THE UNIVERSITY OF
SYDNEY

# Assignment Coversheet – Individual Assignment

Please fill in your details below. Use one form for each assignment.

## Personal Details:

| Family Name | Given Name (s) | Student Number (SID) | Unikey | Signature |
|---|---|---|---|---|
| **Kosidin** | **Aaron Wilbert** | **480011455** | akos8517 | Aaron |

## Assignment Details:

| | | | |
|---|---|---|---|
| Assignment Title | | **Individual Report** | |
| Assignment Number | | **02** | |
| Unit of Study Tutor | | **Henry Bian** | |
| Tutorial ID | | **CC-05** | |
| Due Date | **06 June 2023** | Submission Date | **06 June 2023** |

## Declaration:

1. I understand that all forms of plagiarism and unauthorised collusion are regarded as academic dishonesty by the university, resulting in penalties including failure of the unit of study and possible disciplinary action.
2. I have completed the **Academic Honesty Education Module** on Canvas.
3. I understand that failure to comply with the Academic Dishonesty and Plagiarism in Coursework Policy can lead to the University commencing proceedings against me for potential student misconduct under Chapter 8 of the ***University of Sydney By-Law 1999*** (as amended).
4. This work is substantially my own, and to the extent that any part of this work is not my own I have indicated that it is not my own by acknowledging the source of that part or those parts of the work.
5. The assessment has not been submitted previously for assessment in this or any other unit, or another institution.
6. I acknowledge that the assessor of this assignment may, for the purpose of assessing this assignment may:
   a. Reproduce this assignment and provide a copy to another member of the school; and/or
   b. Use similarity detection software (which may then retain a copy of the assignment on its database for the purpose of future plagiarism checking).
7. I have retained a duplicate copy of the assignment.

| | |
|---|---|
| Please type in your name here to acknowledge this declaration: | Aaron Wilbert Kosidin |

## 1. Introduction

### 1.1 Data set

The dataset I am working with is the Unicorn dataset obtained from Kaggle. This dataset contains information on over 1000 companies, with 13 columns providing various details about each company. The columns in the dataset include the company name, valuation, date joined (in datetime format), country, city, industry, select investors, founded year, total raised amount, financial stage, investors count, deal terms, and portfolio exits.

The company column consists of string values representing the names of the companies. The valuation column contains numeric values representing the monetary valuation of each company. The date joined column is in datetime format, indicating the date on which each company joined the dataset. The country and city columns provide information about the location of the company, expressed as string values.

The industry column represents the industry to which each company belongs, using string values to categorize them. The select investors column contains string values indicating the notable investors associated with each company. The founded year column consists of numeric values representing the year in which each company was founded.

The dataset also includes columns related to financial aspects of the companies. The total raised column represents the amount of money raised by each company, expressed in numeric values. The financial stage column describes the stage of financing for each company, using string values to indicate different stages such as seed, series A, series B, etc. The investors count column provides the number of investors associated with each company, while the deal terms column represents numeric values related to the terms of the investment deals made by each company. Lastly, the portfolio exits column contains numeric values indicating the number of exits from the company's portfolio.

Overall, this Unicorn dataset provides comprehensive information about a wide range of companies, including their valuation, location, industry, financial details, and more. It serves as a valuable resource for exploring and analyzing the characteristics and trends within the unicorn startup ecosystem.

### 1.2 Summary of Contribution

During our group work, my assigned task was to tackle the complex task of visualizing temporal dynamics. I successfully created a visual representation that showcased the evolving relationships and trends over time within our dataset. The temporal dynamic visualization provided valuable insights into the changing patterns and dynamics of the variables we examined.

For my individual assignment, I aim to explore a completely different realm of visualizations that diverges from the group work assignment. By opting for a different type of visualization, I hope to broaden my skills and showcase a diverse range of visualization techniques. This approach will not only demonstrate my versatility in handling various types of data and visualizations but also provide a comprehensive overview of my capabilities in the field of data visualization.

