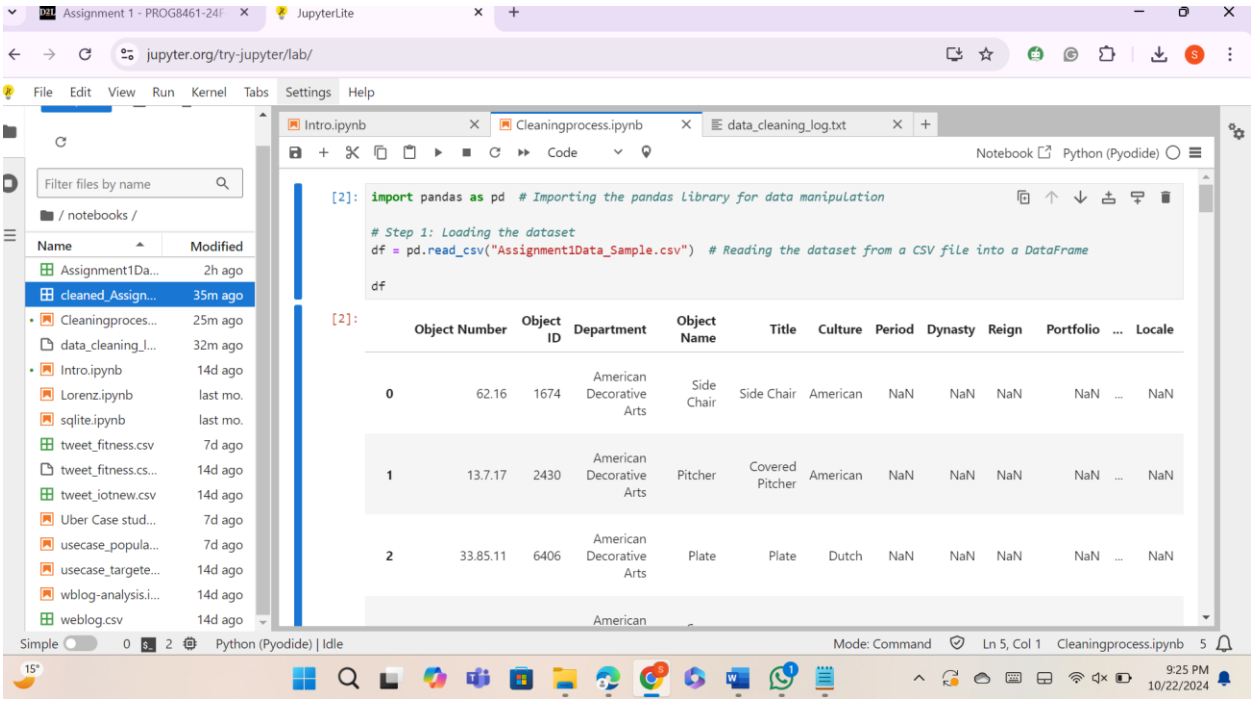
WEB ANALYTICS AND BUSINESS TOOLS

Assignment – 1

TASK 1:

Step 1: I have imported the pandas library and then loaded the dataset into the notebook.



