Startups: How They Do It!

A Data Analysis of Venture Capital Funding

COMP 30780 Data Science in Practice

Jack Gallagher (15508033), Sean Gallen (15500543)

School of Computer Science

University College Dublin

May 11, 2018

Abstract:

Startups are booming! We take a look into the world of everything startups. We examine the key indicators of success, everything from the location of a company, the industry a company operates under, as well as other variables such as where a founder was educated, or whether they had previous experience or not. The amount of money being invested into new companies in this current age is astronomical, and seems to be growing year-on-year. There is an opportunity for anyone to get their foot in the door of this revolution. Throughout this project our main focus was the amount of funding a company receives, and whether or not the factors listed above affect this amount in a positive or negative way. The analysis we performed enabled us to have a look into the world of successful businesses, and potentially how to build them. Throughout our analysis we came across really surprising insights into certain aspects of our research questions. We found some surprising locations that may be ideal to setup startups operating in certain industries. On top of this we established that a founders education is a pivotal factor for a successful startup.

Declaration. We Sean Gallen (15500543) and Jack Gallagher (15508033) declare that this assignment is our own work and that we have correctly acknowledged the work of others. This assignment is in accordance with University and School guidance[1] on good academic conduct in this regard.

[1] See https://www.cs.ucd.ie/sites/default/files/cs-plagiarism-policy_august2017.pdf

1. Introduction

Starting a company is difficult. It takes a huge amount of effort, time, and discipline. It can take years of work with very little reward before hitting it big. It is an exhausting, and very time-consuming process that can take a substantial amount of hours of work a day, only for a sizeable amount of these companies to fail. In recent years there has been a somewhat of an glamorization placed on entrepreneurship, and the world of startups. It can be argued that we are well and truly caught in the age of entrepreneurship. The aim of our work is to develop an understanding of the factors that affect the chances of these companies getting funded, and to what degree. The main factors we investigate in this work are how the correlation between a company's industry and location affect the funding said company receives, the affect a founders education has on the successfulness of their company, and whether founders with previous experience are more successful in their future ventures, versus founders with no previous experience.

This project is mainly aimed at gaining a better understanding of what makes these huge companies that appear in the news everyday become so successful, and why so many companies fail to replicate the success of the very few successful companies. A very daunting and discouraging statistic reported by forbes is that "90% of startups fail" (Neil Patal, 2015). We want to understand what factors affect the other 10% that succeed. This is of huge interest to not only us, but hopefully to all future and current entrepreneurs. This information could provide guidance on where to locate a company, educate yourself, or what university alumni to network among. The capitalistic nature of our current economy seems to be forever growing, and this is having profound effects, both positively and negatively on every aspect of today's society. Due to this we feel that our project is of interest, and potential benefit to a large variety of society.

In order to gain an insight into the factors mentioned above, namely the effect the correlation between a company's industry and location have on the funding said company receives, the affect a founders education has on the successfulness of their company, and whether founders with previous experience are more successful in their future ventures, versus founders with no previous experience, we thought it best to investigate each factor individually. We made the decision to focus on three major countries, Ireland, Great Britain and the United States of America. To begin, we analysed the first factor, the effect the correlation between a company's industry and location have on the funding said company receives. To do this we calculated a normalized value for the dominance of each industry in each location. Our findings were intriguing. We found that in certain industries such as Technology, the locations that we would associate with these industries were not necessarily the locations with the highest chance of getting funded.

Next, we analysed the affect a founders education has on the successfulness of their company. We investigated a founders education in terms of two main aspects: the highest degree obtained by a founder and the institutions they attended. Our exploration and analysis of this research question lead us to the conclusion that both of these aspects impact the funding of their startup. Unlike other research completed on this topic, we decided to evenly split founders into three groups based on the amount of funding their company achieved. By doing this, it enabled us to make comparisons between each of the groups and to explore some intriguing trends further. Our final investigation involved gaining an insight into the differences between an experienced and non-experienced founders. We defined an experienced founder as someone who has founded more than one startup. Alas, we had to deem this final research question as inconclusive due to the small amount of data we had on experienced founders, however, we carried out some very interesting gender analysis of the founders

to compensate. A stark fact about the founders that we discovered, was that there were seven times as many males as there were females present.

The remainder of this report is organised as follows: In the next section, 'Motivations & Objectives' we take a deeper look into the background of our project, the real reason we as a group decided to investigate the questions we did, and ultimately we discuss our motivations behind undergoing such a topical project. Another major part of section two is the 'objectives' of our project, in relation to the objectives we will be listing each of our research questions, discussing in detail all aspects of each of the research questions. The third section of this report is on 'Data Wrangling', in short this section is solely focused on discussing the methods we used in obtaining and preparing our data. Throughout this section we will discuss the tedious process we undertook to retrieve our initial datasets from Crunchbase, as well as the time consuming process of cleaning and preparing our data. In section four, 'Data Analysis and Results' we will delve much deeper into each research question. Having previously described the research questions in section two, section four will be focused on the approach we took to answer these research questions, the results we obtained, and finally a brief discussion on how we interpreted our results. We then move onto the fifth section which is the 'Discussion' section. In the discussion section we will look at two very important aspects of any data science project, reproducibility and the ethical considerations behind the decisions made throughout the project. We will also briefly discuss the limitations we came across as we progressed over the course of our project. In conclusion we finish with our sixth section, 'Conclusions and Future Work'. This section offers a concise summary of all we have done throughout the project, as well as our findings. There are a whole host of opportunities for future work on this project, we will discuss what we would like to do in the future shall the opportunity arise. Like any project report, we have included an appendix and bibliography at the end of the report.

2. Motivations & Objectives

We provide further insight into the motivations behind our reasoning for undertaking this project, as well as our main objectives for this project. After discussing the previous work that has been done on this topic, we will discuss why we chose this project in particular and give an interesting insight into the topic on which this project is based. We will also give an overview of each of the research questions below, describing the features of each of the questions.

2.1. Background & Motivations

Startups are obviously grabbing headlines 'left-right and centre' at the moment and the amount of funding being raised is growing year on year. This of course grabs the attention of students around the world looking for direction in the world after they graduate. Venture capital funds invested $164 billion worldwide in 2017, with an average deal size of over $10 million (PricewaterhouseCoopers and CBInsights, 2017). On top of this, our personal interest was a deciding factor in us pursuing this topic for our research. This is also an area we would like to get involved in the future potentially.

From the outset, we were interested in exploring the extent of which different factors affect the funding of startups. There has been some previous research in this area. This was not a surprise to us as it is a topic a lot of people are curious to learn more about. The impact of digital start-up founders' higher education on reaching equity investment milestones Ratzinger, D., Amess, K., Greenman, A., & Mosey, S. (2017)) is an interesting report focusing on the impact of higher education on start-ups in the digital industry. This was a particular research aspect that we were intrigued by and wanted to examine further. The effects of opportunities and founder experience on new firm performance (Dencker, J. C., & Gruber, M. (2015)) is a very detailed report on the effects of opportunities and

founder experience on new firm performance. However, this report mainly focuses on the previous opportunities of founders in regards to the other jobs they have had. Another area of this topic that we were keen to explore was the location and industry of a startup and the potential implications this has, good or bad.

2.2. Research Questions

2.2.1. RQ1: Does the correlation between a company's location and industry impact the amount of funding a company receives?

The first research question explores whether the location of a company and the industry it is involved impacts the funding achieved by a company. Our aim is to get an insight into the affect the combination of these two factors has on the venture capital raised.

2.2.2. RQ2: Does the founders education affect the funding their company receives?

This research question focuses on discovering whether a correlation exists between a founders education and the funding their company receives. Education of a founder will be analysed in terms of two main aspects. The first of these is the highest degree obtained by a founder. The second aspect related to the institution which a founder attended.

2.2.3. RQ3: Do founders with previous experience command more funding than founders with no previous experience?

For this research question an experienced founder was defined as a founder of more than one startup. We will be investigating the differences in funding that experienced and non-experienced founders receive. For example, exploring if the average funding of an experienced founder differs from that of a founder without experience.

3. Data Wrangling

3.1. Data Acquisition

We acquired our key datasets from a highly regarded startup information database called Crunchbase. The reason why we chose to use data provided by Crunchbase is due to it being updated more frequently than its competitors and due to the comprehensive data it has on startups. Identifying the main aspects of the data needed for our project before pursuing access was very beneficial. Unfortunately, the majority of Crunchbase's data cannot be accessed without a subscription fee. The next step was exploring how to gain research access to their API. Originally, we were not granted such permission, however, after numerous emails with a Crunchbase representative, we successfully were given an API key. There were multiple conditions that we had to uphold in order to use the API. The first of these was that no data accessed using the API was allowed to be used commercially. Following on from that, we had to ensure that this data was only used for our project and is not made publicly available. We ensured that we obeyed these requests and will continue to do so by not making any of the data we used publicly available.

The Crunchbase API was very intuitive to use. We received nine separate datasets in total from this access. Only four of these datasets had the required information that we previously identified. The first of these datasets was a data file called 'organizations.csv'. The data in this file included detailed information about all the organizations on Crunchbase. The organizations present in this dataset were not solely startups but also Venture Capitalists amongst other types. We noted this for our preparation

processes. The important aspects of this dataset include the location where the organization is based, the list of industries which it is involved in and the amount of funding total it received in American dollars. Another crucial aspect of this dataset was a unique identifier given to each organization. This was significant in connecting the multiple key datasets.

The three other key datasets, from the initial nine obtained, focused on information regarding the people involved in the organizations in the first dataset. The first of these datasets called 'people.csv' detailed general information about the people involved in the organizations like their gender and what country they were from. The penultimate key data file called 'jobs.csv' detailed important information on the title of a person at a specific organization. The final key data file called 'degrees.csv' detailed significant educational information on the person at a specific organization like the degree(s) they obtained and the institution they attended. Another important point is that these datasets were able to be linked together using a unique person identifier that was present in all of these datasets that focused on information regarding the people involved in the organizations. A significant detail of this dataset is that there are multiple instances for a founder. A founder has multiple instances if they have founded multiple companies, if they have changed positions in the same company or if they have obtained another degree or changed institution while at the same company.

In conclusion, a huge amount of data was obtained from Crunchbase which had the potential to be very overwhelming. Identifying the key aspects necessary for our project and having an idea of the research questions before pursuing the data from Crunchbase proved to be very beneficial.

3.2. Data Cleaning & Preparation

The nine datasets received from Crunchbase were cleaned and prepared which resulted in two main datasets used for the duration of the project. This process was where most of our time was spent. Two resulting datasets proved be more efficient for the project. The first of these datasets focused on information about the startups. This first dataset was used for the first research question as startups are solely being investigated in this. The second dataset focused on information about the founders of these startups. This second dataset was used for the second and third research question. The second and third research questions were related to gaining an insight on a founder of a startup rather than the startup itself.

3.2.1. Cleaning & Preparation of first dataset

The first dataset that was used was initially based on the "organizations.csv" data file received from Crunchbase. This dataset was used for the first research question which is exploring whether industry and location of startup influences the amount of funding received by a startup. The removal of unnecessary columns was the first step of preparation undertaken. Columns relating to a company's social media URLs are examples of columns deemed unnecessary. Following on from this, all the companies which had no funding value recorded were removed. Crunchbase does not always provide these funding values. The funding value of a founders company was required as it was used as the success metric Similarly, companies with no details on the industry/industries that they are involved with were removed.

company_name	country_code	state_code	funding_total_usd	category_group_list	org_uuid	Technology	Communication	Health Care
ShareRing	AUS	Unknown	3800000.0	Apps,Design,Information Technology,Software	72c415e7-03d6-4f28-bf56-7a6f29d7f1b8	1	0	0
Skyroam	USA	CA	43500000.0	Hardware,Internet Services,Mobile	a2717ae2-d73f-96c7-7177-f0d7d39ec700	1	1	0
SoYoung Technology	CHN	Unknown	110000000.0	Health Care	f586d684-1dc2-879b-b19d-e611c889a304	0	0	1
StepLadder	GBR	Unknown	349550.0	Real Estate	11283bcb-6e6d-427d-b4ff-9c701d53ee0f	0	0	0
Strix Leviathan	Unknown	Unknown	1625000.0	Consumer Electronics,Financial Services,Hardwa...	86286704-cb8d-4567-9ca6-cf74aa9e4c7e	1	0	0

Figure 1 shows a sample of columns from the resulting prepared dataframe of dataset one.

The initial challenge faced while undertaking the preparation process of the first dataset involved transforming a column that detailed the industry of the company. The majority of companies are involved in multiple industries and therefore the industry string values were comma separated. The industry of a company was a key part for the first research question. Transforming the industry list column (‘category_group_list’) resulted in a matrix-like dataframe where each specific industry was a column field. There were 46 different industries in the resulting matrix-like dataframe. However, we grouped similar industries to get a final total of 36 unique industry types which companies can operate under. The first industry we created from the similar industries was Technology. The industries we grouped to get this resulting industry were 'Apps', 'Artificial Intelligence', 'Consumer Electronics', 'Data and Analytics', 'Gaming', 'Information Technology', 'Internet Services' and 'Software'. The second industry created was Finance from the grouping of 'Payments', 'Financial Services' and 'Lending and Investments'. The final industry, Communication, was assigned from 'Messaging and Telecommunications' and 'Mobile'. This matrix-like dataframe was then merged to our original dataframe. A sample of columns from the resulting dataframe can be seen in *Figure 1*. Fields that described location data such as the country, city and state were the next focus of preparation. It was ensured that any ‘NaN’ values in the location data of the companies were replaced with a value of ‘Unknown’.

In conclusion, the first dataset contained ~78,000 rows. The first dataset was used for the first research question. The data in the first dataset detailed information solely about companies that have been founded, like the amount of funding gained, the industries in which it operates and the location that it is based.

3.2.2. Cleaning & Preparation of second dataset

The second dataset used was based initially on the ‘people.csv’ data file received from Crunchbase. This detailed all the individuals involved in a company, not just the founders. However, in order to filter this data to only contain founders of companies, the job data of these individuals had to be added. This jobs data for individuals was contained in the ‘jobs.csv’ data file. There was a unique person identifier (‘person_uuid’) that was in these data files that allowed the merging of the jobs data accurately. Once the merge was successful, the filtering of individuals that were not founders was completed. The next step was adding the educational information of a founder which was contained in

the ‘degrees.csv’ data file. The unique person identifier (‘person_uuid’) present in both dataframes allowed the merging of this educational information accurately.

