Matt Querdasi

QAC 251

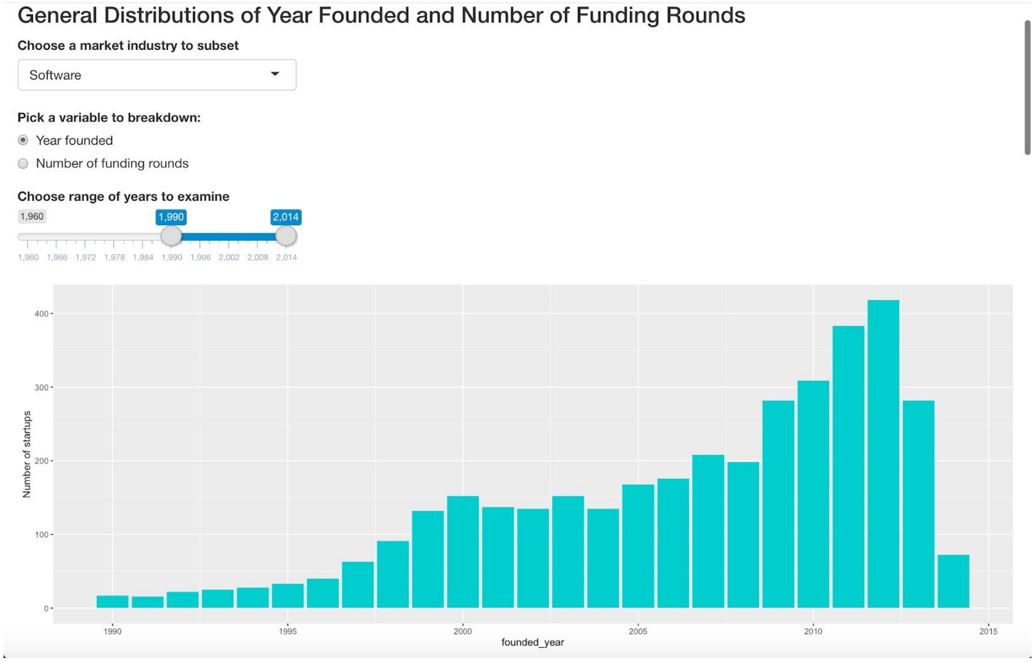
Final Project Report

Introduction

Startup companies are the driving force of innovation within society. They explore new ideas, ask different questions, and are extremely important to how industries form and technology is invented. The behavior of these startups contains valuable insights, and can help us understand more about them. For my final project, I analyzed a Kaggle dataset that contained data on over 50,000 startup companies from countries all over the world. The dataset is originally from Crunchbase, a business information platform that records company and industry data. The dataset I analyzed includes startups from as early as 1902 up until 2014, and includes startups from over 750 unique market industries. With my analysis, I aimed to identify trends using multiple forms of data, and visually display these in a simple, yet informative manner. For my project, I created a Shiny app that allows the user to (depending on the plot) select a market industry of interest, time frame, country, or funding variable to further explore. My analysis utilizes spatial data, time series, and pure numeric and categorical information to clearly display patterns and trends within each category and sub-category of interest.

Shiny App

General Distributions



General Distributions of Year Founded and Number of Funding Rounds

Choose a market industry to subset

Software

Pick a variable to breakdown:

● Year founded

○ Number of funding rounds

Choose range of years to examine

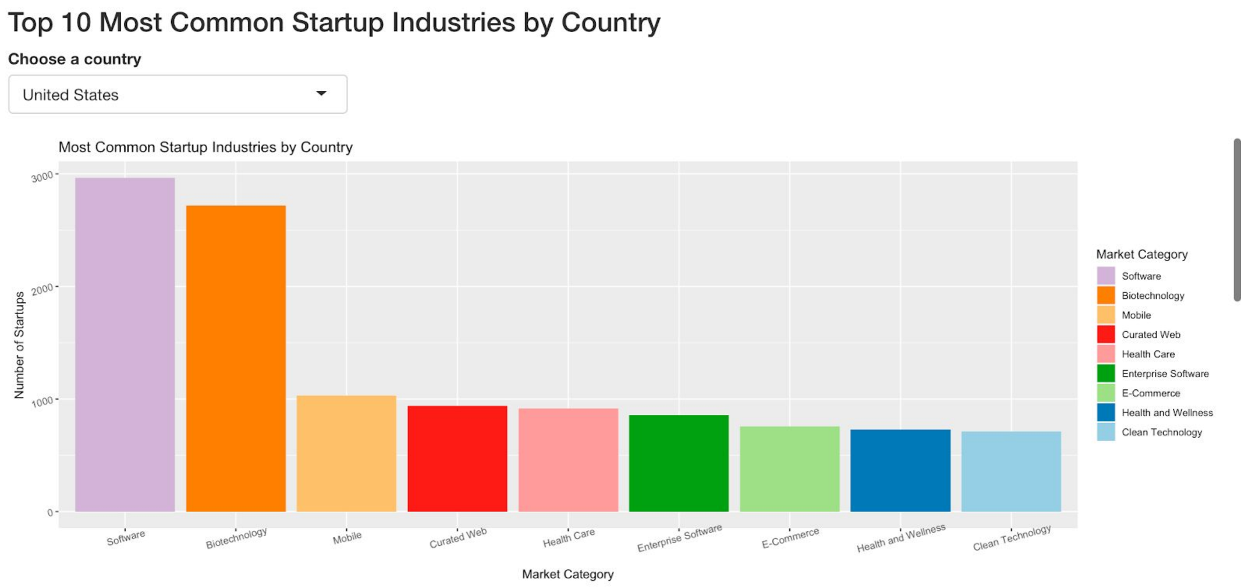
1,960 1,990 2,014

For my first graph in my Shiny app, I display a barplot with three different user options. The first selection is a market industry to subset. This option allows the user to break down the plot for only startups in that market industry. The second user input is choosing the variable to break down, the choices being the year a startup was founded or the number of funding rounds. The last option is simply to allow the user to subset the data to only include a specified range of years.

I intentionally give the user many options to subset by market industry of interest, however there are some specific interesting trends I observed. It can be seen using this display that the number of new Clean Technology startups appear to have peaked around 2007-2009, and after began steadily declining. In addition, the number of new Software startups appear to have doubled from 1997 to 2000, and then evened out for a period of about 6 years, then experienced another doubling from 2008 to 2013. These sharp increases coincide with the tech boom in the late 1990s, and the more recent tech boom in the late 2000s/early 2010s. While each industry has their own unique trend over the years, in comparison to Clean Technology and Software, the number of new Health Care startups had a more steady increase, starting around 1998 and ending around 2013, before steeply declining.

As for findings for the number of funding rounds, for every top 10 industry I chose to display most startups only engage in 1 funding round. For the Software industry this is an extreme: the vast majority of Software startups only engage in 1 round of funding. However, Health Care startups interestingly still mainly engage in 1 funding round, however proportionally much less so. Another interesting feature to note is that most industries had a maximum (i.e. outliers) of about 12-13 funding rounds. That is, a couple startups in each market industry had at most 12-13 funding rounds. However, Enterprise Software had an outlier with 18 funding rounds, about 5 more than most.