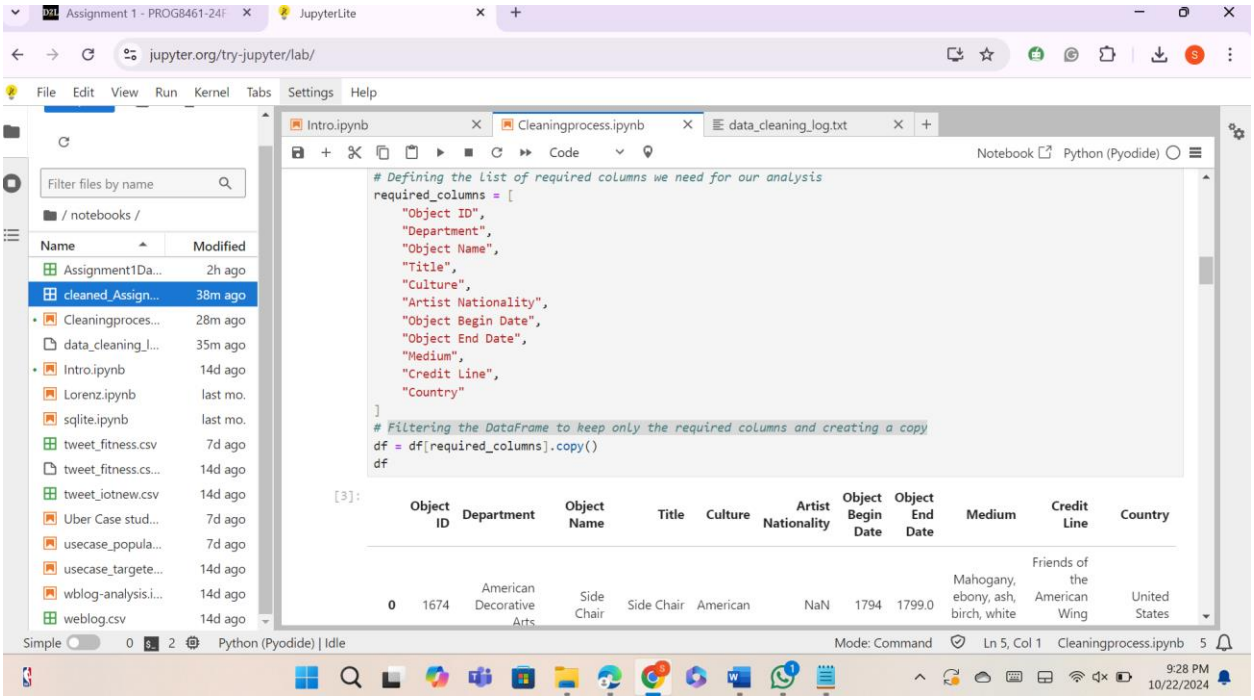
CODE:

```
import pandas as pd  # Importing the pandas library for data manipulation

df = pd.read_csv("Assignment1Data_Sample.csv")

df
```

Step 2: We are removing the unwanted columns and keeping only the required columns. I defined the list of required columns only for our analysis. Filtered the dataframe to keep only the required columns and creating a copy of that to maintain the original data for future reference.



CODE:

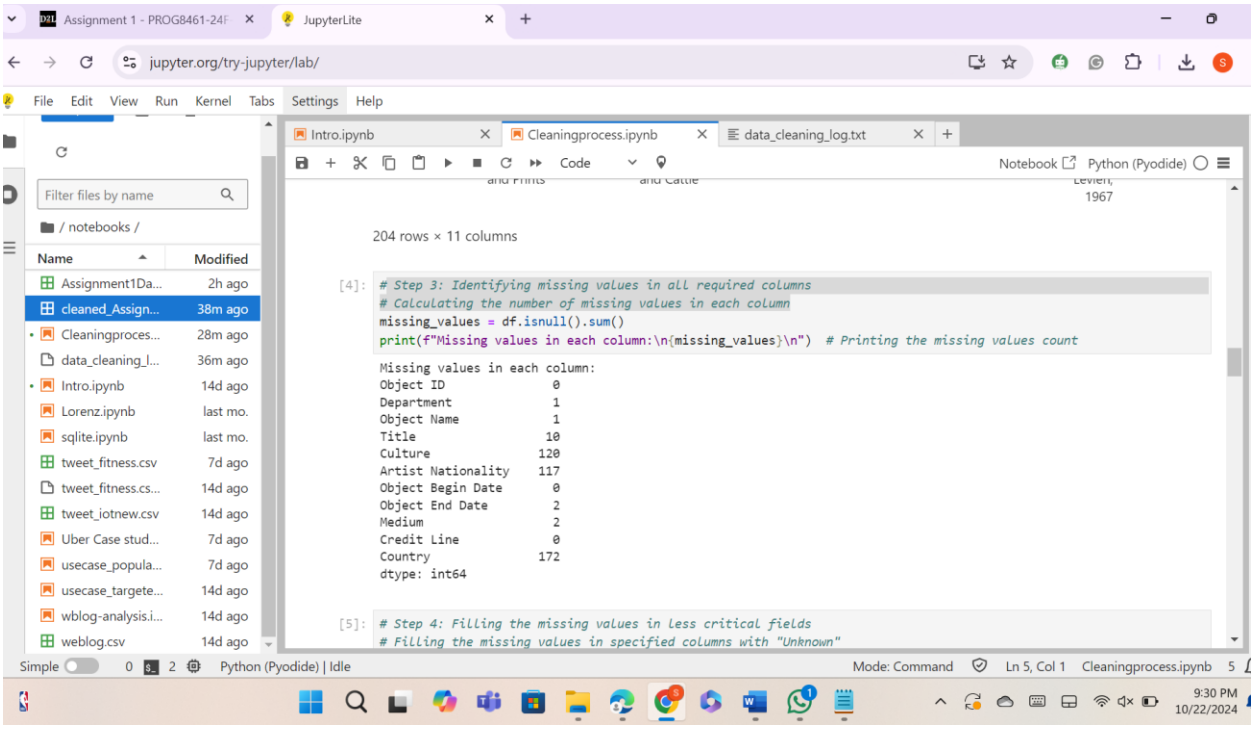
required_columns = [

"Object ID", "Department", "Object Name", "Title", "Culture", "Artist Nationality", "Object Begin Date", "Object End Date", "Medium", "Credit Line", "Country"]

df = df[required_columns].copy()

df

Step 3: In this step I tried to identify the number of missing values present in the dataset.

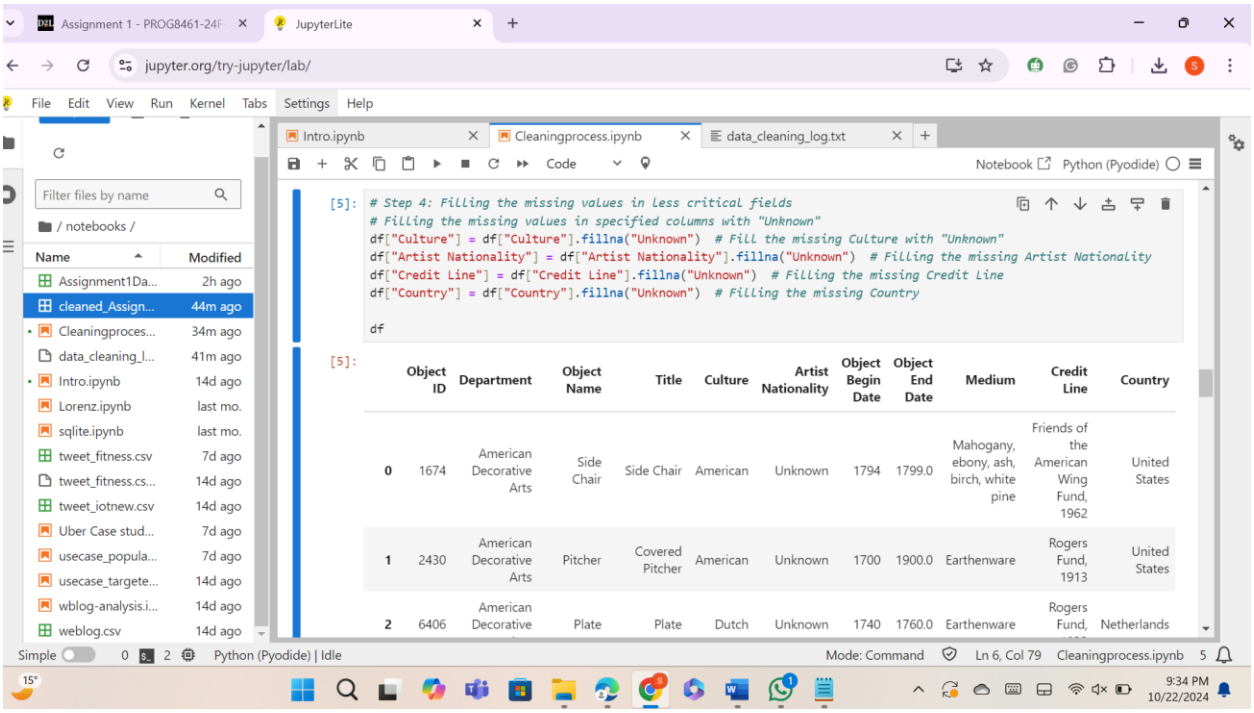


CODE:

```
missing_values = df.isnull().sum()

print(f"Missing values in each column:\n{missing_values}\n")
```

Step 4: Considering those missing values, to make the data consistant I replaced missing values with the value "unknown". Culture, Artist Nationality, Credit Line, missing Country these columns I have replaced with "Unknown".



