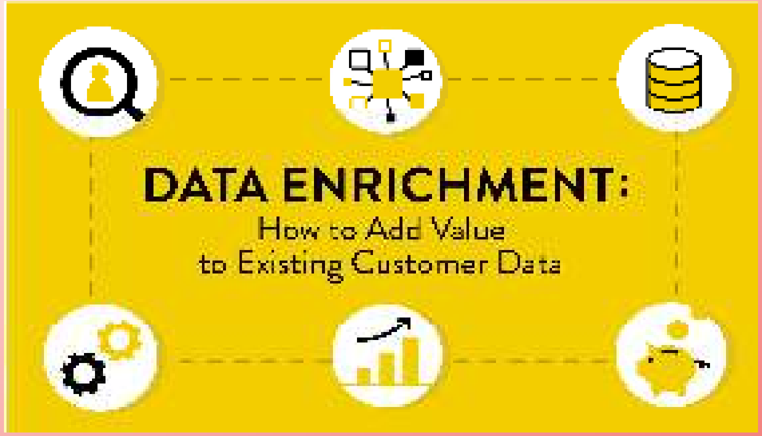


Data Enrichment

Lets take an example here:

Weather can be an important factor in outdoor trips, as it is much less appealing to ride a bicycle on a particularly rainy or windy day.
An analyst might be able to get further insights and draw more well-informed conclusions if you added weather data to the Trip dataset.

To accomplish this, you can use the join function to bring these two datasets together. After loading the Weather data you can see that it is already quite clean and standardized. Next, hitting the join function will allow you to select the Trip dataset that you’ve already cleaned.
You want to join the two datasets using the date value as your join key so that the rows and columns match--however only one date column needs to be incorporated into the output dataset to avoid a duplicate. Once the join is created, this transformed dataset will have both trip and weather data within it

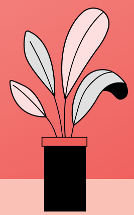


Validation

Checking the quality of data after cleansing, normalizing, denormal zing and enriching the data

Verifying consistency, quality and security of data

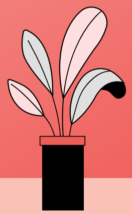
Once the transformations are complete, you can view the Results Summary, which displays detailed statistics of the transformations applied over the entire dataset. You can then export the results of this transformation into the appropriate output format best-fit for your visualization or analytics tool of choice,



Publishing phase

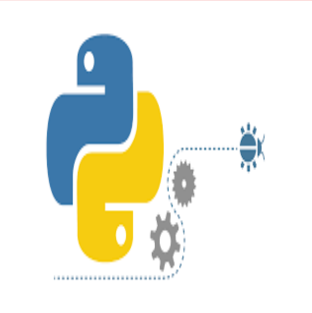
When your data has been successfully structured, cleaned, enriched and validated, it’s time to publish your wrangled output for use in downstream analytics processes. Through the wrangling process, a wider variety of data sources can be used in different statistics, analytics and data visualization applications. This broadens the usage of data throughout the organization and enhances the potential value of data to the business.

Now you’re ready to deliver the output of your data wrangling efforts into the appropriate format for downstream analytic uses. You can publish and save your results as a CSV, JSON, or any other format of your choice.



Tools for Data Wrangling

OPEN
Refine



IBM Watson Studio
Overview

02

Any Questions ?

Activity

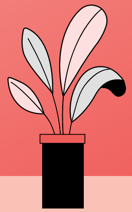
Use Case

Maria is a 25-year-old US Army veteran, newly returned to the civilian workforce. She has recently completed a six-year commitment with the Army. During her time in the Army, she worked in supply management and logistics. She has decided to pursue a degree in Management Systems and Information Technology.