A major part of the second and third research question is comparing the impact a particular attribute of a founder has on the amount of funding they received. It was therefore very important that there was a funding value for the startup of each founder. We filtered out founders that did not have this funding value as it is the success metric to be used for analysis. A significant detail of this dataset is that there are multiple instances for a founder. A founder has multiple instances if they have founded multiple companies, if they have changed positions in the same company or if they have obtained another degree or changed institution while at the same company. We explored the possibility of merging these instances into a single row, however, having these separated instances was more efficient for analysis. The groupby function was a very powerful function used to manipulate these duplicate instances. *Figure 2* demonstrates these reasons with the multiple occurrences of Armas Markkula and Nolan Bushnell.

first_name	last_name	gender	company_name	funding_total_usd	country_code	degree_type	Highest_Degree	institution_uuid
Steve	Wozniak	male	Apple	6.150250e+09	USA	bs	Bachelors	10f9a25b-9675-2281-486e-a52955c706df
Kevin	Harvey	male	Apple	6.150250e+09	USA	bs	Bachelors	c3144da5-8618-2e95-3a13-60417220da5e
Armas	Markkula	male	Apple	6.150250e+09	USA	bs	Masters	867f0af5-a1d0-143d-bbed-5cc252ca40d6
Armas	Markkula	male	Apple	6.150250e+09	USA	ms	Masters	867f0af5-a1d0-143d-bbed-5cc252ca40d6
Kristee	Rosendahl	female	Apple	6.150250e+09	USA	ba	Bachelors	20135206-96eb-8be0-9ac4-670b257e532c
Nolan	Bushnell	male	Atari	2.226000e+07	USA	mba	Masters	20135206-96eb-8be0-9ac4-670b257e532c
Nolan	Bushnell	male	Atari	2.226000e+07	USA	be	Masters	5fed9dd9-f09b-632e-da77-036a077ef5cb

Figure 2 shows a sample of columns from the resulting prepared dataframe of dataset two.

A large part of the preparation undertaken on the second dataset was focused on transforming the founders degree type field (an example value from this field would be ‘B.A’). This was done in order to find the highest degree obtained by a founder. The field describing the degree type of a founder (‘degree_type’) was found to be user-inputted. There were multiple spelling errors in this field which was not ideal. Thus, manipulating the text inside this field and assigning it to an overall degree was very time consuming. Once the degree type field was in an appropriate format, it was transformed to get a resulting matrix-like dataframe. The ‘get_dummies’ function was used for this transformation. The column fields of this dataframe were all the degree types present in the dataset. Degree types were filtered out of the dataset if their count was less than 10. From there, degree types were assigned to an overall degree (e.g ‘B.A’ assigned to ‘Bachelors’). We proceeded next to rank the overall degrees in descending order with a PhD being the highest degree a founder can achieve. The analysis of the highest degree of a founder was mainly focused on the three most popular degrees - Bachelors, Masters and PhD.

The other aspect of education that was explored for the second research question related to the institution a founder attended. As mentioned previously, the name of the institution attended by an individual was anonymized in the data received by Crunchbase. However, we decided to deanonymize the top ten most popular institutions in the dataset to use for analysis. This was the next challenge faced when preparing the second dataset. The method used to deanonymize these top institutions was

to locate a founder who had attended the institution using the unique institution identifier. The name of the institution the founder attended was detailed in their Crunchbase.com profile. The actual name of the institution replaced the unique identifier of the institution. The last part of preparation completed on the second dataset was adding the industry information to the company. This industry information would be used in the analysis of the third research question when comparing the funding total received by experienced and non-experienced for all the top industries. This industry information was simply merged from the resulting first dataset. The unique company identifier ('org_uuid') enabled us to concatenate the industry fields to the second dataset.

In conclusion, the second dataset contained ~44,000 rows. It was used for the second and third research question. The information in this second dataset focused on the founder of a company with details about the company they founded, the highest degree they obtained and the institution they attended.

4. Data Analysis & Results

4.1. RQ1

The first research question explores whether the relationship between a company's industry and location impacts the funding the company receives.

4.1.1. Datasets

For this research question the first dataset (see 3.2.1) was used. This dataset describes the details of the organisation, including their funding amount. There is also includes information on the location in which the company is located, and the industry under which each company operates. The first key field for this research question detailed the categories under which a company operated. The other key fields detailed the country code where the company was located and the state code if it was based in America. The funding total field was very significant as it was used as the success metric. There are roughly 78,000 companies in this dataset. There were 36 unique industries in our data.

4.1.2. Approach

Our focus for the approach was to analyse the combination of industry and location, as opposed to looking at each aspect individually. To enable us to look at how the correlation between the location and industry of a company impacts the funding a company receives, we decided to look at the number of companies operating under each industry, for each location. As well as looking at the total funding of each individual industry, in each location. We will concentrate our analysis on, the United States of America, because it contained the most amount of companies in the dataset, the United Kingdom, due to it containing the second most amount of companies in the dataset, and finally Ireland. We calculated normalized values of the number of companies operating under each industry, as well as the total funding of each individual industry. Normalizing these values ensured that our results would not be skewed in anyway due to the presence of potential outliers.

Our initial analysis focused on the comparison of the dominance of each individual industry in the given location. In the case of the United States we looked at this from the point of view of each individual state, whereas with Great Britain and Ireland, we looked at this on a country wide basis. The aim was to calculate a fractional value for the dominance of each industry when compared with the other industries in the same location. It was decided that the most effective way of calculating this

was to sum the number of companies operating under each unique industry for each individual location, and then divide these values by the total number of companies across the location. This allows us to calculate a value for the dominance of each industry in all of the locations. Although upon completion of this step we found that there were quite a few states with very little companies in certain industries, or in some case across all industries. In order to alleviate any fears of outliers in our results, we thought it best to remove any industries from our data with less than 50 companies operating under that industry in a certain location. We carried out this step strictly on the USA analysis, as in both GBR (Great Britain), and IRL (Ireland) there was quite a small number of companies operating under most industries, it is therefore more accurate to have included the smaller industries in our analysis for these cases.

The next step we undertook in our analysis was the second fractional calculation on the journey to our normalized values. This step was focused on the dominance of each industry, in terms of the amount of funding companies in each industry demanded, in each individual location (e.g. CA in USA, or IRL). The first step of this process involved replacing the binary inputs under the categorisation of each companies industry with the amount of funding received . Then, similar to above, the summation of each industries total funding for each location was calculated. Finally, a fractional value was obtained by dividing the total amount of funding for each industry by the total amount of funding for all industries.

country_code	Administrative Services	Advertising	Agriculture and Farming	Biotechnology	Clothing and Apparel	Commerce and Shopping	Community and Lifestyle	Consumer Goods	Content and Publishing	Design	Education	Energy	Event
IRL	0.000003	0.00008	0.000271	0.001288	6.446384e-07	0.000741	0.000003	0.000008	0.000042	0.000016	0.000538	0.000266	0.000

Figure 3 shows a sample of columns from the normalized dataframe for Ireland.

The third, and final step in analysing the dataframe in preparation for the results was to normalize the above fractional values. We computed a normalized value by multiplying the respective elements from each of the above calculations. *Figure 3* shows a sample of the normalized dataframe for Ireland. The reason why we did this was due to the varying number of companies in each industry. This resulted in a much more accurate result for the correlation between both the location and industry. If the data remained unnormalized, one may interpret what seems to be a thriving industry due to its dominance in terms of the number of companies, as being impactful in terms of the funding on companies in that industry. This is obviously not always the case as certain companies in industries may receive very small amounts of funding. The normalized value takes this into account, when the smaller funding fraction is multiplied by the larger industry dominance fraction, the resulting normalized fraction is more accurately representative of how well an industry is performing in terms of funding.

4.1.3. Results

The most effective way of presenting the results was through heat maps. Heat Maps are visually appealing and allow for optimal interpretation of results in this case. By examining any of the below heatmaps, it is quite straightforward to make comparisons using the shades given in the index.

We came to the conclusion that the correlation between the location and industry of a company does have an impact of the amount of funding a company receives. The visualisations shown below demonstrate our results. There were some very surprising results which will be discussed throughout this section. Such unanticipated results were not expected when it came to the impact the correlation between industry and location had on the funding. To begin, we will discuss the following visualisations.

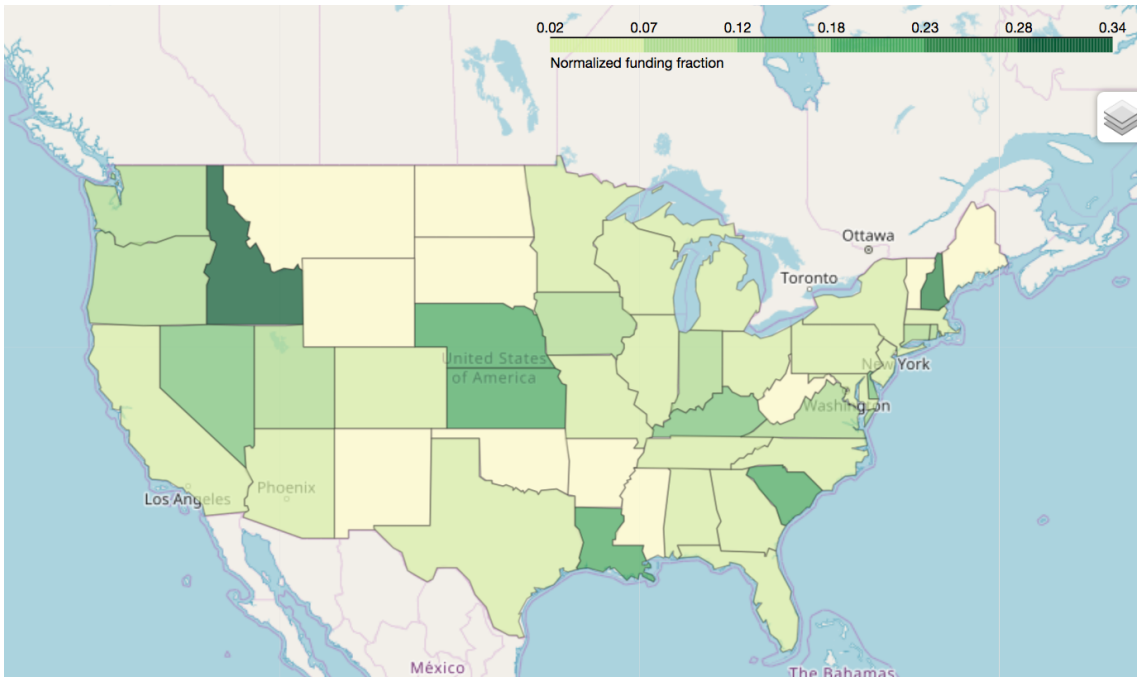


Figure 4(a). Heat map displaying the normalized values for the technology industry across USA.

Figure 4(a) displays the normalized values for each state throughout the technology industry across the United States. The resulting visualisation was quite unexpected. Initially, before analysis of the data one may be inclined to think that the 'best' place to set up a Technology company, in terms of the impact it has on funding would be California. We can see from *Figure 4 (a)*, that this is not the case. The most heavily shaded state on the map, Idaho (ID) is the best place to set up a technology company according to our findings. There are many potential factors as to why this unpredictable result may have occurred, one of which is that there is such a low number of companies operating under other industries in the state, this results in the Technology industry dominating the funding. As a result, locating a Technology company in Idaho (ID) would have a hugely positive impact of the funding a company receives.

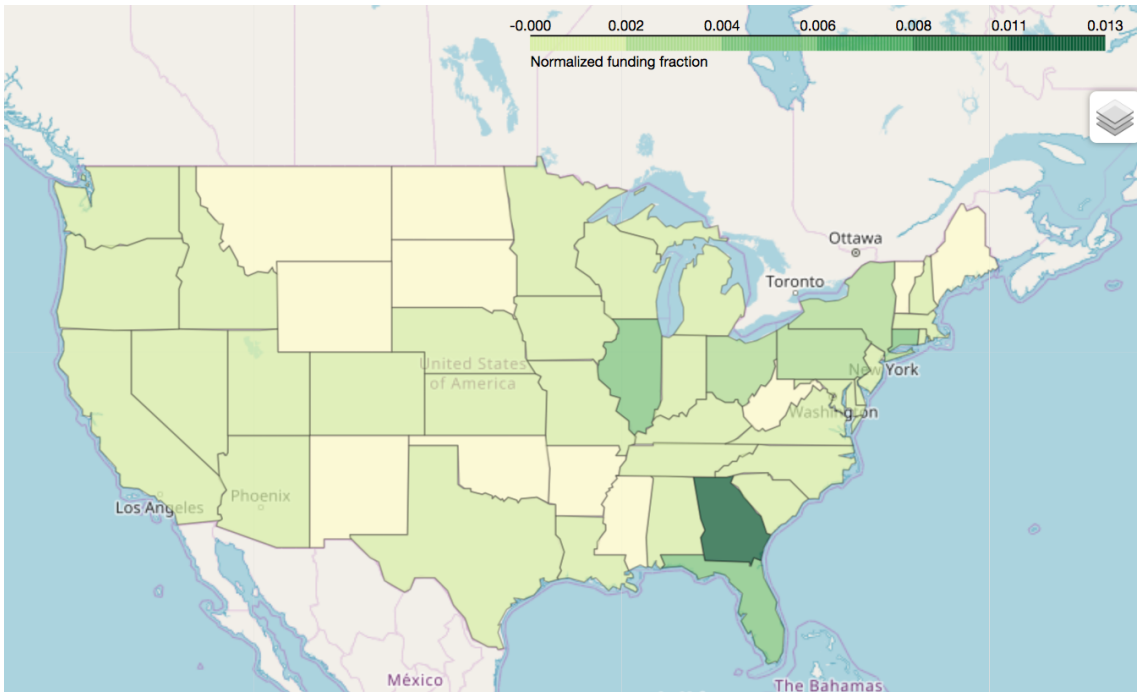


Figure 4(b). Heat map displaying the normalized values for the finance industry across USA.

Our next visualisation, *Figure 4(b)*, is a heat map displaying the normalized values for each state throughout the finance industry across the United States. Once again the resulting visualisation was quite surprising. Once again, before our analysis of this, one may have thought of New York as being the 'hub' of finance in the United States. *Figure 4(b)* demonstrates that this is not the case. Although New York is one of the top 6 most heavily shaded states, the most heavily shaded, and therefore most impactful state in terms of the funding a company in the Finance industry achieves is Georgia (GA). A possible reason behind Georgia being the better state in terms of the impact it has on funding in the finance industry versus New York might because there are a larger variety of industries in New York demanding a larger fraction of the funding in the state.