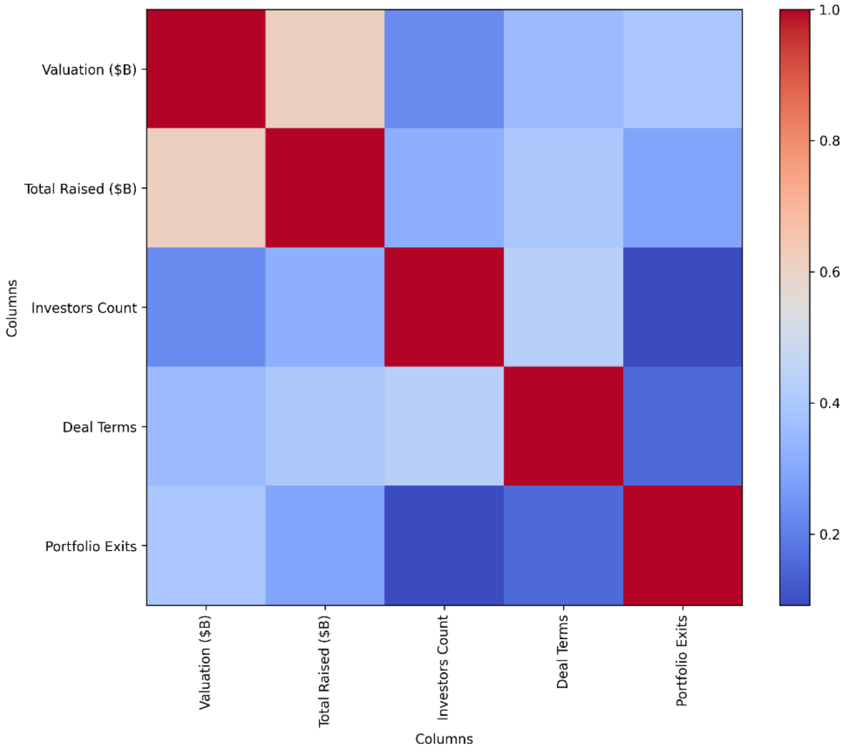
By choosing a distinct visualization approach for my individual assignment, I can delve into new challenges and expand my expertise in visualizing different types of data, relationships, or phenomena. This will ultimately contribute to a well-rounded portfolio of visualizations, highlighting my adaptability and creativity in presenting data visually.

## 2. Design

### 2.1 Tasks

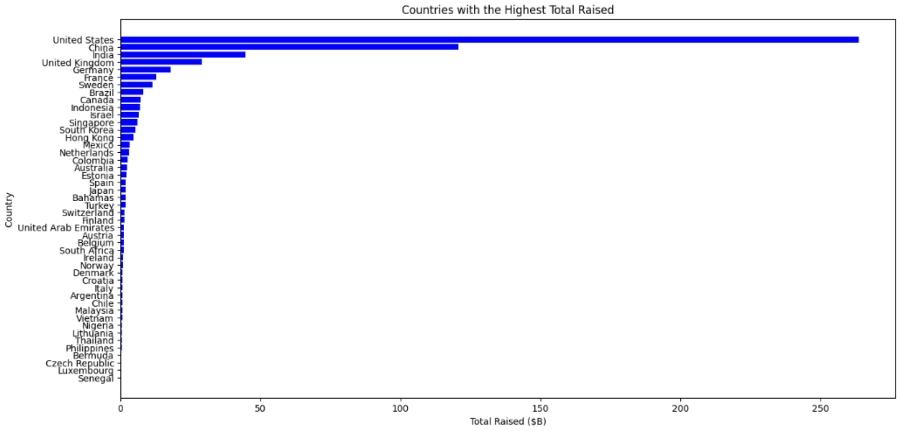
In the middle-level analysis of the Unicorn dataset, Pearson correlations were employed to examine the degree of correlation between numerical variables. This statistical measure allowed for the quantification of the strength and direction of the relationships between different variables. The results provided valuable insights into the interdependencies among the variables, shedding light on their mutual associations. By utilizing Pearson correlations, it was possible to identify the variables that exhibited significant correlations, indicating potential connections and patterns within the dataset. This analysis played a crucial role in uncovering the underlying relationships between numerical variables, contributing to a deeper understanding of the dataset and its dynamics.

The correlation analysis of the Unicorn dataset reveals interesting insights about the relationships between variables. The majority of variables exhibit weak correlations, indicating that they are not strongly associated with each other. However, there is a notable exception in the correlation between "Total Raised ($B)" and "Valuation ($B)". These two variables demonstrate a relatively high correlation, suggesting a strong connection between the amount of funding raised by a company and its valuation. Overall, the analysis indicates that the variables in the dataset have low interdependencies, highlighting the diverse nature of the unicorn companies and the independent factors influencing their success. This is useful for t-SNE plot.