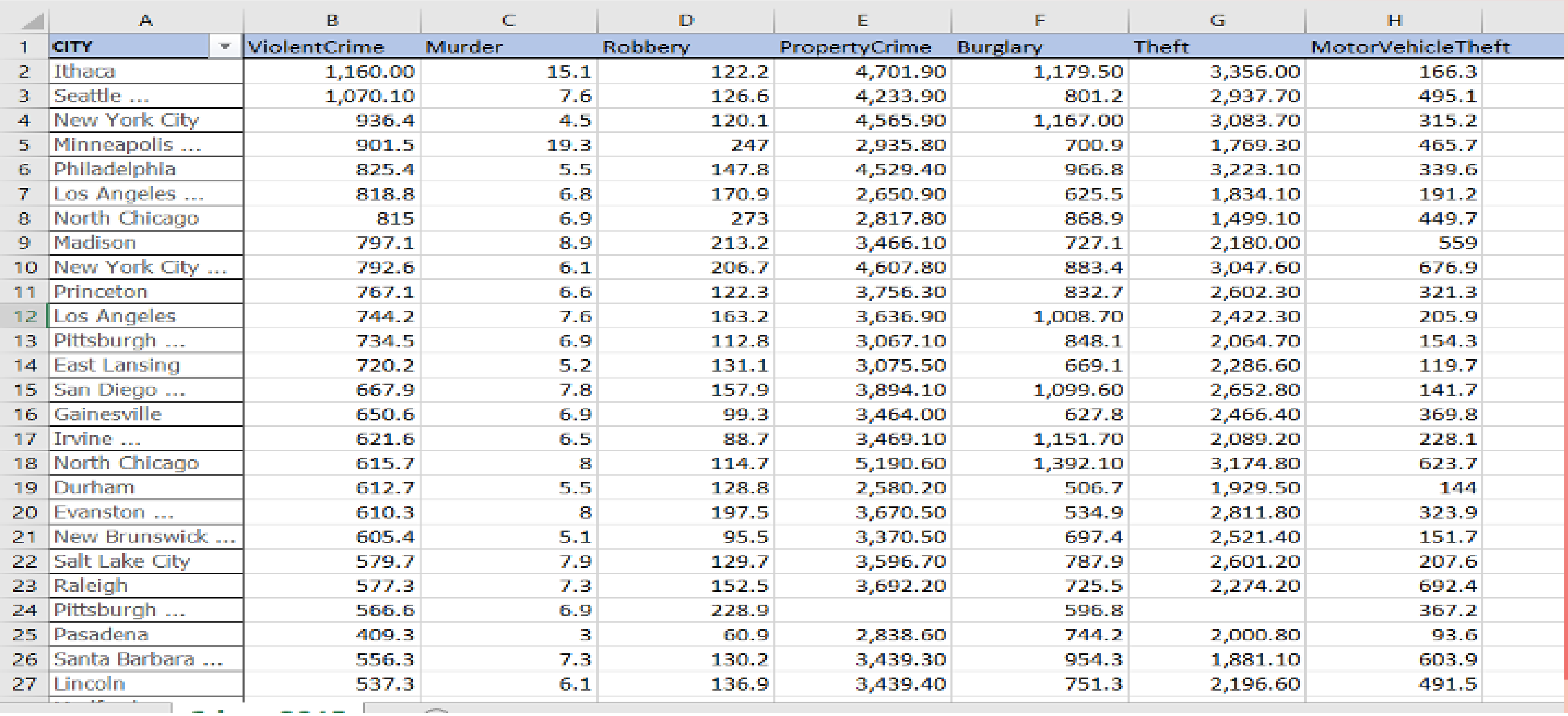
Maria has asked you to use your skill with data to help her search for the best school for her. She is willing to relocate anywhere in the continental United States, but she has a few criteria that her ideal schools must satisfy:

Safety of the city

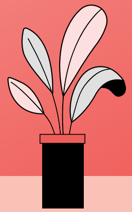
- School should be offering a degree in IT
- Ranking of school.
- .



