Batch: A3 **Roll No.: 16014022050**

Experiment 01

Title: Data Collection and finalizing dataset from problem domain

---

Objective:

1. **To learn how to collect the dataset**
2. **To learn sources of dataset**
3. **To assess the dataset based on Metrics to Measure Data Quality**
4. **To finalize the features of dataset**

Course Outcome:

CO1: Learn how to locate and download datasets, extract insights from that dataand present their findings in a variety of different formats.

Books/ Journals/ Websites referred:

https://www.kaggle.com/

Resources used:

Google & YouTube

Theory:

A Dataset is a set or collection of data. This set is normally presented in a tabular pattern. Every column describes a particular variable. And each row corresponds to agiven member of the data set, as per the given question.

Following points should be written by students

- Problem domain (Healthcare, Ecommerce, Education, Finance, agriculture etc.)
- Motivation for the selected Domain
- Brain stormed features of Dataset (Based on Domain Selected)
- Search for dataset
- Justification for choosing above dataset
- Source of dataset (Link Needs to be given)
- Sample of Finalized dataset (First 5 Records)
- Data Dictionary
- Column wise summary

1. **Finance: global superstore database**
2. **I wanted to analyze the sales and profit of superstores across states in thecountry.**
3.

Table	Total Rows	Total Columns
Orders	51290	24
People	13	2
Returns	1173	3

4. **Database is from kaggle(**https://www.kaggle.com/datasets/shekpaul/global-superstore).
5. **Justification for choosing above dataset: to analyze the sales**
6. **Link:** https://www.kaggle.com/datasets/shekpaul/global-superstore
7. **Screenshot of dataset**

	A	B	C	D	E	F	G	H	I	J
1	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer	Customer Name	Segment	City	State
2	32298	CA-2012-124891	31-07-2012	31-07-2012	Same Day	RH-19495	Rick Hansen	Consumer	New York City	New York
3	26341	IN-2013-77878	05-02-2013	07-02-2013	Second Class	JR-16210	Justin Ritter	Corporate	Wollongong	New South Wales
4	25330	IN-2013-71249	17-10-2013	18-10-2013	First Class	CR-12730	Craig Reiter	Consumer	Brisbane	Queensland
5	13524	ES-2013-1579342	28-01-2013	30-01-2013	First Class	KM-16375	Katherine Murray	Home Office	Berlin	Berlin
6	47221	SG-2013-4320	05-11-2013	06-11-2013	Same Day	RH-9495	Rick Hansen	Consumer	Dakar	Dakar
7	22732	IN-2013-42360	28-06-2013	01-07-2013	Second Class	JM-15655	Jim Mitchum	Corporate	Sydney	New South Wales
8	30570	IN-2011-81826	07-11-2011	09-11-2011	First Class	TS-21340	Toby Swindell	Consumer	Porirua	Wellington
9	31192	IN-2012-86369	14-04-2012	18-04-2012	Standard Class	MB-18085	Mick Brown	Consumer	Hamilton	Waikato
10	40155	CA-2014-135909	14-10-2014	21-10-2014	Standard Class	JW-15220	Jane Waco	Corporate	Sacramento	California
11	40936	CA-2012-116638	28-01-2012	31-01-2012	Second Class	JH-15985	Joseph Holt	Consumer	Concord	North Carolina
12	34577	CA-2011-102988	05-04-2011	09-04-2011	Second Class	GM-14695	Greg Maxwell	Corporate	Alexandria	Virginia
13	28879	ID-2012-28402	19-04-2012	22-04-2012	First Class	AJ-10780	Anthony Jacobs	Corporate	Kabul	Kabul
14	45794	SA-2011-1830	27-12-2011	29-12-2011	Second Class	MM-7260	Magdelene Morse	Consumer	Jizan	Jizan
15	4132	MX-2012-130015	13-11-2012	13-11-2012	Same Day	VF-21715	Vicky Freymann	Home Office	Toledo	Parana
16	27704	IN-2013-73951	06-06-2013	08-06-2013	Second Class	PF-19120	Peter Fuller	Consumer	Mudanjiang	Heilongjiang
17	13779	ES-2014-5099955	31-07-2014	03-08-2014	Second Class	BP-11185	Ben Peterman	Corporate	Paris	Ile-de-France
18	36178	CA-2014-143567	03-11-2014	06-11-2014	Second Class	TB-21175	Thomas Boland	Corporate	Henderson	Kentucky
19	12069	ES-2014-1651774	08-09-2014	14-09-2014	Standard Class	PJ-18835	Patrick Jones	Corporate	Prato	Tuscany
20	22096	IN-2014-11763	31-01-2014	01-02-2014	First Class	JS-15685	Jim Sink	Corporate	Townsville	Queensland
21	49463	TZ-2014-8190	05-12-2014	07-12-2014	Second Class	RH-9555	Ritsa Hightower	Consumer	Uvinza	Kigoma
22	46630	PL-2012-7820	08-08-2012	10-08-2012	First Class	AB-600	Ann Blume	Corporate	Bytom	Silesia
23	31784	CA-2011-154627	29-10-2011	31-10-2011	First Class	SA-20830	Sue Ann Reed	Consumer	Chicago	Illinois
24	21586	IN-2011-44803	02-05-2011	03-05-2011	First Class	JK-15325	Jason Klamczynski	Corporate	Suzhou	Anhui
25	13528	ES-2013-2860574	27-02-2013	01-03-2013	Second Class	LB-16795	Laurel Beltran	Home Office	Edinburgh	Scotland
26	1570	US-2014-133193	31-07-2014	01-08-2014	First Class	NP-18325	Naresj Patel	Consumer	Juárez	Chihuahua
27	3484	MX-2014-165309	05-09-2014	08-09-2014	First Class	VD-21670	Valerie Dominguez	Consumer	Soyapango	San Salvador
28	30191	IN-2011-10286	17-12-2011	20-12-2011	First Class	PB-19210	Phillip Breyer	Corporate	Taipei	Taipei City
29	11645	ES-2011-4699764	14-03-2011	17-03-2011	Second Class	EB-14110	Eugene Barchas	Consumer	Leipzig	Saxony

8. **Column wise summary:**

Row ID: id of rows is provided in this
columnOrder ID: id of all the orders
Order Date: date of orders are given
Ship Date: date when the item has been shipped (dispatched).
Ship Mode: mode of shipment like first class, second class or standard class.Customer ID: id of all the customers.
Customer Name: name of customers.
Segment: example consumer, corporate or home office.
City: example New York City, Paris, Los Angeles.
State: example Berlin, California.
Country: example united states,
Australia.
Postal Code: postal code of all these
countries.Market: example Africa, APAC.
Region: example east, south or
central.Product ID: id of the
product.
Category: technology, furniture or office
suppliesSub-Category: phone, chairs, tables

Conclusion (Students should write in their own words):

Through systematic data collection techniques and a comprehensive analysis of various data sources, we effectively curated a high-quality dataset, ensuring reliability and relevance. The careful assessment of key metrics enabled us to refine the dataset features, laying a strong foundation for future data-driven endeavors.

Post Lab Question:

1. **Explain Role of Data in the Application Design.**
   The first stage of design thinking is to empathize with your users and understand theircontext, challenges, motivations, and goals. Data and analytics can help you gather insights from various sources, such as surveys, interviews, observations, analytics platforms, social media, or customer feedback.

2. **Write different types of Data with Example.**
    - String values: names example Rohan
    - Number/Integer values: example 22, 25
    - Date & Time values: 3/09/94
    - Boolean values: true or false
    - Geographic values:
    - Cluster or mixed values