Unicorn Startup Analysis Dashboard: Identifying Investment Trends and Educational Background of Founders and Investors

Marianella Salinas[1]

I. INTRODUCTION

In today’s highly competitive startup ecosystem, it is crucial to understand the factors contributing to the success and high valuations of ”unicorn” companies. Unicorns are privately held startup companies that are valued at over $1 billion USD. This paper presents a breakdown that analyzes and visualizes patterns and trends in unicorn startups and the educational backgrounds of founders and investors in the startup ecosystem. The primary goal of the dashboard is to identify the types of companies attracting significant investment and high valuations, as well as the countries and industries leading the way. Additionally, the dashboard explores the education network within the startup ecosystem to gain insights into the relationships between institutions, individuals, and their roles. By leveraging various data visualization and analysis techniques, including K-Means and GMM clustering, correlation analysis, scatter plots, heatmaps, stacked bar charts, and network graphs, the paper aims to provide valuable insights into the factors driving the growth of successful startups.

II. DATA AND STRUCTURE

The analysis is based on two datasets: (1) The Unicorn Dataset which contains information on 935 unicorn startups valued at over $1 billion, with features such as company name, valuation, date joined, country, city, industry, and major investors, covering the period from 2011 to 2021, and (2) The Education Dataset which contains information on 21,092 individuals in the startup ecosystem, including CEOs, founders, CTOs, board members, venture capitalists, and more. This dataset contains information on their birthplace, institutions attended, degree types, subjects studied, graduation dates, company affiliations, and job titles, covering the period from 1999 to 2013.

The dashboard is organized into six pages: Overview, Country Exploratory Analysis, Founders and Venture Capitalists, Education Network Graph, Investors and Valuation, and K-Means and GMM Clustering. The Overview page provides general background on the distribution of unicorns across industries and countries, along with their valuation patterns. The Country Exploratory Analysis page highlights the geographical distribution of unicorn startups, exploring trends in valuation and industry. The Founders and Venture Capitalists page examines the educational background and institution affiliations of both Founders and Venture Capitalists in the startup ecosystem. The Education Network Graph visually represents the connections between individuals, institutions, and companies, while the Investors and Valuation page focuses on the relationships between major Venture Capital investors and unicorn valuations. Lastly, the K-Means and GMM Clustering page applies clustering techniques to identify patterns in the data to provide further insights into the factors contributing to high valuations.

Through the use of various visualization techniques and interactive filtering options, the dashboard provides an understanding of the factors driving the growth and success of unicorn startups. By examining patterns in company valuations, industries, countries, and educational backgrounds of key individuals, the dashboard aims to highlight the underlying dynamics of the startup ecosystem.

[1] Marianella Salinas, Class of 2024, Department of Computer Science and Engineering

III. Technologies Used

The main analysis and visualization techniques used for this dashboard are discussed below.

A. Data Manipulation and Preprocessing

This project uses the Pandas library, a powerful Python library for data manipulation and analysis. Pandas offers a convenient DataFrame structure that allows for easy and efficient manipulation of large datasets. For this dashboard, Pandas is used to read CSV files, filter data, perform calculations, and create new columns, among other operations. Choosing Pandas over other libraries was due to its functionality with other libraries like Dash, making it the ideal choice for handling the two datasets efficiently.

B. Visualization Techniques

The dashboard incorporates a range of visualization techniques, using Plotly and Plotly.express libraries for generating interactive and informative plots. Plotly was chosen for its ability to create customizable, interactive, and web-based visualizations directly from Python. This project uses various chart types such as bar charts, violin plots, pie charts, tree maps, and scatter plots, each chosen for their ability to effectively represent different aspects of the data. For example, bar charts and violin plots were used to display valuation distributions across countries and industries, while tree maps and pie charts were employed for visualizing industry counts and investments. Scatter plots were used for clustering and correlation analysis to understand the relationships between features.

C. Network Analysis and Visualization

To explore the educational background of founders and investors in the Startup Ecosystem, the project utilizes the networkx library to generate a network graph. Networkx is a library for the creation, manipulation, and study of complex networks, which made it a great choice for analyzing the connections within the Startup Ecosystem. The dataset used in this analysis contains over 20,000 datapoints, which makes it challenging to display and understand the relationships within the data effectively.

