DATA EXPLORATION

BDAT 1005 – Mathematics for Data Analysis

ABSTRACT

Startup success prediction. Data Exploration

Kifah Owda

200616098

Contents

About the Data Set

Context

Startups are companies in the early stages usually innovate a creative product or solution unique to the market. Startups typically have a very short period to succeed if they are not fast enough they fail because startups start with limited resources and very little to no revenue initially. The startup model is precarious, but with high risks come high returns, attracting many entrepreneurs and risky investors to invest in this model. The startups usually start from the founder's funds like savings, family contributions, or credit cards, later they raise funds from investors in different stages and sources such as early angel investors and VCs (Venture Capital) through multiple rounds of investment that vary with the total fund collected and company evaluations and milestones accomplished. The startup closes and announces bankruptcy if it can't reach milestones, raise more funds, or generate enough revenue to cover its expenses, identified as a failed startup. or it can be sold to another company (acquired) where investors and founders (shareholders) are paid for their shares called acquisition success, a quick and rewarding wealth for founders and investors. Another case, where the company goes IPO and becomes public, this case will not be discussed in this dataset as we will focus on privately held startups which will narrow down the success status to Acquired (succeeded) and closed (failed) to closed.

Motivation

There seems to be no formula to calculate or predict the success or failure of a startup. As an entrepreneur who has been in multiple startups before, I find it exciting to analyze big data about startups trying to find a pattern or correlation between variables and the destiny of a startup company.

Data Dictionary

The data set is for a sample of startup companies in the United States of America (USA), founded between 1985 and 2013. The data set contains 47 variables (columns) and 923 observations (rows). In other words, we have data for 923 startup companies each with 47 data points. The data has 36 qualitative variables and 11 quantitative variables. Some of the columns in the original dataset are calculated columns from original variables that might or might not be presented in this dataset. Please refer to Table 1 below for variable names, statistical types, and descriptions.

Notes: The original dataset contains 49 columns of which 4 are duplicates. The company ID and the state Code. By removing the duplicates we now have 47 unique columns.

Table 1: Data Dictionary

Var.no	Column Name	Data Type	Data Type	Description
1	Count	Nominal	Qualitative	Originally called Unnamed: 0, simple counter for the number of rows/companies
2	Company_ID	Nominal	Qualitative	Unique Company Identifier
3	Company Name	Nominal	Qualitative	Unique Company Name
4	Category_Industry	Nominal	Qualitative	The company's industry
5	City	Nominal	Qualitative	Company address -City
6	Zip_Code	Nominal	Qualitative	Company address - Zip code
7	State_Code	Nominal	Qualitative	Company State
8	Full_Address	Nominal	Qualitative	Company full or partial address
9	Map_latitude	Interval (cont.)	Quantitative	Exact location coordination (latitude) should always be + in USA
10	Map_longitude	Interval (cont.)	Quantitative	Exact location coordination (longitude) should always be - in USA
11	Founded_Date	Ordinal	Qualitative	Date the company was registered
12	Closed_Date	Ordinal	Qualitative	Date the company was closed officially if not acquired
13	First_funding_Date	Ordinal	Qualitative	The date of the first funding round for the company
14	Last_funding_Date	Ordinal	Qualitative	The date of last funding round before closing it sell
15	Age_first_funding_year	Ratio (cont)	Quantitative	The age of the company when it first raised its first funding round (first funding date - founded year)
16	Age_last_funding_year	Ratio (cont)	Quantitative	The age of the company when it raised the last funding round (last funded date- founded year)
18	Number _Milestones	Ratio (disc)	Quantitative	Number of major milestones the company accomplished before closing or sell
17	Age_first_milestone_year	Ratio (cont)	Quantitative	The age of the company when completed first major milestone
19	Age_last_milestone_year	Ratio (cont)	Quantitative	The age of the company when completed last major milestone
20	Relationships	Ratio (disc)	Quantitative	Number of Strong relationship the company has that helped to shape its success
21	Total_funding_rounds	Ratio (disc)	Quantitative	Total number of finance funding rounds (min 1 and max 10)
22	Total_funding_usd	Ratio (cont)	Quantitative	Total funding from all funding rounds
23	has_angel	Ordinal	Qualitative	Yes/ No answer , if the company has an angel investor yes (1) if not (0)
24	has_VC	Ordinal	Qualitative	Yes/ No answer , if the company has VC investor yes (1) if not (0)
25	has_roundA	Ordinal	Qualitative	Yes/ No answer , if the company has raised a round A yes (1) if not (0)
26	has_roundB	Ordinal	Qualitative	Yes/ No answer , if the company has raised a round B yes (1) if not (0)
27	has_roundC	Ordinal	Qualitative	Yes/ No answer , if the company has raised a round C yes (1) if not (0)
28	has_roundD	Ordinal	Qualitative	Yes/ No answer , if the company has raised a round D yes (1) if not (0)
29	Avg_participants	Ratio (cont)	Quantitative	Average number of investors per round
30	is_CA	Nominal	Qualitative	Yes/No answer, if the company in CA state then yes (1) if not (0)
31	is_NY	Nominal	Qualitative	Yes/No answer, if the company in NY state then yes (1) if not (0)
32	is_MA	Nominal	Qualitative	Yes/No answer, if the company in MA state then yes (1) if not (0)
33	is_TX	Nominal	Qualitative	Yes/No answer, if the company in TX state then yes (1) if not (0)
34	is_otherstate	Nominal	Qualitative	Yes/No answer, if the company is not in CA,TX, Ny, MA state then its Other state yes(1)
35	is_industry_software	Nominal	Qualitative	Yes/No answer, if the company in x industry then yes (1) if not (0)
36	is_industry_web	Nominal	Qualitative	Yes/No answer, if the company in x industry then yes (1) if not (0)
37	is_industry_mobile	Nominal	Qualitative	Yes/No answer, if the company in x industry then yes (1) if not (0)
38	is_industry_enterprise	Nominal	Qualitative	Yes/No answer, if the company in x industry then yes (1) if not (0)
39	is_industry_advertising	Nominal	Qualitative	Yes/No answer, if the company in x industry then yes (1) if not (0)
40	is_industry_gamesvideo	Nominal	Qualitative	Yes/No answer, if the company in x industry then yes (1) if not (0)
41	is_industry_ecommerce	Nominal	Qualitative	Yes/No answer, if the company in x industry then yes (1) if not (0)
42	is_industry_biotech	Nominal	Qualitative	Yes/No answer, if the company in x industry then yes (1) if not (0)
43	is_industry_consulting	Nominal	Qualitative	Yes/No answer, if the company in x industry then yes (1) if not (0)
44	is_industry_othercategory	Nominal	Qualitative	Yes/No answer, if the company in x industry then yes (1) if not (0)
45	is_top500	Ordinal	Qualitative	Yes/No answer, if the company was listed in the top 500 companies then yes (1) if not (0)
46	Status	Ordinal	Qualitative	Acquired indicates the company success, closed indicates the company failure
47	Status_Label	Ordinal	Qualitative	Yes/No answer, if the company was acquired then yes (1) if not then it was closed (0)

Ranges and challenges in this dataset:

The table below provides a better understanding of each column (variable) data range and challenges. Please refer to the table 2 below for detailed information

