Batch: C2	**Roll No.: 16010122267**
Experiment 03	

Title: Importing Data and exploring the data

Objective:

1. To learn how to import dataset from various file format

- *Text, csv, pdf, excel, word*

2. To learn how to import dataset from various server

- Example (MySQL, MSSQL, Oracle, DB2, Google spreadsheets, Google drive, AWS, other)
- Minimum One connection with Server (Student choice mentioned in Objective 2)

3. Explore the data over platform

- Live data and Extracted data
- Data types
- Combining two data sources
- View data
- Sort option
- Measures and dimensions
- Splitting the column
- Discrete and continues values
- Drill down and Hierarchies
- Grouping

Course Outcome:

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

Books/ Journals/ Websites referred:

kaggle.com

https://help.tableau.com/current/pro/desktop/en-us/getstarted_buildmanual_ex1basic.htm

Resources used:

Tableau, Excel

Theory (About Data Preprocessing):

1. **Data Cleaning:** Data cleaning is the process of addressing missing values, duplicates, and errors in the dataset. This step ensures that the data is accurate and consistent. Missing values can be filled in, duplicates can be removed, and errors or inconsistencies can be rectified. Data cleaning is fundamental for maintaining the integrity and reliability of your dataset, as errors and omissions can lead to incorrect conclusions in your analysis or models.
2. **Data Transformation:** Data often needs to be transformed to be useful for analysis or modeling. This includes actions like scaling numerical features to a common range, encoding categorical variables into numerical formats (e.g., one-hot encoding), and creating new features through feature engineering. These transformations make the data compatible with the algorithms you intend to use and can reveal underlying patterns or relationships in the data.
3. **Data Splitting:** In order to evaluate the performance of your models, it's crucial to split the dataset into distinct subsets. Typically, this involves creating training, validation, and test sets. Training data is used to train the model, validation data is used to fine-tune it and make decisions about hyperparameters, and the test set is used to assess the model's performance. Data splitting ensures that the model is tested on data it hasn't seen during training, preventing overfitting and providing a realistic evaluation.
4. **Data Normalization:** Data normalization involves standardizing the distribution of your data. This is particularly important when features in your dataset have different scales. Standardizing the data ensures that all features contribute equally to the modeling process and helps improve the performance of algorithms that are sensitive to feature scaling. Common normalization techniques include z-score standardization and Min-Max scaling.
5. **Domain Knowledge:** Domain knowledge refers to expertise in the specific subject area or industry relevant to your data. Incorporating domain knowledge is crucial when making decisions during data preprocessing. It can guide choices related to feature engineering, handling missing values, and dealing with outliers. Domain experts understand the context of the data, which is valuable for making informed preprocessing decisions that align with the problem you're trying to solve or the insights you're seeking to gain. Domain knowledge enhances the relevance and quality of your data preprocessing efforts.

Following points should be written by students

Different approaches of importing dataset:

- Import from various file format (PDF, Excel, .CSV, .txt)
- Import from server

Platform used by the student:

Tableau

Working: (Screenshots of various file format imported in software)

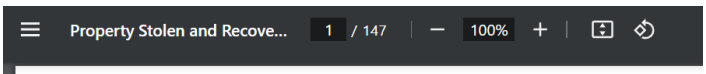
.CSV:

Series_reference	Period	Data_value	Suppressed	STATUS	UNITS	Magnitude	Subject	Group
BDCQ.SF1AA2CA	2016.06	1116.386		F	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)
BDCQ.SF1AA2CA	2016.09	1070.874		F	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)
BDCQ.SF1AA2CA	2016.12	1054.408		F	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)
BDCQ.SF1AA2CA	2017.03	1010.665		F	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)
BDCQ.SF1AA2CA	2017.06	1233.7		F	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)
BDCQ.SF1AA2CA	2017.09	1282.436		F	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)
BDCQ.SF1AA2CA	2017.12	1290.82		F	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)
BDCQ.SF1AA2CA	2018.03	1412.007		F	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)
BDCQ.SF1AA2CA	2018.06	1488.055		F	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)
BDCQ.SF1AA2CA	2018.09	1497.678		F	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)
BDCQ.SF1AA2CA	2018.12	1570.507		F	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)
BDCQ.SF1AA2CA	2019.03	1393.749		F	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)
BDCQ.SF1AA2CA	2019.06	1517.143		F	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)
BDCQ.SF1AA2CA	2019.09	1381.514		F	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)
BDCQ.SF1AA2CA	2019.12	1370.985		F	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)
BDCQ.SF1AA2CA	2020.03	1073.017		F	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)
BDCQ.SF1AA2CA	2020.06	1131.445		F	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)
BDCQ.SF1AA2CA	2020.09	1440.101		F	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)
BDCQ.SF1AA2CA	2020.12	1489.979		F	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)
BDCQ.SF1AA2CA	2021.03	1390.782		F	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)
BDCQ.SF1AA2CA	2021.06	1826.73		F	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)
BDCQ.SF1AA2CA	2021.09	1710.073		F	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)
BDCQ.SF1AA2CA	2021.12	1559.479		F	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)
BDCQ.SF1AA2CA	2022.03	1285.083		F	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)
BDCQ.SF1AA2CA	2022.06	1684.422		F	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)
BDCQ.SF1AA2CA	2022.09	1665.926		F	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)
BDCQ.SF1AA2CA	2022.12	1501.095		F	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)
BDCQ.SF1AA2CA	2023.03	1209.443		F	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)
BDCQ.SF1AA2CA	2023.06	1424.748		F	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)
BDCQ.SF1AA2CS	2016.06	1061.612		R	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)
BDCQ.SF1AA2CS	2016.09	1055.7		R	Dollars	6	Business Data Collection - BDC	Industry by financial variable (NZSIOC Level 2)