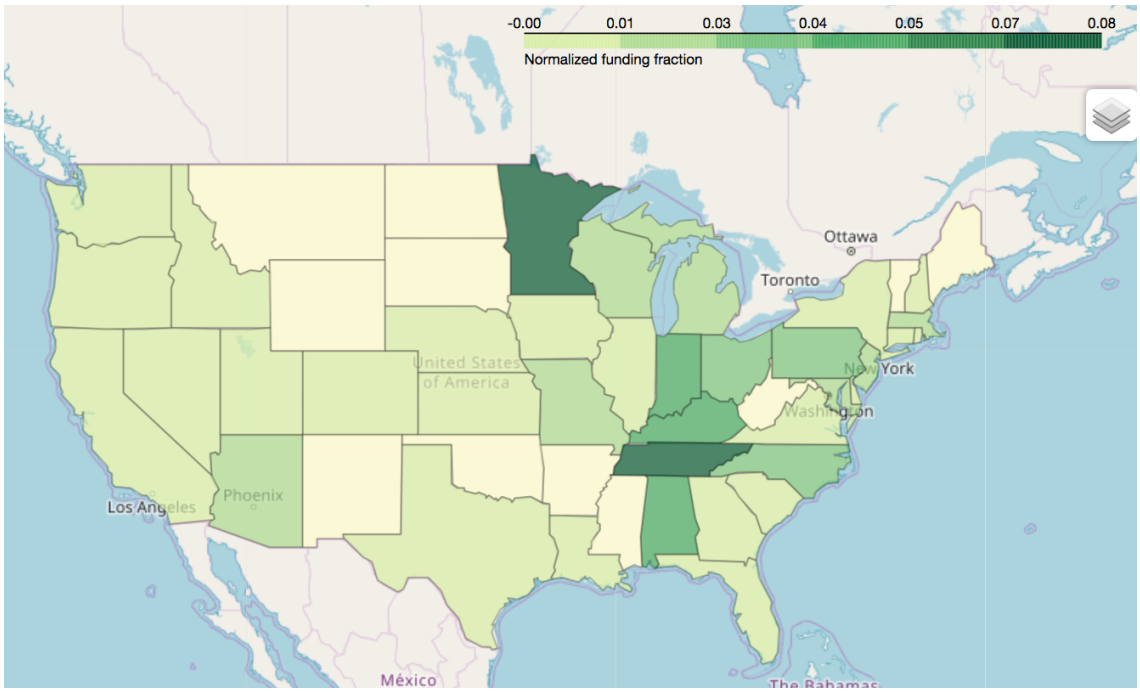


Figure 4(c). Heat map displaying the normalized values for the healthcare industry across USA.

We continued analysing heatmaps, this time in the healthcare industry. *Figure 4(c)* displays the normalized values for each state throughout the healthcare industry across the United States. Looking at the healthcare industry, the East coast seems to dominate. It is remarkable that the major states in terms of the impact on funding in the healthcare industry are all generally located in the East coast, with very little impact being made by the states in the west coast. The two most heavily shaded states, and therefore most impactful in terms of funding were Minnesota (MN) and Tennessee (TN).

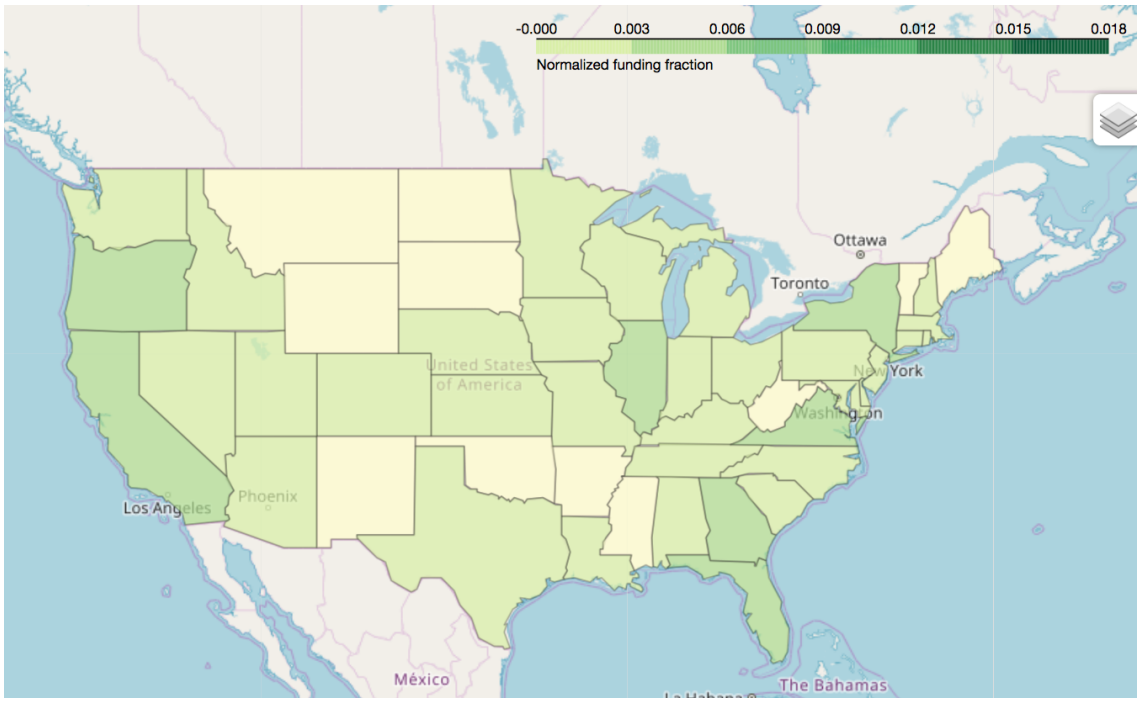


Figure 4(d). Heat map displaying the normalized values for the media and entertainment industry across USA.

Figure 4(d) displays the normalized values for each state throughout the media and entertainment industry across the United States were displayed. Unlike the previous three heat maps, there were no real dominant states in this industry. Less surprisingly, some of the states we may associate with media and entertainment, such as New York (NY) and California (CA) are among the darker shaded states in this heat map.

4.1.4. Discussion

To conclude our findings, we found that the correlation between the location and industry of a company does have an impact on the funding the company receives. Using the normalized values calculated in our analysis, our findings showed a lot more of a relative view in regards to our research question. Our investigation into this subject proved that the correlation between the industry and location most certainly has an impact on the funding received by a company.

We found that in terms of certain industries, such as Technology and Finance it is beneficial to look at setting up a company outside of the usual hubs. The obvious choice of location for certain industries might not always be the best choice in terms of the impact it has on the funding received by the startup. One interesting thing to note is that smaller states that may usually go unnoticed when it comes to choosing a location for a company may actually be thriving states for certain industries, for example, Technology industry in Idaho.

4.2. RQ2

The second research question explores whether a founders education impacts the amount of funding received.

4.2.1. Datasets

For this research question the second dataset was used (see 3.2.2). This dataset describes the details of a founder of startup including their educational information. There is also information on the founded company like the funding total received, the industries it is involved with and the location where it is based. The first key field for this research question detailed the highest degree a founder obtained. The second key field detailed the anonymous unique educational institution identifier that a founder attended. Of course, the funding total field was a very significant feature as it was used as the success metric for a company. There are roughly 26,000 founders in this dataset. There are 4150 unique institutions that are anonymized and 9 different degree types defining the highest degree of a founder.

4.2.2. Approach

A founders education was analysed in terms of two main aspects. The first of these is the highest degree obtained by a founder. The resulting second dataset contains this information. The second aspect related to the institution which a founder attended. We had noticed in our brief initial analysis of our dataset that the majority of founders had attended multiple educational institutions.

To begin with, our initial analysis focused on a founders highest degree by examining them in terms of popularity. There were eight overall degree types in total so this was an important point of analysis to get an insight into whether a certain degree(s) dominated. Founders that had obtained a Bachelor's degree were the most popular, followed by a Masters degree and then there was a steep drop off to a PhD. We expected to see this and discovered that these three degree types represented the highest degree of roughly 95% of founders in our dataset. After some deliberation, it was decided that the presentation of our results for this research question would be done using the Bachelors, Masters and PhD degrees. Another insight into the data showed that a founder which obtained a PhD as their highest degree received the greatest mean funding. Soon after conducting this analysis, we realised that getting the mean of the funding totals of companies was not accurate. There was a huge range difference between the maximum funding total achieved by a startup and the minimum value. Thus, it was decided that splitting the founders into even groups based on the funding they received would improve this.

Another interesting point of analysis that we performed related to the gender of founders in our dataset. We expected to see a rather even gender split of founders initially, perhaps slightly more male founders. A stark fact about the founders in our dataset that we discovered, was that there were seven times as many males as there were females present. In relation to this research question, we investigated whether this skewed gender split continued to be present for founders with different highest degree types. We found that the ratio of males and females was maintained for founders who had achieved a Bachelors or a Masters as their highest degree. There was a slight decrease in the percentage of females obtaining a PhD as their highest degree, however, it was almost negligible difference (roughly 1%).

We chose three specific funding intervals to categorise founders in our dataset. The first group of founders was based on whether they achieved funding between the interval $0 to $1,000,000 inclusive (**Interval One**). The second group of founders was based on whether they achieved funding between the interval $1,000,000 to $10,000,000 inclusive (**Interval Two**). The final group of founders was based on whether they achieved funding greater than $10,000,000 (**Interval Three**). Besides improving the accuracy of our results, these funding intervals allowed us to compare founders in each

of the three evenly split groups. We used these three groups of founders in our analysis and compared the resulting differences across the funding intervals. We will go into more depth of these differences in our results section.

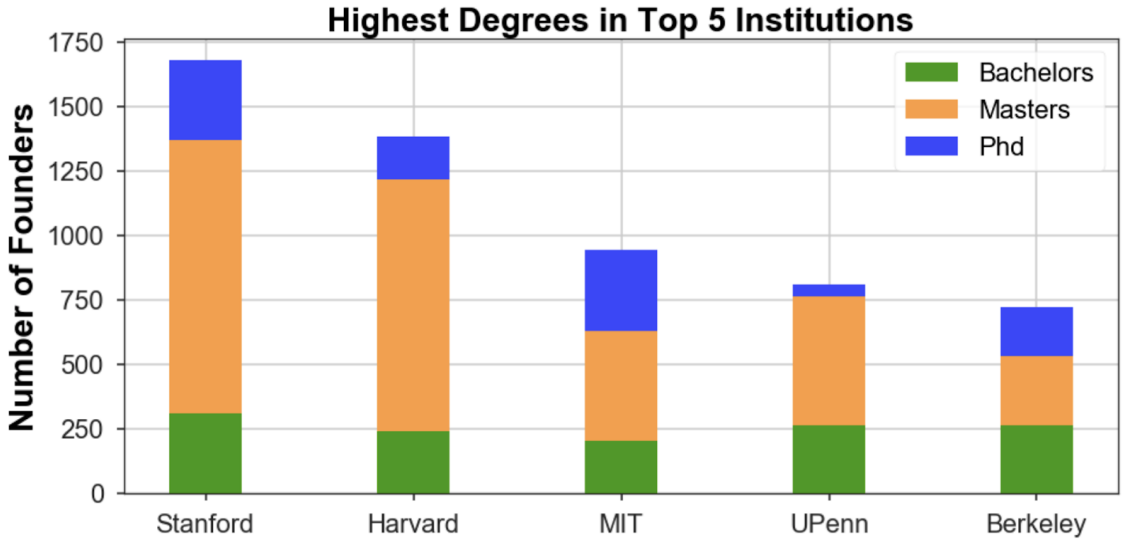


Figure 5 shows the ratio of founders obtaining Bachelors, Masters and PhDs as their highest Degree in the top five institutions.

As mentioned previously, the unique identifiers of the institutions founders attended were anonymized by Crunchbase. Our first point of action was to deanonymize the ten most popular occurrences. We discuss this choice further in our Ethical Considerations. The process of deanonymizing these institutions was very simple. It involved finding a founder who had attended the institution and visiting their Crunchbase profile which is freely available online. We then replaced the unique institution identifier with its actual name. Our initial analysis of these institutions was based on comparing the ratio of Bachelors, Masters and PhD for the five most popular institutions in our dataset. These were Stanford, Harvard, MIT, University of Pennsylvania and University of California, Berkeley. *Figure 5* gives an insight into the ratio of founders obtaining Bachelors, Masters and PhDs as their highest Degree in the top five institutions. We then proceeded to investigate these institutions attended by founders using the funding intervals discussed formerly. The first comparison we explored related to the number of unique institutions present in each of the datasets. We found it very compelling that there were almost nine-hundred more unique institutions being attended by founders who had achieved funding within the first interval compared to founders who had achieved funding between the last interval.

In conclusion to our approach, we decided to investigate a founders education in terms of two educational aspects, highest degree obtained and institution attended. The use of funding intervals was also very important in order to analyze and also visualize any potential trends within the data.

4.2.3. Results

To begin with we will look at our findings of the highest degree obtained by a founder. As previously mentioned, the funding intervals (see section 4.2.2) that we used to separate founders were very effective at showing trends when comparing them. The visualisation shown below demonstrates that the Highest Degree obtained by a founder does impact the amount of funding received. This is shown by the negative trend of the Bachelor Degree as the funding intervals increases. This contrasts hugely to the trend of the higher degrees, Masters and PhD. We did not expect the large percentage of

founders with a Bachelors as their Highest Degree. We grouped the other degree types into ‘Other’ as they accumulated to a very small percentage overall .

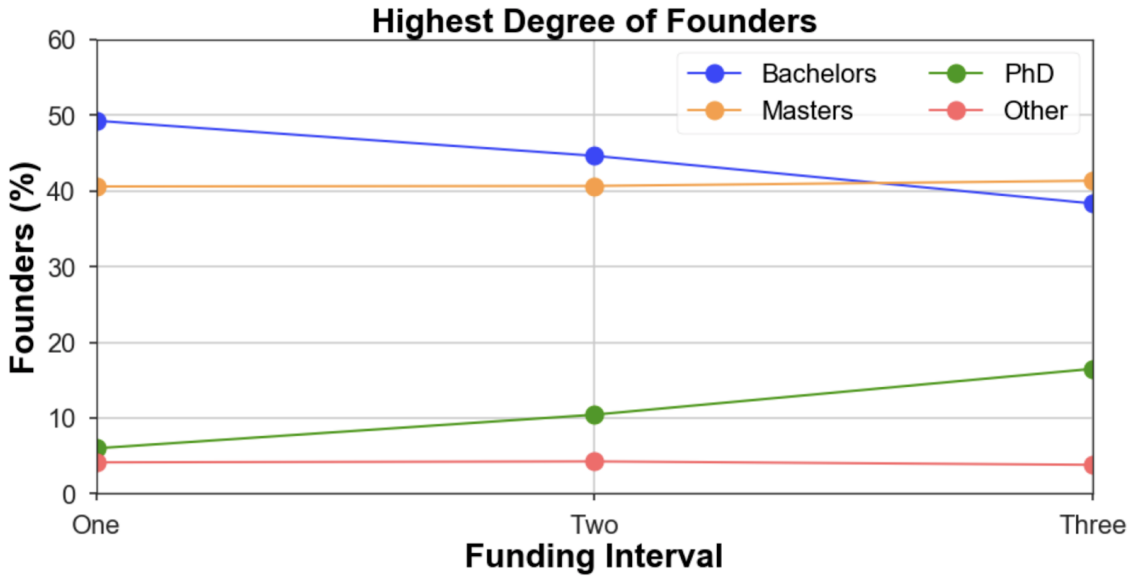


Figure 6. A graph showing the percentage of founders with a certain highest degree across the three funding intervals

We did consider using a stacked bar chart for *Figure 6* however it did not show the trend as effectively in our opinion. A trend that we did expect to see across the funding intervals was the increase in PhDs as the funding amount increased. This was confirmed by our findings shown in *Figure 6*. We concluded that a reason for the low percentage of PhDs could be due to the long period of time it takes for them to be achieved. A Masters degree is seen to have a consistent positive trend across the funding intervals and succeeded in overtaking a Bachelor's degree for the greatest funding interval amount. Thus it can be concluded that it is a beneficial attribute to have when looking for funding greater than ten million USD.