	Object ID	Department	Object Name	Title	Culture	Artist Nationality	Object Begin Date	Object End Date	Medium	Credit Line	Country
0	1674	American Decorative Arts	Side Chair	Side Chair	American	Unknown	1794	1799.0	Mahogany, ebony, ash, birch, white pine	Friends of the American Wing Fund, 1962	United States
1	2430	American Decorative Arts	Pitcher	Covered Pitcher	American	Unknown	1700	1900.0	Earthenware	Rogers Fund, 1913	United States
2	6406	American Decorative	Plate	Plate	Dutch	Unknown	1740	1760.0	Earthenware	Rogers Fund,	Netherlands

CODE:

df["Culture"] = df["Culture"].fillna("Unknown")

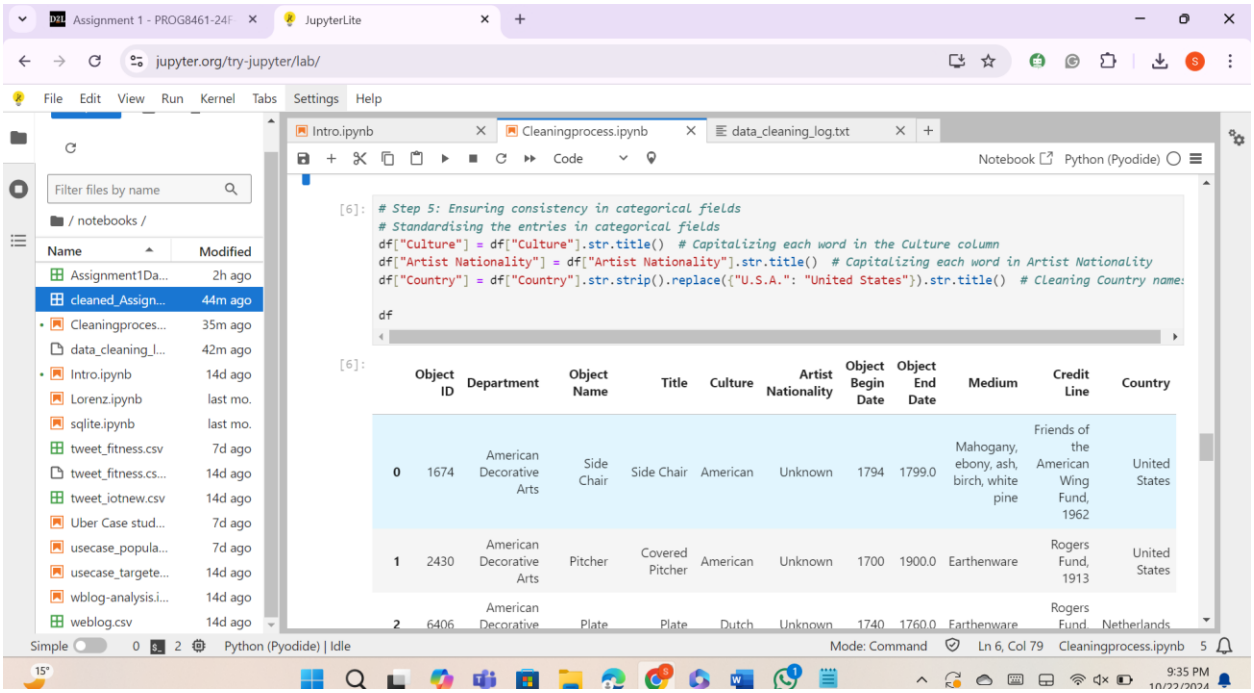
df["Artist Nationality"] = df["Artist Nationality"].fillna("Unknown")

df["Credit Line"] = df["Credit Line"].fillna("Unknown")

df["Country"] = df["Country"].fillna("Unknown")

Df

Step 5: Need to ensure consistency in categorical fields and capitalizing each word in the Culture column, Artist Nationality and cleaning Country names.