To create a more user-friendly interactive graph, a random subset of 300 datapoints was sampled from the dataset. This approach ensures that the generated network graph remains visually appealing and informative, while still providing a representative snapshot of the relationships between individuals, their educational backgrounds, and their affiliations within the ecosystem. This analysis helps to shed light on the importance of education and networking in the startup landscape.

The choice of networkx library was made over other techniques because it has strong compatibility with other Python libraries such as pandas and pyvis, which are used in this project to handle data processing and visualization.

D. Machine learning and Statistical Analysis

The project utilizes the scikit-learn library for clustering analysis, specifically K-Means clustering and Gaussian Mixture Models (GMM), and the scipy library for correlation analysis. The choice of K-Means and GMM techniques was motivated by the ability to identify patterns in the data and create meaningful groupings. These clustering methods help reveal underlying trends and relationships within the dataset that might not be apparent through visualization alone. Scikit-learn was selected for its extensive suite of machine learning algorithms and preprocessing tools, while scipy was used for its comprehensive set of statistical functions.

IV. Analysis of Results and Visualization

A. Overview Page

The global startup landscape is heavily dominated by the United States and China, as evidenced by the bar chart in Figure 1. The United States leads with a startup valuation of $1,600 billion USD, followed by China at $600 billion USD, and the United Kingdom, the third-largest, at $150 billion USD. These figures indicate the significant market concentration in the startup ecosystem, with the United States and China accounting for the vast majority of the global startup valuation. This concentration implies that these two countries are the primary drivers of innovation, investment, and growth in the startup sector.

When examining the industry breakdown, 38% of the startup valuation is attributed to fintech and

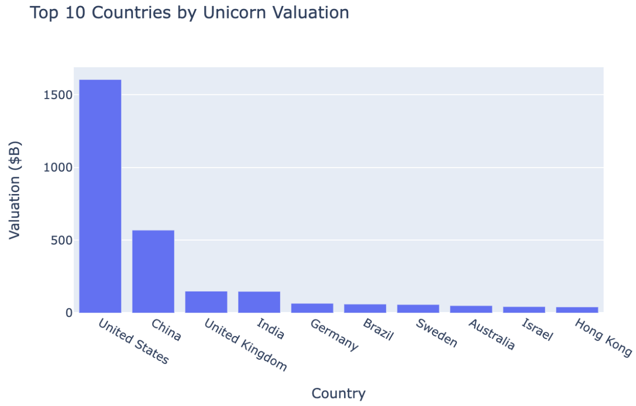

Fig. 1: Startup valuation of the top 10 countries.

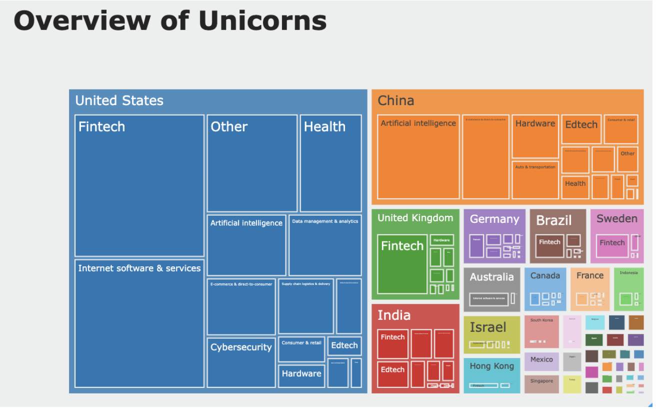

Fig. 2: Treemap visualization of the top industries in each country, emphasizing their respective market concentrations.

software-as-a-service (SaaS) sectors. This concentration of valuation within these two industries suggests that they are central to the growth and success of startups, particularly in the United States and China. These sectors are benefiting from strong investor interest, rapid technological advancements, and increasing market demand. Consequently, the analysis and visualization of data in these industries are essential for understanding the overall trends and factors contributing to the success of startups globally.

To visualize the top industries in countries with an emphasis on market concentration, a treemap is used in Figure 2. Treemaps are an effective visualization technique for this type of data because they can represent hierarchical data in a compact and easily understandable format. The size of each rectangle in the treemap is proportional to the valuation of the industry it represents, allowing for a clear representation of the industry's significance within a given country. Additionally, treemaps can display the relative valuation of various industries across countries, facilitating a comparison of market concentration and industry dominance between nations. By using a treemap to visualize this data, the patterns and trends in the startup ecosystem can be uncovered to gain insights into the industries that are driving growth and attracting investments in each country.