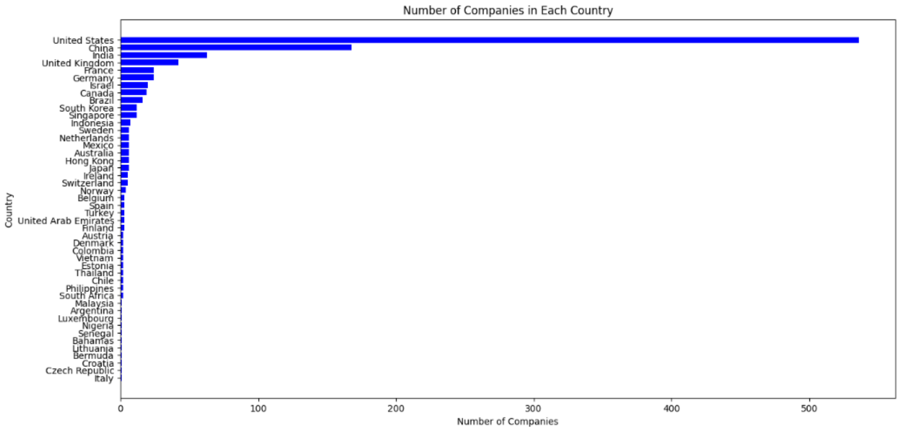


In the complex task, a horizontal bar graph was employed to compare the total raised values in billion dollars across different countries. The visualization clearly indicates that the United States stands out as the country with the highest total raised values. The bar corresponding to the United States is positioned at the top of the graph, signifying its dominant position in terms of fundraising among all the countries in the dataset. This visual representation effectively highlights the significant contribution of the United States in the realm of unicorn companies and showcases its prominence in attracting investment and generating substantial funding.

In addition to the United States, the horizontal bar graph also reveals that China and India secure the second and third positions, respectively, in terms of total raised values in billion dollars. These countries demonstrate strong performance and significant investment activity, making them noteworthy contenders in the unicorn landscape. The visualization underscores the global nature of unicorn companies, with China and India emerging as key players alongside the United States. This information provides valuable insights into the distribution of unicorn companies and highlights the prominence of these countries in terms of fundraising and investment opportunities. This is useful visualization for Ranking Heatmap as dynamic data.



In another complex task, I analyzed the number of companies in each country and visualized the results using a bar graph. As expected, the United States claimed the top spot with the highest number of unicorn companies. Following closely behind were China in second place and India in third place. This finding shed light on the correlation between the number of companies and the total raised value, providing further insight into why the United States dominates in terms of fundraising. The visualization reinforces the notion that the concentration of unicorn companies in certain countries contributes to their success in securing substantial funding.



## 2.2 Data Processing

To analyse high dimensional data, I focused on extracting information related to the country named United States. From the dataset, I created a subset that includes key variables such as Valuation ($B), Total Raised ($B), Portfolio Exits, Investors Count, Deal Terms, and Industry. This subset allows for a more focused analysis on the specific aspects of the United States' data.

To examine graph data, I decided to extract the top 50 companies based on their "Total Raised ($B)" metric. This subset specifically includes the variables "Company" and "Country". By narrowing down the dataset to these top companies, I can gain insights into their performance and contributions within their respective countries.

To explore the temporal dynamics of the dataset, I introduced a new column called "Year Joined" by extracting the year information from the "Date Joined" column. This allows for a closer examination of the dataset over time. Additionally, I calculated the rankings based on the "Total Raised ($B)" metric for each country and year. This ranking system provides insights into the performance and fundraising activities of companies within their respective countries and across different years. By considering the rankings, we can identify the top-performing companies and track their progress over time, revealing interesting patterns and trends in the dataset.

## 2.3 Analysis

When dealing with high-dimensional data, it is important to standardize the variables to ensure they are on a consistent scale. In our dataset, variables such as "Total Raised ($B)" and "Valuation ($B)" have large values, while others have small values. To address this, we apply standardization, which transforms the data to have a mean of 0 and a standard deviation of 1. This standardization process ensures that each variable contributes equally to the analysis, regardless of its original scale. Additionally, to tackle the challenge of high dimensionality, we employ the t-SNE (t-Distributed Stochastic Neighbor Embedding) method, which allows us to reduce the dimensions of the data while preserving its underlying structure. By reducing the dimensions, we can visualize and analyze the data more effectively.