Table 2: Data Ranges & Challenges

Var.no	Column Name	Range	Limitations	Notes
1	Count	[1 - 1153]	Broken serial number/ counter	Not continues counter, indicating that some rows has been deleted or the counter is wrong.
2	Company_ID	Unique IDs	-	Not clear how the ID was generated but mostly it is linked to the company listing on websites like TechCrunch or sim
3	Company Name	Unique Names	-	-
4	Category_Industry	35 different categories	-	-
5	City	City names	-	-
6	Zip_Code	Zip codes	Invalid Zip codes format	Some zip codes need validation
7	State_Code	All states in USA	-	All in USA with higher frequency for (MA,CA,NY,TX)
8	Full_Address	Addresses	Blanks	Missing Data or duplicates info when looking to other address columns. Mostly wont be needed in our case unless in
9	Map_latitude	[25.75 - 59.33]	-	Find the min and max for the USA to validate the data range
10	Map_longitude	[-122 - 18.075]	Invalid data point (+ value)	One positive value, need it be validated
11	Founded_Date	1984-2013		
12	Closed_Date	2001-2013	Blanks	Many Null values , this variable has data only when the company was closed and not acquired
13	First_funding_Date	2000-2013	-	This value could be before or after the founding date
14	Last_funding_Date	2001-2013	-	This value should be less than the closed date never later
15	Age_first_funding_year	[-9 - 22]	Large Ngetive values might be invalid	Seem like a calculated column, the values are a combination of -,+, 0. while small negative values are acceptable la
16	Age_last_funding_year	[-9 - 22]	Negative values invalid	Needs to be always positive otherwise data is not valid
18	Number_Milestones	[0 - 8]	-	Ranges from (0,8) is a very narrow range meaning the milestone reported must have a high requirement to be accom
17	Age_first_milestone_year	[-14 - 28]	Blanks & large negative values	Values have negative might mean that companies have reached milestones before official founding a company, als
19	Age_last_milestone_year	[-7 - 25]	Blanks & large negative values	Some negative values indicates error or invalid data, it is not possible to reach last milestone before founding compa
20	Relationships	[0 - 63]	-	Relationship in this data didn't have a clear explanation but we assume it relates to the number of relationship the co
21	Total_funding_rounds	[1 - 10]	-	Needs to make sense with the funding round categories like series A,B,, etc.
22	Total_funding_usd	[11k - 5.7b]	Outliers	some outliers, minimum of $11K, and max are $510 Millions and $5.7Billion
23	has_angel	0/1	-	Early stage investors - usually refers to pre-seed round of investment
24	has_VC	0/1	-	VC investors indicating pre-seed or seed round of investment
25	has_roundA	0/1	-	After seed round the company funding rounds are named by A,B,C and usually are Big rounds ranging from couple of
26	has_roundB	0/1	-	after seed round the company funding rounds are named by A,B,C and usually are big rounds ranging from couple of
27	has_roundC	0/1	-	After seed round the company funding rounds are named by A,B,C and usually are Big rounds ranging from couple of
28	has_roundD	0/1	-	after seed round the company funding rounds are named by A,B,C and usually are big rounds ranging from couple of
29	Avg_participants	[1 - 16]	-	No clear definition for the avg participants or why the average was introduced rather than the total number, but since
30	is_CA	0/1	-	Calculated column - Category
31	is_NY	0/1	-	Calculated column - Category
32	is_MA	0/1	-	Calculated column - Category
33	is_TX	0/1	-	Calculated column - Category
34	is_otherstate	0/1	-	Calculated column - Category
35	is_industry_software	0/1	-	Calculated column - Category
36	is_industry_web	0/1	-	Calculated column - Category
37	is_industry_mobile	0/1	-	Calculated column - Category
38	is_industry_enterprise	0/1	-	Calculated column - Category
39	is_industry_advertising	0/1	-	Calculated column - Category
40	is_industry_gamesvideo	0/1	-	Calculated column - Category
41	is_industry_ecommerce	0/1	-	Calculated column - Category
42	is_industry_biotech	0/1	-	Calculated column - Category
43	is_industry_consulting	0/1	-	Calculated column - Category
44	is_industry_othercategory	0/1	-	Calculated column - Category
45	is_top500	0/1	-	Calculated column - Category
46	Status	Acquired/failed	-	Calculated column - Category
47	Status_Label	0/1	-	Calculated column - Category

In this section I am going to discuss the most 4 challenging variables and our assumptions:

1. Age at first milestone:

The data set has multiple blank variables and many negative values. Logically the company could have reached major milestones before official founding like a working prototype milestone, but it shouldn't be a large value, -1 to -2 which means 1 or 2 years before the founding date is acceptable other than this the data is hard to interpret.

2. Age at last milestone:

Any negative value here is not acceptable and we need to validate the data, maybe wrong data entry. However, any negative age is not appropriate.

3. Relationships:

The original data dictionary doesn’t describe the “relationship” variable it only indicates it as a numeric variable. Since the range is from 0-16 it is hard to consider this as a category. We can safely assume that this variable indicates the number of strong relationships the startup was able to create during its lifetime. These relationships could be industry experts, advisors, early bird customers, and industry partners who provide significant value for each startup. Networking and relationships are important factors for the success or failure of any startup.

4. Avg-participants

The original data dictionary does not provide a description for the "avg-participants" variable. We can assume that this variable represents the average number of participants in each financing round. Typically, each round of investment includes one lead investor and multiple other investors. The "avg-participants" column is listed after the financing rounds columns and it is a numeric variable.

Data Source

The data has been downloaded as .csv from Kaggle for a project named “Startup Success Prediction”. The owner is Manish KC (Owner). (KC, 2020)

The Owner has acknowledged the data source to its original owner Ramkishan Panthena, who listed a machine learning project predicting the success of startup at GitHub. (Panthena, 2018)

Finer Questions

We have identified 2 primary questions, each with multiple support questions that further break down the topic for investigation. All questions are specific to the USA and for companies established or closed between 1985 and 2013.

1. **How the raised fund influence the success of the startup company and what factors have influenced the total fund?**
    a. How did the total funds raised and the total number of rounds correlate with the startup's success?
    b. What fund round had the most influence on the company’s success?
    c. How did being recognized in media lists like the top 500 startups affect the total fund? Any correlations with the startup's success?
    d. How did the number of relationships affect the total fund? Any correlations with the startup's success?

2. **How did the startup industry influence a startup's success and failure rates?**
    a. What industries had the highest fund rates?
    b. What industries had the highest success rate?
    c. Does a and b correlate together?
    d. How did the location of the startups correlate with the startup's success ?

Univariate analysis

This section address the understanding and cleaning of each variable a lone including, removing duplicates, fill or remove null values, finding and addressing outliers for each variable.

In step 1 we will find the statistical description for all quantitative variables.

Table 3: Descriptive analysis - numerical variables

	Age_first_funding_year	*Aage_last_funding_year*	*Number_Milestones*	*Age_first_milestone_year*	*Age_last_milestone_year*	*Relationships*	*Total_funding_rounds*	*Total_funding_usd*	*Avg_participants*
Mean	2.235630011	3.931455796	1.841820152	3.055353048	4.754422568	7.710725894	2.310942579	25419749.09	2.838586132
Standard Error	0.08263242	0.097689943	0.043534963	0.107216094	0.115681214	0.23915593	0.045782746	6241891.138	0.061703242
Median	1.4466	3.5288	2	2.5205	4.4767	5	2	10000000	2.5
Mode	0	0	1	0	4.0027	3	1	10000000	1
Standard Deviation	2.51044854	2.967909847	1.322632023	2.977057143	3.212107156	7.265776	1.390921713	189634364.5	1.874600949
Sample Variance	6.30235187	8.808488858	1.749355469	8.862869232	10.31763238	52.79150088	1.93466321	3.59612E+16	3.51412872
Kurtosis	10.06065679	3.155693004	0.260326441	5.639304918	2.080202209	8.634078622	2.264505874	872.3579133	5.069727712
Skewness	2.104000712	1.092075357	0.577378064	0.944467666	0.711933788	2.329961225	1.35691706	29.15246057	1.767553697
Range	30.9425	30.9425	8	38.8548	31.6904	63	9	5699989000	15
Minimum	-9.0466	-9.0466	0	-14.1699	-7.0055	0	1	11000	1
Maximum	21.8959	21.8959	8	24.6849	24.6849	63	10	5700000000	16
Sum	2063.4865	3628.7337	1700	2355.6772	3665.6598	7117	2133	23462428412	2620.015
Count	923	923	923	771	771	923	923	923	923
Confidence Level(95.0%)	0.162169452	0.191720446	0.085439119	0.210470513	0.227087963	0.469353144	0.089850482	12249962.72	0.121095097
Outliers									
Q1	0.5726	1.6685	1	1	2.3836	3	1	2700000	1.5
Q3	3.5808	5.5616	3	4.6904	6.7534	10	3	24750000	3.8
IQR	3.0082	3.8931	2	3.6904	4.3698	7	2	22050000	2.3
Lower bound (Q1-1.5IQR)	-3.9397	-4.17115	-2	-4.5356	-4.1711	-7.5	-2	-30375000	-1.95
Upper bound (Q3+1.5IQR)	8.0931	11.40125	6	10.226	13.3081	20.5	6	57825000	7.25

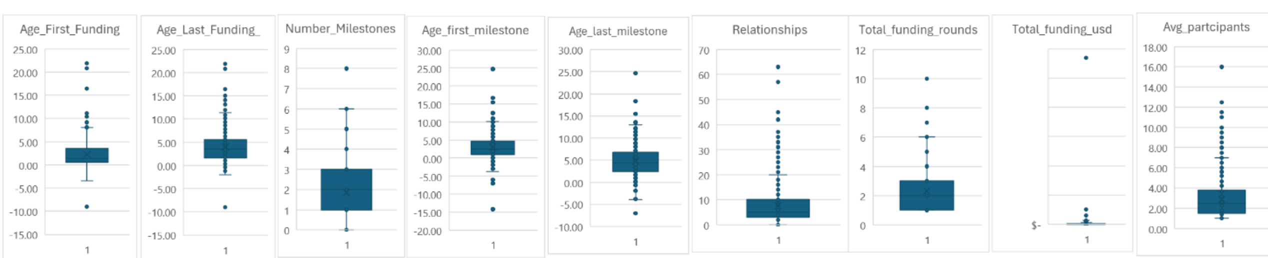


Figure 1: Box Plot for each qualitative variable before cleaning

From this analysis we can find multiple points to work on for every variable:

- Age_first_funding year,Age_last_funding_year: difference between mean and median suggesting outliers and highly skewed. Also, large negative values don't make sense. Further investigation is needed. Data cleaning like removing the negative value is considered.

