|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **StudentID** | **URL for Word/PDF file** | **For Marking**  **(Yes/No)** |
| Hirenkumar Patel | 9029986 | [**Assignment 1.docx**](https://github.com/hpatel9986/Web_Analytics/tree/main/Assignment_1) | No |
| Sreejith Ajithkumar Nair | 9045575 | [**https://drive.google.com/drive/folders/1OoxVE\_baL\_p2-j20fkm0OnkEod6j3NxB?usp=sharing**](https://drive.google.com/drive/folders/1OoxVE_baL_p2-j20fkm0OnkEod6j3NxB?usp=sharing) | Yes |

**Task 1**

**1.** Data Overview  
data = pd.read\_csv("Assignment1Data\_Sample.csv")

data.head()

A screenshot of a computer

Description automatically generated

data.shape

data.info()

A screenshot of a computer

Description automatically generated

**2.** Data Selection

df = data[["Object ID","Department","Object Name","Title","Culture",

"Artist Nationality","Object Begin Date","Object End Date",

"Medium","Credit Line", "Country"]]

df.head()

A screenshot of a computer

Description automatically generated

df.shape

df.info()

A screenshot of a computer

Description automatically generated

**3.** Uniqueness Check

df['Object ID'].value\_counts()

df.drop\_duplicates(inplace=True) #dropping duplicates to get unique "Object ID"

df.shape

df['Object ID'].value\_counts() #"Object ID" is unique

A screenshot of a computer

Description automatically generated

**4**. Data Type Conversion

df.Department = df.Department.astype('category')

df.Culture = df.Culture.astype('category')

df['Artist Nationality']=df['Artist Nationality'].astype('category')

df['Credit Line'] = df['Credit Line'].astype('category')

df.Country = df.Country.astype('category')

df['Object End Date'] = df['Object End Date'].astype('Int64')

print(df.dtypes)

A screenshot of a computer

Description automatically generated

**5**. Missing Values Analysis

df.isna().sum()

df.Department.value\_counts(dropna=False)

df.Department = df.Department.cat.add\_categories('Unknown')

df.Department.fillna('Unknown', inplace =True)

df.Department.value\_counts(dropna=False)

df['Object Name'].value\_counts(dropna=False)

df['Object Name'].fillna('Unknown', inplace =True)

df['Object Name'].value\_counts(dropna=False)

df.Title.value\_counts(dropna=False)

df.Title.fillna('Unknown', inplace =True)

df.Title.value\_counts(dropna=False)

df.Culture.value\_counts(dropna=False)

df.Culture = df.Culture.cat.add\_categories('Unknown')

df.Culture.fillna('Unknown', inplace =True)

df.Culture.value\_counts(dropna=False)

df['Artist Nationality'].value\_counts(dropna=False)

df['Artist Nationality'] = df['Artist Nationality'].cat.add\_categories('Unknown')

df['Artist Nationality'].fillna('Unknown', inplace =True)

df['Artist Nationality'].value\_counts(dropna=False)

df.Medium.value\_counts(dropna=False)

df.Medium.fillna('Unknown', inplace =True)

df.Medium.value\_counts(dropna=False)

df.Country.value\_counts(dropna=False)

df.Country = df.Country.cat.add\_categories('Unknown')

df.Country.fillna('Unknown', inplace =True)

df.Country.value\_counts(dropna=False)

A screenshot of a computer

Description automatically generated

**6**. Year Conversion

df['Object End Date'] = df['Object End Date'].fillna(df['Object Begin Date'])

def convert\_to\_bc(year):

if year < 0:

return f"{abs(year)} BC"

else:

return str(year)

df['Object Begin Date'] = df['Object Begin Date'].apply(convert\_to\_bc)

df['Object End Date'] = df['Object End Date'].apply(convert\_to\_bc)

A screenshot of a computer

Description automatically generated

**7**. Row Dropping

df = df.drop(df[(df['Object Begin Date'] == '0') & (df['Object End Date'] == '0')].index)

df.shape

A screenshot of a computer

Description automatically generated

**8.** Data Consistency Checks

df.Department.value\_counts(dropna=False)

df['Object Name'].value\_counts(dropna=False)

df['Title'].value\_counts(dropna=False)

df['Culture'].value\_counts(dropna=False)

def clean\_culture(culture):

culture = culture.replace('possibly ', '').replace('probably ', '')

primary\_culture = culture.split(',')[0].strip()

return primary\_culture

df.Culture = df.Culture.apply(clean\_culture)

df['Culture'].value\_counts(dropna=False)

df['Culture'] = df['Culture'].replace('Japan','Japanese').replace('China','Chinese').replace('Iran','Iranian').replace('India (Guler','Indian').replace('Italic','Italian')

df['Culture'].value\_counts(dropna=False)

df['Artist Nationality'].value\_counts(dropna=False)

def clean\_nationality(nationality):

nationalities = nationality.replace('?', '').replace(', ', '|').split('|')

return ', '.join(sorted(set(nationalities)))

df['Artist Nationality'] = df['Artist Nationality'].apply(clean\_nationality)

df['Artist Nationality'].value\_counts(dropna=False)

df.Medium = df.Medium.replace('[no medium available]','Unknown')

df.Medium.value\_counts(dropna=False)

df['Credit Line'].value\_counts(dropna=False)

df.Country.value\_counts(dropna=False)

def clean\_country(country):

country = country.replace('present-day ', '')

return country

df.Country = df.Country.apply(clean\_country)

df.Country.value\_counts(dropna=False)

