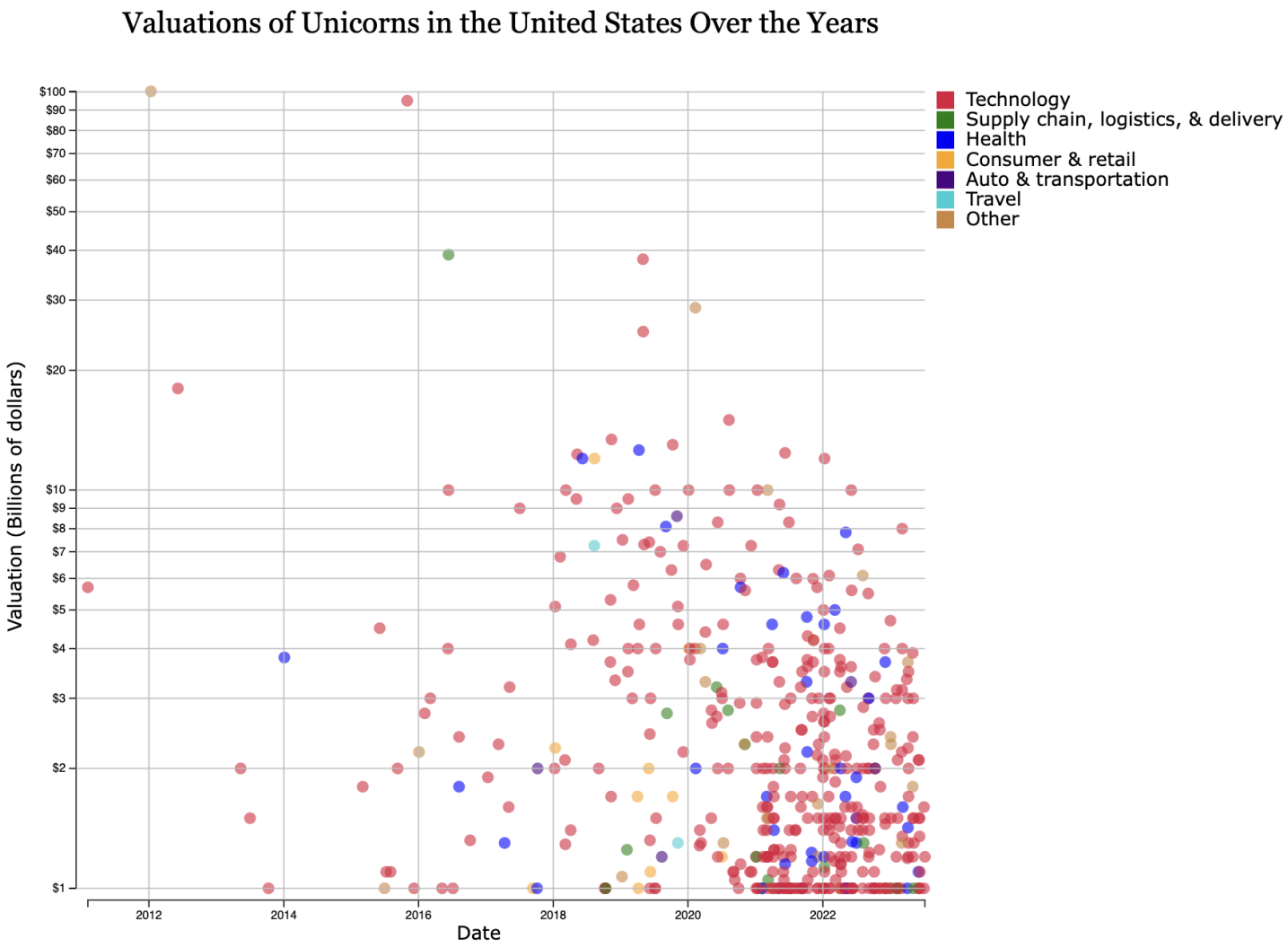
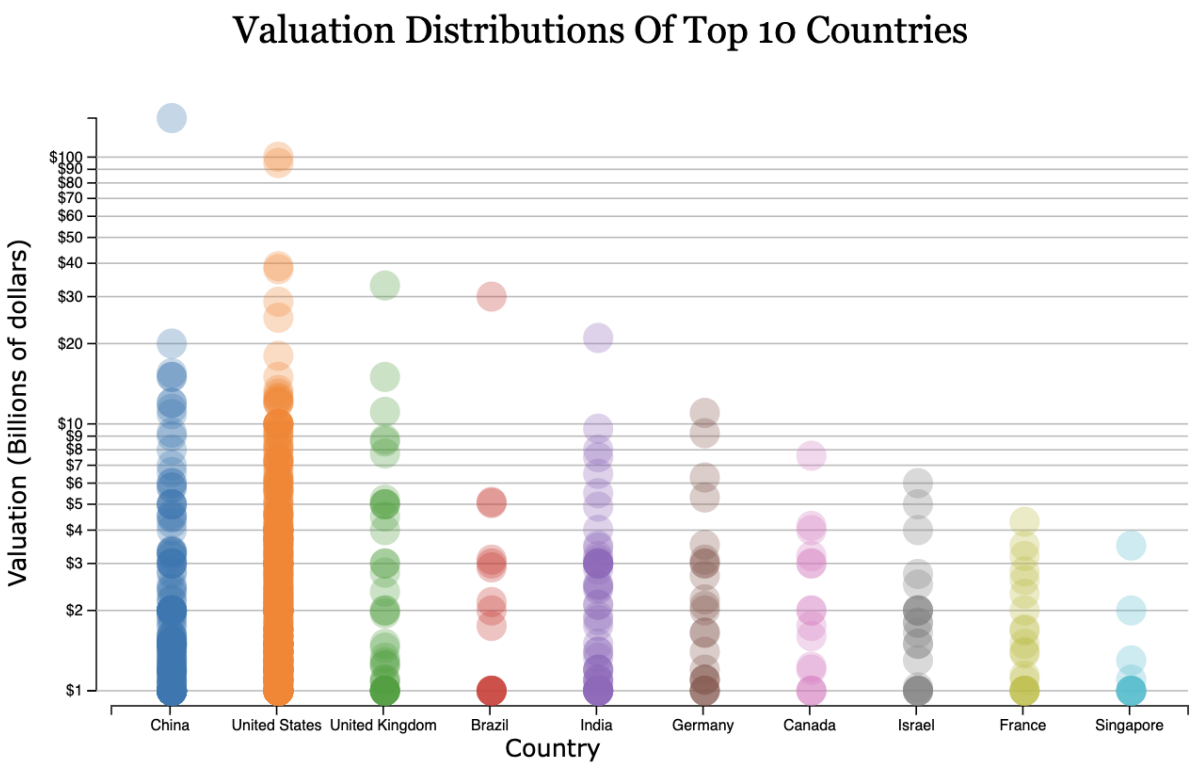
Isuru Herath, Alexander Weseley, Milan Patlikh