- Age_first_milestone, Age_last_mislestone: the difference between mean and median suggesting outliers and high skewness. Also, large negative values don't make sense. Further investigation is needed. data cleaning like removing the negative value is considered. The counts for those 2 variables are less than all other counts, suggesting blank values. After investigation, we found that when Number_Milsetone=0, the value at Age_first_milestone and the value at Age_last_mislestone are blanks. Accordingly, we will replace these blank values with zeros. These 2 variables will be ignored in our analysis as that data doesn't explain what is considered a milestone.
- Relationships, mean and median are very different, the variable is highly skewed and the maximum is far away from the upper bound.
  For the purpose of the study, we will exclude some companies where the number of relationships is equal to or larger than 30. These outliers will be excluded from calculations involving the relationship variable. The company IDs that will be excluded for the relationships variable are c:63, c:4, c:10054, c:726, c:10176, c:23125, c:109, c:7299, c:23756, c:1285, c:3288, c:48640, c:1388, c:4453, c:23139, c:1088, c:142.

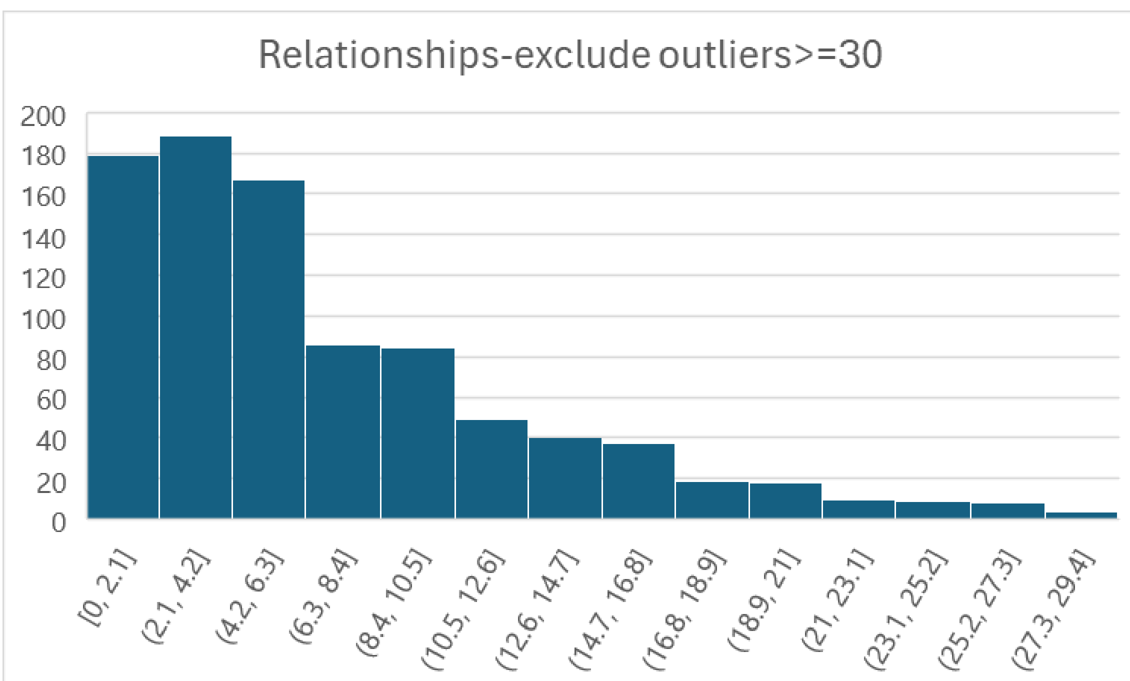
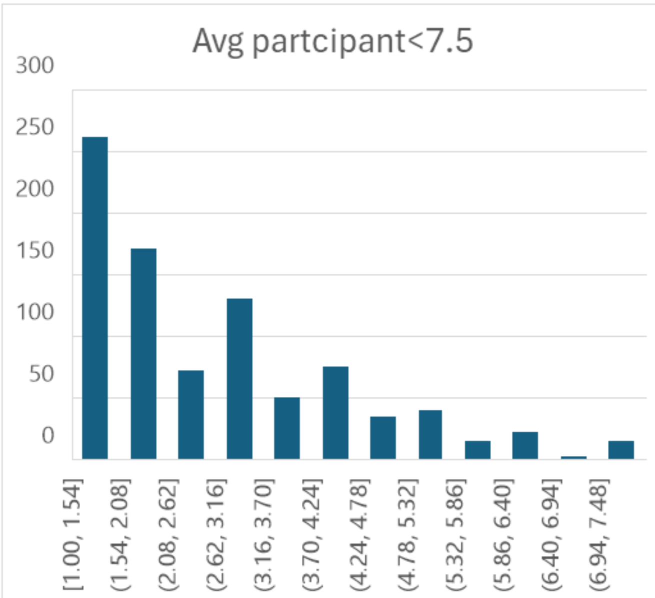


Figure 2: Relationships variable after excluding the outliers for normal distribution purposes.

We will accept the normal distribution of the Relationships variable assumption when we exclude the outliers. The kurtosis 1.68 and the skewness is 1.35 both less than the absolute value of 2.

- Avg_partcipent data seem normal, however, the outliers and range calculations suggest that some data points might be considered outliers. We will exclude all relationships that are larger than 7.5 to ensure normal distribution of this variable.



avg partcipant<7.5	
Mean	2.641977
Standard Error	0.05008
Median	2.3333
Mode	1
Standard Deviati	1.493193
Sample Variance	2.229627
Kurtosis	0.201235
Skewness	0.879305
Range	6.5
Minimum	1
Maximum	7.5
Sum	2348.717
Count	889

- Total_fudnig_usd, shows a large difference between mean and median, highly skewed and large positive kurtosis value. However, Mode and Median are equal, we anticipate outliers that cause this up normality, we will consider logarithmic representation of the total fund.

In step 2: find the frequency and distribution of categorical data.

- Company name and ID – the frequency table shows that every name has one unique ID, only Redwood Systems has 2 Identical ID, suggesting a duplication in our Data which will be deleted.

Company Name and ID	
Company Name	**Count of Company_ID**
Redwood Systems	2
Wheelz	1
TravelMuse	1
Tapioca Mobile	1
Power Analog Microelectronics	1
Virtual Computer	1
Powerset	1
ZettaCore	1
Predictify	1
threadsy	1
Prenova	1
TwoFish	1
Producteev	1
PostPath	1
Provade	1
XGraph	1
Provige	
Swaptre	
PublicEarth	1

Figure 3: Data duplication

- Full Address variable contains 493 blank records, the data in this column is duplicates of the zipcode, city, and province variable so we decided to remove this variable because it doesn't add any value to our dataset.
- Category_Industry frequency table shows the top industries the startups operates in, and that's explain the category coding that the data owner has established (is_software),(is_mobile),,etc

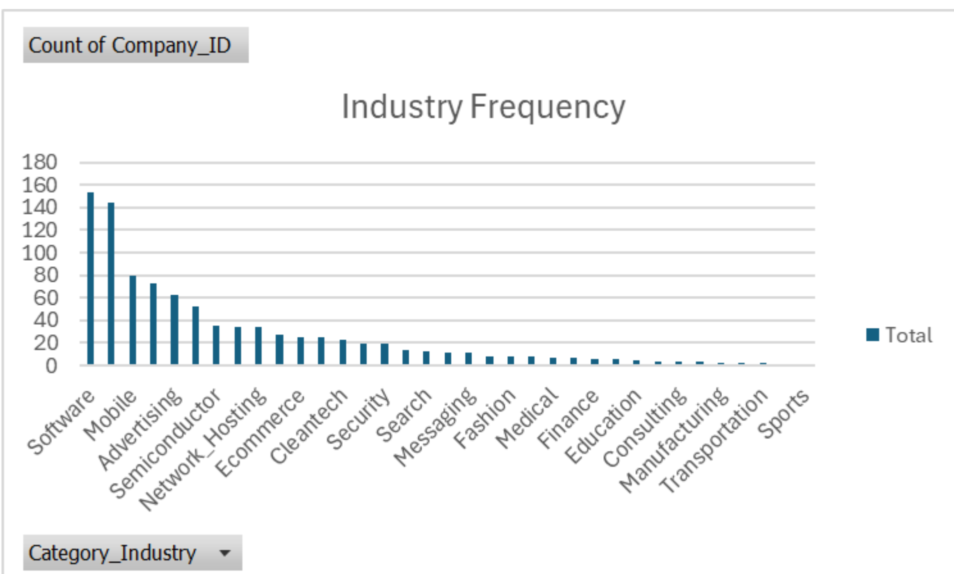


Figure 4:Industry Category frequency distribution

- State_code, the frequency table, and charts show that some states have higher number of startups than others.