A screenshot of a computer

Description automatically generated

**Task 2**

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

The 3 Vs of Data refer to Volume, Velocity, and Variety.

**Volume**: The volume refers to the amount of data generated globally, signifying the magnitude of data produced, stored, and processed continuously. With billions of connected devices, such as smartphones and IoT sensors, enormous volume of data is generated every second. Traditional database which are designed for smaller datasets, often cannot accommodate this scale of data, which leads to adoption of distributed storage systems and cloud platforms. Major sources contributing to this high volume include social media interactions, e-commerce transactions, IoT data streams, and extensive scientific research datasets. [1]

**Velocity**: It denotes the speed at which data is generated and made available for analysis. For a business that rely on real-time information, the ability to process and respond to data instantaneously becomes crucial. This rapid influx of data, driven by digital transactions, real-time analytics, social media updates, and IoT devices, allows organizations to make swift decisions and adapt to changing circumstances. However, this velocity also raises concerns related to privacy and data security, necessitating robust strategies for managing real-time information. [1]

**Variety**: It highlights the diverse range of data types that organizations must handle. Today’s data landscape encompasses structured data (such as databases), semi-structured data (like XML and JSON files), and unstructured data (including social media posts, videos, and emails). This diversity presents significant challenges, as traditional data management systems often struggle to process unstructured data effectively. To address this, new storage solutions and analytical tools have emerged, enabling organizations to capture, store, and analyze the wide array of data formats.[1]

1. **List capabilities of Business Intelligence systems.**

* **Data Management**: BI systems should offer robust data management capabilities, allowing seamless connectivity to a variety of data sources, including cloud services and internal systems like ERP and CRM.
* **Data Lineage Tracking**: Effective BI tools enable users to trace the origin of data, ensuring its reliability by documenting its journey and transformations throughout its lifecycle.
* **Interactive Data Visualizations**: BI systems should provide interactive visualizations that allow users to create engaging presentations and easily interpret data, making complex insights more accessible.
* **Cloud Deployment**: The ability to deploy BI solutions in the cloud, on-premises, or through a hybrid model is essential for flexibility and scalability, meeting diverse business needs.
* **Customization and Integration**: BI tools should offer customization options and integration capabilities, allowing them to connect with various external and internal systems for a comprehensive view of data.
* **Reporting**: BI solutions must facilitate the creation and distribution of reports, both ad hoc and standard, with features that support data exploration and interactivity.
* **Dashboard Management**: A strong BI platform should provide dashboard solutions that unify data visualization and workflow management, presenting insights in an easily digestible format.
* **Collaboration**: Effective BI systems encourage collaboration by allowing teams to work together on dashboards and reports, facilitating quicker decision-making and insight sharing.
* **Predictive Analytics and Modeling**: Incorporating predictive analytics capabilities enables organizations to forecast trends and navigate uncertainties, helping to inform strategic decisions based on data-driven insights.[2]

1. **Different types of data with example for each type.**

**Qualitative Data**: Qualitative Data refers to data that describes characteristics or qualities and cannot be measured numerically. It is often used to categorize items or individuals.

It can be classified into two:

* **Nominal Data**: Nominal Data are those data that labels variables without any quantitative value and has no inherent order. Example: Colors of cars (red, blue, green).
* **Ordinal Data**: Data that involves order or ranking, where the categories have a meaningful sequence but the distances between them are not quantifiable are known as Ordinal Data. Example: Education levels (high school, college, university).

**Quantitative Data**: Quantitative Data is numerical and can be measured, often used to quantify variables and answer questions related to "how many" or "how much."

* **Discrete Data**: Definition: Data that represents countable values, limited to specific whole numbers and cannot be subdivided.

Example: Number of students in a classroom.

* **Continuous Data**: Definition: Data that can take any value within a given range, often representing measurements that can vary infinitely.