.XLSX

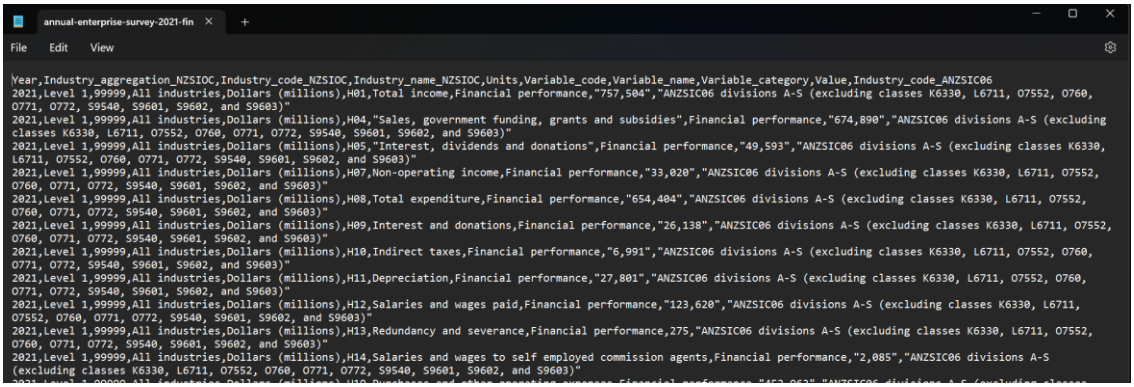
Area_Name	Ye	Group_Name	sub_group_name_fixed	Cases_Property_Recovere	Cas
Andaman & Nicobar Islands	2001	Burglary - Property	Burglary	27	
Andhra Pradesh	2001	Burglary - Property	Burglary	3321	
Arunachal Pradesh	2001	Burglary - Property	Burglary	66	
Assam	2001	Burglary - Property	Burglary	539	
Bihar	2001	Burglary - Property	Burglary	367	
Chandigarh	2001	Burglary - Property	Burglary	119	
Chhattisgarh	2001	Burglary - Property	Burglary	1169	
Dadra & Nagar Haveli	2001	Burglary - Property	Burglary	10	
Daman & Diu	2001	Burglary - Property	Burglary	7	
Delhi	2001	Burglary - Property	Burglary	642	
Goa	2001	Burglary - Property	Burglary	65	
Gujarat	2001	Burglary - Property	Burglary	1124	
Haryana	2001	Burglary - Property	Burglary	1257	
Himachal Pradesh	2001	Burglary - Property	Burglary	93	
Jammu & Kashmir	2001	Burglary - Property	Burglary	161	
Jharkhand	2001	Burglary - Property	Burglary	186	
Karnataka	2001	Burglary - Property	Burglary	1768	
Kerala	2001	Burglary - Property	Burglary	815	
Lakshadweep	2001	Burglary - Property	Burglary	0	
Madhya Pradesh	2001	Burglary - Property	Burglary	3526	
Maharashtra	2001	Burglary - Property	Burglary	3595	
Manipur	2001	Burglary - Property	Burglary	3	
Meghalaya	2001	Burglary - Property	Burglary	50	
Mizoram	2001	Burglary - Property	Burglary	188	
Nagaland	2001	Burglary - Property	Burglary	21	
Odisha	2001	Burglary - Property	Burglary	1388	
Puducherry	2001	Burglary - Property	Burglary	34	
Punjab	2001	Burglary - Property	Burglary	721	
Rajasthan	2001	Burglary - Property	Burglary	1926	
Sikkim	2001	Burglary - Property	Burglary	16	
Tamil Nadu	2001	Burglary - Property	Burglary	3227	