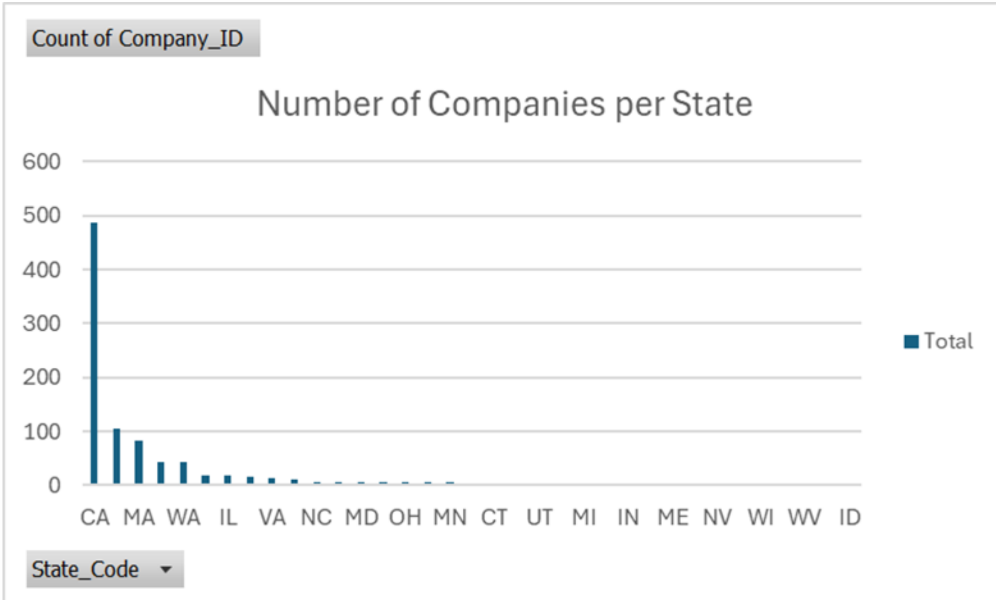


Figure 5: States distribution

- Founded Date analysis
  The founded date distribution shows , normal distribution with abnormal behavior around the year 2000. This will be investigated after cleaning the data.

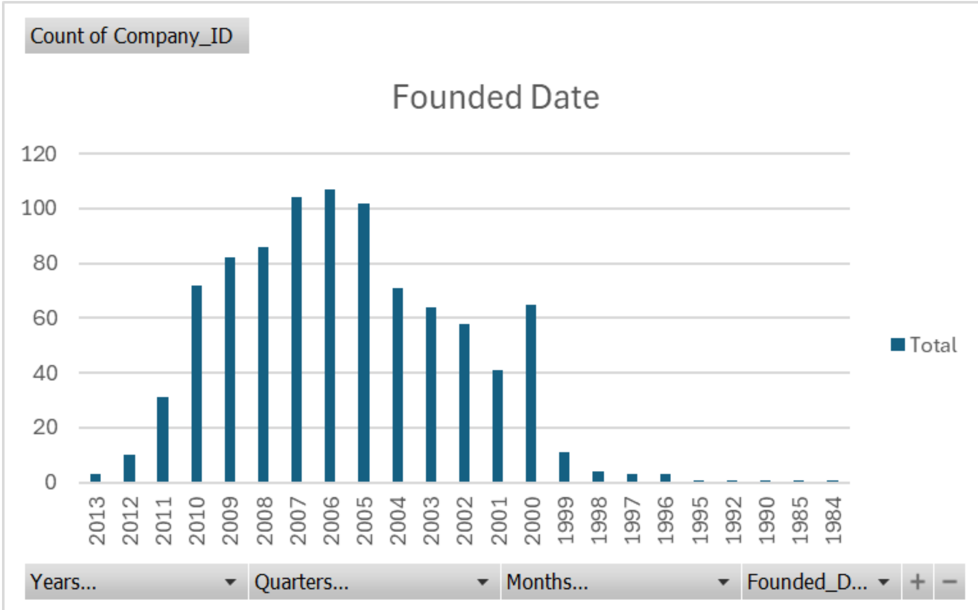


Figure 6: Founded Date before data cleaning

- Closed Date Analysis
  Closed Date has over 50% of the records (587 records) as blank that is due, when the company is acquired not closed, the company has no closed date.

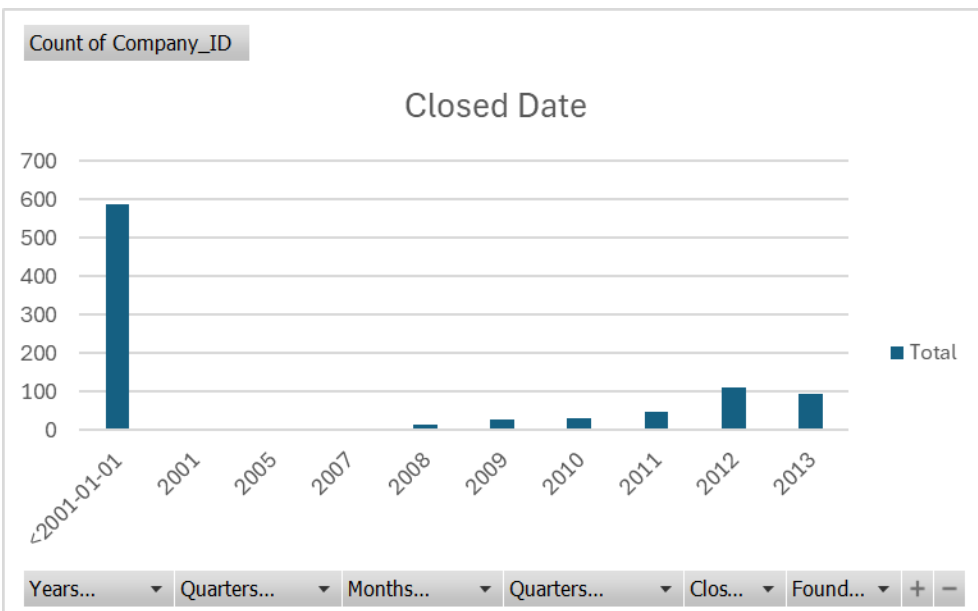


Figure 7: Closed date with blank dates

Filtering the closed date by excluding the blank values shows that 2012 and 2013 has the highest amount of company closings.

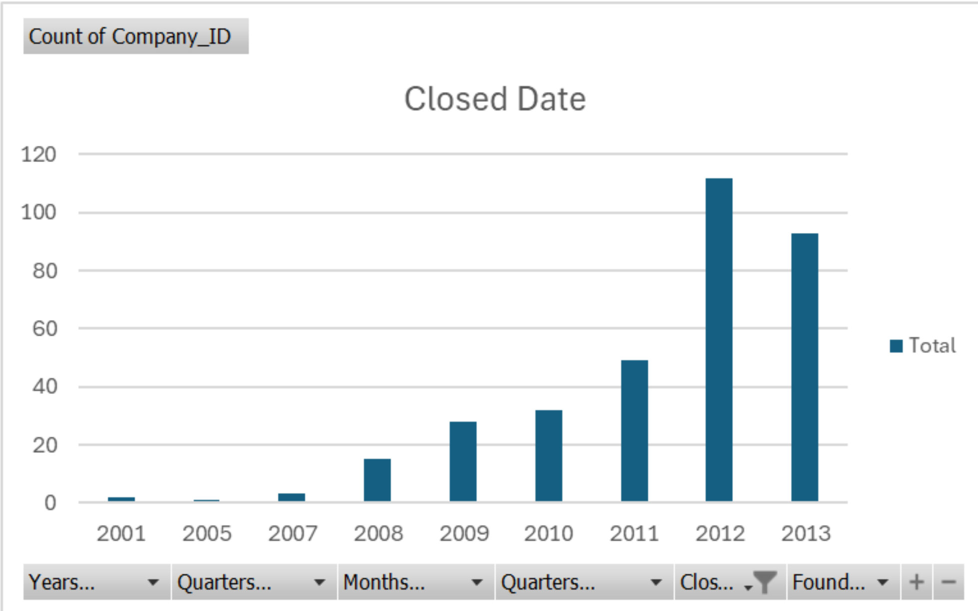


Figure 8: Closed Date after excluding Blank Dates

Dealing with missing closed dates, we decided to add a placeholder 9999-12-31, this date means the company didn't close but was acquired, this date will be excluded when calculating the average closing/failure age for companies. We haven't imputed the data, as this would show significant closing events in the year 2013. It will be hard to separate it from actual companies closing in 2013.

Table 4: Closed Date by Year(9999 year indicates acquired companies with no closed date)

Row Labels	Count of Company_ID
2001	2
2005	1
2007	3
2008	15
2009	28
2010	32
2011	49
2012	112
2013	93
9999	587
Grand Total	922

the closed Date distribution by exactly excluding the 9999, acquired companies with no closed date.

