

# Task-1

## 1. Definition of Data Analytics :

Data analytics is the process of examining raw data to find meaningful patterns, trends, and insights that help in making informed decisions. It involves collecting, cleaning, analyzing, and visualizing data to understand past behavior, predict future outcomes, and support strategic planning.

## 2. Two Historical Milestones That Shaped the Field:

### **Development of Relational Databases (1970s):**

The invention of the relational database by Edgar F. Codd revolutionized how data was stored and accessed. It allowed structured data to be stored in tables with relationships, making it easier to query and analyze large datasets efficiently.

### **Rise of Big Data and Hadoop (2000s):**

As data grew exponentially, traditional tools couldn't handle the volume. The introduction of **Hadoop**, an open-source framework, enabled distributed storage and processing of massive datasets across clusters of computers, making large-scale data analysis feasible and cost-effective.

## 3. Three Current Trends in Data Analytics:

### **Artificial Intelligence (AI) and Machine Learning:**

AI and ML are transforming analytics by enabling systems to learn from data and make predictions without being explicitly programmed. This is widely used in recommendation systems, fraud detection, and predictive maintenance.

### **Cloud-Based Analytics:**

Organizations are increasingly using cloud platforms (like AWS, Azure, or Google Cloud) to store and analyze data. Cloud analytics offers scalability, remote access, and powerful computing capabilities without large infrastructure costs.

### **Real-Time Data Processing:**

Real-time analytics allows businesses to analyze data as it is generated, leading to faster decisions. Examples include stock market monitoring, live traffic updates, and instant fraud detection in banking.

## Task-2

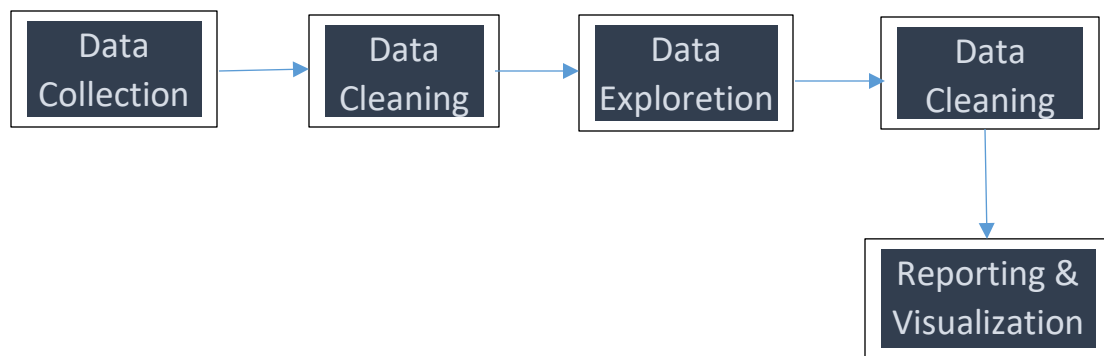


Figure: End to End Data Analysis process

### **Data Collection**

Gathering raw data from various sources like databases, APIs, files, surveys, or IoT devices. Quality analysis depends on having the right and complete data. This stage ensures all necessary data is captured for further steps.

### **Data Cleaning (Data Preparation)**

Removing errors, duplicates, missing values, and inconsistencies. Transforming data into a usable format. Dirty or inaccurate data leads to wrong insights. Cleaning ensures data quality and reliability.

### **Data Exploration (EDA - Exploratory Data Analysis)**

Understanding the data's structure, summary statistics, distributions, and relationships through graphs and plots. Helps identify patterns, anomalies, and insights early. Guides the direction of deeper analysis or model building.

### **Data Analysis (Modeling & Interpretation)**

Applying statistical methods, machine learning models, or business logic to uncover insights or make predictions. This stage turns raw data into actionable knowledge, helping answer specific business or research questions.

## **5. Reporting & Visualization**

Presenting findings using dashboards, charts, or reports using tools like Excel, Power BI, or Tableau. Effective communication ensures stakeholders understand insights and can make data-driven decisions.