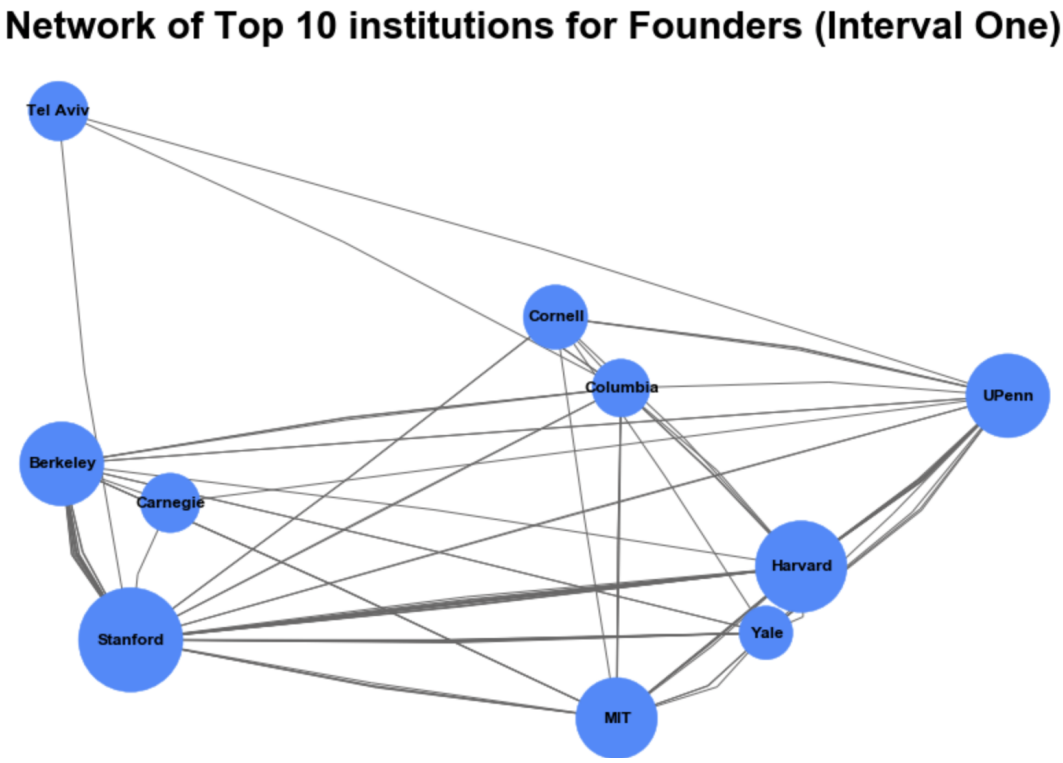


Figure 7(a). A network graph showing the connections between the institutions founders who have achieved funding within the first interval.

We will continue to use the funding intervals when comparing the institutions that founders attended. We had noticed in our analysis that majority of founders had attended multiple institutions. We believed that visualising the connections between these institutions would be most effective using a network graph. A node in our network graphs below represents an education institution attended by a founder. The node size is dependent on the number of founders attending these institutions, and an edge represents a founder who has attended multiple institutions e.g Bachelor's in Harvard, Master's in Stanford. As alluded to before, we used the top ten institutions that we deanonymized when visualising these networks. Observing *Figure 7(a)*, the educational institution nodes are small, there are not a lot of distinct paths besides those from the dominant universities like Harvard University and Stanford University. An interesting point relating to this funding interval is it has the most unique number of institutions in comparison to the other two intervals.

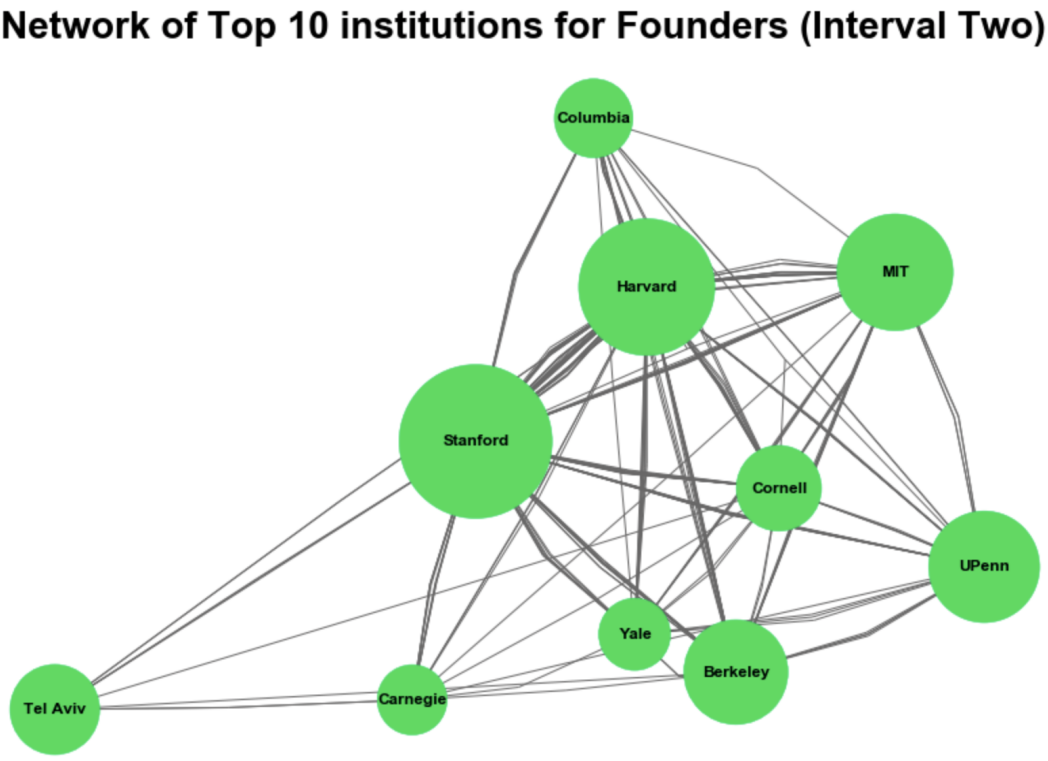


Figure 7(b). A network graph showing the connections between the institutions founders who have achieved funding within the second interval.

If we compare *Figure 7(a)* to *Figure 7(b)*, we can instantly see that there are a lot more connections between the nodes in the latter. Node sizes have also increased, specifically the top five institutions - Stanford, Harvard, MIT, UPenn and Berkeley. The number of distinct paths between these top institutions has grown. Tel Aviv is an Israeli educational institution which has interestingly appeared in the top institutions founders attended. The connections this university has between the likes of Harvard and Stanford are very intriguing. Carnegie Mellon University is also a very intriguing inclusion in the top universities as it is a private research institution.

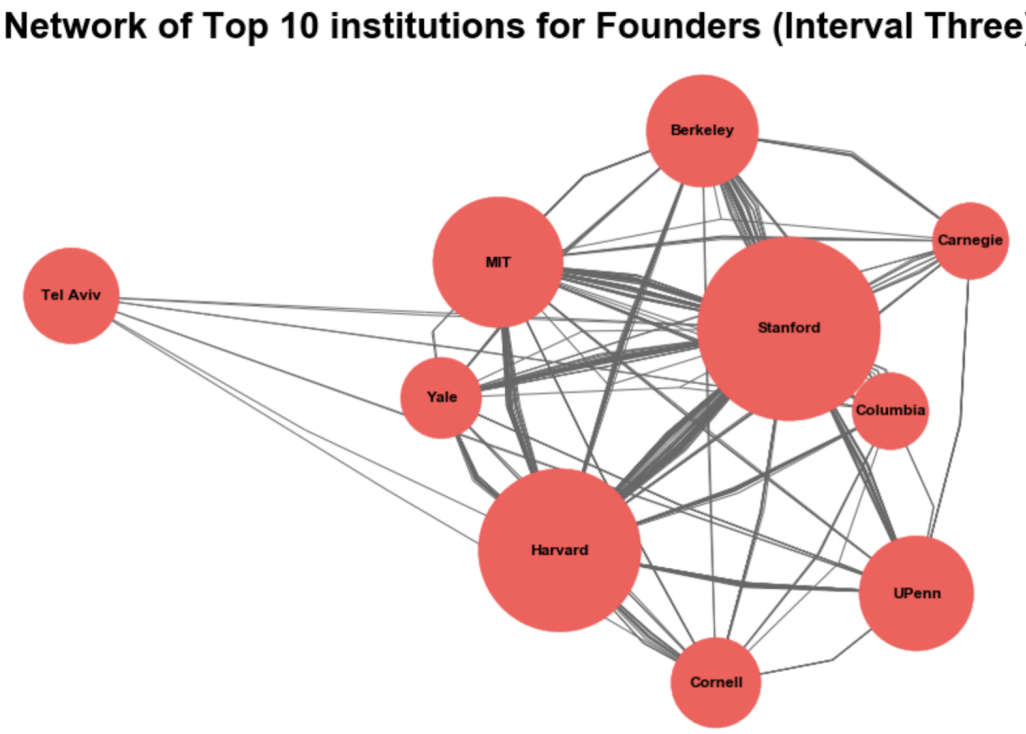


Figure 7(c). A network graph showing the connections between the institutions founders who have achieved funding within the third interval.

Finally, comparing *Figure 7(c)* with *Figure 7(a)* and *Figure 7(b)*, shows the greatest growth of node size. The number of distinct paths between these top universities has also increased drastically. Stanford and Harvard dominate the visualisations. This indicates that founders which are alumni from these top institutions receive a greater amount of funding. These visualisations show us that as the interval of funding received increases, the number of founders that attended these top institutions grows. This suggests that having attended one of these top institutions could be an attractive attribute to have when searching for funding from investors. Using numbers to emphasise this point, only 1271 companies of the first funding interval had a founder that attended one of the top ten institutions. This contrasts hugely to the 3800+ companies of the third funding interval that had a founder that attended these top institutions.

4.2.4. Discussion

To conclude our findings, we found that a founders education does significantly impact the amount of funding a startup receives.

Firstly, our investigation into the educational aspect of the highest degree of a founder proved that there was a difference in funding being received based on which degree they had obtained. Even though there was a drastic decrease in the percentage of founders having a highest degree of a Bachelor's as the funding interval grew, we concluded that it was still an effective degree to have in terms of time spent achieving it. We would have liked to have done more analysis in relation to the number of years a founder spends achieving a degree. This could potentially be added to further analysis on this topic in the future. We found it fascinating to see that there was a drastic increase of founders with a highest degree of a PhD over the funding intervals. This is probably due to it being a very highly regarded degree. It takes a long period of time to achieve a PhD and if the degree area is

relevant to the industry area which the startup is involved with, then investors might consider investing more.

Secondly, in terms of the institutions attended by founders, we found there was a correlation between attending the top universities and the amount of funding achieved. We determined the top universities as those most frequently attended by founders in the whole dataset. This indicates that the institution a founder attends has an impact on the funding gained. It was also very fascinating to see that the number of unique institutions decreased as the funding total increased. However, there was still a significant number of different institutions that a founder of the third funding interval attended.

4.3. RQ3

The third research question explores whether founders with previous experience command more funding than founders with no previous experience.

4.3.1. Datasets

For this research question, the second dataset was used (see 3.2.2). This dataset describes the details of a founder of startup. There is information present in the dataset describing the funding total received by their startup, the industries which the startup is involved with and the location where it is based. As previously stated, there are multiple instances of a founders which must be taken into account when grouping and analysing the data. It is very important that no founders are double counted in this research question. The unique person and organization identifiers mean that there are no complications with startups that have the same company name or a person with the same name. Of course, the funding total field was a key feature of this dataset as it was used as the success metric. There are ~26,000 founders in this dataset. As well 4,150 unique institutions that are anonymized, and 9 different degree types defining the highest degree of a founder.

4.3.2. Approach

The first point of action when approaching this research question was to define the difference between an experienced and non-experienced founder. An experienced founder was characterized as someone that founded more than one company. Therefore, someone who has only founded one company would be characterized as non-experienced. This allowed us to begin our analysis on these two groups of founders. The first approach to this research question was to get an insight into the basic differences between experienced and non-experienced founders. Gender analysis of experienced founders showed that males dominated both experienced and non-experienced founders. This was not surprising to see giving the dominance of males in our dataset.

Secondly, we explored the average funding received by experienced versus non-experienced founders. Following on from this we decided that we would identify whether there are differences of an experienced founders first startup compared to a non-experienced founders only startup. We were interested to see whether experienced founders potentially received more average funding for their first startup compared to that of non-experienced founders.

4.3.3. Results and Discussion

We soon realized from our analysis that the sample size of the experienced founders in our dataset was roughly ten times smaller than the data of the non-experienced founders. The reason why this impacts the accuracy of this analysis is due to the large range in funding values that we have in our dataset. Some founders achieve funding of 1000 USD whereas others achieve well over 10 Million USD in funding. Getting the average of the large range in funding values does not give a fair representation of

the data as a whole. We combatted this previously using funding intervals (see section 4.2.2) to categorize founders into separate groups for analysis.

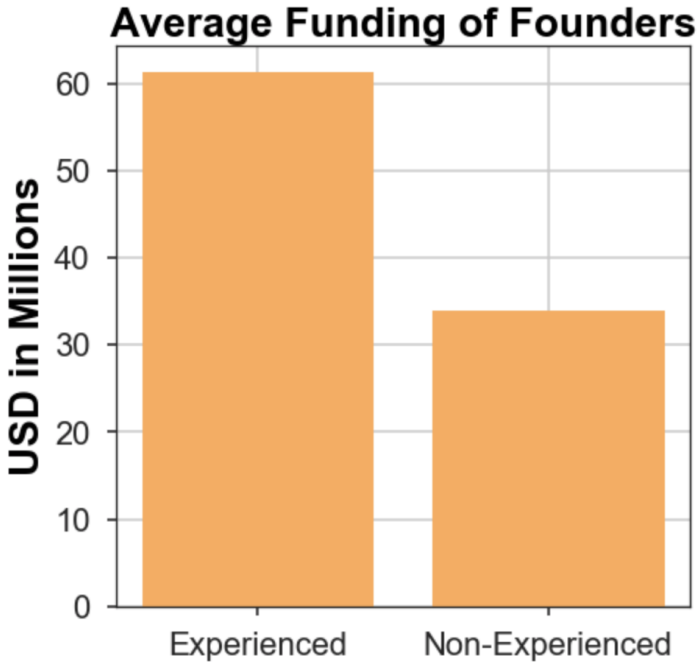


Figure 8 Shows the average funding received by experienced and non-experienced founders

Our result from our exploration of the average funding received by experienced versus non-experienced founders is shown by *Figure 8*. The number of experienced founders in our data was ~2,600 compared to ~24,000 non-experienced founders. This could be a contributing factor as to why there is such a large difference in funding shown in the visualisation. We attempted to use the funding intervals to investigate this difference further. Our results from this analysis suggest that the main difference between average funding received by experienced and non-experienced founders occurs within the third funding interval (shown in Appendix *Fig. M*)

A limitation that we discovered while doing this research question was not being able to classify experienced founders startups as ‘first startup’, ‘second startup’ etc. The reason why we could not do this is due of the fact a founder may have had a startup removed from the dataset if it had no funding value. Classifying the founders startups based on the order they were founded, proved to be impossible due to potential missing values. Combining this reason with the fact there were no founding dates of companies in our dataset, we were very restricted in our pursuit of exploring experienced and non-experienced founders in-depth. We decided to deem this question inconclusive based on these reasons. While discovering these limitations, we have also identified potential adjustments that would allow for future work. For this particular research question, we focused solely on the previous experience of a founder in regards to the previous companies they have founded, which limited our sample size, however, total job experience of a founder could be taken into consideration instead for future work. Founders that have been involved with companies previously (e.g CEO, Manager etc) were present in the data Crunchbase provided, this was mentioned in our Data Acquisition section (Section 3.1). This data could be used to form a metric or score of previous experience of a person.