In the case of graph data, our focus is on identifying the top-performing companies based on their "Total Raised" metric. We extract the top 50 companies with the highest total raised amount. Next, we group these companies based on their "Country" attribute to identify clusters of companies that share similar geographic locations. This analysis helps us gain insights into the distribution of successful companies across different countries, allowing us to observe any patterns or trends that may exist within the dataset.

For dynamic data analysis, we leverage the "Year Joined" information extracted from the "Date Joined" column. By aggregating the "Total Raised ($B)" data and grouping it by both "Year Joined" and "Country," we can examine the total fundraising amount for each country in each year. To further understand the relative performance of countries, we introduce the concept of ranking. The ranking is calculated based on the total raised amount, allowing us to determine which countries have achieved the highest levels of fundraising in billion dollars. This ranking system provides valuable insights into the temporal dynamics of fundraising activities across different countries and years. Additionally, we identify the top 10 countries in each year to highlight the leaders in terms of total raised funds.

### 2.4 Visualisation

For high-dimensional data, I employ t-SNE (t-Distributed Stochastic Neighbor Embedding) as a dimensional reduction technique to visualize the similarities between different industries. By applying t-SNE, we can effectively reduce the dimensions of the data while preserving the underlying structure. This allows us to gain insights into the relationships and clusters within the industry landscape. The advantage of using t-SNE is its ability to capture complex patterns and non-linear relationships in the data, making it suitable for visualizing high-dimensional datasets.

In my individual assignment, I opted for a graph data visualization approach that diverges from the group work assignment. I decided to utilize a connectivity matrix to represent the relationships between the top 50 companies based on their total raised value. This matrix visualizes the connections among the companies by assigning colors to the cells based on their shared country. When two companies share the same country, the corresponding cell is colored blue, indicating their connection. On the other hand, if the companies have different countries, the cell is colored white, indicating no direct connection based on country. This connectivity matrix offers a comprehensive overview of the geographic distribution and relationships among the top companies in terms of total raised value. I chose this visualization method to overcome challenges such as edge crossing or node overlapping, particularly given the size of the dataset.

For dynamic data analysis, I adopt a heatmap approach with a ranking-based visualization. The x-axis of the heatmap represents the years in which companies joined, while the y-axis represents the ranking, with 1 at the top and 10 at the bottom. Each cell in the heatmap corresponds to a specific ranking for a given year and country. The color of each cell distinguishes different countries, allowing us to identify which country achieved the highest total raised value in a particular year. Additionally, the country label is included within each cell, providing a clear indication of the leading countries in terms of fundraising in each year. This heatmap visualization offers insights into the temporal dynamics and relative performance of different countries in the fundraising landscape.

## 3. Implementation

Prior to visualization, I performed data cleaning on the "Unicorn_Companies.csv" dataset. This involved removing the dollar sign symbol "$" from the "Total Raised" and "Valuation" columns. Additionally, I made modifications to the "City", "Industry", and "Select Investors" columns to correct misplaced values and resolve any inconsistencies in the data. To enhance the analysis, I extracted the year from the "Date Joined" column and created a new column named "Year Joined".

In developing the Visual Analytics System, I utilized Visual Studio Code as the primary tool to create HTML webpages. Each webpage within the system is designed to serve a specific purpose, showcasing different visualizations, and providing distinct insights. The system comprises four main webpages, each equipped with four buttons to navigate between them seamlessly. These webpages offer a diverse range of visualizations, including an overview page to provide a holistic view of the data, a page dedicated to exploring high-dimensional data, a graph visualization page to analyze connectivity and relationships, and a dynamic data page to observe temporal patterns and changes over time. By employing Visual Studio Code, I was able to develop an interactive and user-friendly interface that facilitates effective data exploration and analysis within the Visual Analytics System.