Project 1 Report

a.



Isuru Herath, Alexander Weseley, Milan Patlikh
Project 1 Report



b. The dataset we chose to analyze was the worldwide unicorn start-ups dataset. We chose it due to its richness in the types of data provided for each data point. There are both categorical and continuous numerical variables. This allows us to create interesting and varied visualizations from it. The categorical variables included the names of the start-ups, investors, names of the countries and cities in which the start up is based, and the industry it's in. More specifically, these were nominal, unordered variables. The numerical, or more specifically the quantitative ordinal variables were the dates each start-up achieved unicorn status (greater than or equal to $1 billion valuation), and their valuations. In all the dataset contained information about 936 start-ups.

For the first visualization, we filtered out all start-ups that are not based in the United States. We chose to focus on American start-ups because including all 936 points made the scatterplot very cluttered, and it was difficult to glean any insights or trends from this visualization. Additionally, American start-ups accounted for over half of all start-ups in the datasets at 477. We also combined the "Artificial intelligence", "Artificial Intelligence," "Fintech," "Finttech," "Internet software & services," "Data management & analytics," "Edtech," "E-commerce & direct-to-consumer," "Hardware," "Cybersecurity," and "Mobile & telecommunications" industries into one "Technology" category. This was done to prevent there from being an overwhelming number of different categories, and consequently different colors on the plot. Originally, there were 15 different industry categories, this narrowed them down to seven. This way we have fewer more distinguishable colors that are easier for viewers to read, and again we can see more trends from this arrangement. For the second visualization, we filtered the dataset to only include the ten countries that had the most unicorn star-ups in the dataset. This decision was made because the data contains start-ups from 47 different countries and this would be overwhelming to include on one axis of a dot plot. The aim of this plot was to emphasize the differences between the valuation distributions of unicorns across countries and this would not have been effective with that many countries.

c. The first visualization is the result of iteration and decision making. Our initial design contained every data point in the data set, however, we quickly realized that the scatter plot we had made was nearly incomprehensible. There were far too many data points, so it was difficult to notice any patterns, trends or insights from the data. To add to this,

there were some outliers, so most points in our linearly-scaled scatter plot were clustered together at the bottom. Similarly, we had color coded the data points by industry, of which there were too many, leading to a salad of indistinguishable colors. This led to several key decisions. First, we decided to focus only on US-based startups as they accounted for over half the data points. Second, we categorized all tech companies together under the assumption that this would reveal some valuable patterns and insights. Next, we scaled the y-axis (valuation in dollars) using a logarithmic scale, which helped to visualize the differences in valuations, large and small. Lastly, we made the pixel opacity 0.6 to make the colors distinguishable while also seeing overlapping points. In the end, we mapped each data point as described above to a circle (the mark), using horizontal and vertical aligned positions as visual channels. We additionally used color as a visual channel to distinguish data between technology and non-technology companies.

