**Batch: H1-3 Roll No.: 16010122083**

**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/datasets/tahzeer/unicorn-startup-companies-july-2023?resource=download

# Resources used:

Google

Kaggle

MS Excel

# Theory:

The Problem domain is based on the Finance and Data of Unicorn Start-ups in July 2023.

I have been working upon the case study and analysing start-ups and large, mid and low cap companies for a long time. So I have decided to analyse the dataset of the unicorn start-ups of various origin and showing the name of the investors.

The dataset when analysed we can find out the joining data and the company valuation achieved in the how many years and months. The analysis to show countries with the maximum number of unicorns in it. Find out the Sector with maximum number of Unicorn.

The dataset was found on Kaggle, it is list of all the unicorns stating their evaluation, data of joining, country of origin, industry of unicorn and their investors.

The dataset was perfect for use because the data is comprehensive and contains al the categories that will be needed for the analysis as well as enough number of entries for the case study.

The Dataset was obtained from:.

<https://www.kaggle.com/datasets/tahzeer/unicorn-startup-companies-july-2023?resource=download>

This of First 5 data value of the dataset:



Sr\_no: Serial Number

Company: Names of the Unicorn companies

Valuation: Stating the company evaluation of the unicorn company

Date\_joined: The date when the company got registered

Country: The Country of Origin of the unicorn.

City: The City of Origin

Industry: Industrial sector of the Unicorn

Investors: Investor names of the given Unicorn.

Column Stats:

Country with maximum number of Unicorns: United States (54%)

Enterprise industry has the highest stack in the list of unicorns. (30%)

Maximum unicorns have been register in 2020-2021 (554 out of 1216)

Following points should be written by students

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

# Motivation for the selected Domain

# Brain stormed features of Dataset (Based on Domain Selected)

# 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

# Conclusion (Students should write in their own words):

# I learnt how to study data from a large dataset and analyze and find useful data from it. Also add the analysis to my case study for completing my problem statement.

# Post Lab Question:

# Explain Role of Data in the Application Design.

# Data plays a pivotal role in application design, shaping its functionality, efficiency, and user experience. It serves as the foundation upon which the application operates, influencing every decision from feature implementation to user interactions. Data informs the design of databases, data models, and storage mechanisms, determining how information is structured, stored, retrieved, and updated.

# Incorporating the right data allows applications to deliver accurate and meaningful insights, driving informed decision-making. Effective data management ensures data integrity, security, and scalability, ensuring the application's performance remains optimal as user and data loads increase. Furthermore, data-driven design enables personalized user experiences, tailoring content and functionality to individual preferences and behaviors.

# Overall, data-centric design empowers applications to provide relevant, responsive, and reliable solutions, enhancing user satisfaction and enabling businesses to leverage insights for innovation and growth. Therefore, meticulous consideration of data is fundamental for crafting successful and impactful applications.

# Write different types of Data with Example.

# Nominal: Colour of hair (Blonde, red, Brown, Black, etc.)

# Ordinal: Letter grades in the exam (A, B, C, D, etc.)

# Discrete: Age in years

# Continuous: Temperature

# Boolean data: True and False

# Time-Series Data: Stock market prices over the course of a month, Temperature readings taken every hour in a day , Website traffic data over a week.