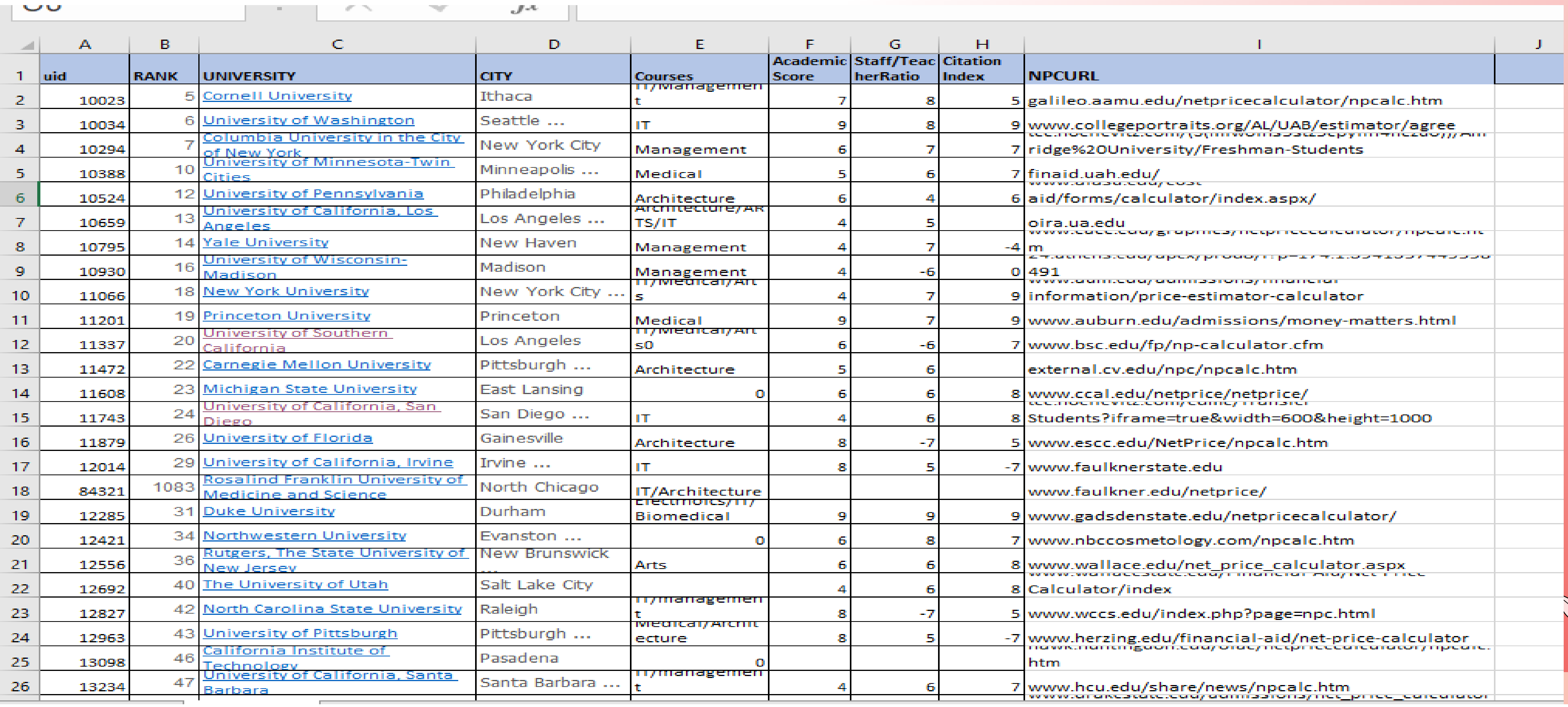
CRIME DATA



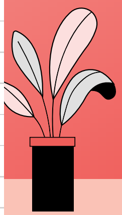
	A	B	C	D	E	F	G	H
1	CITY	ViolentCrime	Murder	Robbery	PropertyCrime	Burglary	Theft	MotorVehicleTheft
2	Ithaca	1,160.00	15.1	122.2	4,701.90	1,179.50	3,356.00	166.3
3	Seattle ...	1,070.10	7.6	126.6	4,233.90	801.2	2,937.70	495.1
4	New York City	936.4	4.5	120.1	4,565.90	1,167.00	3,083.70	315.2
5	Minneapolis ...	901.5	19.3	247	2,935.80	700.9	1,769.30	465.7
6	Philadelphia	825.4	5.5	147.8	4,529.40	966.8	3,223.10	339.6
7	Los Angeles ...	818.8	6.8	170.9	2,650.90	625.5	1,834.10	191.2
8	North Chicago	815	6.9	273	2,817.80	868.9	1,499.10	449.7
9	Madison	797.1	8.9	213.2	3,466.10	727.1	2,180.00	559
10	New York City ...	792.6	6.1	206.7	4,607.80	883.4	3,047.60	676.9
11	Princeton	767.1	6.6	122.3	3,756.30	832.7	2,602.30	321.3
12	Los Angeles	744.2	7.6	163.2	3,636.90	1,008.70	2,422.30	205.9
13	Pittsburgh ...	734.5	6.9	112.8	3,067.10	848.1	2,064.70	154.3
14	East Lansing	720.2	5.2	131.1	3,075.50	669.1	2,286.60	119.7
15	San Diego ...	667.9	7.8	157.9	3,894.10	1,099.60	2,652.80	141.7
16	Gainesville	650.6	6.9	99.3	3,464.00	627.8	2,466.40	369.8
17	Irvine ...	621.6	6.5	88.7	3,469.10	1,151.70	2,089.20	228.1
18	North Chicago	615.7	8	114.7	5,190.60	1,392.10	3,174.80	623.7
19	Durham	612.7	5.5	128.8	2,580.20	506.7	1,929.50	144
20	Evanston ...	610.3	8	197.5	3,670.50	534.9	2,811.80	323.9
21	New Brunswick ...	605.4	5.1	95.5	3,370.50	697.4	2,521.40	151.7
22	Salt Lake City	579.7	7.9	129.7	3,596.70	787.9	2,601.20	207.6
23	Raleigh	577.3	7.3	152.5	3,692.20	725.5	2,274.20	692.4
24	Pittsburgh ...	566.6	6.9	228.9		596.8		367.2
25	Pasadena	409.3	3	60.9	2,838.60	744.2	2,000.80	93.6
26	Santa Barbara ...	556.3	7.3	130.2	3,439.30	954.3	1,881.10	603.9
27	Lincoln	537.3	6.1	136.9	3,439.40	751.3	2,196.60	491.5