Most Common Startup Industries by Country



Top 10 Most Common Startup Industries by Country

Choose a country

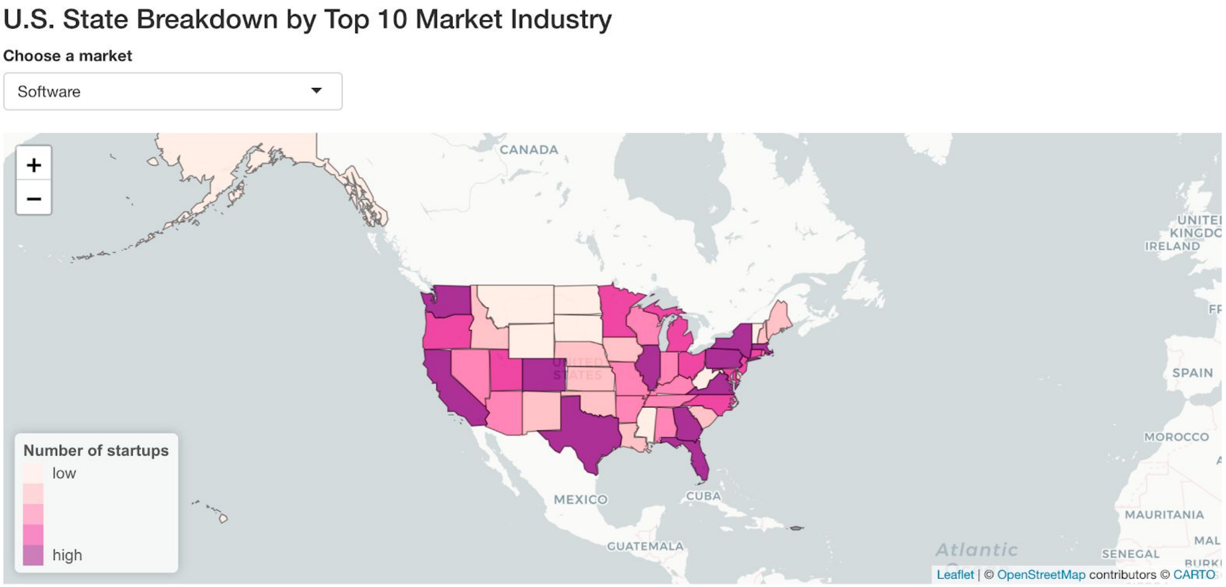
United States

For my second visualization, I give a simple overview of the top 10 most common startup industries for the top 20 countries with the most amount of startups. While not particularly important, selections for the countries are ordered by amount of startups, so the United States has the most startups and South Korea the least.

I found from my visualization that most countries follow very similar trends. For practically all countries, Software or a similar technology based industry (Mobile, Curated Web, etc.) was the most common type of startup. For the United States, Software and Biotechnology were by far the most common type of startup, with the next most common being Mobile. For the U.S., there appeared to be practically two tiers of top 10 startups: Software and Biotechnology both at around 2700 startups, and then all other categories in the 700-1000 range.

It was also interesting to see different apparent specializations for each country. For example, in countries like the U.S. and China Clean Technology startups are not very common--the 9th most common startup in the U.S. and unlisted (not in the top 10) for China. However, for startups in Canada, Clean Technology is the 4th most common type of startup.

