Assignment-02

Name: Suraiya Islam Lira Student ID: NA-E08827

Part-A: Conceptual Understanding

Task-1: Data Analytics Foundations

♦ Define data Analytics in your own words.

Ans: Data analytics is the process of examining raw data to find patterns, trends, or insights and draw conclusions. It is useful to make better decisions. It involves collecting data, organizing it, and using tools or techniques like statistics or visualizations to understand what the data is telling us.

♦ Discuss two historical milestones that shaped the field.

Ans: Here two key historical milestones that significantly shaped field of data analytics.

1. Invention of Database (1970s)

● The creation of the relational database by Edgar F. Codd made it possible to store, organize, and retrieve large amounts of structured data easily.

● This laid the foundation for modern data analysis.

● Before invention of database data was often stored in physical files and spreadsheets.

2. Rise of Big Data and Cloud Computing (2010s)

● The explosion of digital data from social media, smartphones, and sensors created the need for big data tools like Hadoop, Spark, and cloud platforms.

● This allowed organizations to analyze massive datasets in real time, leading to smarter business and scientific decisions.

These two milestones database and big data technologies together shaped data analytics into the powerful decision-making tool it is today.

♦ Explain three current trends in data analytics (e.g., AI, cloud analytics, real-time data processing).

Ans: Here Explain three current trends in data analytics:

1. Data Democratization

● Explanation:

Data democratization means giving everyone in an organization regardless of technical background access to data and easy-to-use analytics tools. Instead of relying only on IT or data specialists, everyday employees can analyze information and make data-driven decisions.

● Example:

Sales teams using tools like Tableau or Power BI to create their own reports without needing a data analyst.

It empowers more people to use data, which speeds up decision-making and encourages a data-driven culture.

2. Augmented Analytics

● Explanation:

Augmented analytics uses artificial intelligence (AI) and machine learning to help users discover insights automatically. It reduces the need for manual analysis by highlighting patterns, suggesting visuals, or even writing reports for you.

● Example:

A tool that scans your sales data and tells you automatically that sales dropped in a specific region and suggests possible causes.

It saves time, improves accuracy, and helps non-experts uncover insights they might have missed.

3. Data Privacy and Ethical Analytics

● Explanation:

As the use of data grows, so does the need to ensure that it's collected and used ethically, securely, and in line with privacy laws (like GDPR). This trend focuses on protecting people's information and ensuring fair, unbiased analytics.

● Example:

Companies investing in privacy-preserving technologies or ensuring that AI models are free from discrimination.

Trust is essential. Ethical analytics helps organizations avoid legal issues, protect reputation, and treat data subjects fairly.

These three trends Data Democratization, Augmented Analytics, and Ethical Analytics are driving how data analytics is evolving today.