B. Country Exploratory Analysis Page

In the Country Exploratory Analysis page, Figure 3 showcases unicorn concentration by industry and country using a heatmap visualization. The dominance of the United States and China in the startup ecosystem becomes evident in this visualization, making it difficult to distinguish values for other countries. The smaller squares representing other countries are overshadowed by the larger squares representing China and the United States. This disproportionate representation affects the visualization process by hindering insights and trends from the data of smaller players in the startup ecosystem.

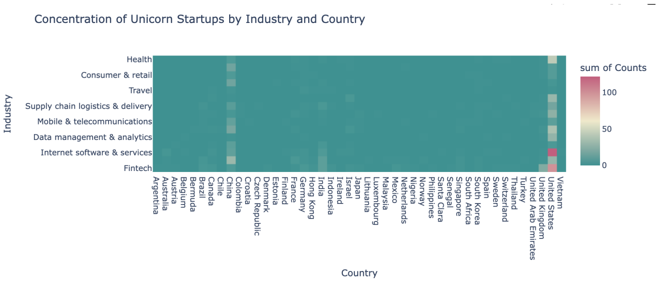

Fig. 3: Heatmap showing the concentration of unicorns by industry and country, highlighting the dominance of the United States and China in the startup ecosystem.

Figure 4 explores the locations and cities with the most startups, often referred to as startup hubs. The map visualization highlights that North America, specifically San Francisco, has a high concentration of startups. The green circle, indicating the count and size of startups, covers almost the entire North American region. The overwhelming presence of startups in San Francisco and its surrounding areas reinforces that certain cities are more adapted to the growth and development of startups. This could be attributed to factors such as access to resources, funding, and talent.

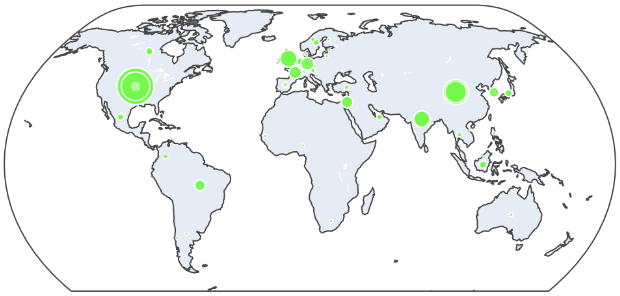

Fig. 4: World map visualizing the top startup cities. The green circles represent the count and size of startups, emphasizing the significance of certain locations as startup hubs.

C. *Founders and Venture Capitalists Page*

In the analysis of degrees and graduation years of founders and investors, interactive visualizations such as pie charts, PDF graphs, and heatmaps provide valuable insights into the differences between these two groups. For example, the pie chart illustrates the overall distribution of degree types, while the PDF graph shows the overall distribution of graduation years for both founders and venture capitalists. The heatmap reveals the overall distribution of top institutions with degrees that appeared in the education dataset (Figure 6).

The data reveals that founders are, on average, significantly younger than venture capitalists, with an average graduation year of 2007 compared to 1995 for venture capitalists. Additionally, founders hold MBA degrees at a rate of 13%, as opposed to 21% of venture capitalists. This trend indicates that more formal business education is valued among venture capitalists. A strong correlation is also observed between top universities and the individuals in the education dataset, suggesting that networking and alumni connections play a crucial role in the startup ecosystem. Attending a prestigious university may increase one's chances of success in the industry due to the strong support network and resources these institutions provide.

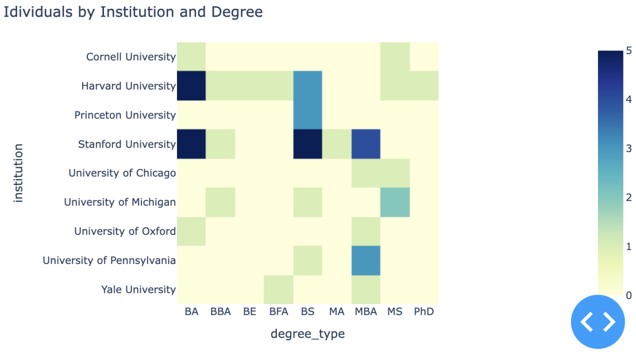

Fig. 5: Heatmap of institutions by degree type for startup founders and CEOs.