COLLEGE SCORE



	uid	RANK	UNIVERSITY	CITY	Courses	Academic Score	Staff/TeacherRatio	Citation Index	NPCURL
2	10023	5	Cornell University	Ithaca	IT/Management	7	8	5	galileo.aamu.edu/netpricecalculator/npcalc.htm
3	10034	6	University of Washington	Seattle ...	IT	9	8	9	www.collegeportraits.org/AL/UAB/estimator/agree
4	10294	7	Columbia University in the City of New York	New York City	Management	6	7	7	[illegible]ridge%20University/Freshman-Students
5	10388	10	University of Minnesota-Twin Cities	Minneapolis ...	Medical	5	6	7	finaid.uah.edu/
6	10524	12	University of Pennsylvania	Philadelphia	Architecture	6	4	6	www.alasu.edu/cost aid/forms/calculator/index.aspx/
7	10659	13	University of California, Los Angeles	Los Angeles ...	Architecture/ARTS/IT	4	5		oira.ua.edu
8	10795	14	Yale University	New Haven	Management	4	7	-4	www.cacc.edu/graphics/netpricecalculator/npcalc.htm
9	10930	16	University of Wisconsin-Madison	Madison	Management	4	-6	0	24.athens.edu/apex/prod/f?p=174:1:[illegible]491
10	11066	18	New York University	New York City ...	IT/Medical/Arts	4	7	9	www.aum.edu/admissions/financial information/price-estimator-calculator
11	11201	19	Princeton University	Princeton	Medical	9	7	9	www.auburn.edu/admissions/money-matters.html
12	11337	20	University of Southern California	Los Angeles	IT/Medical/Arts0	6	-6	7	www.bsc.edu/fp/np-calculator.cfm
13	11472	22	Carnegie Mellon University	Pittsburgh ...	Architecture	5	6		external.cv.edu/npc/npcalc.htm
14	11608	23	Michigan State University	East Lansing	0	6	6	8	www.ccal.edu/netprice/netprice/
15	11743	24	University of California, San Diego	San Diego ...	IT	4	6	8	[illegible]Transfer Students?iframe=true&width=600&height=1000
16	11879	26	University of Florida	Gainesville	Architecture	8	-7	5	www.escc.edu/NetPrice/npcalc.htm
17	12014	29	University of California, Irvine	Irvine ...	IT	8	5	-7	www.faulknerstate.edu
18	84321	1083	Rosalind Franklin University of Medicine and Science	North Chicago	IT/Architecture				www.faulkner.edu/netprice/
19	12285	31	Duke University	Durham	Electronics/IT/Biomedical	9	9	9	www.gadsdenstate.edu/netpricecalculator/
20	12421	34	Northwestern University	Evanston ...	0	6	8	7	www.nbccosmetology.com/npcalc.htm
21	12556	36	Rutgers, The State University of New Jersey	New Brunswick ...	Arts	6	6	8	www.wallace.edu/net_price_calculator.aspx
22	12692	40	The University of Utah	Salt Lake City		4	6	8	www.wallacestate.edu/Financial Aid/Net Price Calculator/index
23	12827	42	North Carolina State University	Raleigh	IT/management	8	-7	5	www.wccs.edu/index.php?page=npc.html
24	12963	43	University of Pittsburgh	Pittsburgh ...	Medical/Architecture	8	5	-7	www.herzing.edu/financial-aid/net-price-calculator
25	13098	46	California Institute of Technology	Pasadena	0				hawk.huntingdon.edu/[illegible]/netpricecalculator/npcalc.htm
26	13234	47	University of California, Santa Barbara	Santa Barbara ...	IT/management	4	6	7	www.hcu.edu/share/news/npcalc.htm



Your Task

1)Clean the data:

Remove any duplicates

Missing Values

Inconsistent values

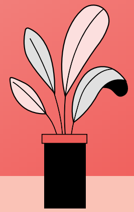
2) Data Enrichment:

Calculate School Ranking

Calculate overall crime rate

3) Structure the data:
Merge the tables and produce the dataset which must have :

1. Top 5 schools on rankings for It college
2. Be in a city that is below 50th percentile in overall crime
3. Remove unnecessary columns



IT USED TO BE THAT NOBODY ON THE INTERNET KNEW THAT I WAS A DOG. NOW, BECAUSE OF BIG DATA, EVERYBODY KNOWS THAT I AM A 15-YEAR OLD LABRADOODLE WHO SECRETLY LIKES CAT FOOD.

KDNuggets.com • cartertoons.com

Thank You