5. Discussion

This is the penultimate section of our report, in this section we will take the opportunity to discuss various aspects of our project that we feel are hugely important for any data science project, as well as discussing the limitations we came across while undertaking this project.

5.1. Ethical Considerations

At the beginning of this project we had very little knowledge of ethics within data science, and initially didn't really take them into account. As the duration of the project progressed we placed more and more emphasis on the ethical aspect of our project, especially as we learned more about the need for stricter regulation and ethics throughout the industry. The recent Cambridge Analytica data scandal has resulted in the topic of data ethics being highlighted significantly, and has ultimately called for a major change in the emphasis put on the ethical aspect of data throughout the industry. With the GDPR (General Data Protection Regulation) being imposed from the 25th of May 2018, this could have a hugely positive effect for data ethics throughout the industry as it is causing companies to rethink how they deal with their huge amounts of data on a daily basis. Due to the very public nature of the data we are using for our project we fortunately didn't have a large number of ethical issues to deal with, although we did have some concerns that we had to consider.

5.1.1. Obtention of Our Data:

The conditions we obtained the data under were as follows: The data may not be shared, or made publically available. We can also not use the data for commercial purposes without further discussion with Crunchbase.

This is one of the main ethical concerns we had throughout our project, and one we aim to withhold moving forward. It is very important that we neither release any of the data made available to us, nor use it for commercial purposes. Not only could legal action ensue in the case of us breaking either of the two conditions mentioned above, but it would also be irresponsible as data scientists to break the trust of a data provider.

5.1.2. Institution IDs:

The next major ethical concern that we were faced with was whether or not it was ethical to deanonymize the institution ID of a group of founders in our dataset. When we received the data from Crunchbase, the institution ID of each of the founders throughout the data was anonymized. This suggested to us that this may be seen as quite sensitive information, and it could be a issue we take into consideration in terms of ethics. In the case of a general person in society, the release of their institution ID might be seen as a very sensitive and invasive piece of information to reveal. Although because of the nature of the people throughout our dataset we reconsidered whether or not it would be unethical to reveal their institution IDs.

The institutions that we wanted to deanonymize the institution IDs were the top ten institutions in our dataset, the people in our dataset are all founders of public companies throughout the world. As we are only deanonymizing the institution IDs for the top ten institutions, we deemed it as ethical to release the names of these top ten institutions. The main reason behind this decision is because it is more likely that the more prominent and well known founders in our dataset attended one of these top ten institutions. Regardless of this fact, because of the public nature of these companies the founders information is freely available across the web, on Crunchbase, LinkedIn, as well as other large

websites. We feel we have provided enough justification to deem the deanonymization of the institution IDs as being ethical.

5.1.3. Founders Names:

The final ethical concern that we had to take into account was whether or not it was ethical to leave the founders names deanonymized throughout the data. Unlike the institution IDs, when we received the data from Crunchbase, the founders names throughout the data were not anonymized. We were quite confused by this decision from Crunchbase. We believe it is a lot more unethical to release a person's name when compared with the release of the institution they attended. For example, if it were the other way around and only the institution ID was deanonymized it would be impossible to make the link back to a founder strictly from the institution ID.

Although, regardless of the above we ultimately found it ethical to leave the founders name deanonymized throughout the data. A similar argument stands from the reasons for deanonymizing the institution IDs above, because of the public nature of these companies the founders information is freely available across the web, on Crunchbase, LinkedIn, as well as other large websites. For example if we only provided the company names, it would be simple for someone to make the link to a founder. Some of the larger founders in our data appear in the news on a frequent basis, the vast majority of company founders are very much so in the public eye. We feel we have provided adequate justification to allow for the retention of deanonymized founder names throughout our data.

5.2. Reproducibility

Reproducibility is not only hugely important across data science, but across every aspect of science in general. Once again, similar to our ethical considerations, initially we had very little knowledge or understanding for how important of an aspect reproducibility is for any data science project. As the duration of the project progressed, we slowly gained a better understanding of the importance of reproducibility and the techniques we should incorporate to ensure our project is a reproducible as possible. Reproducibility is the ability of an entire experiment or study to be replicated, either by the same data scientist or by someone else working independently. Reproducibility gives others the opportunity to verify that the data and methods used are accurate. It also allows others to build upon our work, it essentially allows for people to focus on the content of the data analysis, rather than a lengthy report.

For our project in particular there were a number of steps we undertook to ensure full reproducibility. To begin, we documented each step of the project in the markdown cells of each of the notebooks, making it as straightforward as possible to understand each and every decision we made throughout our analysis. This will also enable future researchers to understand what is happening in each of our notebooks. We used Github for version control, which allowed us to be constantly in sync with all the key changes being made throughout the duration of our project. We also provided a requirements file detailing python packages necessary. As alluded to previously, we obtained the datasets used under a number of conditions, one of which being that we don’t release the data in any situation. This is obviously a major block in terms of reproducing this project. The data can be purchased from the Crunchbase website which opens up reproducibility to researchers willing to purchase the data.

5.3. Limitations

We had four major limitations throughout this project, the first being the trouble we encountered while trying to obtain our data, as well as unforeseeable time constraints, and finally a lack of knowledge in the field of analysis, which also ties in with statistical analysis. We will discuss each of these limitations below:

Cannot Release Datasets:

Due to the conditions we agreed with Crunchbase on obtention of the datasets, it is not possible for us to release our datasets to anyone. This can be seen as a major limitation in terms of reproducibility and potential future work. Although this can be seen as a limitation, it is quite straightforward to overcome.

Not being able to release the datasets is obviously a major block in terms of reproducing the project, and for any possible future work on this project. Although to overcome this limitation, the data can be purchased from the Crunchbase website which opens up reproducibility to researchers willing to purchase the data.

Lack of Knowledge:

Having no experience undertaking a technical project has played as a minor limitation moving forward through our project. It is a huge feat to have completed such a comprehensive project in such a short period of time, with such little experience.

Unforeseeable Time Constraints:

Time is something that needs to be taken into account, and managed throughout the duration of a project. Unfortunately, as mentioned in the 'Obtaining our data' section above, a duration of one week was lost from the outset of this project. Another aspect that placed quite a bit of constraint on the timing of this project was the unexpected length of time taken to prepare and clean the data. Due to the time constraints mentioned previously, the depth of analysis that occurred was limited to a slight extent. If the time limitations mentioned did not occur, given more time we would have had the opportunity to undertake some of the analysis from our future work section below (see section 6).

6. Conclusions & Future Work

Although having a successful company has its many benefits, mainly in terms of monetary gain, as well as self fulfilment, a huge amount of new companies fail. We have shown throughout or analysis that there are many factors that have an affect on the funding of a company. However the results weren't always as we anticipated. We have shown that the 'obvious' choice of location for a company might not always be the best choice in terms of the funding the company aims to receive. On top of this we established that the highest degree obtained and institution attended by a founder are pivotal factors for a successful startup.

There is a huge amount of future work that can be undertaken in continuation of our analysis. Predictive analysis into how to build a successful startup would be fascinating. It would be interesting to be able to find the best location for an each industry, the best institution to hire from for every industry, as well as the best venture capital fund to approach for funding. Another very topical and captivating piece of future analysis is a more in depth analysis into how the gender of a founder

effects a startup. In our piece of gender analysis we found that there a lot more male founders, but does this mean that men have a higher chance of startup success? Finally, potential future work could be done in terms of total experience of a founder that is not solely based on founding experience. Founders that have been involved with companies previously (e.g CEO, Manager etc) were present in the data Crunchbase data provided as we mentioned in our Data Acquisition section. This data could be used to form a metric or score of previous experience of a founder.

7. Bibliography

Ratzinger, D., Amess, K., Greenman, A., & Mosey, S. (2017). The impact of digital start-up founders' higher education on reaching equity investment milestones. *The Journal of Technology Transfer*, 760–778.

Dencker, J. C., & Gruber, M. (2015). The effects of opportunities and founder experience on new firm performance. *Strategic Management Journal*, *36*(7), 1035-1052.

Patel, N. (2015). 90% Of Startups Fail: Here's What You Need To Know About The 10%. *Forbes.[online] Available at: http://www. forbes. com/sites/neilpatel/2015/01/16/90-of-startups-will-fail-heres-what-you-need-to-know-about-the-10/print/[Accessed 11 Oct. 2015]*.

CB Insights Research. (2018). *Venture Capital Funding Report 2017*. [online] Available at: https://www.cbinsights.com/research/report/venture-capital-q4-2017/ [Accessed 11 May 2018].

8. Appendix

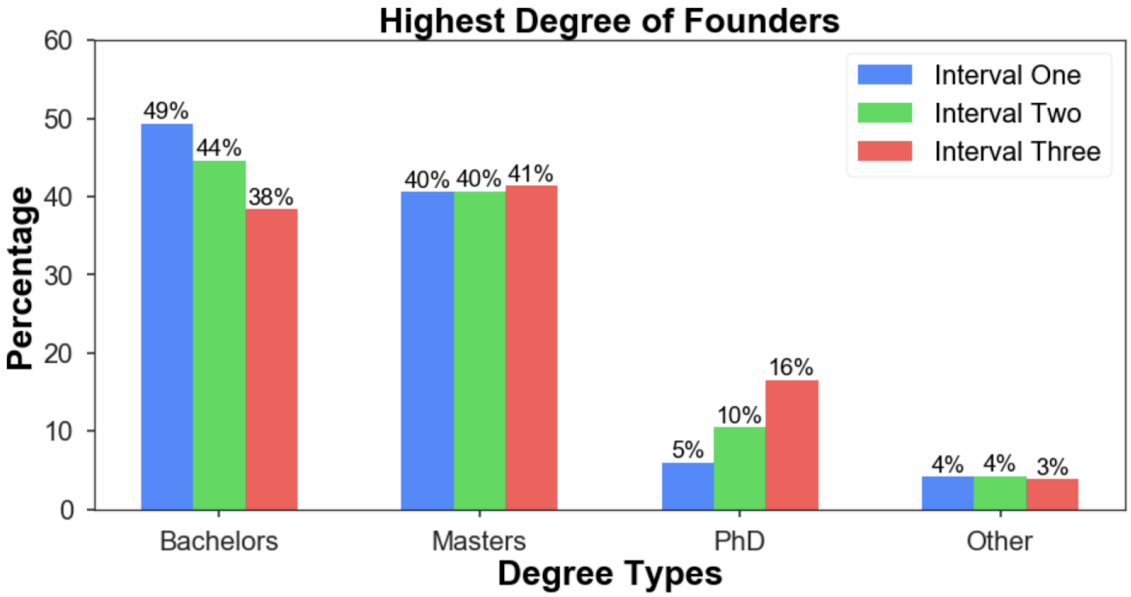


Fig. A The figure above was originally how we represented the Highest Degrees of founders before deciding to use Figure X.

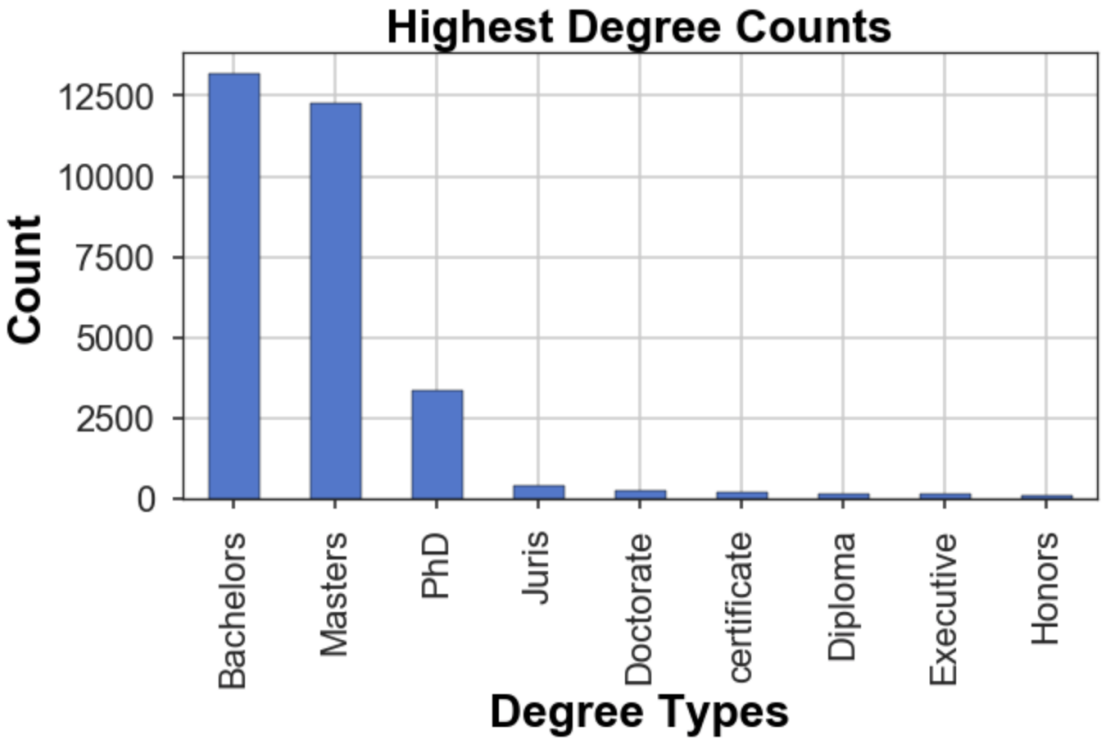
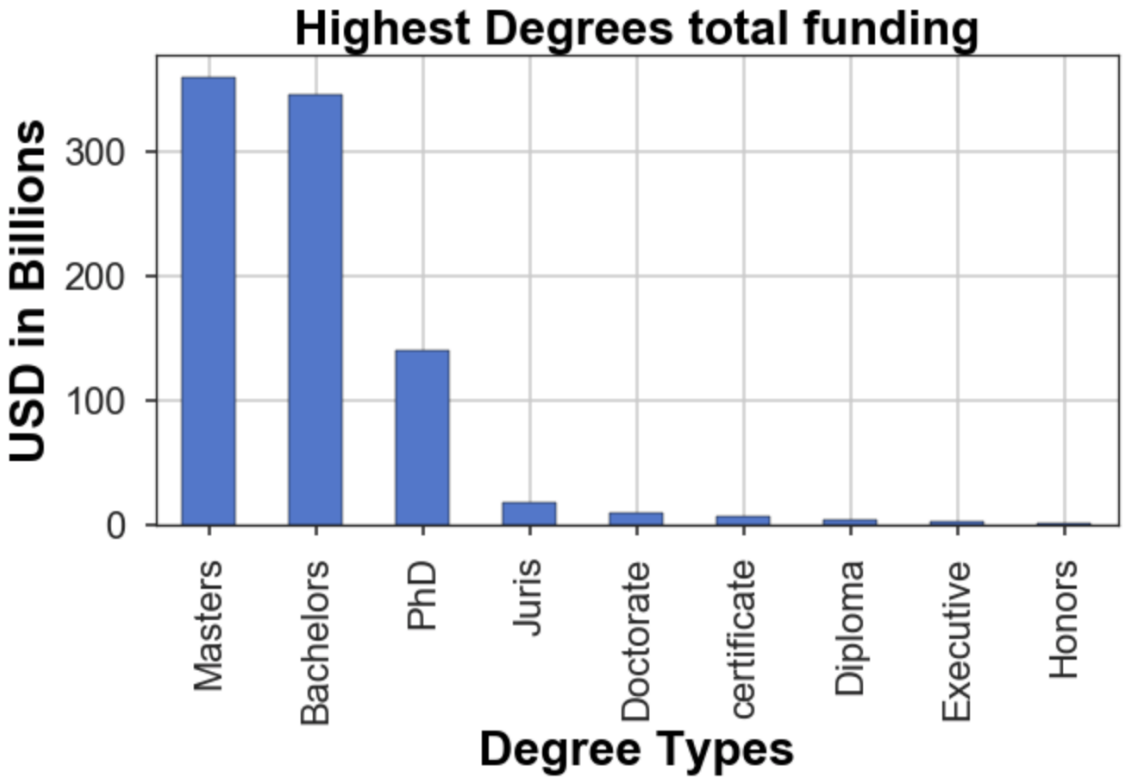


Fig. B. above shows the most popular degree types that a founder had obtained as their highest.