D. *Education Network Page*

The presence of an individual in the startup ecosystem appears to be related to their education, as observed in the network graphs (Figures 6 and 7). In each random sample generation, large clusters are formed around two prominent universities: Harvard and Stanford. This trend suggests that these institutions play a significant role in producing individuals who go on to be involved in the startup ecosystem. Their strong presence in the data (Figure 6 - network graph close-up) implies that the education and networking opportunities provided by these universities contribute to their graduates' success in the industry.

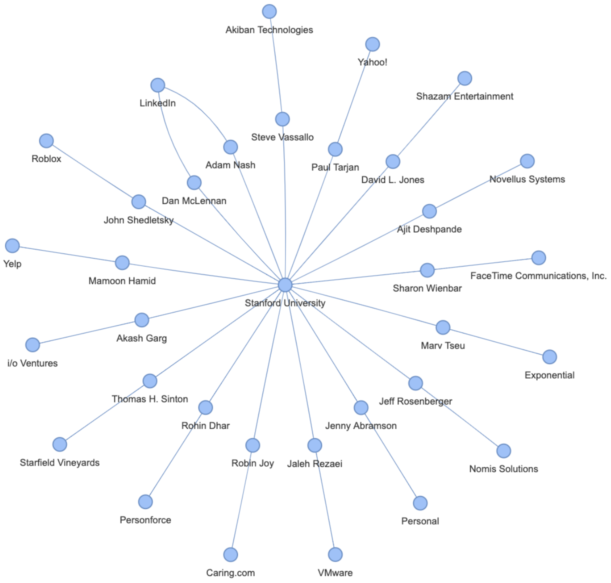

Fig. 6: Close up of the Network Graph.

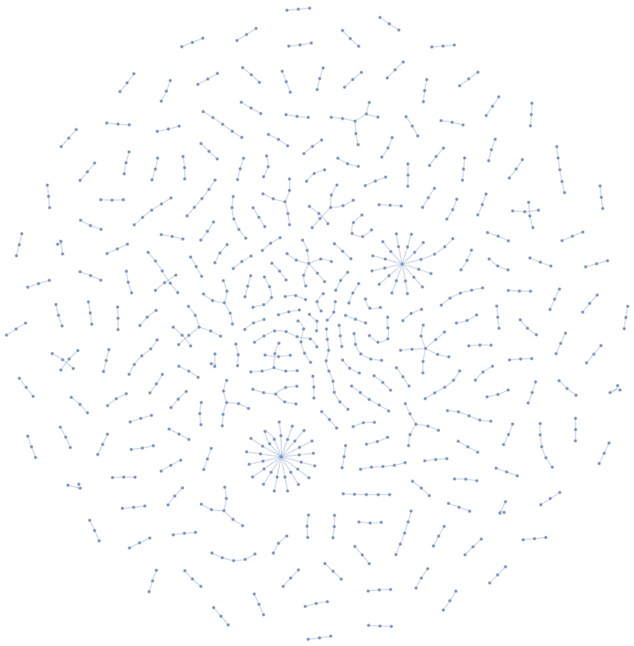

Fig. 7: Network Graph of Startup Ecosytem.

By randomly selecting 300 individuals from the education dataset of 21,000 features, the resulting graph remains visually appealing and informative while still providing a representative snapshot of the relationships between individuals, their educational backgrounds, and their affiliations within the ecosystem. Overall, Figure 7 analyzes the complex relationships within the startup ecosystem and highlights the importance of education and networking in this landscape, even when examining a reduced sample size.

E. K-Means and GMM Clustering Page

The clustering of different industries by year and valuation helps to reveal underlying trends and patterns in the startup ecosystem. The user is able to use a dropdown menu to select the specific industry and radio buttons to select the number of clusters. In this analysis, K-Means and Gaussian Mixture Model (GMM) clustering techniques are utilized to group similar data points together based on their features which are the year and valuation. The K-Means algorithm attempts to find the best partitioning of the data into 'k' clusters, where each data point belongs to the cluster with the nearest mean, while GMM clustering uses a probabilistic approach (Figure 8).