State Breakdown by Industry



U.S. State Breakdown by Top 10 Market Industry

Choose a market

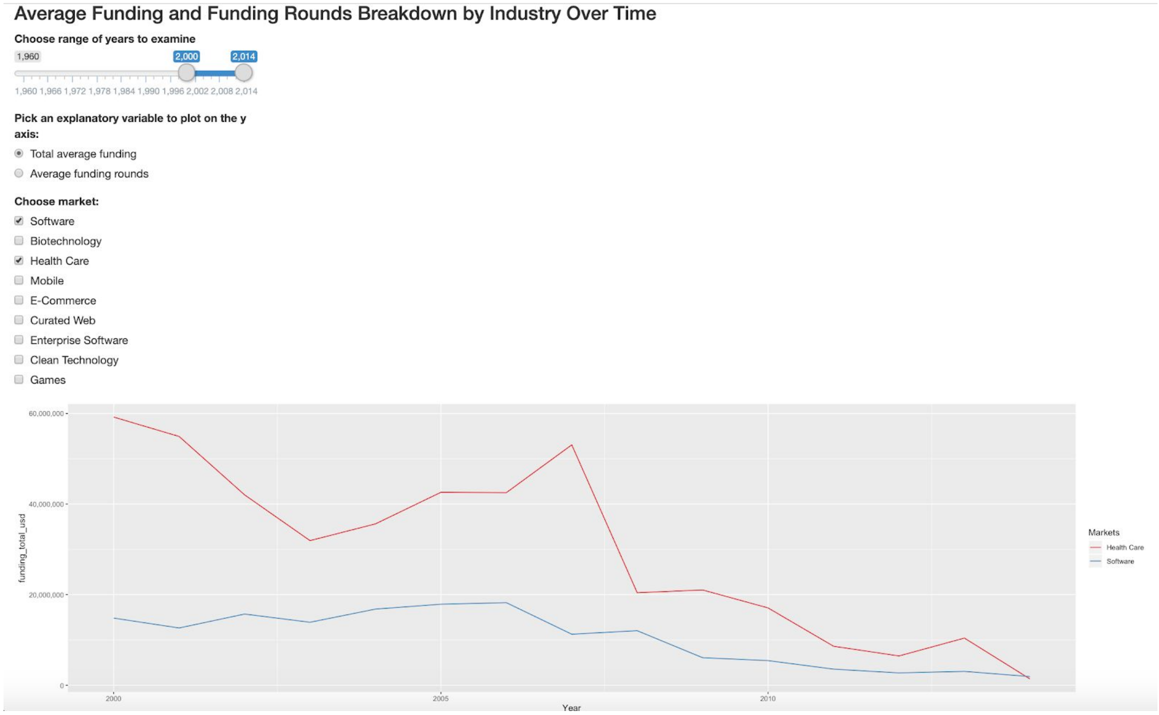
Software

For my third visualization, I display a breakdown for a selected market industry for each State in the United States. I created visualization using Leaflet and include a pop-up with the State name and the exact number of startups for the selected industry in that State. Users can select a market from a list of the 10 most common startups in the U.S., and the resulting Leaflet will fill according to the number of that type of startup in each state.

The findings from this visualization were really interesting. There is a very strong overall trend that most startups are concentrated towards the coasts, or select midwest states such as Illinois or Minnesota. For Software startups, the coastal concentration was especially apparent, with especially concentrated States including California, Washington, Texas, New York, and Massachusetts. For Health Care startups the distribution was more evenly spread, and not explicitly coastal. The major Health Care startup state is California with 291 startups, however the Tri-state and New England regions are also very concentrated. In addition, Minnesota, Texas, and Florida also contained a large amount of Health Care startups.

Another interesting finding came after examining Clean Technology startups. The vast majority of these types of startups are on the Westcoast, however there is a dense clustering around New York and Pennsylvania. Unsurprisingly, but perhaps the most important finding: for practically all of these startup categories, California contained the most amount of startups.

Total Average Funding and Average Funding Rounds



Average Funding and Funding Rounds Breakdown by Industry Over Time

Choose range of years to examine

1,960 — 2,000 – 2,014

Pick an explanatory variable to plot on the y axis:

● Total average funding

○ Average funding rounds

Choose market:

☑ Software

□ Biotechnology

☑ Health Care

□ Mobile

□ E-Commerce

□ Curated Web

□ Enterprise Software

□ Clean Technology

□ Games

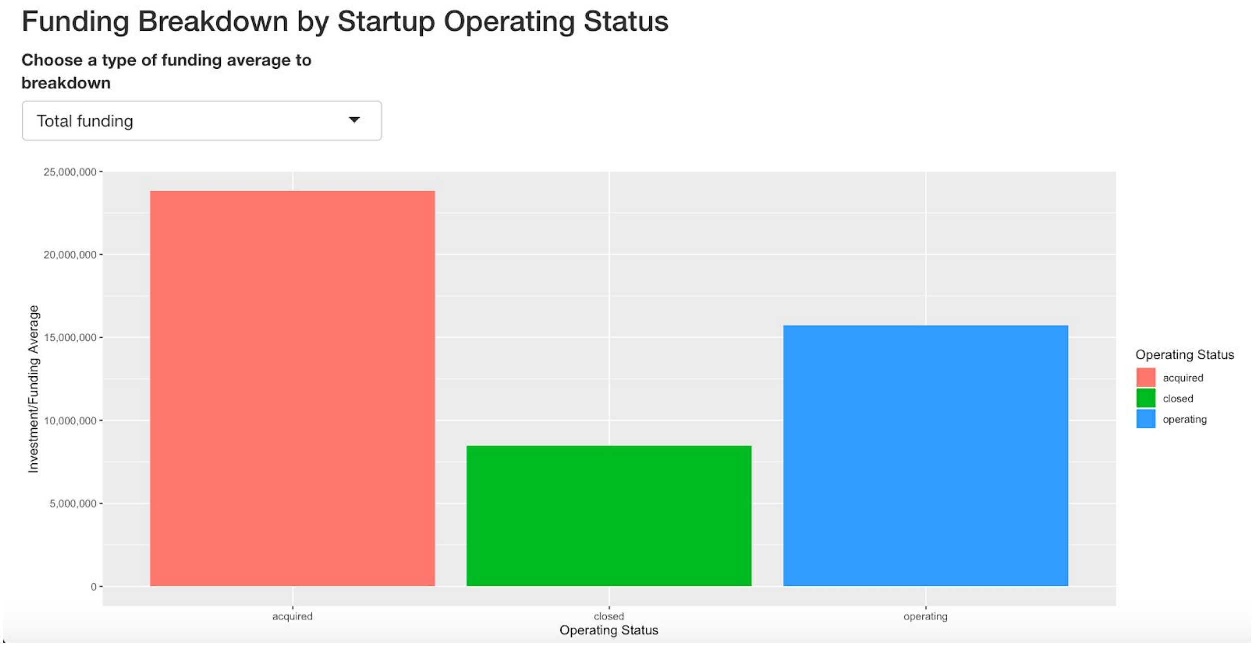
For my fourth visualization, I display a line plot over time explaining a user chosen variable of either the total average funding or the average number of funding rounds. I allow the user to choose the time frame they would like to examine to make it simpler to observe trends, and to allow users to observe a specific market industry, or compare multiple. Using these features, users can see trends in funding over time by industry, or see how the average number of funding rounds have changed over time.

