US Startups 2017

CS3300 Project 2
Angela Wu, Rong Hu, Val Mack

Story

This project highlights “startup” companies in the United States with the intention of showing meaningful information about the country’s most successful entrepreneurial ventures. A startup is a young company, commonly characterized by being privately owned by a few employees, having an innovative business idea, not generating profit, and being supported by venture capital and/or angel investments [1]. The data for this visualization comes from the website AngelList (https://angel.co/), a platform for startups to raise money online, recruit employees, and apply for funding.

After finding this data, we thought it would be good to visualize because startups are extremely popular, and our primary audience (college students) will be familiar with the companies we feature and compare. We found the data to have enough features such that our project would be non-trivial, and we knew that there were multiple types of visualizations that we would be able to create from the dataset. Ultimately, it ended up being interesting to visualize because the data indicates patterns within the ‘world’ of startups that give insight into how entrepreneurship manifests in U.S. society today. The goal is encourage the audience to formulate follow up questions as to why startups express such patterns, and to make them want to learn more.

For Figure 1, we analyzed the startups within three dimensions: Locations, Funding and Market Focus. We showed how fundings are distributed among different states in each key market. It is clear from our visualization that California earns the majority (> 50% in most markets) of funding in almost every market, while New York is the state that have the second greatest funding. But the 3rd and 4th largest funding states are different in different markets.

Figure 2 presented our startup data in three dimensions. The first dimension is the composition of the total startups based on total money raised. It is clear from the graph that the number of startups with funding less than 10 million is the largest, and that the number of startups goes rapidly down with the funding range going up. The second dimension shows startups’ connection to top U.S. universities and companies based on where a startup’s team members come from. For example, of the 28,971 startups with funding less than 10 million, there are 1194 startups that have team members graduating from Stanford. An obvious pattern here is that the larger the total number of startups in a certain funding range is, the more diverse those startups’ human resources are. The third dimension is the type of startups’ team sources. Not only we divided it into universities and companies, but in each subdivision, we further distinguished them using either general location (east coast and west coast for university) or industries (IT, consulting, financial for companies) to provide more information. We can see that Stanford and UC Berkeley are the top entrepreneurial universities, and many team members of startups previously worked at Google or Microsoft. In each funding category, the four sources are always linked with the largest numbers of startups.

For Figure 3 we examined two dimensions of the top 100 startups in terms of amount of money raised. The goal of the visualization was to indicate the 'optimal' number of employees for a successful startup, by showing how many startups have a given number of employees. Each circle represents an employee size range, for example, 1-10, which sits at the center. The radius of each circle represents the difference in the size ranges, with the largest being 5000+ employees. When the user hovers over a size range, the area of the circle becomes filled with a solid color and the user sees a number displayed to indicate how many startups have employment numbers within that size range. This visualization allows the audience to question why most successful startup companies gravitate to being a certain size.

Data Logic

For Figure 1, the data is drawn from https://angel.co/companies. Due to limitation of exporting data from this platform, we got top 100 startups of each key market for visualization. Startups are ranked based on the amount of fund they raised. The original data contains location info in city level but we need startup's geographic info in state level. So we do an inner join on the columns of city in the original dataset with the mapping dataset of city and state. But we then found the city data contains cities of different levels, and there are some cities with the same name belonging to different states. So after the inner join, we have to verify the city and state mapping manually for those special cases. Then we calculate the total number of the startups in each state and their total funding. The final data table for each market contains state's name, number of startups in this state and total funds raised by startups of this state.

For figure 2, the data is manually searched and collected from https://angel.co/companies. First we search the total number of startups. Then we restrict the total money raised and obtain the numbers of startups for each funding category. In each funding category, we add the condition of team member source and obtain the number of startups with team members from a certain source. These data are all compiled into two csv files with fields of source, target, and value. Data for the funding categories are complete, since there are only 9 funding categories. However, the number of universities and companies are extremely large. So we only choose to present the top ones in each subgroup.

For Figure 3, the data was collected in a .csv file from https://angel.co/companies. The data represents the top 100 startups in the United States based on the amount of capital raised. For this visualization, the relevant column from the dataset was number of employees. A D3 nest function was then used to provide counts for the number of companies with each employment size range. One of the difficulties of using this data was converting the size ranges into numbers that could be used within a scale in order to display differences in the radius of each circle. This was done by creating artificial groupings based on the highest number in each range, and also creating real groupings based on the ranges themselves.

Visualization Logic

For Figure 1, we build US map with path element found from the internet [2], and state data are used to separate the map's shapes. The colors in each state represents the total funding raised by the startups in this state. Note that we only consider the startups in the top 100 startup list in each key market. Color is scaled with linear scale. For each state, when user move the mouse on the state shape, a tooltip shows up with detailing information about the number of startups in this state and the funding they raised. When user hover mouse to a specific state, a pie graph will show up if the state has a least one startup in the ranking list of top 100 startup in these four markets. The pie graph describes the startup distribution in different markets of this state.

For figure 2, we allow users to choose which source they want to view. After loading the csv file, we rearrange the data into objects of nodes and links, and use functions from D3's sankey plugin to generate a sankey diagram showing relationships between various nodes. The height of each rectangle and the width of each link are scaled to represent the number of startups. The nodes that are positioned in the same column are ranked descendingly according to its value. In the "team from" section, colors are used to indicate different locations /industries of universities/companies. Color is scaled using ordinal scale. We also use university/company logos for the "team source" nodes to make them easier to tell by a viewer. When user hover on a node or a link, a tooltip shows up to provide detailed information. A link also becomes highlighted if hovered.

For Figure 3 we visualize the data with circles to indicate the massive size of the companies. The circles are not complete, and extend outside of the bounds of the div in order to emphasize the enormity. Very plain and consistent colors are chosen because we want the focus to be on the size of the elements, and because each element lies in the same context as its neighbors (meaning they are all about number of employees). We used highlighted area to represent ranges. We also feel the user might be interested in know what specific companies are being represented, so we list them as the user hovers over each circle. We feel that this way of representing the data would not only be clear, but also be creative, as opposed to using a traditional bar graph, etc. A chromatic scale was used to show the amount of companies in each section.

Team Members' Contributions

Angela Wu:

1. Responsible for Figure 1 design and implementation.
2. Filtered and integrated datasets for Figure 1.
3. Led the discussion and moved the project forward.
4. Contributed to and edited the reports.

Rong Hu:

1. Responsible for Figure 2 design and implementation.
2. Collected and integrated data for Figure 2.

3. Participated in the discussion and listened to others’ feedback.
4. Contributed to and edited the reports.

Val Mack:

1. Responsible for Figure 3 design and implementation.
2. Designed theme for the web page.
3. Maintained Github repo.
4. Contributed to and edited the reports.

Sources

[1] Fontinelle, 2017. http://www.investopedia.com/ask/answers/12/what-is-a-startup.asp

[2] http://sites.bxmc.poly.edu/~guanchensong/MER/?cat=8