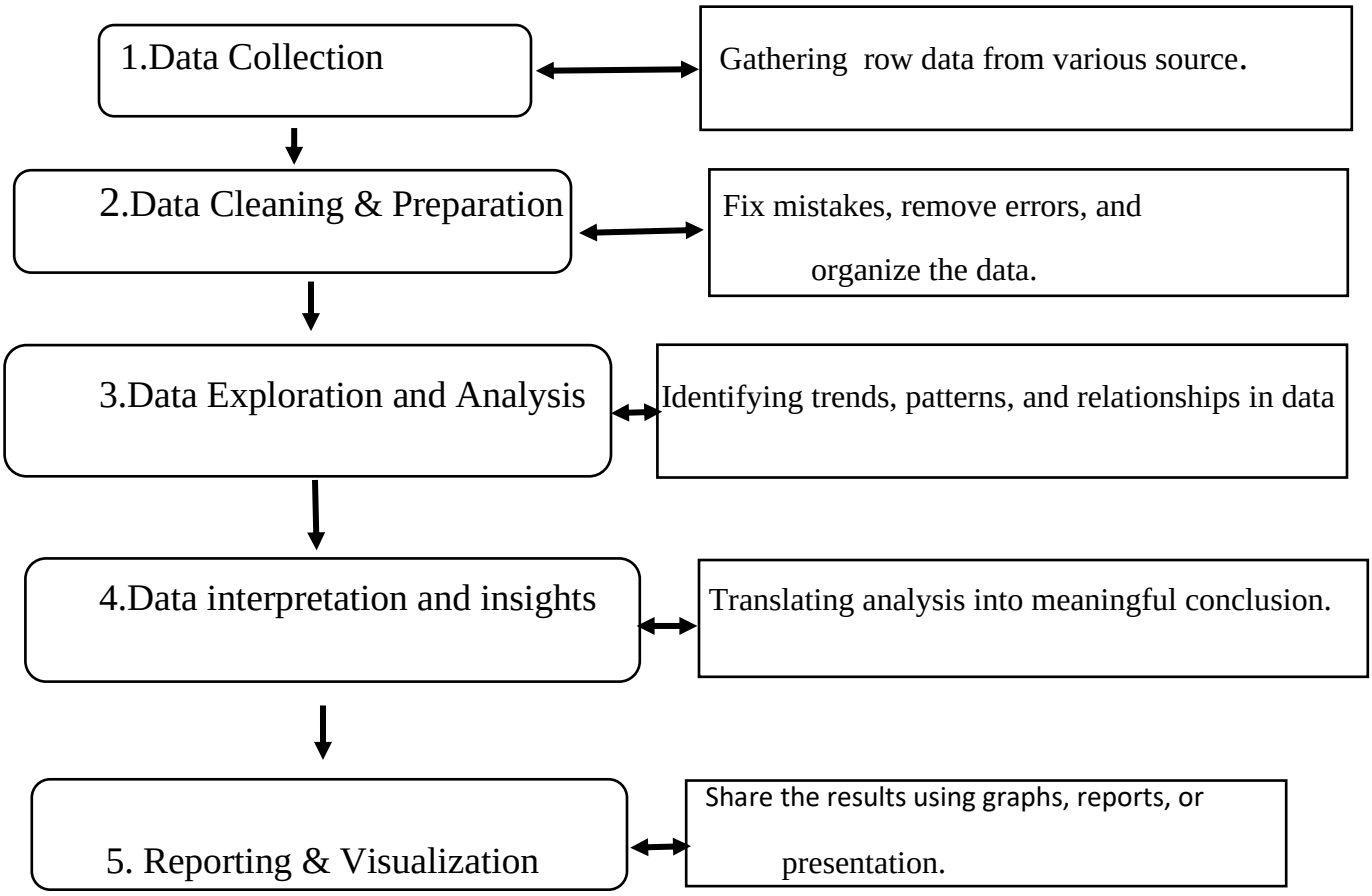
Task 2: Analytics Process Mapping

Draw and describe the end-to-end data analytics process using a diagram. For each stage (e.g., data collection, cleaning, exploration, analysis, reporting), briefly explain:

● What happens in that stage

● Why it is important

Ans: Here is a simple diagram of the end-to-end data analytics process:



Here's a brief and clear explanation for each stage of end-to-end data analytics process , with what happens and why each step is important:

1. Data Collection

What happens:

In this stage, data is gathered from different sources such as company databases, websites, social media, surveys, sensors, or business systems. The goal is to collect all the relevant information needed to answer a specific question or solve a problem.

Why it is important:

Without proper data collection, there would be no information to analyze. High-quality, relevant data is the foundation of any good analysis.

2. Data Cleaning and Preparation

What happens:

The collected data is checked for errors, missing values, duplicates, and inconsistencies. It is then organized into a usable format for analysis, such as correcting mistakes, filling gaps, or removing irrelevant data.

Why it is important:

If the data is messy or incorrect, the results of the analysis will also be wrong. Cleaning ensures the data is accurate, which leads to reliable and trustworthy insights.

3. Data Exploration

What happens:

The prepared data is explored using charts, graphs, tables, and basic statistical summaries. This helps to understand the overall structure of the data, identify patterns, spot outliers, and form ideas about what to analyze deeper.