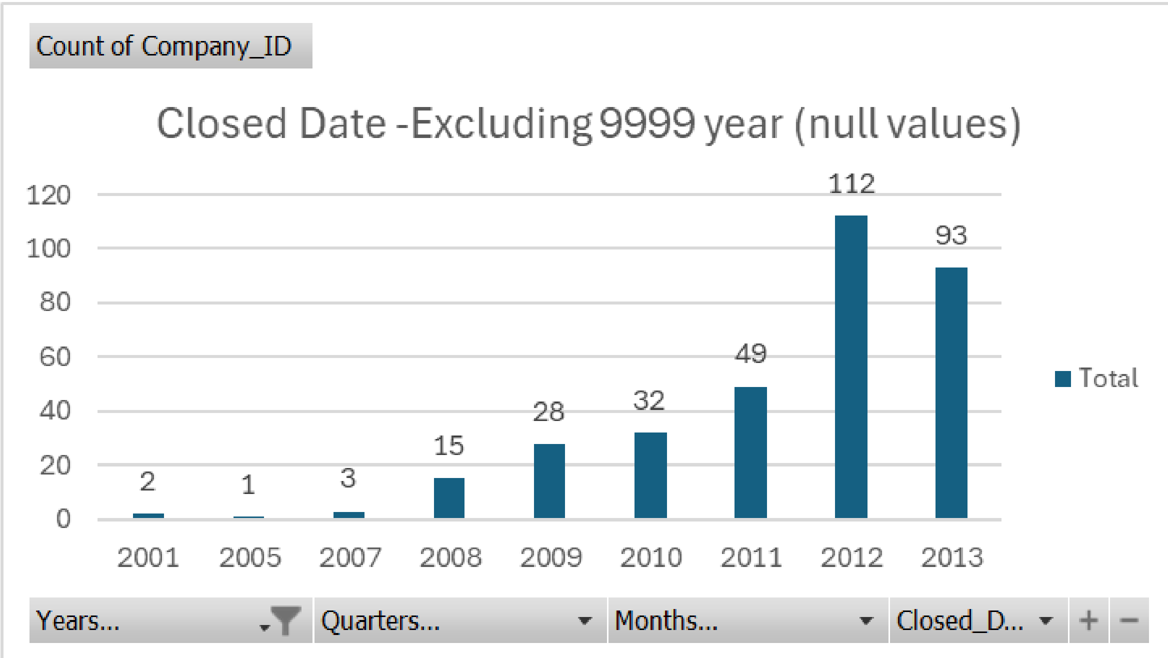


Figure 9: Closed Date Excluding Acquired Companies

- Double-check previously coded fields such as the company status:

We have done multiple tests to check if the counts match between the coded variables and the original variable and it did.

Table 5: comparing the coded column status labels

Row Labels	Count of Status
0	326
1	596
Grand Total	**922**

Row Labels	Count of Company Name
closed	**326**
acquired	**596**
Grand Total	**922**

In step 3: from 1 & 2 we can decide how to clean our data, our strategy is:

- Replace blank values in first and last age milestones to Zeros (linked to a number of milestones_zero)- 152 records have been updated for each variable.
- Deal with the negative values at age_last_fund and age_last milestons.
- After cleaning the data decide about the outliers.

Handle Negative values in variables age first and last fund and milestones.

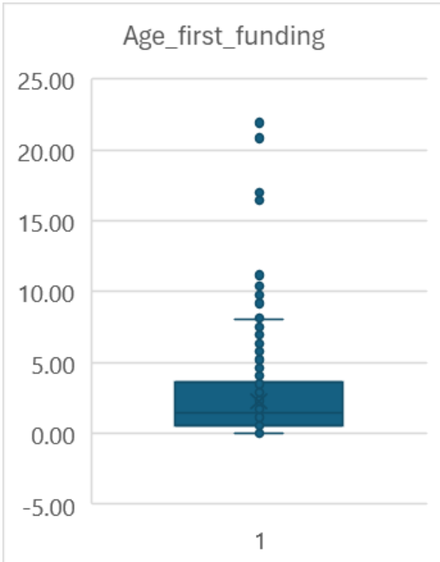
1. Some negative values are a reason for a wrong data entry. The date_founded was placed for the date_closed and vise versa, confirmed this by a Google search. affected 3 records. c:280611, c:170, c:2245.

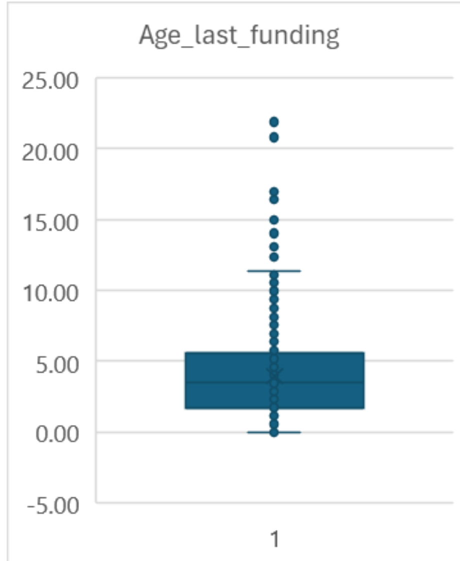
2. Comparing the First_funding_date and last_funding_date , 8 records showed that the last fund date was less than the first fund date. After research, we found an error pattern where the first and last funds were placed in the wrong column. Confirmed by Google search. the 8 records were updated:
   LaunchGram, Lookery, Stickybits, Moprise, Trooval, ReTel Technologies, Challeng, Games, and Zymetis.

3. Negative values in Age_first_funding—all values <-1—were initially sourced from Google searches, but have since been updated using data from Crunchbase. For values >-1 and less than 0, we have set founded_date equal to first_fund_date. This decision was made because these numbers likely reflect minor differences, often in days, between the two dates. Upon review in Crunchbase, we found that most investment events are recorded with precise day, month, and year details, whereas founded_date typically includes only the year. Therefore, our assumption is valid and should not impact any calculations, as our primary interest lies in the age in years, rather than fractions such as months or days.

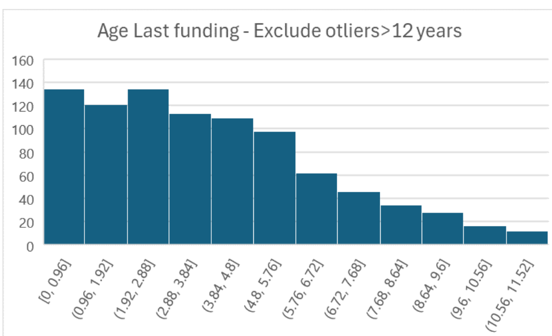
ge_first_funding_year		*Age_last_funding_year*	
Mean	2.278888921	Mean	3.933221
Standard Error	0.081148412	Standard Error	0.096282
Median	1.46575	Median	3.4945
Mode	0	Mode	0
Standard Deviation	2.455996621	Standard Deviation	2.914017
Sample Variance	6.031919403	Sample Variance	8.491497
Kurtosis	10.62498525	Kurtosis	3.122632
Skewness	2.364648288	Skewness	1.245217
Range	21.8986	Range	21.8986
Minimum	-0.0027	Minimum	-0.0027
Maximum	21.8959	Maximum	21.8959
Sum	2087.462252	Sum	3602.83
Count	916	Count	916

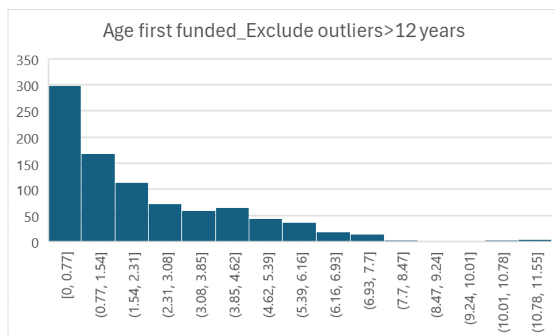
Figure 10: Age first and last funding after cleaning the data





In any analysis where the Age_first _funded and Age_last_funded will be used, we will exclude the values over 12 years. Accordingly, the following company ID's will always be excluded c:44064, c:57923, c:1753, c:24040, c:25383, c:46070, c:3562, c:23692, c:14993.





After excluding the outliers we can treat these variables as normally distributed since the absolute values of kurtosis and skewness are less than 2.

	Age_last_funding_year	*Age_first_funding_year*
Kurtosis	-0.184976254	1.923560312
Skewness	0.668885616	1.352164454

4. Regarding Age_first_milestone_year and Age_last_milestone_year, we will not address negative values because we lack specific information about the dates of these milestones, making it impossible to calculate from the available data. Therefore, these two variables will be removed from our dataset and will not be included in our research questions.

5. Working with the total fund raised, clearly the c:13219, Clearwire with 5.7 billion USD is an outlier. c:51023. We have also deleted all the values 200-300 million dollars so all the values left are from range 11,000 – 162 million. The Graph is less skewed without the removed outliers. However, we won't remove all outliers from the dataset, it might be excluded in some calculations.