.PDF:



Area_Name	Year	Group_Name
Andaman & Nicobar Islands	2001	Burglary - Property
Andhra Pradesh	2001	Burglary - Property
Arunachal Pradesh	2001	Burglary - Property
Assam	2001	Burglary - Property
Bihar	2001	Burglary - Property
Chandigarh	2001	Burglary - Property
Chhattisgarh	2001	Burglary - Property
Dadra & Nagar Haveli	2001	Burglary - Property
Daman & Diu	2001	Burglary - Property
Delhi	2001	Burglary - Property
Goa	2001	Burglary - Property
Gujarat	2001	Burglary - Property
Haryana	2001	Burglary - Property
Himachal Pradesh	2001	Burglary - Property
Jammu & Kashmir	2001	Burglary - Property
Jharkhand	2001	Burglary - Property
Karnataka	2001	Burglary - Property
Kerala	2001	Burglary - Property
Lakshadweep	2001	Burglary - Property
Madhya Pradesh	2001	Burglary - Property
Maharashtra	2001	Burglary - Property
Manipur	2001	Burglary - Property
Meghalaya	2001	Burglary - Property
Mizoram	2001	Burglary - Property
Nagaland	2001	Burglary - Property
Odisha	2001	Burglary - Property
Puducherry	2001	Burglary - Property
Punjab	2001	Burglary - Property
Rajasthan	2001	Burglary - Property
Sikkim	2001	Burglary - Property
Tamil Nadu	2001	Burglary - Property
Tripura	2001	Burglary - Property
Uttar Pradesh	2001	Burglary - Property
Uttarakhand	2001	Burglary - Property
West Bengal	2001	Burglary - Property

.txt:



MySQL:

```sql
CREATE TABLE Person (
    Name VARCHAR(50),
    City VARCHAR(50),
    UID INT
);

-- Insert data for the three individuals
INSERT INTO Person (Name, City, UID)
VALUES
    ('Atharva Upare', 'Panvel', 1001),
    ('Agney Komath', 'Belapur', 1002),
    ('Atharva Bhosale', 'Dadar', 1003);
```

Person

Name	City	UID
Atharva Upare	Panvel	1001
Agney Komath	Belapur	1002
Atharva Bhosale	Dadar	1003

Conclusion (Students should write in their own words):

Through this experiment, we leant to import various data files of different formats and from data servers to Tableau.

Date: _____ **Signature of faculty in-charge**

Post Lab Question:

1. **List down types of data Tableau (any other you use) can import?**

Ans:

1. **Flat Files:** Tableau can import data from flat files such as CSV, Excel, and text files. These are among the most common data sources used in Tableau.
2. **Databases:** Tableau supports various database systems, including SQL databases like MySQL, PostgreSQL, Microsoft SQL Server, and Oracle, as well as NoSQL databases like MongoDB.
3. **Cloud-Based Data Sources:** Tableau can connect to cloud-based data sources, such as Amazon Redshift, Google BigQuery, and various data warehouses on cloud platforms.
4. **Web Data Connectors:** Tableau has web data connectors that allow you to connect to data from various web-based sources and APIs, including web services and online platforms.
5. **Live Data Connection:** You can establish live connections to data sources, enabling real-time or near-real-time data analysis and visualization.
6. **Data Cubes:** Tableau can connect to multi-dimensional data sources and OLAP (Online Analytical Processing) cubes for more complex data analysis.
7. **Big Data:** Tableau integrates with big data platforms like Hadoop and Spark, allowing you to analyze and visualize large-scale data.
8. **Geospatial Data:** Tableau has built-in support for geospatial data, making it easy to work with location-based data, geographic information, and mapping.
9. **Web Data:** You can scrape data from websites and use it for analysis and visualization in Tableau.
10. **Data Extracts:** Tableau allows you to create data extracts (Tableau Data Extract - .hyper files) for faster data processing and visualization, especially with large datasets.
11. **Statistical Files:** Tableau can import statistical files, including SAS, SPSS, and R data files.
12. **Custom Data Sources:** Tableau provides options for creating custom connectors to import data from sources that aren't natively supported.

2. **What is significance of Measures and Dimensions in dataset stored in Tableau(any other you use)?**

Ans:

In Tableau and similar tools, measures are essential for quantitative data, enabling calculations and aggregation of numerical values, such as sales revenue or profit margins. Dimensions, on the other hand, provide context and structure by categorizing data into discrete groups or categories, like product names or geographic locations. This differentiation empowers effective data visualization, filtering, and analysis, ensuring that data insights are meaningful and actionable for informed decision-making. By correctly identifying measures and dimensions, you can create impactful visualizations that deliver valuable insights from your dataset.