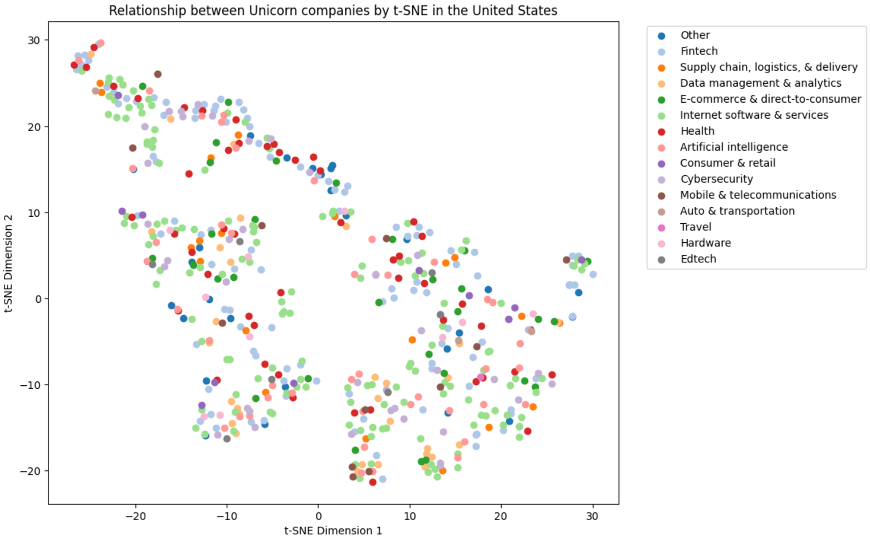
For the analysis and visualization tasks, I utilized Python along with the seaborn (sns) and matplotlib (plt) libraries. These powerful tools allowed me to generate visually appealing and informative visualizations. Specifically, I employed t-SNE visualization to explore the high-dimensional data, created a connectivity matrix to uncover relationships between companies based on shared countries, and utilized a ranking heatmap to showcase the top countries based on total raised values. The combination of seaborn and matplotlib offered flexibility and versatility in creating these visual representations of the data.

# 4. Evaluation

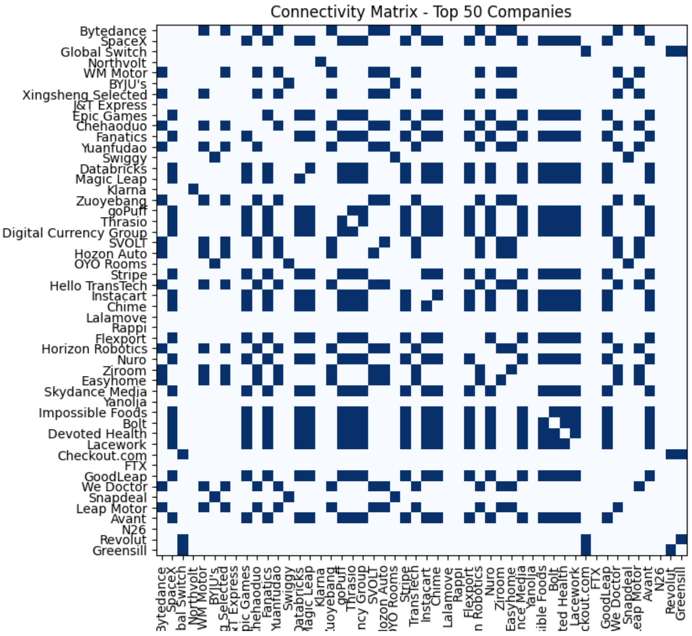
## 4.1 Results

### 4.1.1 Visualisation

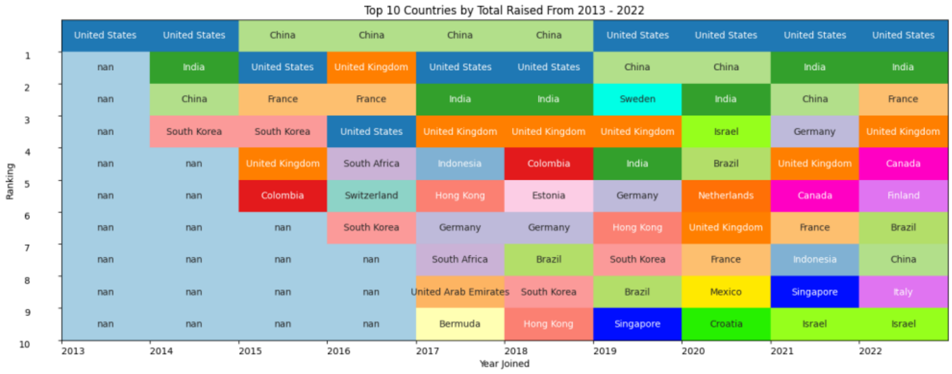
This is visualization of High Dimensional Data which is t-SNE plot.