### Task-3

Term	Definition
<b>Data types</b>	Categories that specify the type of data (e.g., integer, float, string) a variable can hold.
<b>Variables</b>	Named storage locations in programs or datasets that hold values which can change.
<b>Metrics</b>	Quantitative measurements used to assess, compare, and track performance or behavior.
<b>KPIs</b>	Key Performance Indicators; critical metrics that measure the success of business goals.
<b>Structured data</b>	Organized data that follows a fixed schema, usually in tables with rows and columns.
<b>API</b>	Application Programming Interface; a set of rules that allows one software to communicate with another.
<b>SQL</b>	Structured Query Language; used to manage and query relational databases.
<b>Data Warehouse</b>	A centralized repository for storing and analyzing large volumes of structured data from multiple sources.
<b>Pandas</b>	A Python library used for data manipulation and analysis, especially with tabular data.
<b>tidyverse</b>	A collection of R packages designed for data science, including tools for data wrangling and visualization.

## Task-4

## What is the source and format of the data?

The dataset was collected internally from a franchise operation (Well Food, Qatar branch) and maintained in Excel format (.xlsx). It consists of daily records for sales and wastage across various product categories like Sweets, Cake, Bakery, Arabian, Cookies, and Hot Items.

## Is it structured, semi-structured, or unstructured?

The dataset is structured.

It is organized into rows and columns, where each row represents a daily record and columns represent variables such as date, product categories, sales, and wastage. This makes it easy to analyze using tools like Excel, Python (Pandas), or Power BI.

## What are the ethical considerations in using this data?

- **Confidentiality:**  
Since the data comes from a business operation, it may include sensitive business performance information. It should not be shared publicly without permission.
- **Data Accuracy:**  
Care must be taken to ensure the data is accurate and not manipulated, as this could lead to misleading conclusions or poor business decisions.
- **Purpose Limitation:**  
The data should be used strictly for educational or internal analysis purposes, not for any unauthorized commercial or competitive advantage.
- **Anonymity (if applicable):**  
If customer or employee details were included, their identities must be anonymized. However, this dataset seems to be product-level only, so privacy concerns are minimal.

Description Code	Opening	Receiving	Total	Return Fa	Wastage	Shira	Closing	Total sale
Sweets S-1	10,169	14,274	24,443		154		12,683	11,606
Dry Sweet S-1	9,290	7,953	17,243		200		13,450	3,593
Hot Bread B-1		6,130	6,130		990			5,140
Bun B-2	165	5,565	5,730				1,300	4,430
Plain Cake C-1	305	1,375	1,680				145	1,535
Chanachur B-3	390	2,650	3,040				310	2,640
Semai B-4	18,665		18,665				17,985	680
Arabian A	62,560		62,560				62,170	390
Pastry & F-2	5,465	5,580	11,045		80		7,015	3,950
Cookies Co-3	5,155	600	5,755				1,570	4,185
Bela & Toi B-5	48,260		48,260				46,780	1,480
Chocolate O-1	24,180		24,180				24,060	120
Vell Cook Co-2	7,415		7,415				6,875	540
Custard C C-3		-	-				-	-
Noodles O-2	75	75	-				75	-
Jam & Jeli O-3		-	-				-	-
Nutri Pow-O-5	1,980		1,980				1,800	180
Sauce O-5	1,440		1,440				1,280	160
<b>Total</b>	<b>195,454</b>	<b>44,097</b>	<b>239,551</b>	<b>-</b>	<b>1,424</b>	<b>-</b>	<b>197,498</b>	<b>40,629</b>

Row Labels	Tot Sum of Wastage	Sum of Total sale
Arabian	80	3950
<b>Grand Total</b>	<b>80</b>	<b>3950</b>

Row Labels	Tot Sum of Wastage	Sum of Total sale
Hot	990	5140
<b>Grand Total</b>	<b>990</b>	<b>5140</b>

Row Labels	Tot Sum of Wastage	Sum of Total sale
Dry Sweets	200	3593
Sweets	154	11606
<b>Grand Total</b>	<b>354</b>	<b>15199</b>

Row Labels	Tot Sum of Wastage	Sum of Total sale
Bela & Toast	120	
Bread	4430	
Bun	1535	
Chanachur	680	
Semai	390	
<b>Grand Total</b>	<b>7155</b>	