CODE:

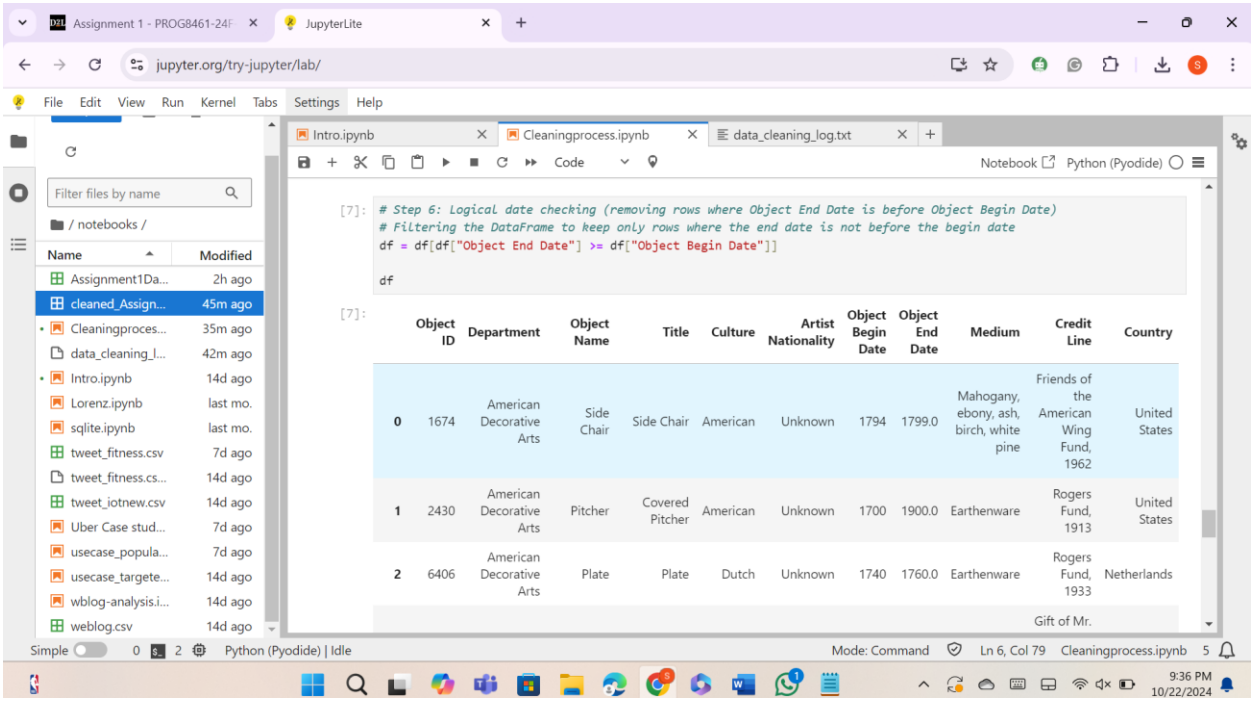
df["Culture"] = df["Culture"].str.title()

df["Artist Nationality"] = df["Artist Nationality"].str.title()

df["Country"] = df["Country"].str.strip().replace({"U.S.A.": "United States"}).str.title()

Df

Step 6: Removing the rows where Object End Date is before Object Begin Date and filtering the DataFrame to keep only rows where the end date is not before the begin date.

