Startup Scraper:

Unveiling the secrets of the startup/VC industry through publicly accessible data

By: Ian Liu, Evan Grossman-McKee, Nicholas Alioto-Pier

Abstract:

Hypothesis

We utilized publicly available, reputable sources to analyze the **biases and considerations of top venture capital firms (VCs) in providing funding** and analyze the **effects of VC funding amounts on startup success**. Our specific hypotheses are: **(1).** Pre-pandemic, minority-founded startups received a **different amount of funding** than non-minority-founded at Y-Combinator (YC). **(2).** Pre-pandemic, fintech startups **received more funding** than non-fintech startups at YC. **(3).** The initial total funding of a startup at YC is **positively associated with its employee count in June 2020**. **(4).** Total funding amount and a brief company description can be used to accurately **predict if a YC startup is live or dead** in June 2020.

Data

We chose to analyze companies from **Y-Combinator's publicly accessible startup directory** **https://www.ycombinator.com/.** For context, Y-Combinator is widely considered the most successful VC in the world, helping launch over 4,000 companies since 2005.

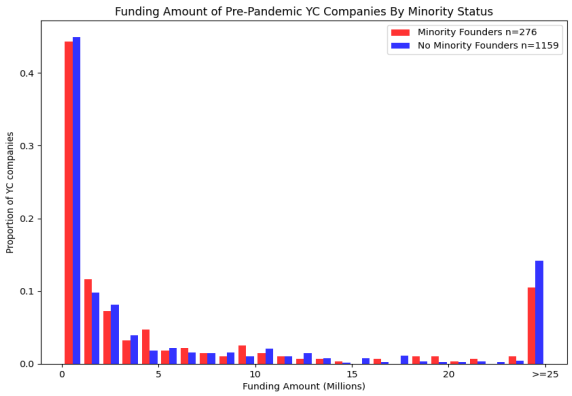
We first scraped data from the official Y-Combinator website, which contains info on company descriptions, industry, and minority-founder status. We combined this with scraped data from an unofficially affiliated, but reputable, website: **Y-Combinator database** **https://www.ycdb.co/,** which contains info on total funding amounts and company status through June 2020.

Funding amounts are rightly skewed with extreme outliers: most YC companies receive $100k or less in funding, while some unicorns receive over $6B.

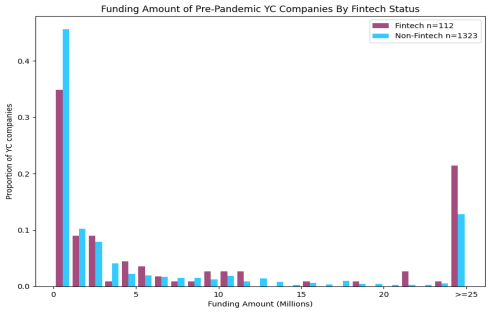
Findings

Claim #1: Pre-pandemic, minority-founded startups did not receive a different amount of total funding than non-minority-founded startups at YC.

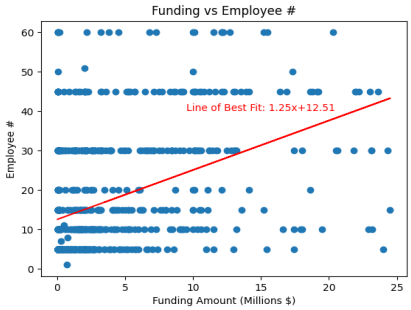
Support for Claim #1: We used a 2-sided Mann-Whitney's U-Test with a significant value of $\alpha = 0.05$ to compare the funding distributions of minority-founded startups with non-minority-founded startups. We found that there were no significant differences in funding distributions ($p = 0.452$). While mean funding for non-minority-founded startups ($34.83M) is almost double the mean funding for minority founded startups ($18.98M), the median funding for non-minority founded startups ($1.7M) is higher than minority-founded startups ($1.6M).



Claim #2: Pre- pandemic, fintech startups received more total funding than non-fintech startups at YC.
Support for Claim #2: We used a 1-sided Mann-Whitney's U-Test with a significant value of $\alpha = 0.05$ to see if the funding distributions of fintech startups is greater than non-fintech startups. We found that there is a significant difference in funding ($p = 0.001$). Mean funding for fintech startups ($44.79M) is almost 1.5x the mean funding of other startups ($30.68M) and median funding for fintech startups ($2.7M) is 1.8x the median funding of non-fintech startups ($1.5M).



Claim #3: The initial total funding of a pre-pandemic YC startup is positively associated with its employee count in June 2020.
Support for Claim #3: We used a multi-variate OLS regression with a significant value of $\alpha = 0.05$ to find to see if total funding amount is positively correlated with employee count in June 2020, controlling for years since YC. With the control, our linear regression model returned a p-value of $p = 9.69e\text{-}40$ and an R-value of $r = 0.419$. In other words, we discovered a weakly positive, but highly statistically significant, correlation between pre-pandemic YC startup funding amount and employee count in June 2020.



Claim #4: A pre-pandemic YC company's description and total funding amount is a relatively accurate predictor of whether it is live or dead in June 2020.
Support for Claim #4: We created a KNN model that would predict whether a company is live or dead in June 2020 based on its brief company description and funding amount. We used Google's Word2Vec model trained on a corpus consisting of only YC company descriptions to convert each company description into a numeric vector that represents similarities across YC descriptions. We then added funding as roughly 3x more important than company description in KNN. Finally, we randomly under-sampled the # of live companies to be equal to the # of dead companies. Our KNN model resulted in an average cross-fold validation accuracy of 0.737. Our model was not biased towards live or dead predictions as shown by the confusion matrix below.