This is visualization of graph data which is Connectivity Matrix.



This is a visualization of dynamic data which is Ranking in Heatmap.



Top 10 Countries by Total Raised From 2013 - 2022

| 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 |
|---|---|---|---|---|---|---|---|---|---|
| United States | United States | China | China | China | China | United States | United States | United States | United States |
| nan | India | United States | United Kingdom | United States | United States | China | China | India | India |
| nan | China | France | France | India | India | Sweden | India | China | France |
| nan | South Korea | South Korea | United States | United Kingdom | United Kingdom | United Kingdom | Israel | Germany | United Kingdom |
| nan | nan | United Kingdom | South Africa | Indonesia | Colombia | India | Brazil | United Kingdom | Canada |
| nan | nan | Colombia | Switzerland | Hong Kong | Estonia | Germany | Netherlands | Canada | Finland |
| nan | nan | nan | South Korea | Germany | Germany | Hong Kong | United Kingdom | France | Brazil |
| nan | nan | nan | nan | South Africa | Brazil | South Korea | France | Indonesia | China |
| nan | nan | nan | nan | United Arab Emirates | South Korea | Brazil | Mexico | Singapore | Italy |
| nan | nan | nan | nan | Bermuda | Hong Kong | Singapore | Croatia | Israel | Israel |

### 4.1.2 Visual analysis / Storytelling

The t-SNE plot provides an intriguing visual representation of the data, revealing that there are no distinct clustering patterns among the variables. This suggests that the correlation between the variables is relatively low, indicating a lack of strong linear relationships. However, upon closer inspection, an interesting observation emerges - there appears to be a line-like pattern where most industries are aligned. This finding implies that there is no significant relationship between the industries themselves, as they do not exhibit distinct clustering or grouping tendencies.

Moving on to the connectivity matrix, we can observe that many companies in the dataset share common countries. However, the matrix does not provide specific information about which countries these companies belong to. For instance, the top company, Bytedance, is connected to several other companies such as Leap Motor, We Doctor, Easyhome, Ziroom, and more, indicating shared countries among them. It is worth noting that most companies in the connectivity matrix exhibit similar country affiliations. However, there are two exceptions - N26 and Northwolt - which do not share countries with any other companies mentioned in the matrix. This suggests that these two companies have connections with other companies not included in the matrix, emphasizing the complexity and interconnectedness of the dataset.

Lastly, the ranking displayed in the heatmap reveals interesting insights into the distribution of unicorn companies over time. In 2013, there was only one country with unicorn companies. However, as the years progressed, the number of countries housing unicorn companies increased. The heatmap illustrates this growth, demonstrating that more countries began to have unicorn companies in subsequent years. By 2017, there were more than 10 countries with unicorn companies. One particularly noteworthy observation is that three countries - United States, China, and India - consistently occupy the top three positions in terms of the number of unicorn companies each year. The United States consistently holds the highest number of unicorn companies, while China and India consistently follow closely behind. This finding highlights the dominance of these countries in the unicorn landscape and their consistent presence as leaders in the emergence of successful startups.

### 4.1.3 Pros / Cons

Here are pros and cons of t-SNE visualization from high dimensional data.

**Pros:**

- **Non-linear Relationships:** t-SNE captures non-linear relationships between data points, allowing for the visualization of complex patterns that might be missed by linear techniques.
- **Local Structure Preservation:** t-SNE emphasizes the preservation of local structures, making it useful for identifying clusters or groups of similar data points.
- **Effective Dimensionality Reduction**: t-SNE reduces the dimensionality of high-dimensional data while preserving its structure, making it easier to visualize and interpret.

**Cons:**

- **Lack of Global Structure:** t-SNE primarily focuses on local structures, which means that the global structure of the data may not be accurately represented in the visualization. Care must be taken to interpret the t-SNE plot in the context of the specific local structures it highlights.
- **Parameter Sensitivity:** t-SNE has parameters that significantly impact the resulting visualization. Choosing appropriate parameter values can be challenging and may require experimentation to achieve the desired visualization.
- **Computational Complexity:** t-SNE can be computationally expensive, particularly for large datasets. It may require substantial computational resources and time to generate the t-SNE plot for datasets with a large number of data points.

