PRINCE KHENI

DATA ANALYSIS AND VISUALIZATION

HOMEWORK – 5

❖ ABOUT THE DATASET (Source, Number of variables, Quality of data, Limitations):

➢ **SOURCE:** The dataset has been taken from Kaggle and the link is attached below.

Link: https://www.kaggle.com/datasets/deepcontractor/unicorn-companies-dataset

➢ **NUMBER AND QUALITY OF VARIABLES/DATA:**

- A unicorn company is a private company with a valuation of more than $1 billion, and today there are over 1,000 unicorn companies around the world!
- This dataset contains a csv table with 1,074 records, one for each company.
- There are 10 columns whose description is given in the data dictionary below.
- Each record contains details on the company's current valuation, total funding, country of origin, industry, select investors, and the years they were founded and became unicorns.
- Some of the values in the dataset were missing, and data preprocessing is being done on the dataset.
- We also created a new columns namely Investor Counts, with help of website ‘Crunchbase’.

- **LIMITATION:**

- It was tough to get all the exact investor count on excel file from 'Crunchbase' website. We used some fetching techniques to do it.
- Also, there are many missing values in the dataset, and these are the limitations of our dataset.

Data Dictionary:

Field	Description	Data Type
Company	Name of the company	String
Valuation	Company valuation in Billions (B) of dollars	Integer
Year Joined	The year in which the company reached $1 billion in valuation	Date
Industry	Industry/Sector in which company is established	String
City	City the company was founded in	String
Country	Country the company was founded in	String
Continent	Continent the company was founded in	String
Year Founded	Year the company was founded	Date
Funding	Total amount raised across all funding rounds	Integer

	in Billions (B) or Millions (M) of dollars	
Select Investors	Top 4 investing firms or individual investors (some have less than 4)	String
Investor Count	Number of Investor the company/startup has	Integer
Valuation to Funding Ratio	Ratio of Valuation divided by Funding	Float

- **Technical Approach**

From this dataset, we are trying to create a story where we will be following the following technical steps on our data in order to achieve the best insights.

1. Data Collection
2. Data Exploration
3. Data Analysis
4. Data Visualization
5. Reporting and Presentation

Since, this are the major steps and we have not elaborated these as of now. We can increase the number of steps as we begin to work on the project.

- **Analytical Softwares that will be used on this project:**

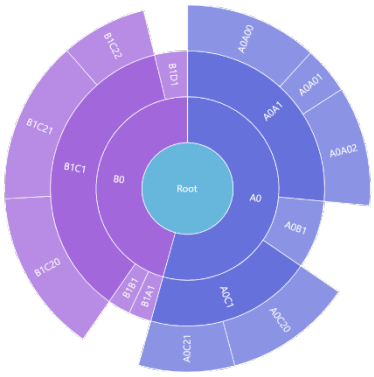
We will be using mostly three software that are listed below. These are not limited, and we can use any other software as well, if required.

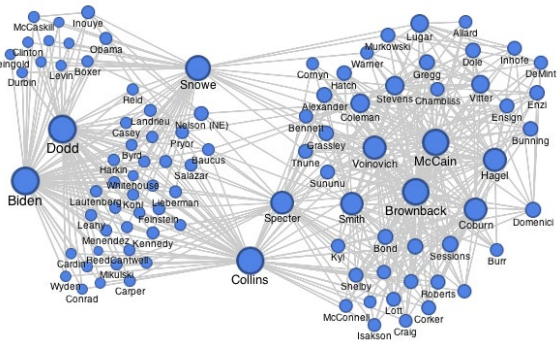
1. Tableau => Mainly for visualization purposes.
2. Python => Mainly for Network Graphs.

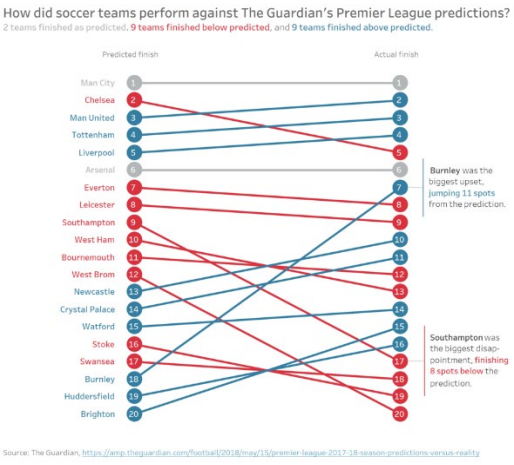
3. R => Mainly for data cleaning and engaging with variables.

❖ **Types of visualization that we'll be creating:**

There are approximately 5 to 6 charts we will be creating which will include Pie Chart, Bar Chart, Network Chart, Slope Chart, and many others. We are not yet fixed with the exact number of charts, and for that we will stick to the number below 10. Some of the visual are attached below for reference. We will try our best to replicate this visual in our project.







❖ **Exploratory Data Analysis:**

- We found out that there are a few companies which are bankrupt, but their data is still there. So, it means that the dataset contains data of all companies which are either still unicorns or bankrupt.
- We made a new column named 'Investor Count' using the website 'Crunchbase', which will be useful to create some interesting visualization.
- There are many EDA that are yet to be done. We'll be attaching those in our final report and presentation.

❖ **Possible Challenges:**

- There are few challenges like, we don't have much numerical data. So, it is hard to create interesting visuals on it.
- Also, there are many missing values in the dataset, so it will be a tedious job to clean it before visualizing.
- Other than that, we have decided to create a network graph, and it would be challenging to make it.
- There will be many other difficulties which would arise, and we will try our best to make out project interesting.

❖ **Questions we'll be trying to answer:**

- Which industry has more unicorns so far and how long does it take to become a unicorn?
- Who are the major investors in the top startup companies?
- Which sector is raising more money and are those sectors/industries most values by their valuation?

- We are still trying to create some interesting hypothesis which will help us to create some unique and interesting visualization.

!!!THANK YOU FOR READING THIS BRIEFING!!!