Row Labels	Tot Sum of Wastage	Sum of Total sale
Custard Cake	0	
Pastry & Ready Cake	4185	
Plain Cake	2640	
<b>Grand Total</b>	<b>6825</b>	

Row Labels	Tot Sum of Wastage	Sum of Total sale
Chocolate	0	
Jam & Jelly	0	
Noodles	0	
Nutri Powder	180	
Sauce	160	

Row Labels	Tot Sum of Wastage	Sum of Total sale
Cookies	1480	
Vell Cookies	540	
<b>Grand Total</b>	<b>2020</b>	

## Task-5

### Data Profiling

Attribute	Value
Total Rows	120
Total Columns	16
Key Columns	Date, Month, Day, Sweets, Cake, Bakery, Hot, Cookies, Arabian, Wastage, Total Sell
Data Types	Mixed (Date, Object, Int, Float)
Null Values Found	No Null value
Structured Format	Yes (organized tabular data)

## Task 6: Data Cleaning

### Handling Missing Values

- **Column:** Others had **1 missing value**
- **Strategy:** Replaced the missing value with **0**, assuming no sale occurred for that day in that category (imputation).
- **Result:** No missing values remain in the dataset.

### Removing Duplicates

- The dataset originally had **no duplicate rows**, but a check and cleaning step was applied to ensure data integrity.

Actually I collect the data on my own for this reason the error in the dataset almost null. When I collect this I carefully collect the data . It is 4 month sells & wastage data in my

	A	B	C	D	E	F	G	H	I	J	K	L
	Date	Month	Day	Sweets	Hot	Bakery	Cake	Arabian	Cookies	Others	Total Sell	Wastage
	1/1/2025	Jan	Wed	24,944	6,385	15,120	44,320	4,400	1,240	-	96,409	167874
	1/2/2025	Jan	Thu	14,232	6,705	4,400	13,790	5,020	3,630	160	47,937	81642
	1/3/2025	Jan	Fri	12,298	6,245	4,025	6,060	2,505	1,740	-	32,873	53448
	1/4/2025	Jan	Sat	10,103	4,470	5,290	5,290	3,580	3,145	-	31,878	53653
	1/5/2025	Jan	Sun	20,715	4,200	3,925	8,445	1,380	2,275	-	40,940	61165
	1/6/2025	Jan	Mon	15,865	4,990	4,950	6,120	2,185	2,655	480	37,245	58625
	1/7/2025	Jan	Tue	9,041	4,300	7,695	4,440	2,225	3,030	-	30,731	52421
	1/8/2025	Jan	Wed	13,183	4,080	4,850	7,600	2,475	3,445	-	35,633	58083
	1/9/2025	Jan	Thu	16,555	4,650	2,870	4,265	4,560	3,670	-	36,570	56585
	1/10/2025	Jan	Fri	18,406	4,915	4,630	3,745	7,020	-	-	38,716	59026
	1/11/2025	Jan	Sat	16,571	4,115	2,755	6,295	-	-	-	29,736	42901
	1/12/2025	Jan	Sun	12,915	5,815	5,695	5,375	4,260	-	-	34,060	55205
	1/13/2025	Jan	Mon	14,893	3,880	5,995	3,690	4,085	1,395	-	33,938	52983
	1/14/2025	Jan	Tue	23,871	4,340	5,460	8,280	4,305	1,080	160	47,496	71121
	1/15/2025	Jan	Wed	19,573	3,880	3,580	4,105	5,465	995	160	37,758	55943
	1/16/2025	Jan	Thu	20,343	4,960	5,095	6,650	5,095	1,000	160	43,303	66263
	1/17/2025	Jan	Fri	32,305	5,270	6,615	6,680	5,085	1,495	160	57,610	82915
	1/18/2025	Jan	Sat	13,574	4,950	5,290	5,525	4,225	1,530	-	35,094	56614
	1/19/2025	Jan	Sun	18,789	8,660	5,065	5,925	3,530	1,220	-	43,189	67589
	1/20/2025	Jan	Mon	16,383	4,260	4,090	5,935	4,260	680	800	36,408	56433
	1/21/2025	Jan	Tue	11,455	5,645	5,670	13,070	2,790	770	-	39,400	67345
	1/22/2025	Jan	Wed	10,653	2,075	5,685	8,185	2,155	1,630	-	32,383	55013
	Sells	Sheet8	SvW	Wastage	S	W	Sheet5	Dashbord	+			