Why it is important:

Exploration helps the analyst get familiar with the data, uncover hidden patterns, and avoid mistakes in the later stages. It sets the direction for more detailed analysis.

4. Data Interpretation and Insights

What happens:

The actual analysis is done using statistical methods, business logic, or predictive models to find answers to key questions. The results are then interpreted to generate insights that explain what the data is revealing.

Why it is important:

This step provides the meaningful information that businesses need to make decisions. Without proper interpretation, the analysis would be just numbers without understanding.

5. Reporting and Visualization

What happens:

The findings are presented through reports, charts, dashboards, or presentations so that non-technical audiences can easily understand the insights.

Why it is important:

Clear reporting ensures that the insights are communicated effectively, so decision-makers can take the right actions based on the data.

Each step builds on the previous one, and together they help turn raw data into useful business knowledge.

Task 3: Terminology and Tools

- Match the following terms to their correct definitions:

(Data types, Variables, Metrics, KPIs, Structured data, API, SQL, Data Warehouse, Pandas, tidyverse)

- Create a glossary table

Ans:

Terms	Definition	Limitation
Data types	The kind of data stored (like numbers, text, or dates).	Wrong data type can cause errors in analysis.
Variables	Names used to store values that can change (like sales, age, or name).	Incorrect or inconsistent variables can lead to wrong results.
Metrics	Numbers used to measure something (like sales, revenue, or clicks).	Numbers used to measure something (like sales, revenue, or clicks).
KPIs	The most important metrics that show if goals are being met (like profit margin, customer retention).	Choosing the wrong KPIs can lead to bad business decisions.
Structured Data	Neatly organized data in rows and columns (like in Excel or databases).	Hard to handle unstructured information like images or videos.
API	A way for different software to talk to each other and exchange data.	APIs can break if not updated or managed well.
SQL	A language used to work with databases (to find, add, or change data).	Not ideal for handling unstructured or huge real-time data.
Data Warehouse	A big storage system that holds data from different sources for analysis.	Expensive and sometimes slow for real-time analysis.
Pandas	A Python tool to help clean, organize, and analyze data easily.	Can get slow with very large datasets (big data).
tidyverse	A set of tools in R language for cleaning, analyzing, and visualizing data	Limited mostly to R users, not as popular in other programming languages.
Data Format	The way data is saved or arranged, such as text files (.txt), CSV, JSON, or XML.	Some formats are harder to work with or don’t support complex data (e.g., plain text).

NumPy	A Python library for handling numerical data and large arrays efficiently.	Limited for working with labeled or structured data (like tables); mainly for numeric analysis.
dplyr	An R package used to manipulate and summarize data easily, part of tidyverse.	Only works with R language; may not handle very large datasets efficiently.
Business Objectives	The specific results a company aims to achieve (like increasing sales or reducing costs).	Without clear objectives, it's hard to measure success or select the right KPIs.

Part B:Practical Tasks(Hands-on-Application)

Task 4: Data Collection Exploration

Choose one real dataset to work with from any of the following sources:

♦ Kaggle

♦ Google Dataset Search

♦ Government portals (e.g., data.gov)

♦ Or collect primary data via a Google Form

Answer the following:

● What is the source and format of the data?

➡ https://www.kaggle.com/c/titanic/data?select=train.csv

I got the data from Kaggle, a website where people can download datasets for learning and practicing data analysis. The Titanic data is in CSV format, which means it is like a table with rows and columns, similar to an Excel file.

● Is it structured/semi-structured/unstructured?

➡ The Titanic data is structured because everything is organized in a clear table. Each column has a label like Name, Age, Fare, Survived, and each row shows the details of a passenger. It is easy to read and use for analysis.

● What are the ethical considerations in using this data?

➡ When collecting secondary data we should consider the followings:

♦ Make sure you have the right to use the data.

♦ Protect privacy, even if you didn't collect the data.

♦ Use the data fairly and honestly.