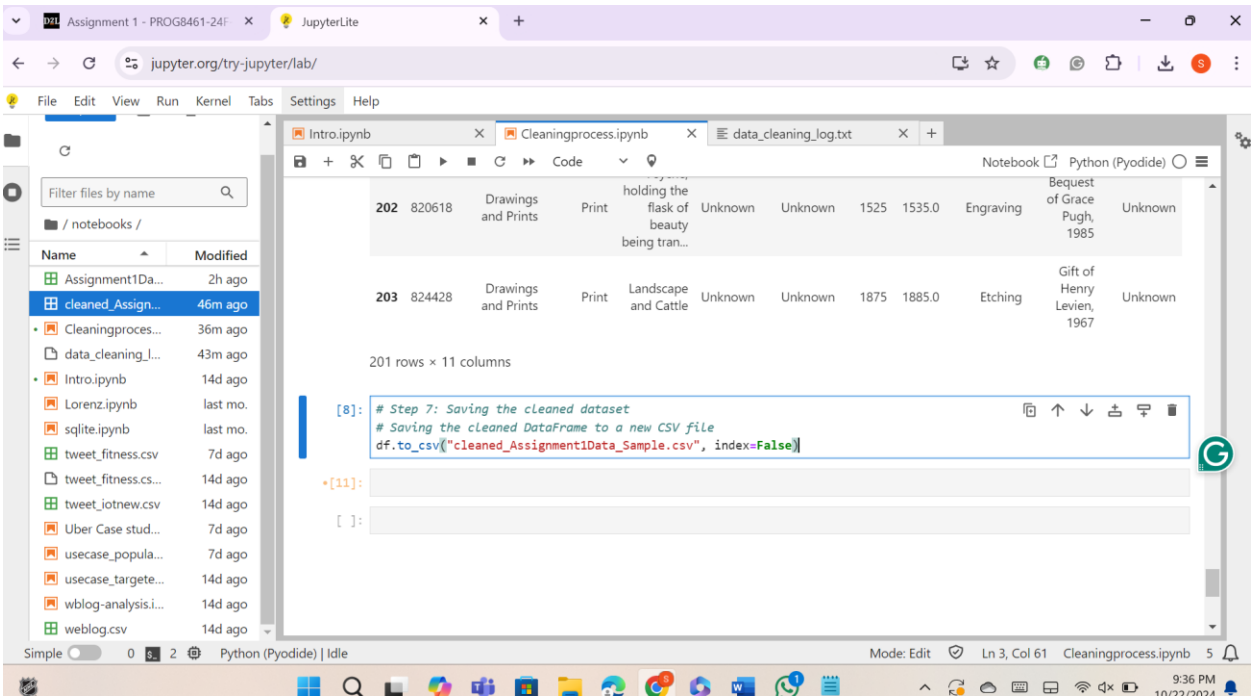


CODE:

df = df[df["Object End Date"] >= df["Object Begin Date"]]

Df

Step 7: At last I have saved the cleaned dataset to a new CSV file.



CODE:

df.to_csv("cleaned_Assignment1Data_Sample.csv", index=False)

TASK 2:

1) What are the 3 Vs of Data and explain each one in detail?

Ans:

- Volume: refers to the amount of data being generated, stored, and processed. Nowadays, organizations are dealing with large amounts of data from various sources such as social media, sensors, devices, transactions, etc. Petabytes of data collected and stored.
- Velocity: refers to the speed at which data is generated, processed, and analyzed. It emphasises the real-time or near-real-time nature of data. With data being generated continuously from sources like social media streams, financial markets, or IoT devices, there is a need to process and analyze data quickly to gain timely insights.
- Variety: refers to the different types or formats of data that are collected, stored, and analyzed. Traditionally, data was mostly organized in rows and columns in relational databases, text documents, images, videos, emails, social media posts.

2) List capabilities of Business Intelligence systems.

Ans: Business Intelligence systems provide organizations with tools and technologies to make data-driven decisions by transforming raw data into meaningful insights. The capabilities of BI systems:

- Data integration and ETL: collect data from multiple sources, including databases, spreadsheets, cloud services. They perform ETL processes to gather data, clean it, transform it into a consistent format, and load it into a data warehouse or a central repository.
- Reporting and dashboards: tools enable automated reporting and interactive dashboards that provide visualizations like charts, graphs, and heat maps that allow users to explore data insights in a user-friendly way.
- Key performance indicators (KPIs): systems enable organizations to define and track KPIs, helping monitor business performance.