Date	Month	Day	Sweets	Hot	Bakery	Cake	Arabian	Cookies	Others	Wastage
1/1/2025	Jan	Wed	0	0	0	0	0	0	0	0
1/2/2025	Jan	Thu	154	0	300	0	0	0	0	454
1/3/2025	Jan	Fri	0	1265	0	0	0	0	0	1265
1/4/2025	Jan	Sat	1882	910	0	2800	500	0	0	6092
1/5/2025	Jan	Sun	2233	790	50	1200	0	0	0	4273
1/6/2025	Jan	Mon	1200	0	230	0	400	0	0	1830
1/7/2025	Jan	Tue	2245	0	0	0	380	0	0	2625
1/8/2025	Jan	Wed	144	220	0	1840	0	0	0	2204
1/9/2025	Jan	Thu	1627	0	0	0	0	0	0	1627
1/10/2025	Jan	Fri	0	185	0	1940	560	0	0	2685
1/11/2025	Jan	Sat	0	1120	570	0	0	0	0	1690
1/12/2025	Jan	Sun	0	0	915	850	60	0	0	1825
1/13/2025	Jan	Mon	1650	0	0	300	0	0	0	1950
1/14/2025	Jan	Tue	1782	0	0	1170	0	0	0	2952
1/15/2025	Jan	Wed	324	0	0	0	0	0	0	324
1/16/2025	Jan	Thu	1828	0	0	0	0	0	0	1828
1/17/2025	Jan	Fri	1142	0	0	0	0	0	0	1142
1/18/2025	Jan	Sat	0	170	90	280	550	0	0	1090
1/19/2025	Jan	Sun	0	75	0	770	0	0	0	845
1/20/2025	Jan	Mon	1252	0	0	0	0	0	0	1252
1/21/2025	Jan	Tue	1423	0	0	0	360	0	0	1783
1/22/2025	Jan	Wed	0	150	0	0	250	0	0	400
	Sells	Sheet8	SvW	Wastage	S	W	Sheet5	Dashbord	+	

## Task 7: Tool Evaluation & Reflection

### 1. Why did you choose this tool?

I chose **Microsoft Excel** because it is widely available, user-friendly, and well-suited for handling structured business data. As a beginner-friendly tool, Excel allowed me to perform essential data analytics tasks like cleaning, filtering, calculations, and visualizations without needing advanced programming skills.

### 2. Strengths and Limitations Experienced

#### Strengths:

- **Ease of Use:** Excel's interface is intuitive, especially for working with tabular data.
- **Built-in Functions:** SUM, AVERAGE, IF, and PivotTables helped in summarizing and analyzing data quickly.
- **Basic Visualizations:** I used line charts and bar graphs to highlight trends in sales and wastage.
- **No Setup Required:** Excel required no extra installation or programming knowledge.

#### Limitations:

- **Limited Automation:** For repetitive tasks like filtering or comparing months, I had to do things manually.
- **Scalability Issues:** Excel can slow down or crash when working with very large datasets.
- **No Advanced Analytics:** Tasks like statistical modeling or machine learning can't be done easily in Excel.