Fig, C. above shows the cumulative funding for each degree type (in terms of highest degree obtained by founders)

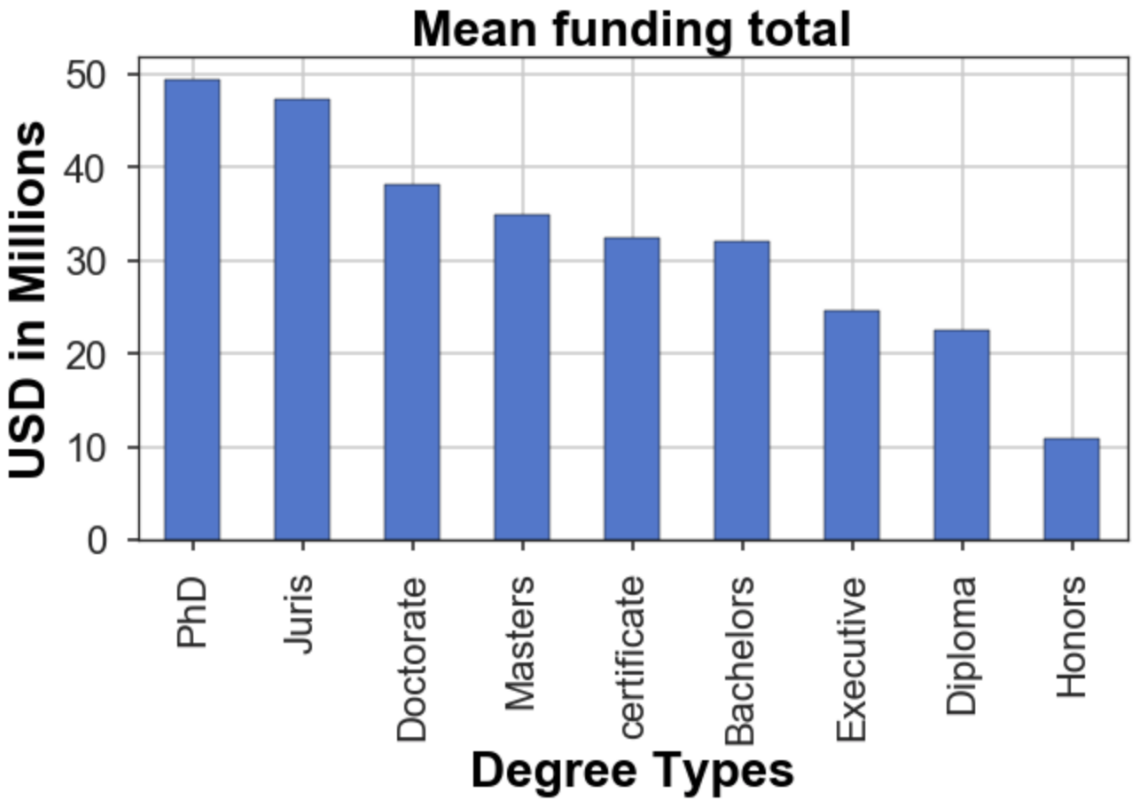
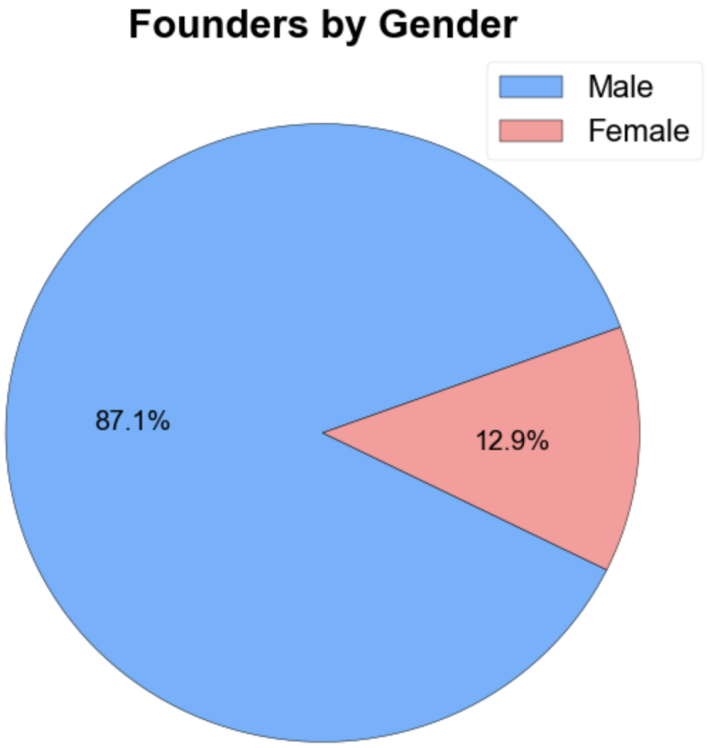


Fig. D. above shows the mean funding for each degree type (in terms of highest degree obtained by founders)



Fig, E above shows the ratio of founders in our whole dataset

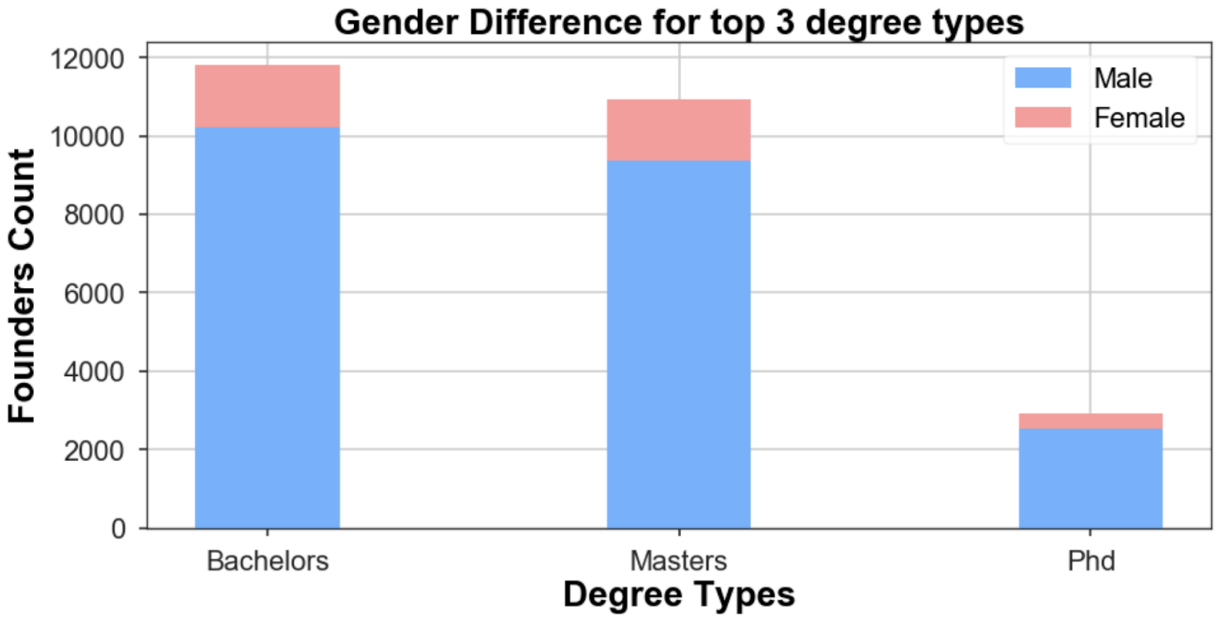


Fig. F above shows the gender ratios for the top 3 degree types (in terms of highest degree obtained by founders)

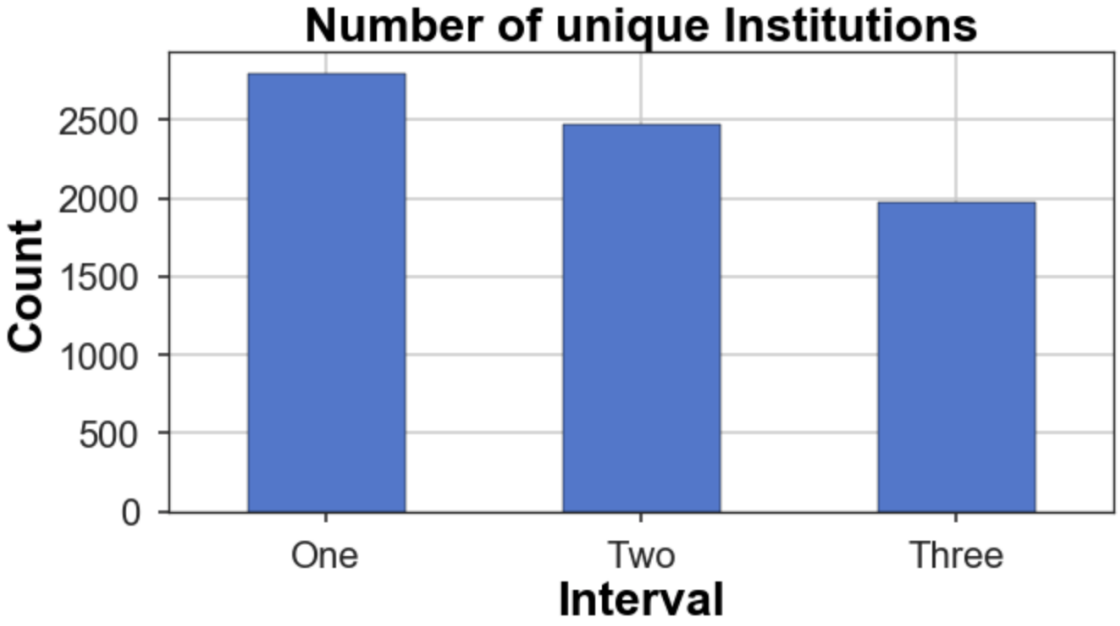


Fig. G. above shows the decrease of unique institutions as the funding interval increases.

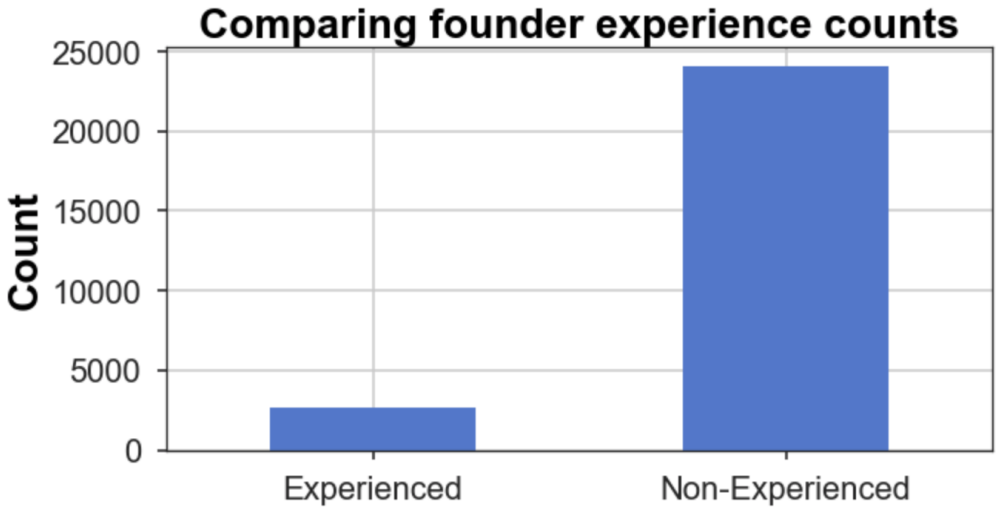


Fig. H. above shows the counts of experienced vs non-experienced in our data.

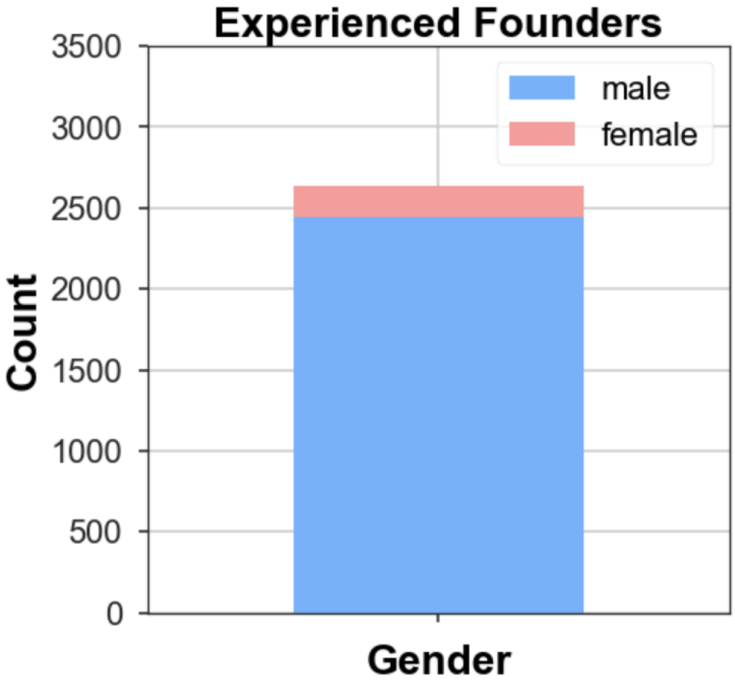


Fig. I. above shows the gender ratio of experienced founders

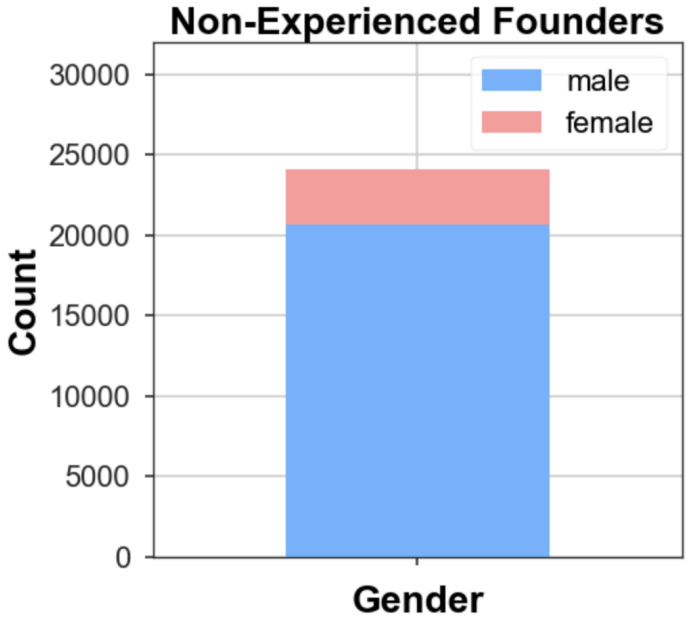


Fig. J. above shows the gender ratio of non-experienced founders.