Here are pros and cons of connectivity matrix from graph data.

**Pros:**

- **Clear Representation of Connections:** The connectivity matrix provides a clear and concise representation of connections between nodes in a graph. It allows for a quick visual assessment of which nodes are connected and which are not.
- **Compact Visualization:** The matrix format allows for a compact visualization, especially when dealing with large graphs. It avoids the issue of node overlapping or edge crossings that can occur in other graph visualizations.
- **Easy Identification of Patterns:** By examining the matrix, patterns such as clusters or groups of connected nodes can be easily identified. It enables the detection of relationships and similarities among nodes based on their connectivity.

**Cons:**

- **Lack of Graph Layout Information:** The connectivity matrix does not provide explicit layout information such as node positions or edge routes. It only represents the connections between nodes, making it challenging to visualize the overall structure of the graph.
- **Limited Visual Encoding:** The matrix format primarily relies on binary or color-based encoding to represent connections. This may limit the ability to convey additional attributes or properties of nodes and edges that could be useful for analysis.

- **Difficulty with Large Graphs:** As the size of the graph increases, the connectivity matrix can become visually overwhelming and difficult to interpret. The matrix cells may become too small to discern individual connections, leading to information loss.

Here are pros and cons of ranking heatmaps from dynamic data.

**Pros:**

- **Clear Hierarchical Structure:** Ranking heatmaps provide a clear representation of hierarchical structures based on the ranking of values. They allow for easy identification of high-ranking and low-ranking items, facilitating quick comparisons and insights.
- **Effective Highlighting of Patterns:** Heatmaps use color gradients to highlight patterns and trends in the ranking data. This makes it easier to identify clusters, outliers, and other patterns that may not be immediately apparent in raw numerical data.
- **Efficient Comparison across Multiple Categories:** Ranking heatmaps enable efficient comparison of rankings across multiple categories. By organizing the data in a matrix format, it becomes easier to identify variations in rankings between different groups or dimensions.

**Cons:**

- **Limited Quantitative Precision:** Heatmaps primarily rely on color gradients to represent rankings, which can be subjective and lack precise quantitative information. This can limit the ability to make accurate numerical comparisons between individual rankings.
- **Potential for Misinterpretation:** The use of colors in heatmaps can sometimes lead to misinterpretation, especially if the color scale or legend is not properly defined or understood. It is important to provide clear explanations and context to avoid misreading or misrepresenting the data.
- **Sensitivity to Data Range and Scale:** The effectiveness of ranking heatmaps can be influenced by the range and scale of the data. If the range is too narrow or the scale is not properly calibrated, it can result in misleading or distorted representations of rankings.

### 4.2 Discussion: Summary, Limitation

In summary, the Unicorn dataset provided valuable insights into the world of unicorns, with visualizations offering a deeper understanding of the data. The t-SNE plot revealed the absence of strong correlations between industries, while the connectivity matrix showcased shared country affiliations among companies. The ranking heatmap demonstrated the temporal dynamics of unicorn companies across different countries, with the United States, China, and India consistently emerging as key players. These visualizations provided a comprehensive overview of the dataset, highlighting industry relationships, country connections, and temporal trends within the unicorn ecosystem.

One limitation of the Unicorn dataset is that it may not provide a comprehensive representation of the entire unicorn landscape, as it contains information on a specific subset of companies. This limitation could impact the generalizability of the findings and insights derived from the dataset. Additionally, the dataset lacks detailed information about the specific countries mentioned in the connectivity matrix, making it difficult to discern the exact country connections between companies. Furthermore, the t-SNE plot, while useful for visualizing similarities between industries, may not capture the complete complexity and nuances of the underlying data. It is important to acknowledge these limitations when interpreting the results and consider the dataset as a snapshot rather than a complete depiction of the entire unicorn ecosystem.

Another limitation of the Unicorn dataset is the lack of clarity and precision in the Valuation and Total Raised columns. The dataset does not provide explicit information about the currency used or whether the values are in billions, millions, or other units. This ambiguity can introduce uncertainty and make it challenging to compare and analyze the financial aspects of companies accurately. It is crucial to exercise caution when interpreting and comparing the Valuation and Total Raised values, considering the potential inconsistencies and lack of standardized units in the dataset.

## 5. Appendix

### 5.1 Individual notes

Each week, we had meetings with other members and each member did each task weekly.