**Batch:H2-4 Roll No.:16010122257**

**Experiment 01**

|  |
| --- |
| **Title: Data Collection and finalizing dataset from problem domain** |

# Objective:

# 1. To learn how to collect the dataset

# 2. To learn sources of dataset

# 3. To assess the dataset based on Metrics to Measure Data Quality

# 4. To finalize the features of dataset

# 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:

# <https://www.kaggle.com>

# <https://data.gov>

# <https://data.world/datasets/health>

# Resources used:

# <https://www.kaggle.com/datasets/kaggleprollc/suicides-in-india-data-collection>

# Theory:

First,we’ve to determine the type of and,the specific,data we require.Then,select suitable methods/tools to gather the information in an effective manner.Thorough validation of the collected data is to be done to ensure its accuracy.At last,address any missing or incomplete data points in an appropriate manner.

# Following points should be written by students

# Problem domain (Healthcare, Ecommerce, Education, Finance, agriculture etc.)

# Motivation for the selected Domain

# Problem Statement

# Brain stormed features in problem statement (Based on Domain Selected) and its importance

# Search for dataset

# Justification for choosing above dataset

# Source of dataset (Link Needs to be given)

# Sample of Finalized dataset (First 5 Records)

# Data Dictionary

# Column wise summary

# Problem domain:Healthcare

# Motivation for the selected domain: Suicide is an emerging problem in India. In 2021, 1.64 lakh people took their own lives, according to the NCRB. The suicide rate jumped to 11.3 in 2020 and was at a record high at 12 in 2021. According to the WHO estimates, India has the 41st highest suicide rate globally, as of 2019. Thus,I selected this topic,because it’s a rising problem and needs to be addressed.

# Problem Statement: To analyze the types and causes of death by suicide in India.

# Dataset features:Causes behind suicide,ranging from bankruptcy and drug abuse to professional and career problems,categorized by age ranges below 18,18-30,30-45,45-60 , 60 and above years and total,of three different genders-male,female,transgender,along with the grand total.

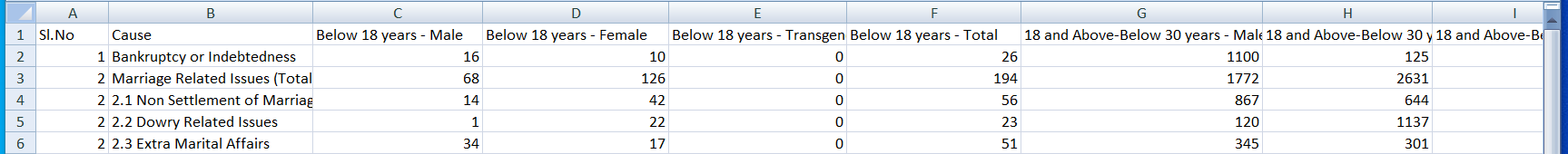
# Dataset search:Explored sites containing various datasets like Kaggle,Data World, US gov. etc.

# Justification:Relevant for analyzing and addressing the issues plaguing the people and helping them out,on a large scale(helplines,relaxing norms in educational institutes,etc.).

**Dataset source:** [**https://www.kaggle.com/datasets/kaggleprollc/suicides-in-india-data-collection**](https://www.kaggle.com/datasets/kaggleprollc/suicides-in-india-data-collection)

**Data dictionary:**Clarification of the various healthcare terminologies and variables related to no. of deaths,deaths by categories ,etc.

**Column-wise summary:**Causes of deaths,deaths by range of age and gender.

**Sampleofdatasets: **

# koil.png

# koil2.png

# Conclusion (Students should write in their own words):To conclude,I learnt how to choose an appropriate dataset from proper and genuine sources,and visualize,brainstorm ideas to obtain valuable insights from the raw data.

# Post Lab Question:

# Explain Role of Data in the Application Design.

# Proper data handling and management are essential for creating effective and efficient applications, create a better user experience and influence user behavior more effectively.

# 1.Data Representation and Storage: One of the primary roles of data in application design is to represent and store information. Different types of data, such as user input, configuration settings, and application state, need to be structured and stored appropriately to facilitate easy access and retrieval.

# 2.Functionality and Behavior: The data defines how an application functions and behaves. The logic and algorithms within the application often manipulate and process data to generate meaningful results. The processing of data ultimately determines the application's behavior and output.

# 3.User Interface (UI) and User Experience (UX): Data influences the design of the user interface and user experience. How the data is presented to users, how they interact with it, and how they input information all impact the usability and appeal of the application.

# To summarise,data plays a big role in how an app works.It helps decide the organization of things and how users interact with the app.Data helps in making the app smart,like understanding what users want.It acts like the app’s brain,making it run well,keeping information safe,and helping people use it easily.

# Write different types of Data with Example.

# There are 4 Types of Data: Nominal, Ordinal and Discrete,Continuous,grouped under Qualitative and Quantitative Data,respectively.

# Nominal: Nominal Data is used to label variables without any order or quantitative value. The color of hair can be considered nominal data, as one color can’t be compared with another color. With the help of nominal data, numerical tasks can’t be performed and not possible to give an order to sort the data. These data don’t have any meaningful order; their values are distributed into distinct categories.

# Examples- Nationality (Indian, German, American)

# Gender (Male, Female, Others)

# Eye Color (Black, Brown, etc.)

# Ordinal: Ordinal data are naturally ordered,where a number is present in some kind of order via their position on the scale. These data are used for observation like customer satisfaction, happiness, etc., but arithmetical tasks can’t be performed on them.Ordinal data are somewhat ordered compared to nominal data,which don’t have any meaningful order.

# Examples- Letter grades in the exam (A, B, C, D, etc.)

# Position of teams in a tournament(First,Second,Third,etc.)

# Economic Status (High, Medium, and Low)

# Discrete: Discrete data is a count that involves only integers. The discrete values cannot be subdivided into parts i.e. discrete data can take only certain values. The data variables cannot be divided into smaller parts.

# Examples-Set Theory,Relations:Uses

# The number of students in a class.

# The number of questions answered correctly in a test.

# Continuous:This is the type of data that can meaningfully be divided into finer levels.It can be measured on a scale and can have almost any numeric value-integer,floating point,etc.Continuous data can be recorded at many different measurements-width,temperature,time,etc.A good rule for defining a data as continuous or discrete is that,if the point of measurement can be reduced in half and still make sense, the data is continuous.

# Example-The amount of time required to complete a project.

# The height of a tree.

# The length of a snake.

# The speed of cars.

# Qualitative Data:Nominal data,Ordinal data

# Represents non-numeric information like audio,images,symbols or text.E.g.Customer reviews,social media posts,etc.

# Quantitative Data:Discrete data,Continuous data

# Can be expressed in numbers and used for statistical analysis.E.g.Age of students in a class,and their respective heights,etc.