Since we are using secondary data (data that was already collected and shared by others), we need to make sure we are using it only for the right reasons, like for learning, research, and analysis.

The Titanic dataset is public and anonymous, so there are no serious privacy concerns. However, we still need to be careful not to misuse the data or draw wrong or unfair conclusions from it.

Task 5: Data Profiling & Quality Checks

Using your dataset, perform:

● Data profiling (summary of number of rows, columns, data types, null values)

Ans:

Number of rows: 891 passengers(each row=one person)

Number of columns: 12 columns(Features like PassengerId, Survived, Pclas, Name ,Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked)

Data types:

- Object(text or mixed data): Name, Sex, Ticket, Cabin, Embarked.
- int(whole numbers): PassengerID, Survived, Pclass, Sibsp, Parch,
- float(numbers with decimal): Age, Fare

891 rows × 12 columns

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null    int64  
 3   Name         891 non-null    object 
 4   Sex          891 non-null    object 
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64  
 7   Parch        891 non-null    int64  
 8   Ticket       891 non-null    object 
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object 
 11  Embarked     889 non-null    object 
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

Null values: It means that some data is missing or not available in a dataset. Null Values can affect analysis and calculations.

We need to handle them by either:

1.Filling with a value(like mean,median)

2. Removing rows/columns with too many nulls

```
df.isnull().sum()
```

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

- Check for duplicates, missing values, and outliers

Ans:

Check duplicates: My data set has no duplicates. If duplicates are found in a dataset remove them cause keeping duplicates can mislead the analysis by counting the same information multiple times.

```
df.duplicated().sum()
```

```
0
```

Check for missing values: I found some missing values in the columns Age(177), Cabin(687), and Embarked(2).

Here two cases arise:

1.If missing few values like ‘Age’ and ‘Embarked’ fill using mean, median, or mode.

2. If missing too many values like ‘Cabin’ Drop the column.

```
df.isnull().sum()
```

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

Missing values can cause errors in analysis or models.

Check for outliers: We can find outliers in many different way like in pandas using ‘describe()’ function which shows the min, max, mean and quartiles for each column. If the max values is far away from the 75% quartile value the it’s mean a outliers. Or using boxplot in seaborn we can find outlies and also Using IQR method we can find outliers.

In Pandas I am using describe() function to find outliers. I find outliers in Age and Fare column as their max values is far away from 75%(Q3) values.

Outliers can skew result and make analysis less reliable. There is three situation arises:

1.Some times keep them if they are real and important

2. OR remove them if they are data entry errors or too rare to matter.

3. Reduces extreme values using IQR method

In My dataset ‘Fare’ often has outliers some people paid extremely high prices.

‘Age’ may also have very young babies and very old(80+) as outliers.

So I keep the outliers of age and I can reduces large values of Fare.

Check for outliers in all columns which contain numeric values.

```
[15]: numeric_columns=df.select_dtypes(include=['number'])
      print(numeric_columns.describe())
```

```
       PassengerId    Survived      Pclass         Age       SibSp  \
count   891.000000  891.000000  891.000000  714.000000  891.000000
mean    446.000000    0.383838    2.308642   29.699118    0.523008
std     257.353842    0.486592    0.836071   14.526497    1.102743
min       1.000000    0.000000    1.000000    0.420000    0.000000
25%     223.500000    0.000000    2.000000   20.125000    0.000000
50%     446.000000    0.000000    3.000000   28.000000    0.000000
75%     668.500000    1.000000    3.000000   38.000000    1.000000
max     891.000000    1.000000    3.000000   80.000000    8.000000

            Parch        Fare
count  891.000000  891.000000
mean     0.381594   32.204208
std      0.806057   49.693429
min      0.000000    0.000000
25%      0.000000    7.910400
50%      0.000000   14.454200
75%      0.000000   31.000000
max      6.000000  512.329200
```

Task 6: Data cleaning

Clean your dataset by:

♦ Handling missing values (mention chosen strategy: removal, imputation, etc.)