3) Different types of data with examples for each type.

Ans:

- Structured data: data that is highly organized and formatted in a way that makes it easily searchable in databases. It fis organised into rows and columns and can be stored in relational databases.

  Examples: relational databases, spreadsheets.

- Unstructured data: data that does not have a predefined structure and is not organized in a manner that fits relational databases. It is in form of text, images, or videos and requires more processing to extract meaning.

  Examples: text documents, images, videos, and social media posts.

- Semi-structured data: data that does not reside in a traditional database but has some organizational properties that make it easier to analyze. Semi-structured data often uses tags or markers to define hierarchy and structure.

  Examples: XML/JSON files, NoSQL databases.

4) Define data visualization.

Ans: Data visualization means the graphical representation of data and information. It involves using visual elements such as charts, graphs, maps, and diagrams to present data in a way that makes it easier to understand and analyze patterns, trends, and insights. The primary goal of data visualization is to help users quickly grasp complex data, make informed decisions, and identify correlations or outliers that might not be evident in the raw data.

Characteristics: simplifies complex data, enhances decision-making, communicates insights effectively analytical knowledge.

5) What is a KPI and provide an example.

Ans: A KPI is a measurable value that demonstrates how effectively an organization or individual is achieving a specific objective or goal. KPIs are used to track progress, measure success, and help organizations make informed decisions. They are essential for monitoring performance in areas critical to business success, such as sales, customer satisfaction, and operational efficiency.

Characteristics of a good KPI: specific, measurable, achievable.

<u>Example:</u> let us say the KPI is specific (website traffic), measurable (a 20% increase), and time-bound (monthly), making it an actionable performance indicator.

KPI: increase monthly website traffic by 20%.

Measurement: tracks the number of unique visitors to the website each month using web analytics tools (e.g., Google Analytics).

Objective: drives more traffic to the website to improve brand awareness and lead generation.

6) What is a BI system?

Ans: A business intelligence system is a technology driven platform that helps organizations collect, integrate, analyze, and present business data to facilitate decision-making. BI systems transform raw data into meaningful insights that can drive strategic, tactical, and operational decisions. These systems integrate data from multiple sources and use analytical tools to provide a comprehensive view of business performance.

key functions of a BI System:

- Data collection and integration: these systems gather data from various sources like databases, cloud services, spreadsheets, and external sources.
- Data storage: the data is often stored in a data warehouse or other structured repositories where it is organized and cleaned for analysis.
- Data analysis: these systems offer analytical tools that allow users to perform in-depth analyses, including reporting, data mining, and predictive analytics.

7) What are the 5 C's of Data for data preparation and the purpose of each?

Ans: The 5 C's of data in data preparation are key steps that ensure data is ready for analysis. Each step plays a crucial role in making sure that data is accurate, clean, and suitable for generating meaningful insights.

- Collection: gathering data from various sources, including databases, external APIs, spreadsheets, and real-time data streams.
- Cleaning: removing or fixing errors, inconsistencies, duplicates, and inaccuracies in the data.
- Combining: merging data from multiple sources to create a unified dataset.
- Conforming: standardizing data to ensure consistency in formats, units, naming conventions, and data types.
- Consolidating: aggregating or summarizing data to generate meaningful insights or prepare it for further analysis.

8) What are some key success factors of a successful BI program and explain each factor?

Ans: A successful business intelligence program relies on several key success factors (KSFs) to ensure that it delivers value, insights, and supports data-driven decision-making across the organization. Here are some of the critical success factors:

- Clear business objectives: identifying clear use cases and specific business questions ensures that the BI tools deliver insights that support decision-making.
- Strong executive sponsorship: successful BI programs need buy-in from leadership. Having the executives or key decision-makers support the program ensures that it receives the necessary funding, resources, and organizational focus.
- Data governance and quality: high-quality data is essential for accurate insights.
- User training and adoption: end users need proper training to effectively use BI tools. This includes training on how to access reports, create custom dashboards, and interpret data visualizations.
- Cross-functional collaboration: a BI program often requires input from multiple departments. Collaboration ensures that the right data is captured and that insights address the needs of various business units.