---

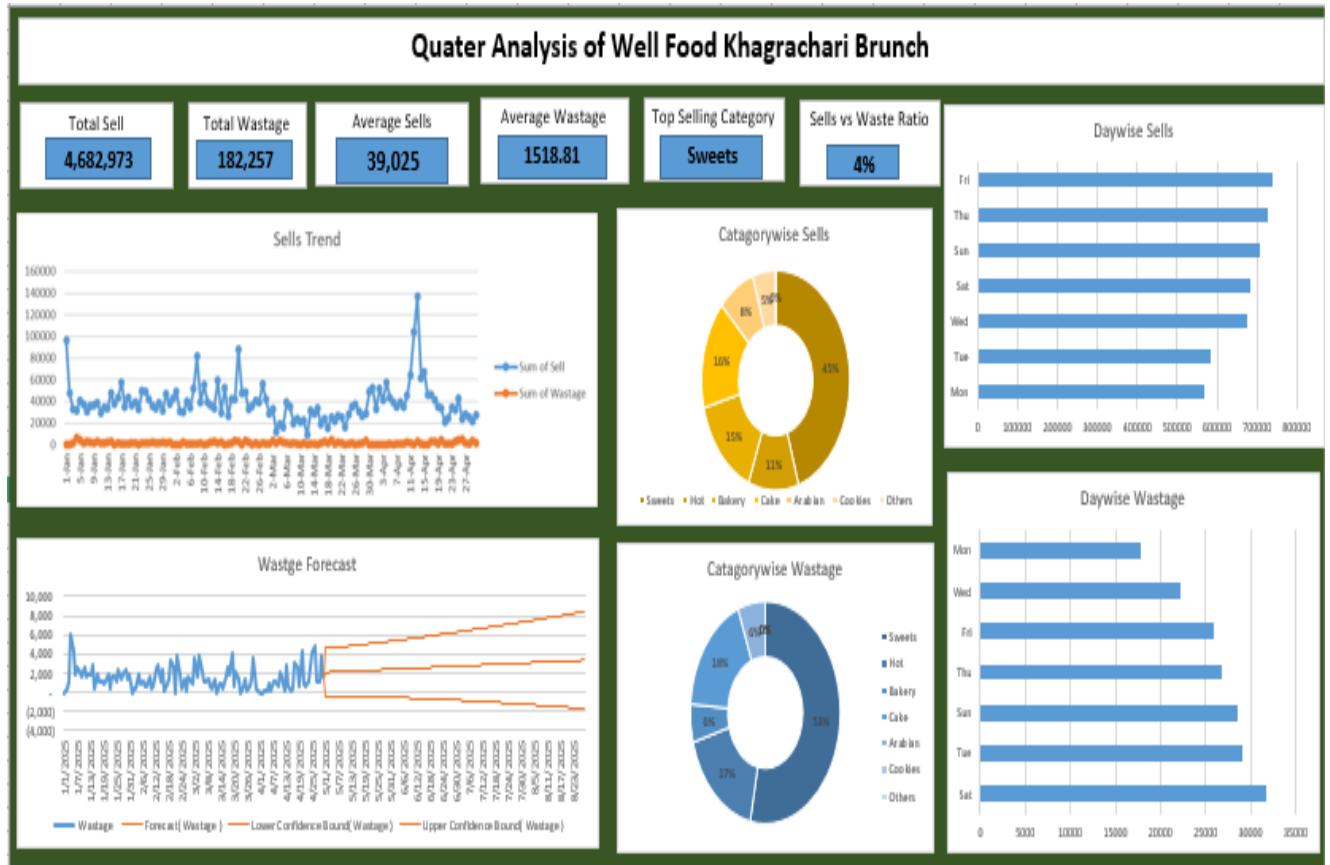
### 3. Would you consider using another tool for future analytics projects?

Yes. For more complex or large-scale analytics, I would consider tools like:

- **Power BI** for interactive dashboards and automated reporting
- **Python (Pandas/Matplotlib)** for advanced cleaning and analysis
- **SQL** for querying larger databases

These tools offer more flexibility, automation, and scalability than Excel. However, for quick and medium-sized analyses, Excel remains a great starting point.

The Dashbord I made with my data



It is the report about 4 month data report.



## **Task 8: Mini Case Study – Application of Data Analytics**

**Domain:** *E-commerce*

### **Real-World Application:**

#### **Personalized Product Recommendations**

In the e-commerce sector, data analytics is widely used to power **recommendation systems** on platforms like Amazon, Daraz, and Alibaba. These systems analyze customer behavior to suggest products that users are more likely to buy.

#### **Type of Data Used:**

- **Browsing history** (pages visited, time spent)
- **Past purchases**
- **Click-through rates**
- **Customer demographics** (age, location, gender)
- **Ratings and reviews**

#### **Decisions Made Using the Data:**

- Suggest products that are similar to past purchases or popular among similar users.
- Customize homepage content based on user preferences.
- Send targeted email promotions and discount offers.
- Adjust pricing strategies based on user interest and product performance.

#### **Impact:**

This use of data analytics helps increase conversion rates, boost customer satisfaction, and drive sales growth by making shopping experiences more relevant and personalized.