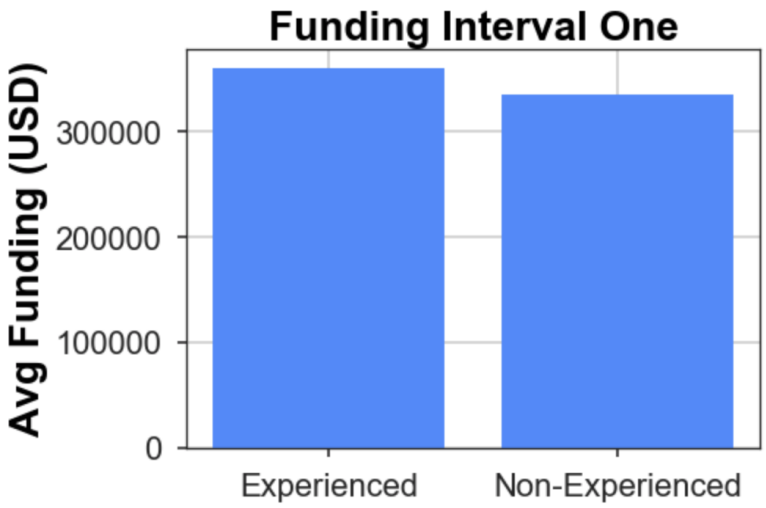


Fig. K. above shows the average funding of experienced and non-experienced founders of the first interval

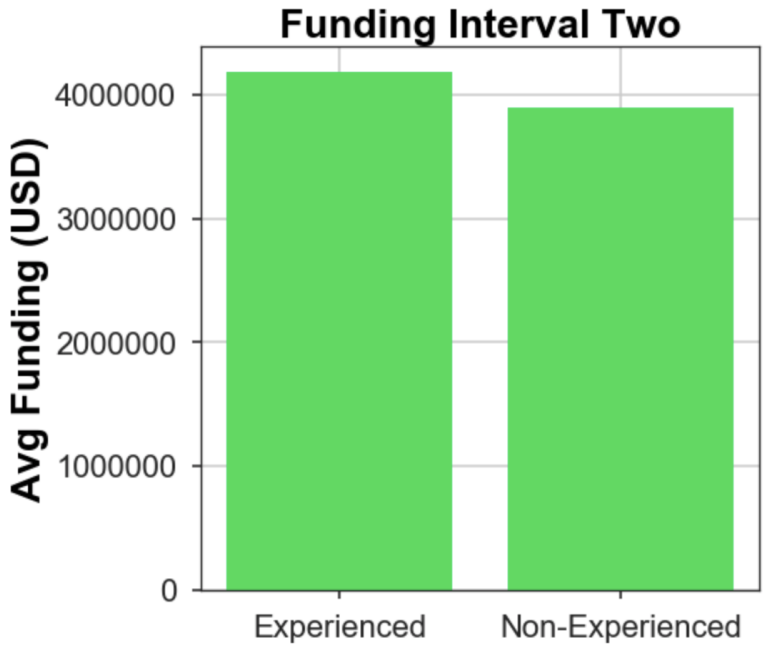


Fig. L. above shows the average funding of experienced and non-experienced founders of the second interval

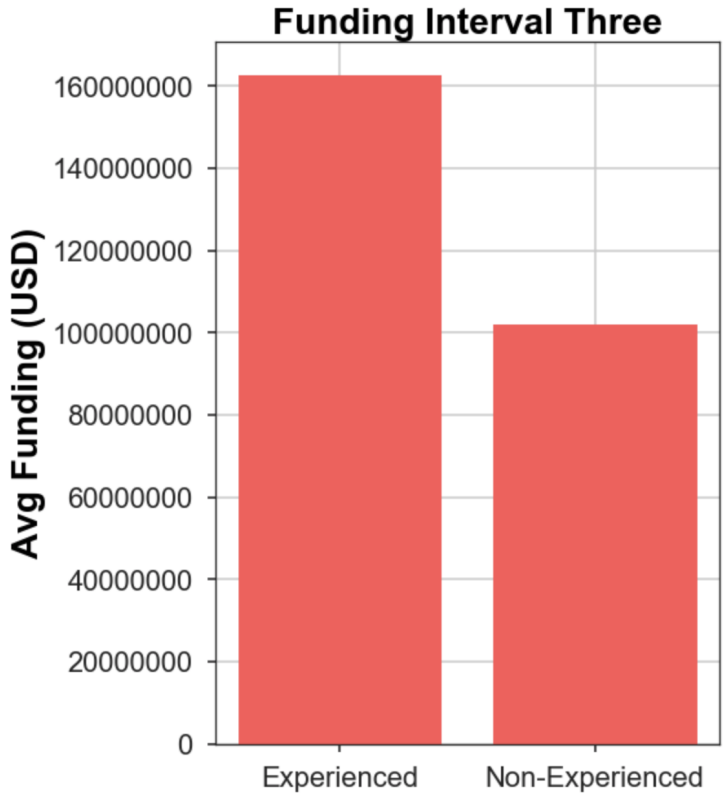


Fig. M. above shows the average funding of experienced and non-experienced founders of the third interval

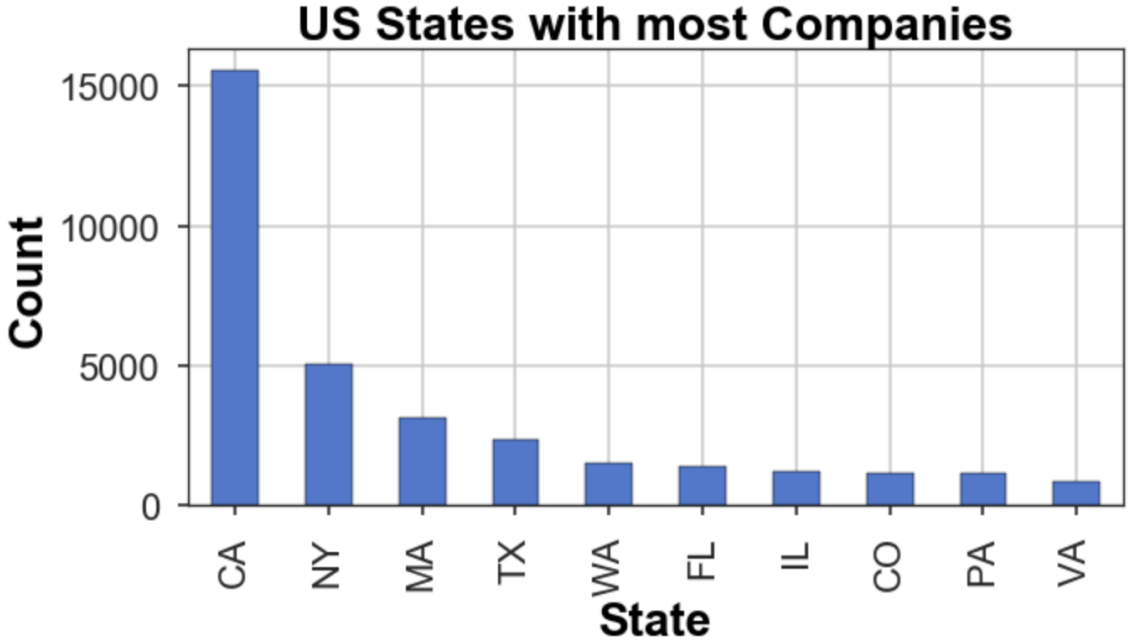


Fig. N. Shows the top 10 states with the most amount of companies in USA.

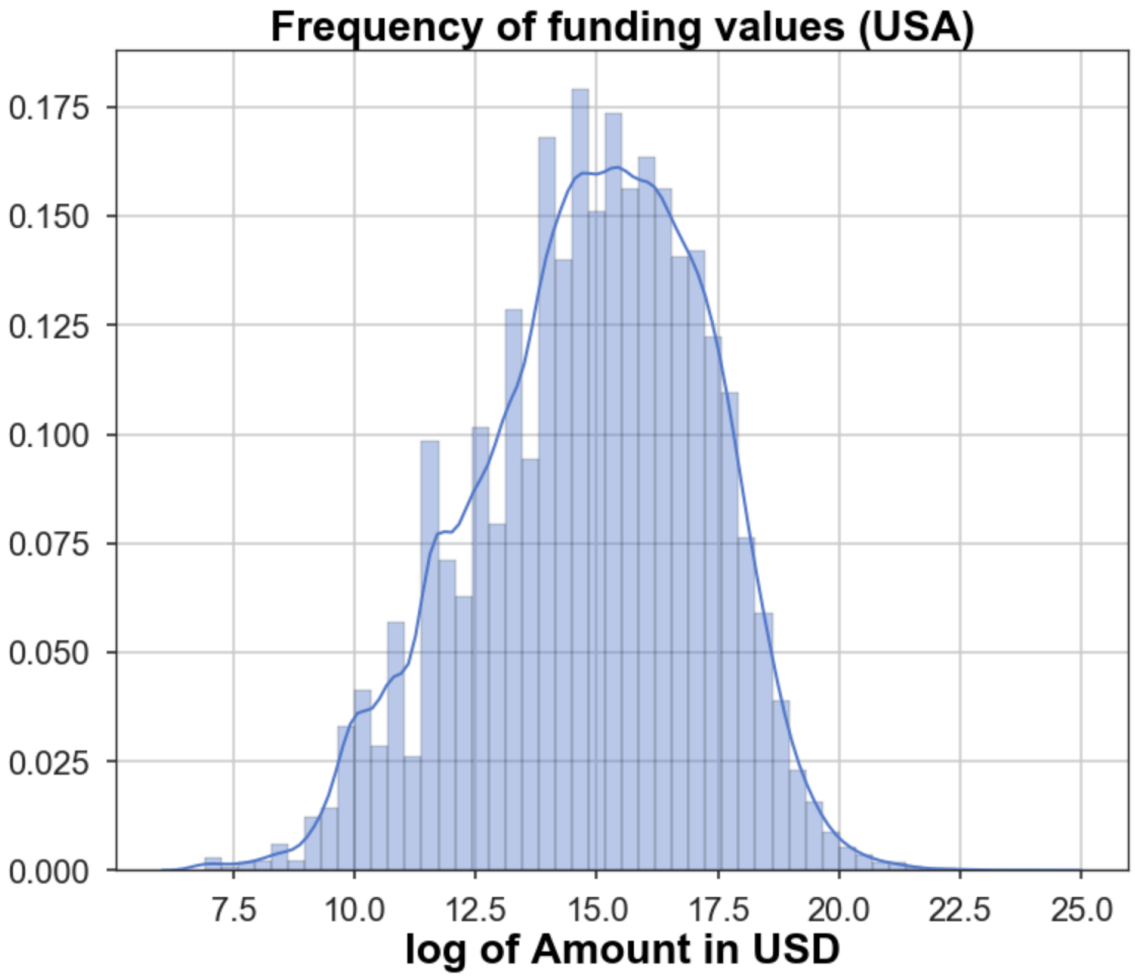


Fig. O. Shows the distribution of funding in terms of the number of companies achieving each funding value displayed in USA.

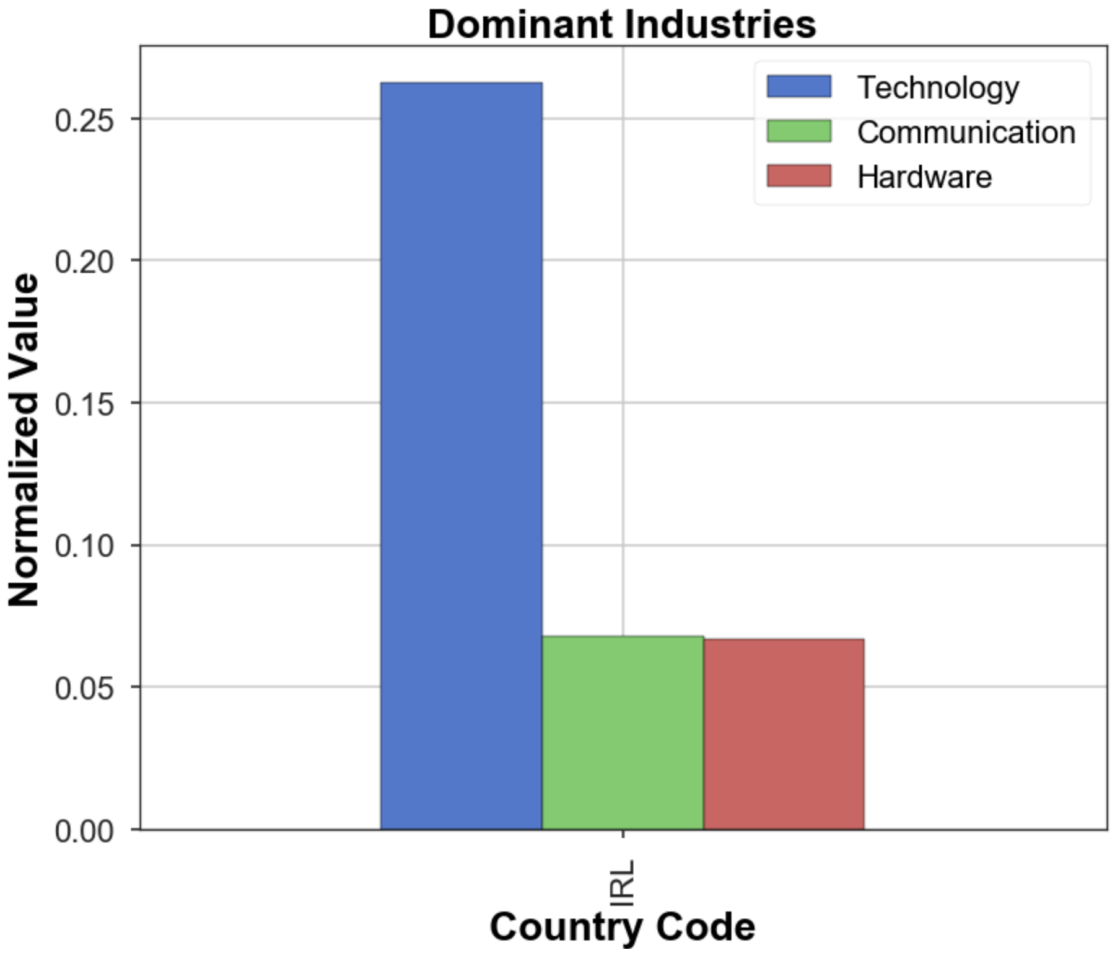


Fig. P. Shows the top 3 industries in terms of the amount of companies in each industry across IRL.