Similar to the first visualization, the second visualization comparing different countries also did not read very well at first. By comparing countries, we decided it was best to only include the top 10 countries. By quickly running pandas code in Jupyter Notebook, we found that over 800 out of 936 unicorns came from these top 10 countries. By expanding the graph with more countries, it would space the visualization more out without adding much useful information. With country names at the bottom, some of the longer names overlapped with normal graph size distributions. We chose to space out each country more to prevent this clutter. Since the x-axis was categorical, no scale was needed. Additionally, like the first visualization, the dots were very concentrated at the bottom with a linear y-scale. Instead, a logarithmic scale for the y-axis helped to space

out the majority of the dots from $1-$10 billion, while still highlighting the big outliers that were much greater than $10 billion. To account for the high volume of unicorns between $1-$10 billion, the dots were also given a very low opacity of 0.3. This makes it easier to recognize that there are many data points in an area, since the dots overlap each other and have a greater total opacity. Lastly, a color scale was added to each individual country to make the graph more appealing. This change had no effect on reading the data, as it is already divided by country, but by making each country a different color, the graph showed a more captivating differentiation between each 'column' of dots. The marks used were circles, while the visual channels were horizontal and vertical aligned position. Lastly, we used a different color for each country to make it easier to distinguish between countries, as well as to make the visualization more visually appealing.

d. The visualizations we created tell individual stories when looked at separately, however, when put together they paint a picture about the world of startups. The first visualization uses data from just American startups, split into technology and non-technology companies. The data tells us that unicorn companies have become more common in recent years. Importantly, it appears that starting around the year 2021 companies achieved unicorn status with a much higher frequency than before. The reason for this is not obvious, but it may have been an effect of the overall market bull run that followed the initial COVID-19 pandemic scares. Many of these companies are valued at exactly $1 billion dollars, just enough to be considered unicorns. An assumption can be made that companies were raising money at these valuations solely for the sake of joining the exclusive unicorn club. Another important part about the first visualization is that most

companies are technology companies. We wanted to highlight this fact in our visual because we thought it was key to understanding the world of venture capital today. Investors were pouring money into technology companies in hopes of achieving high returns. It would be interesting to see how this has changed with the recent market downturn, especially as it has affected the valuations of technology companies.

In the second visualization, seeing the comparisons of unicorns by country is especially interesting because it is evident that the United States and China have the most data points. While China has the highest unicorn out of them all, the United States seems to have many more unicorns overall, and additionally has many more above $10 billion in valuation. Throughout the top 10 countries, the countries' highest valued unicorn seems to have a positive correlation with the total number of unicorns in the country. The one outlier is Brazil, which has one very highly valued unicorn, but not that many other unicorns. Overall, the graph depicts that China and the United States have the most and highest valued unicorns, with the incrementally decreasing amounts for the other countries in the top 10. From the audience point of view, it may be surprising to see that there are a large amount of high valued unicorns in other parts of the world, since the audience is most likely more exposed to the start-ups in the United States and a few other countries like the United Kingdom and Canada.

Contributions: Isuru did preliminary drafts of various visualizations of the dataset using Matplotlib in Jupyter notebook, which helped us arrive at a visualization of a scatterplot of time vs. valuation. We were originally considering a bubble plot with time on the

x-axis and industry on the y-axis. This preliminary visualization showed that there were no clear trends from this plot. Isuru also created the axes, axes labels, gridlines, and data points for the first visualization. He also created the plot titles for both visualizations. All of this took a total of approximately 5-6 hours to complete distributed over the duration of the assignment.

Milan worked on the second visualization. We figured this visualization was slightly less work than the other one, so figured only one person needed to do the bulk of creating the dot plot. After exploring the best ways to compare the countries by looking at numbers and quick graphs on Jupyter Notebook, the importing and visualization of data was not too difficult. We made a new csv file in Jupyter Notebook that only included the unicorns from the top 10 countries, which made it easy to translate the entire dataset quickly. Other than that, the most tedious tasks involved choosing a logarithmic scale and figuring out how to best map points with categorical variables.

Alex did exploratory data analysis on several data sets, realizing that the data set on unicorn startup companies could lend itself well to visualizations. Along with Isuru, Alex helped create the first visualization (scatter plot), helping make key design decisions. Importantly, he found a way to add a legend to the visualization. Additionally, Alex helped write the Design Rationale and Story aspects of the write up. The most time consuming aspect of this project was the process of iterating our visual design and figuring out what story we wanted to tell with it.