Total_funding_usd - removed outliers	
Mean	17726053.84
Standard Error	747707.9479
Median	10000000
Mode	10000000
Standard Deviation	22629748.84
Sample Variance	5.12106E+14
Kurtosis	8.04690529
Skewness	2.478601005
Range	162253126
Minimum	11000
Maximum	162264126
Sum	16237065313
Count	916

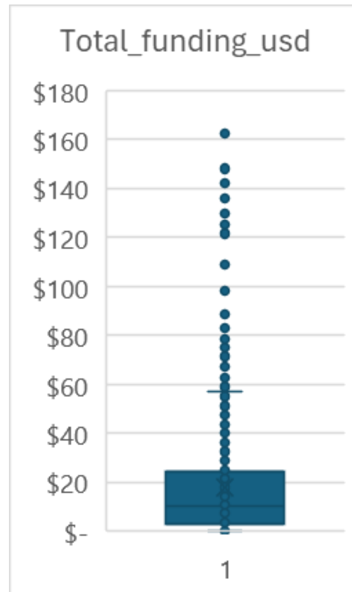
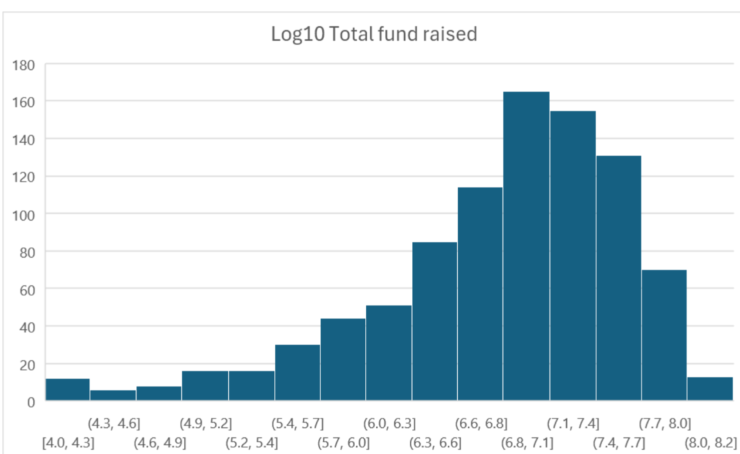


Figure 11: Total revenue Descriptive statistics after removing significant outliers

We also will be using the logarithmic representation of the total revenue to explain some relationships. The graph is negatively skewed but we can accept it as normally distributed since the the absolute values of the skewness and kurtosis are less than 2.



Log10 (total fund)	
Mean	6.825548
Standard Error	0.025629
Median	7
Mode	7
Standard Deviation	0.77566
Sample Variance	0.601648
Kurtosis	1.217544
Skewness	-1.08488
Range	4.16883
Minimum	4.041393
Maximum	8.210223
Sum	6252.202
Count	916

Figure 12 Logarithm, base 10 function of the total revenue raised

Outliers handling

In our analysis of numerical variables (e.g., Relationships, Age_first_funded, Age_last_funded, total_funding_rounds, Avg_participants), we addressed outliers identified as values beyond 1.5 times the interquartile range (IQR) above and below the quartiles. These outliers caused non-normal distribution of the variables. Instead of removing the records with outliers, we excluded them only for specific tests where their presence might distort results. This approach allows us to utilize the same records across different analyses.

To evaluate the normality of a variable, we use kurtosis and skewness with the criterion that their absolute values should be less than 2. As a result, some outliers may fall outside the 1.5 IQR range but still result in an acceptable normal distribution for our purposes.

Data Coding

Total fund category

We have also categorized the total fund raised into 2 column categories, one displays the range and the other is a numerical scale. The category haven't been used in our calculations and we have used the normal logarithmic Log10 of the total fund as quantitative variable.

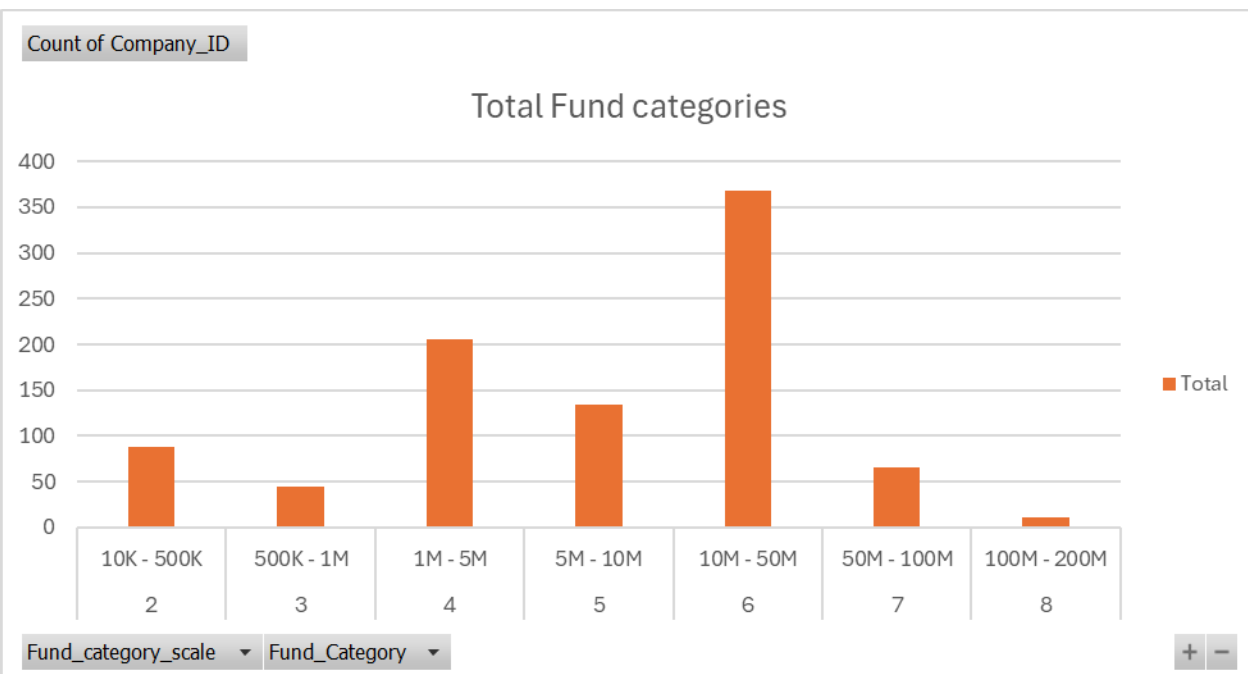


Figure 13: Total Raised fund Categories

Top_industries

We have created a new variable called Top_industries where we categorized the top 6 industries separately and grouped all other industries into other_industries.

State_Category

We have created new variable called State_Category, which categorizes all states other than (MA, CA, IL, WA, NY, TX) as”other states”

Univariate analysis Highlights

After carefully ensuring our data is accurate and reliable, we've taken the following action to clean it:

1. We fixed issues with founding, funding, and closing dates, especially correcting negative age values. This included updating how we calculate the ages of companies' first and last funding rounds.
2. We standardized how we handle missing closed dates, using '9999-12-31' to indicate companies that were acquired, which made our analysis clearer.
3. Our analysis revealed that the distribution of total funds raised was not normal. To address this, we implemented a logarithmic transformation to better handle the skewed data distribution. Additionally, we introduced three coded variables to enhance our analytical approach in future projects.
4. After removing duplicates and significant outliers, our sample size reduced to 916 companies from the original 923.
5. We are aware of potential outliers in other numerical variables. These outliers were retained in the dataset and will be excluded from each test separately.
6. Due to challenges verifying the age of first and last milestones, which often resulted in negative values, we made the decision to remove these variables from our dataset.

Hypothesis testing

In this section, we will be conducting a statistical to test our hypothesis about our research finer questions. For this research, we will structure our analysis by statistical tests. So we will conduct 5 tests, and link our finer questions to each section.

Odds Ratio Test

We chose the odds ratio (OR) test because we have a sample dataset, not a population. tests to help us answer the 2 supporting questions, as both are about 2 dichotomous variables below:

What fund round had the most influence on the company's success?

To help us find and answer these questions, we need to conduct 6 odds ratio tests, considering a confidence level of 95%.

	Variable 1	Variable 2	OR	CI
OR 1	has_angel	Status	0.711340206	[0.5245 - 0.9647]
OR 2	has_VC	Status	0.78339471	[0.5881 - 1.0436]
OR 3	has_roundA	Status	2.19123506	[1.6609 - 2.8907]
OR 4	has_roundB	Status	2.612863135	[1.9396 - 3.5197]
OR 5	has_roundC	Status	2.500534759	[1.73622 - 3.6012]
OR 6	has_roundD	Status	3.310679612	[1.8413 - 5.9523]

For each odds ratio calculation table please refer to Odds Ratio sheet in the attached Excel workbook. The odds ratio calculations suggest a potential positive association between raising higher rounds of investment (A, B, C, and D) and the success of the company in being acquired. For instance, startups that raised round D are 3.3 times more likely to be acquired successfully.