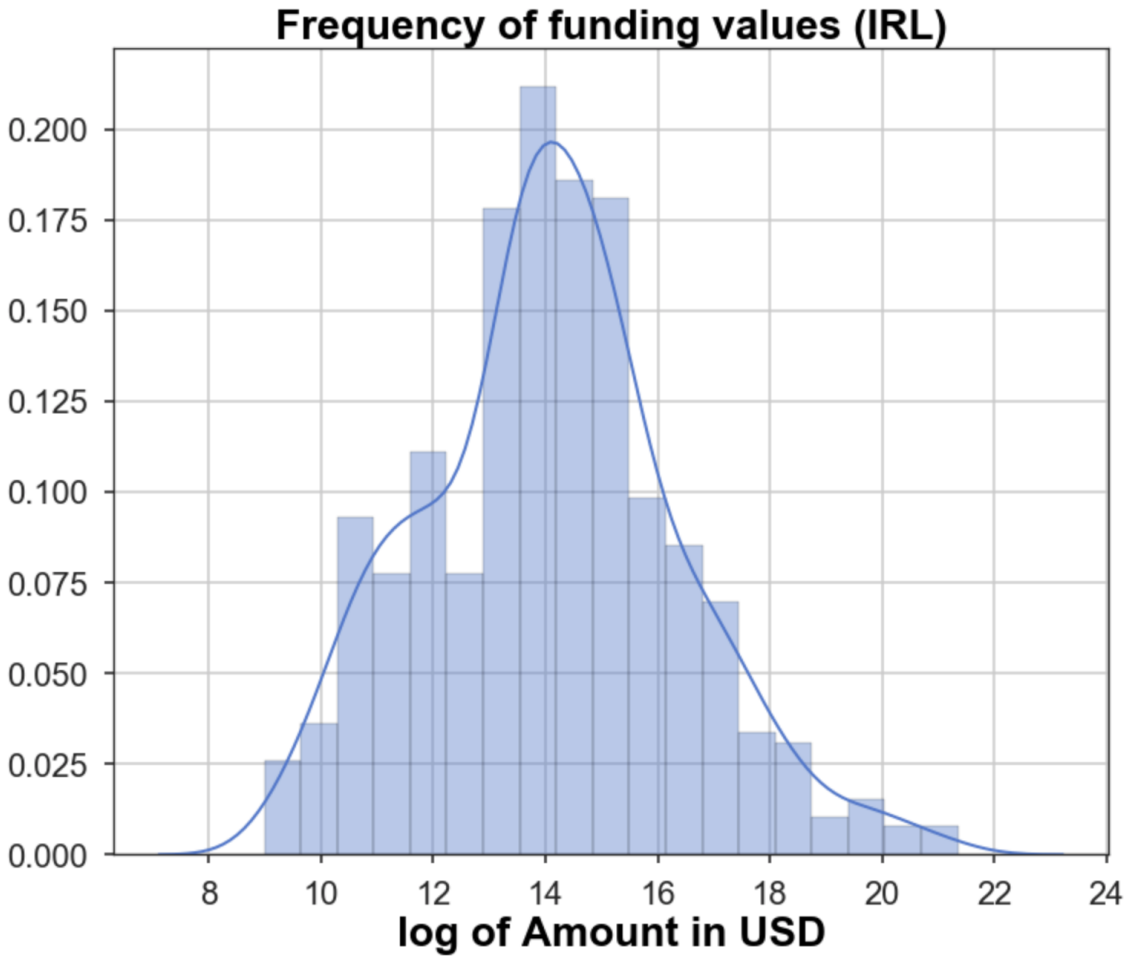


Fig.Q . Shows the distribution of funding in terms of the number of companies achieving each funding value displayed in Ireland.

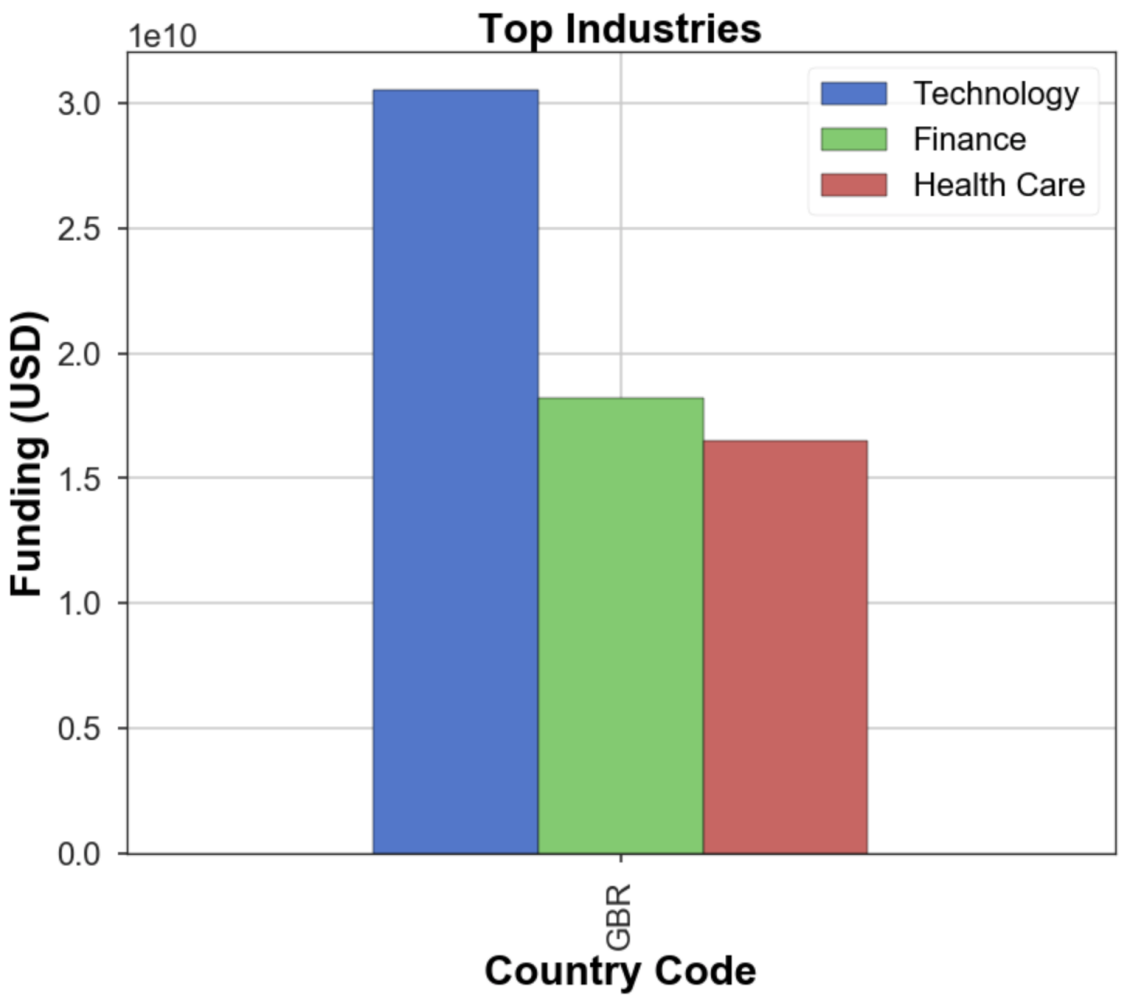


Fig. R. This bar chart shows the top 3 industries with the most total funding across Great Britain.

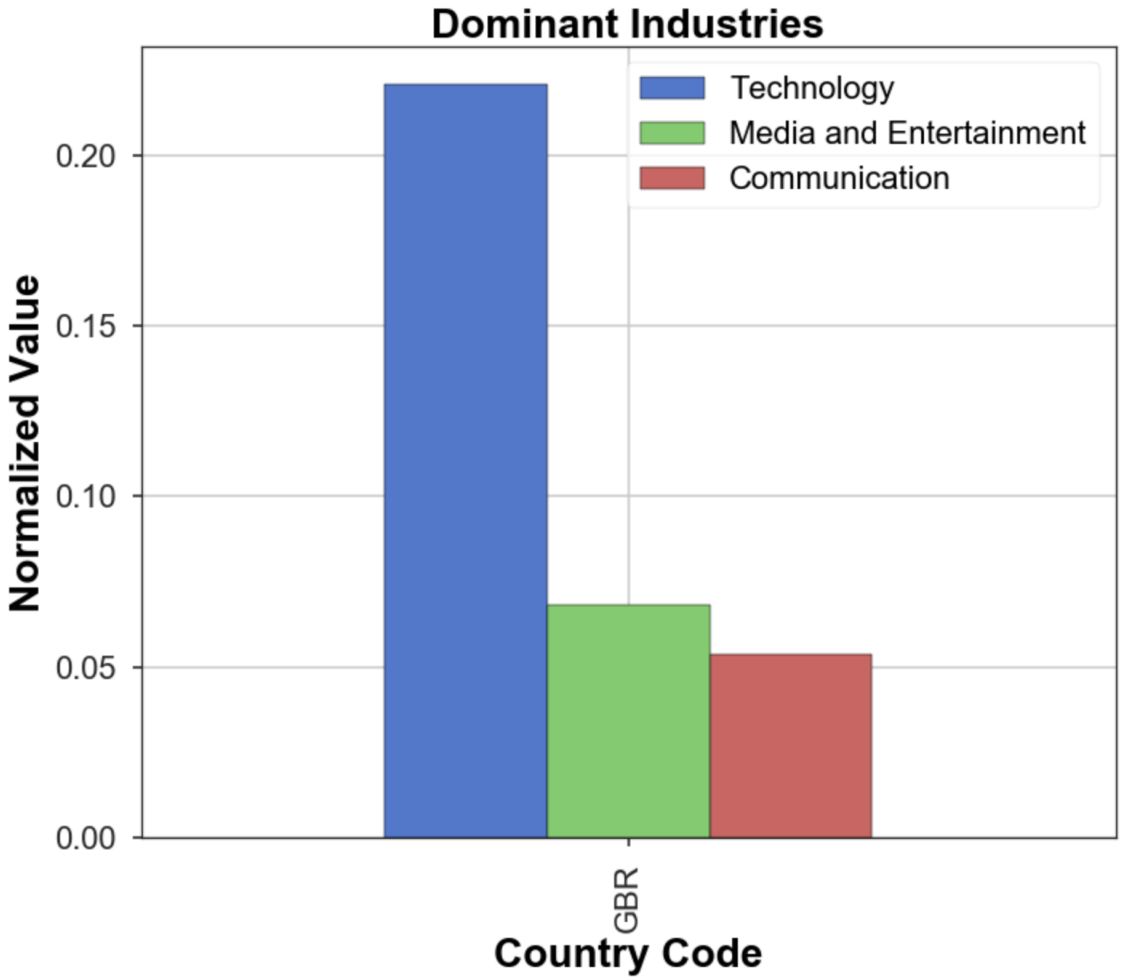


Fig. S. Shows the top 3 industries in Great Britain (GBR) based on the amount of companies in each industries.

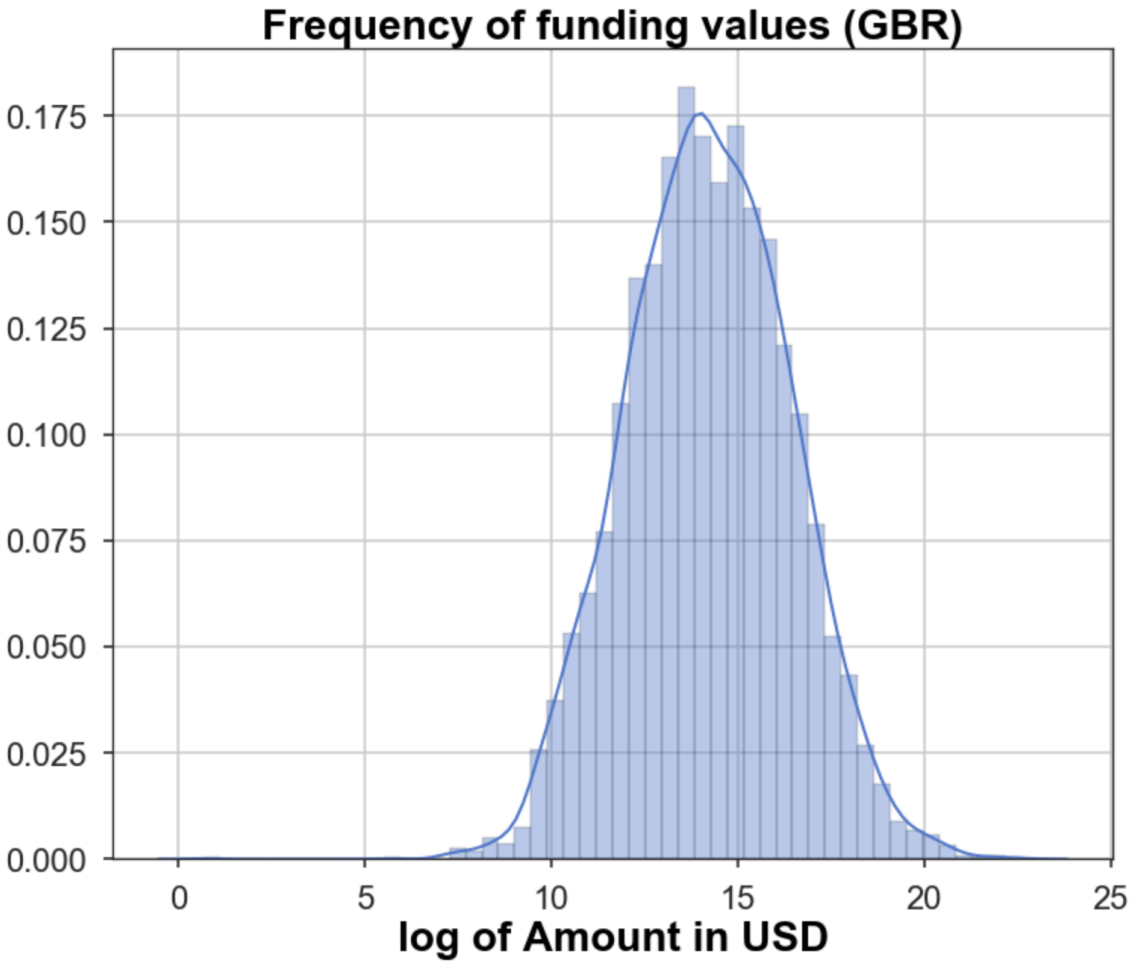


Fig.T . Shows the distribution of funding in terms of the number of companies achieving each funding value displayed in Great Britain.