One interesting finding that I observed using this visualization was a sharp decline in average total funding right around 2008-2009. It seems that most industries either experienced a sharp decline in average funding, or a slower decline that started around 2008. I would assume this was because of the 2008 housing crisis, as funding was harder to come by during the economic turmoil.

As for more market specific observations, it is once again interesting to compare Software and Health Care. Funding for Health Care startups is consistently almost double Software startups funding. This seems to indicate that Health Care startups typically need a lot of funding, while Software startups are typically on the lower end. This assumption is also reinforced by the sheer amount of Software startups seen in previous visualizations. One possible explanation coming from this visualization (and the previous) is that Software startups typically require much less funding to get started, and thus more Software startups are attempted.

It's also interesting that this relationship between Software startups and Health Care startups is present for the average funding rounds as well. Software startups typically have significantly fewer funding rounds than Health Care startups. This may perhaps be because Health Care startups generally require more funding, or perhaps because many Software startups don't make it past the second round of funding. In addition, another general trend is that the average number of funding rounds across all industries seems to be decreasing. This may also be because of the 2008 crisis, as it appears 2008 is the start of the decline for most industries.

Funding Breakdown



Funding Breakdown by Startup Operating Status

Choose a type of funding average to breakdown

Total funding ▾

For my final visualization, I breakdown funding and investment by the operating status of the startup. The operating status' are "acquired" for when a startup is bought by another company, "operating" for when the startup is still operating independently, and "closed" for when a startup closes/shuts down. In my visualization, I give the user the option of choosing what funding/investment type they want to examine.

While many observations that come from this visualization may be thought of as general knowledge, I believe they are important to note with statistical evidence. For example, startups that have been acquired on average receive $7-8 million more in total funding than companies that are still in operation, and approximately $15 million more than companies that have closed down. It is also interesting to note that typically startups that are still operating receive the most amount of seed money. This could perhaps be because of the intended structuring of the company that the founders have, i.e. those who receive more seed money have more of an incentive to stay independent (i.e. grow their startup more) rather than to sell their startup.

While each investment type displays a unique trend, another particularly interesting finding came from the angel investment breakdown. Startups that have closed actually have the highest average angel investment, an extremely surprising fact (at least to me). Another interesting funding type is the grant category. Perhaps unsurprisingly, startups that are still operating by far receive the most money from grants. A possible explanation for this may be because grants often come from government entities, and therefore companies receiving grants may intend to not be bought out, and remain independent. Regardless, this visualization provides many additional variables to observe.

Questions and Future Implications

My intention with my analysis was to provide a broad overview of the observations in the Crunchbase dataset and to identify basic geographic and quantitative trends for these startups. While I believe I largely accomplished my goal, I believe further examination could prove extremely interesting as well. I believe it would be exciting to go more in depth with the geographic analysis that I started. One of the great features of the dataset is that it provides information on the country, state, and even city where the startup was founded. I would ideally like to investigate these apparent state breakdowns on a city level. For example, in my analysis I found that California by and large has the most startups for practically every startup industry. However, does this finding represent California as a whole? Or are these startups concentrated in a couple of cities? Or perhaps a region? I'm sure everyone has heard of "silicon valley", but perhaps there are other regions, not necessarily in California, where specific industries are concentrated. In addition, I would like to further delve into my last visualization breaking down funding by operating status. I think it would be extremely valuable to identify more traits that successful startups share, and try to identify potential reasons why other startups don't succeed. While I was able to provide a basic funding overview for operating status with my last visualization, in future work I would like to go deeper and perhaps see if things like location, number of nearby startups, and year founded play a role in the success of a startup.