Has Round_D vs Status			
Count of Company_ID	**Column Labels**		
Row Labels	**acquired**	**closed**	**Grand Total**
1	77	14	91
0	515	310	825
Grand Total	**592**	**324**	**916**
OR	3.310679612		
CI	[1.8413-5.9523]		

However, the odds ratio indicates a negative association between angel and VC rounds and the company's success, which is not a favorable interpretation. A more accurate description would be that companies that only raised angel rounds are 1.4 times more likely to fail, whereas those that raised angel and/or VC rounds exclusively are 1.2 times more likely to fail.

Has angel vs status			
Count of Company_ID	**Column Labels**		
Row Labels	**acquired**	**closed**	**Grand Total**
1	138	97	235
0	454	227	681
Grand Total	**592**	**324**	**916**
OR	0.711340206	1.406	

How did being recognized in media lists like the top 500 startups affect the total fund? Any correlations with the startup's success?

This question has 2 parts to it, the first one is categorical with quantitative data and the second one is 2 dichotomous variables. So we will be calculating the odds ratio for variable

Is_top_500 vs Status			
Count of Company_ID	**Column Labels**		
Row Labels	**acquired**	**closed**	**Grand Total**
1	532	208	740
0	60	116	176
Grand Total	**592**	**324**	**916**
OR	4.944871795		
CI	[3.4829-7.0204]		

is_top_ 500 and Status.

Variable 1	Variable 2	OR	CI
is_top_500	Status	4.944871795	[3.4829-7.0204]

The odds ratio calculations suggest a potential positive relationship between being listed as one of the top 500 startups and the likelihood of being acquired. Startups that were listed as one of the top 500 are 5 times more likely to be successfully acquired.

Chi-Square Test

From our FINER questions we have identified 2 questions were we can test our hypothesis using the Chi-square test. (one polytomous variable and one dichotomous variable)

What industries had the highest success rate?

We need to conduct a chi-square test between the "status" variable and the "industry" variable to test our null hypothesis: There is no statistically significant difference between startup success status and the industry in which they operate. After removing industries with fewer than 5 entries in each status category (acquired or closed), we are left with 785 total observations and 15 industries.

Chi-Square Test industry vs status

Count of Company_I	Column															
Row Labels	**Software**	**Web**	**Mobile**	**Enterprise**	**Advertisin**	**Games_Vi**	**Semicond**	**Network_I**	**Biotech**	**Hardware**	**Ecommerc**	**Public_Re**	**Cleantech**	**Analytics**	**Security**	**Social**
acquired	100	92	51	56	45	31	24	24	21	11	11	10	9	16	15	8
closed	52	51	27	17	17	21	11	10	12	16	14	15	11	3	4	6
Grand Total	**152**	**143**	**78**	**73**	**62**	**52**	**35**	**34**	**33**	**27**	**25**	**25**	**20**	**19**	**19**	**14**

Excluding all stetes with less than 5 entries

Row Labels	Software	Web	Mobile	Enterprise	Advertising	Games_Vi	Semicond	Network_I	Biotech	Hardware	Ecommerc	Public_Rel	Cleantech	Social	Search	Grand Tota
acquired	100	92	51	56	45	31	24	24	21	11	11	10	9	8	7	500
closed	52	51	27	17	17	21	11	10	12	16	14	15	11	6	5	285
Grand Total	152	143	78	73	62	52	35	34	33	27	25	25	20	14	12	785

Expected

Row Labels	Software	Web	Mobile	Enterprise	Advertising	Games_Vi	Semicond	Network_I	Biotech	Hardware	Ecommerc	Public_Rel	Cleantech	Social	Search	Grand Tota
acquired	96.81529	91.0828	49.68153	46.49682	39.49045	33.12102	22.29299	21.65605	21.01911	17.19745	15.92357	15.92357	12.73885	8.917197	7.643312	500
closed	55.18471	51.9172	28.31847	26.50318	22.50955	18.87898	12.70701	12.34395	11.98089	9.802548	9.076433	9.076433	7.261146	5.082803	4.356688	285
Grand Total	152	143	78	73	62	52	35	34	33	27	25	25	20	14	12	785

ChiSquare

Row Labels	Software	Web	Mobile	Enterprise	Advertising	Games_Vi	Semicond	Network_I	Biotech	Hardware	Ecommerc	Public_Rel	Cleantech	Social	Search
acquired	0.10476	0.009236	0.03499	1.942295	0.768672	0.135827	0.130708	0.253698	1.74E-05	2.233378	1.522367	2.203567	1.097354	0.09434	0.054145
closed	0.18379	0.016204	0.061386	3.407535	1.348547	0.238293	0.229312	0.445084	3.05E-05	3.918207	2.670819	3.865907	1.925182	0.165509	0.094992
															29.15615

Chi-Square	29.15615
P-Value	0.009954

Variable 1	Variable 2	Chi-square	p-value
Status	Industry	29.151615	0.009954

The chi-square value of 29.151615 suggests a statistically significant difference between startup success and the industries they operate. With a p-value of 0.009954, which is less than the significance level of 0.05 typically used for a 95% confidence level, we can confidently REJECT the null hypothesis that there is no significant difference between the industry and status.

How did the location (state) of the startups correlate with the startup's success?

We conducted a Chi_square test, excluding all states with less than 5 entries for closed or acquired companies. . After removing states with fewer than 5 entries in each status category (acquired or closed), we are left with 833 total observations and 10 states.

Excluding all stetes with less than 5 entries											
Row Labels	CA	NY	MA	TX	WA	IL	CO	PA	VA	GA	Grand Tota
acquired	330	77	64	22	23	9	14	6	7	6	558
closed	155	29	19	19	18	9	4	11	6	5	275
Grand Total	485	106	83	41	41	18	18	17	13	11	833

Excpected

Row Labels	CA	NY	MA	TX	WA	IL	CO	PA	VA	GA	Grand Tota
acquired	324.886	71.006	55.59904	27.46459	27.46459	12.05762	12.05762	11.38776	8.708283	7.368547	558
closed	160.114	34.994	27.40096	13.53541	13.53541	5.942377	5.942377	5.612245	4.291717	3.631453	275
Grand Total	485	106	83	41	41	18	18	17	13	11	833

chi square

Row Labels	CA	NY	MA	TX	WA	IL	CO	PA	VA	GA	Grand Tota
acquired	0.0805	0.505985	1.269377	1.08728	0.725754	0.775365	0.3129	2.549045	0.33511	0.254178	
closed	0.163343	1.026691	2.575681	2.20619	1.47262	1.573286	0.634902	5.172245	0.679968	0.51575	
Grand Total											23.91617

Chi-Square	23.91617
P-Value	0.004436

Variable 1	Variable 2	Chi-square	p-value
Status	State_Code	23.91617	0.0044

The chi-square value of 23.91617 suggests a statistically significant difference between startup success and the state where the startups operate. With a p-value of 0.0044, which is below the typical significance level of 0.05 used for a 95% confidence level, we confidently reject the null hypothesis that there is no significant difference between startup success and the state they operate from.

This observation is supported by the frequency table (pivot table), which shows that the highest number of acquired companies are located in California (CA). This can be reasonably attributed to Silicon Valley, a business hub for technology companies in the United States. Alternatively, it could indicate that the sample is not fully representative of the entire startup population across states.

T-Test

The t-test is used to analyze questions involving a quantitative variable categorized into two groups based on a dichotomous variable. We have identified three specific hypotheses that require a t-test for testing.

How did startups being recognized in media lists like the top 500 startups affect the total fund? Any correlations with the startup's success?

The question has 2 parts, the first part is the relationship between being recognized as a top 500 startup and the total fund, this is where we need to test our hypothesis using the t-test. The second part has been tested earlier using the odds ratio.

Step1: check the normal distribution assumption of each group

For the first part, we tried to conduct the t test between the total fund and the status column, however, the total fund distribution for companies acquired is not normal same as the with companies closed. So we couldn't conduct a t-test.