➡ In My dataset there is three columns(Age, Cabin, Embarked) where some values are missing.

Now I follow the strategies:

● Age: Some values(177) are missing .So I fill missing values with median age.(imputation)

● Cabin: Many values(687) are missing. So I drop the column as too much missing data.(removal)

● Embarked: Only two missing values. So fill missing values with most common values.

Handle missing values

```
# Fill age with median
x=df['Age'].median()
df.fillna({'Age':x},inplace=True)
# Drop cabin(too much missing values
df.drop('Cabin',axis=1,inplace=True)
# Fill Embarked with most common frequent value
y=df['Embarked'].mode()[0]
df.fillna({'Embarked':y},inplace=True)
```

```
df.isnull().sum()
```

```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Embarked       0
dtype: int64
```

♦ Removing duplicates

➡ Duplicates can mislead analysis. But my dataset has no duplicate value. If exist remove duplicates using drop_duplicates() function in pandas.

♦ Correcting formatting issues or inconsistent entries.

➡ The Titanic dataset has no wrong format issues.

But if there is any wrong formatting issues in a dataset then using astype() function to fix it.

Wrong formatting or inconsistent entries gives wrong counts, wrong average and misleading insights. Correct format makes filtering and grouping reliable.

Inconsistent Entries: If there is inconsistent entries then set it to NaN(value) or NaT(date) then drop the rows using dropna() function or fix it.

In my dataset there might be inconsistent entries in Sex and Embarked column for lower case(male, female) and uppercase (S ,Q,C).

Now I check it using value_count() function.

```
[ ]: df['Sex'].value_counts()

[ ]: Sex
     male      577
     female    314
     Name: count, dtype: int64

[ ]: df['Embarked'].value_counts()

[ ]: Embarked
     S    646
     C    168
     Q     77
     Name: count, dtype: int64
```

So there is no wrong formatting issues and inconsistent entries.

Task 7: Tool Evaluation & Reflection

Write a reflection on the tool you used:

♦ Why did you choose this tool?

➡ I chose Pandas because it is a popular and easy-to-use Python library for data analysis. It helps me to quickly read, clean, and explore datasets like the Titanic data.

♦ What were the strengths and limitations you experienced?

➡ **Strengths:**

- It makes data cleaning and summarizing simple.
- It works well with large datasets.
- There is lots of help and tutorials online.

Limitations:

- It needs coding knowledge, so small mistakes can cause errors.
- It is not a visual tool; I need to use other libraries to make graphs.

♦ Would you consider using another tool for future analytics projects?

➡ Yes, I would still use Pandas for most projects because it is very powerful and flexible. But for quick visual reports or when coding is not needed, I might use Excel or Power BI.

Task 8: Mini Case Study – Application of Data Analytics

Pick one domain (e.g., healthcare, education, e-commerce, sports, finance) and:

● Describe one real-world application of data analytics in that domain.

➡ From The above domain I pick sports.

Real World Application:

In cricket ,data analytics is used to improve team Performance and match strategies. Coaches and analysts Study player and match data to decide batting order ,Bowling Strategies, and field placement based on opponent's strengths and weaknesses.

● Mention what kind of data is used and what decisions are made from it.

The kind data is used:

♦ Player statistics (runs, strike rate, bowling economy, wickets)

♦ Match conditions (pitch type, weather)

♦ Opposition analysis (how certain players perform against specific bowlers or in different conditions)

♦ Fitness and injury data of players

Decisions are Made:

- ♦ Selecting the best players for upcoming matches.
- ♦ Choosing batting and bowling orders based on data.
- ♦ Setting fielding positions to reduce scoring chances.
- ♦ Managing player workload to avoid injuries.

Case Study: England Cricket Team's Data-Driven Strategy

- Uses data analytics and AI models to analyze players' form, fitness, and opposition patterns.
- Helps in selecting the best playing XI, setting field placements, and deciding batting and bowling orders.
- Contributed to England's success in winning the 2019 ICC Cricket World Cup.