Example: Height of individuals (measured in centimeters).[3]

1. **Define data visualization.**

Data visualization is the process of representing data and information graphically. Utilizing visual elements such as charts, graphs, and maps, data visualization tools offer a clear and accessible means to identify trend, outliers, and patterns within the data.[4]

1. **What is a KPI and provide an example**

A Key Performance Indicator (KPI) is a measurable value that demonstrates how effectively an organization is achieving its key business objectives. KPIs are crucial for assessing performance and guiding strategic decisions. They provide objective measurements of critical aspects of a business, enabling stakeholders to evaluate success against specific targets.

Example of a KPI: Net Promoter Score (NPS): This KPI measures customer loyalty by asking customers how likely they are to recommend the company to others on a scale of 0 to 10. A higher NPS indicates stronger customer loyalty and a greater likelihood of referrals.

1. **What is a BI system?**

A Business Intelligence (BI) system refers to the processes, technologies, and tools required to transform data into meaningful information, convert that information into actionable knowledge, and ultimately develop plans that facilitate profitable business actions. BI includes elements such as data warehousing, analytical tools, and content or knowledge management systems.

1. **What are the 5 C's of Data for data preparation and the purpose of each?**

**Clean**: This involves removing or correcting incomplete, inaccurate, or irrelevant data points to ensure the integrity and accuracy of the dataset.

**Consistent**: Establishing standard relationships between data points is crucial. This consistency allows for reliable comparisons and analysis across the system.

**Conformed**: Standardizing the protocols for data analysis ensures uniformity in how data is interpreted and assessed throughout the system, enhancing clarity and comparability.

**Current**: It is essential to define the timeliness of data availability, ensuring that new data is accessible quickly to facilitate prompt decision-making.

**Comprehensive**: This ensures that all necessary data is available to support informed decision-making, covering all aspects required for thorough analysis.[5]

1. **What are some Key Success Factors of a Successful BI Program and explain each factor?**

* **Create a Company-Wide BI Strategy**

A comprehensive BI strategy anticipates and scopes the needs of all relevant business areas (e.g., finance, sales, IT). This ensures that the BI initiative is aligned with the overall business objectives and fosters organization-wide understanding and support.

* **Build Slowly**

Implement the BI program iteratively by defining specific segments aligned with business requirements. Starting small allows teams to build knowledge progressively and ensures that the solution effectively meets the actual needs of the users.

* **Achieve a Whole of Business View**

Integrating disparate data sources into a single BI system provides a unified view of the organization. This "single source of truth" allows for better decision-making and helps organizations turn data into actionable insights.

* **Get Fast Access to Your Data**

Fast access to data through dashboards and visualizations enhances user experience and facilitates timely decision-making. Self-service capabilities empower users to derive insights quickly without needing to navigate complex spreadsheets.

* **Capitalize on Real-Time Data**

Providing real-time insights allows users to monitor performance continuously and make informed decisions on the spot. This fosters user engagement and supports proactive management of business operations.

* **Collaborate**

Encouraging collaboration among stakeholders ensures that the BI solution meets diverse needs. Involving key customers in the project from the start enhances user engagement and generates a sense of ownership over the solution.

* **Ensure Business Needs Drive Technology**

The technology should be driven by actual business needs rather than the other way around. Understanding user requirements ensures that the BI tools deployed solve real problems and deliver actionable insights.

* **Build Competitive Advantage**

A successful BI strategy provides organizations with a competitive edge by revealing insights that can lead to strategic decisions. By utilizing data effectively, organizations can identify opportunities and address issues proactively, enhancing operational efficiency and profitability.[6][7]

**Reference**

1. The Knowledge Academy. *3 Vs of big data with example*. Retrieved from <https://www.theknowledgeacademy.com/blog/big-data-3v/>
2. Adair, B. (2021, February 11). *Business intelligence capabilities | BI capabilities list for 2021*. SelectHub. Retrieved from, <https://www.selecthub.com/business-intelligence/list-bi-capabilities/>
3. Tomar, J. (2024, September 21). *The 4 types of data in data science*. GUVI Blogs. Retrieved from <https://www.guvi.in/blog/types-of-data-in-data-science/>
4. *What Is Data Visualization? Definition, Examples, And Learning Resources.* (2024). *Tableau*. Retrieved from, <https://www.tableau.com/visualization/what-is-data-visualization>
5. Five CS of data. (2023). Kottmannconsulting.com. Retrieved from, [https://kottmannconsulting.com/Talking%20Technology/five\_cs\_of\_data.cfm](https://kottmannconsulting.com/Talking%20Technology/five_cs_of_data.cfmn)
6. Team, Y. (2024, October 2). *Top 5 keys to a successful Business Intelligence program: Macquarie Uni*. Yellowfin BI. <https://www.yellowfinbi.com/blog/top-5-keys-to-a-successful-business-intelligence-program-macquarie-uni>
7. Phocas Software. *7 keys to a successful business intelligence strategy*. Phocas Software. Retrieved from, from <https://www.phocassoftware.com/resources/blog/7-keys-to-a-successful-business-intelligence-strategy>