The elbow method is employed to determine the optimal number of clusters ('k') for the K-Means algorithm. This method involves plotting the within-cluster sum of squares (WCSS) against the number of clusters, and then identifying the point where the curve starts to flatten, which resembles an elbow. The elbow point signifies the optimal number of clusters, where adding more clusters would not significantly improve the model's performance for the data by industry.

Filtering by industry allows for more focused analysis and identification of trends and outliers within specific sectors. Outliers can affect the results of clustering algorithms, potentially causing clusters to be biased towards the outliers or causing more clusters to be formed. In K-Means clustering, the centroids of the clusters can be largely influenced by outliers, which might lead to suboptimal partitions of the Unicorn data. GMM clustering is less sensitive to outliers because of its probabilistic approach.

In Figure 9, the dataset does not have many outliers, so Figure 8 and 9 can be compared in terms of industry and time. It can be concluded that date joined and valuation do not have much of a correlation. Being aware of the impact of outliers on clustering results is crucial for interpreting the visualizations and understanding the trends in the data, especially in the context of the startup industry, where valuations can vary greatly.

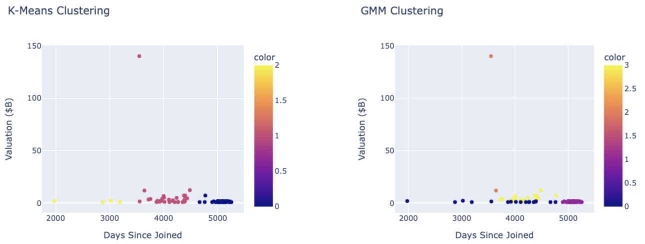

Fig. 8: K-Means and GMM Clustering of the Artificial Intelligence Industry with 3 clusters.

V. Conclusion

This project aimed to analyze and visualize the patterns and trends in unicorn startups and explore the role of education and alumni networks within the startup ecosystem. Using various data visualization and analysis techniques, this paper

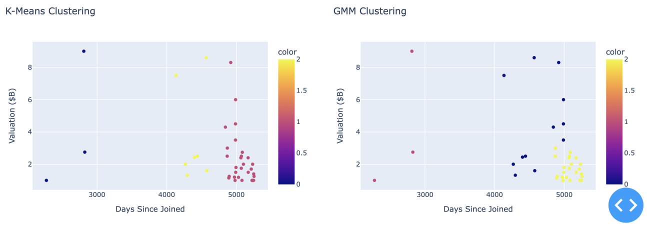

Fig. 9: K-Means and GMM Clustering of the Cybersecurity Industry with 3 clusters and less outliers.

reveals valuable insights into the factors that drive the growth of successful startups and identify the countries and industries leading the way.

If given the opportunity to revisit the project, I would focus on expanding the education dataset to encompass a broader range of educational backgrounds, industries, and countries. Additionally, I would navigate outliers better, specifically with the heatmap. Regarding Figure 3, the heatmap does not effectively display correlations between smaller countries in the startup ecosystem. In future iterations, I would consider alternative visualizations that allow for a more balanced representation of the data, enabling users to understand the differences across different countries and industries. Specifically, I would create a heatmap that excludes the US and China to provide a fairer representation.

One area for further investigation is the comparison of the distribution of investments made by venture capitalists and investors from different educational backgrounds, specifically those with business degrees versus technical degrees. By using a p-test, I could test the null hypothesis to determine if there is no significant difference in the distribution of investments made by these two groups.

Overall, the project successfully demonstrated the importance of education and networking in the startup landscape, shedding light on the factors that contribute to the success of unicorn companies. The findings show the need for aspiring entrepreneurs and investors to be aware of the potential advantages of educational backgrounds and connections, and to leverage these resources effectively in the competitive global venture capital environment.

References

[1] https://www.kaggle.com/datasets/ramjasmaurya/unicorn-startups

[2] https://www.kaggle.com/datasets/justinas/startup-investments.

[3] https://www.entrepreneur.com/en-ae/growth-strategies/the-role-of-academia-in-the-world-of-startups/423629.

[4] https://pitchbook.com/news/articles/unicorn-startups-list-trends