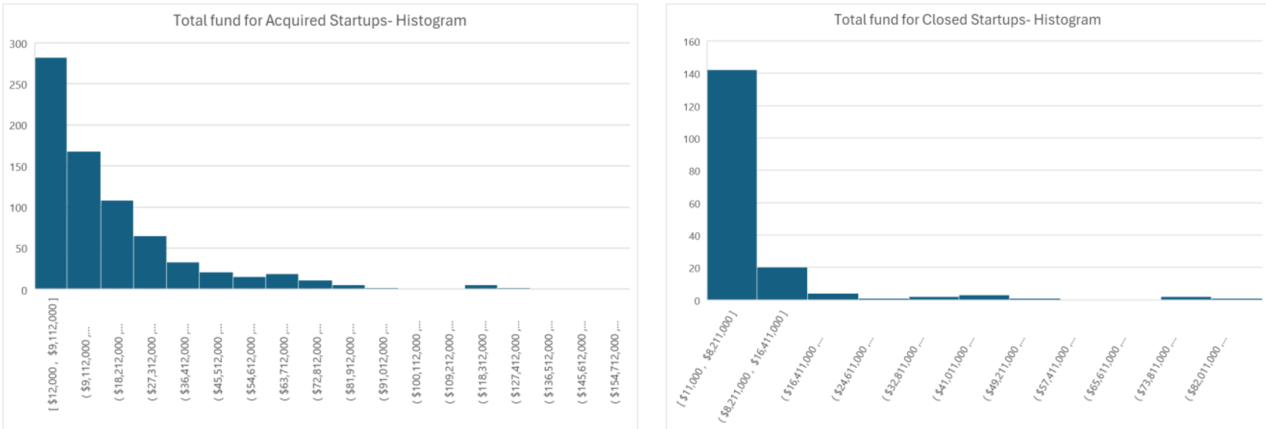
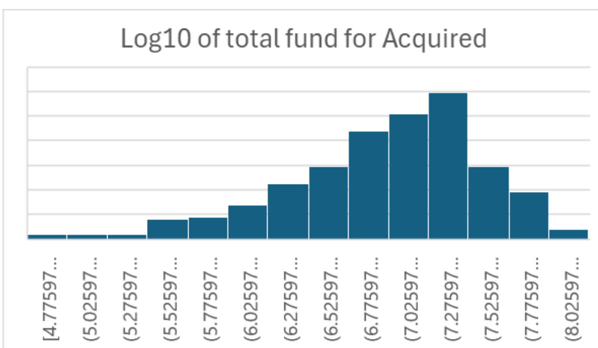


Figure 14: Nonnormality for total funds grouped by closed and acquired companies

We have transformed the total fund variable using log10 and grouped it by status. Upon inspection, both groups exhibit acceptable normality distributions, with skewness and kurtosis absolute values less than 2.



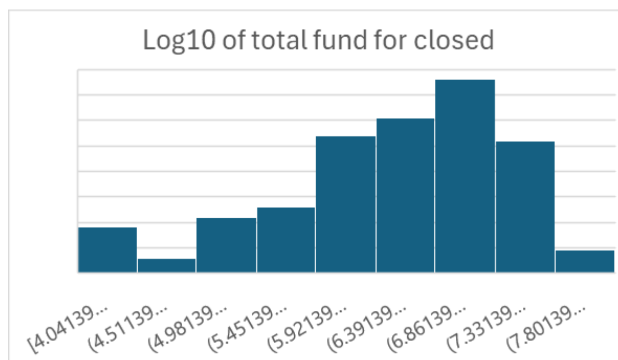


Table 6: Normal distribution for the grouped logarithmic total fund

Log10 of total funds for acquired companies		*Log 10 of the total find for closed companies*	
Mean	7.000162739	Mean	6.506499025
Standard Error	0.025086077	Standard Error	0.051660754
Median	7.096910013	Median	6.694583042
Mode	7	Mode	6.602059991
Standard Deviation	0.610370607	Standard Deviation	0.929893579
Sample Variance	0.372552278	Sample Variance	0.864702068
Kurtosis	0.688123066	Kurtosis	-0.03963853
Skewness	-0.829760185	Skewness	-0.74286128
Range	3.376314013	Range	4.16882983
Minimum	4.775974331	Minimum	4.041392685
Maximum	8.152288344	Maximum	8.210222515
Sum	4144.096341	Sum	2108.105684
Count	592	Count	324

Step2: Test the null hypothesis using t-test

Our null hypothesis, is the difference in the logarithmic total fund mean between group A, acquired companies and the group B closed companies happened by chance.

t-Test: Two-Sample Assuming Equal Variances		
	acquired	*closed*
Mean	7.000162739	6.506499025
Variance	0.372552278	0.864702068
Observations	592	324
Pooled Variance	0.546473922	
Hypothesized Mean Difference	0	
df	914	
t Stat	9.663447832	
P(T<=t) one-tail	2.11293E-21	
t Critical one-tail	1.646522472	
P(T<=t) two-tail	4.23E-21	
t Critical two-tail	1.962562849	

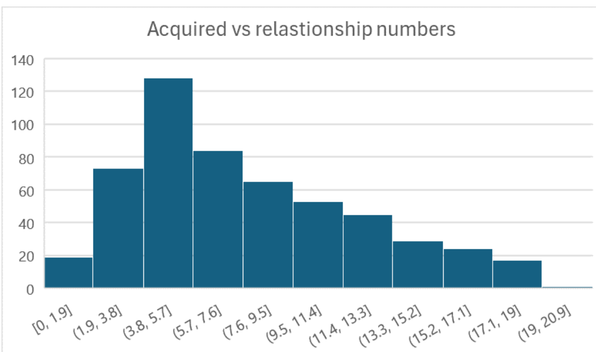
With a t-value of 1.96 and, we reject the null hypothesis, indicating a statistically significant difference in the mean logarithmic total fund values between acquired and closed startup companies. The positive t-value suggests that, on average, the logarithmic total funds raised are higher for acquired companies compared to closed companies. Therefore, we conclude that acquired companies tend to have higher average fund-raising amounts, demonstrating a significant statistical difference in company status based on total fund.

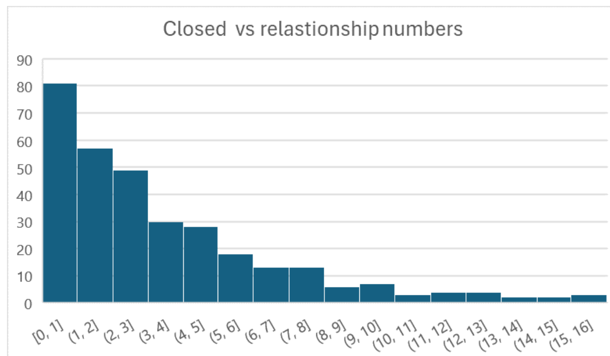
How did the number of relationships affect the total fund? Any correlations with the startup's success?

We will conduct a t-test to examine the relationship between the number of relationships (quantitative) and startup success (dichotomous).

Step1: we need to assess the normality within each group.

The number of relationships is not normally distributed, and consequently, the grouped data do not exhibit normality. For the purpose of this study, we will exclude outliers from both groups, specifically any values over 20 relationships, to ensure normality. A total of 58 outliers have been excluded from this analysis due to exceeding the upper bound of 20. After excluding the outliers the 2 groups show acceptable normality with absolute values of Kurtosis and skewness less than 2.





	acquired	*closed*
Kurtosis	-0.379044132	1.851226
Skewness	0.687975073	1.383951

Step 2: test the null hypothesis

Our null hypothesis: the difference in the total relationships mean between group A, acquired companies and group B closed companies happened by chance.

t-Test: Two-Sample Assuming Equal Variances		
	acquired	*closed*
Mean	7.708178439	3.85625
Variance	20.50127377	11.49024
Observations	538	320
Pooled Variance	17.14319102	
Hypothesized Mean Diff	0	
df	856	
t Stat	13.17816258	
P(T<=t) one-tail	1.5264E-36	
t Critical one-tail	1.646635671	
P(T<=t) two-tail	3.05281E-36	
t Critical two-tail	1.962739186	

With a p-value of 1.05E-36, significantly smaller than 0.05, and a t-critical value of 1.96, we reject the null hypothesis of no significant difference in the average number of relationships between successful and closed companies. The positive t-value indicates that successful companies have on average, a higher number of relationships compared to closed companies.

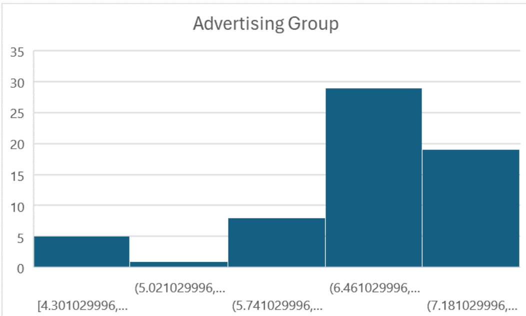
ANOVA (Analysis of Variation)

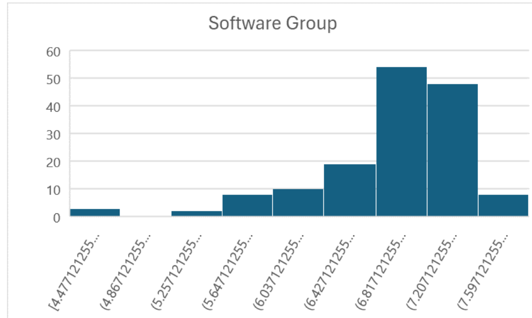
We will use the ANOVA testing to test our hypothesis for quantitative variables grouped in more than 2 categories. This test will help us to test hypothesis to answer the following FINER questions.

What industries had the highest fund rates?

We need to test our null hypothesis that there is no significant difference in the mean total funds between different industries, assuming all industries have the same average fund. using the Top_industries coded variable and the log10 of the total fund

Step1: We assessed the normality of each group by calculating skewness and kurtosis,





accepting normality when the absolute values of these measures were less than 2. Please refer to the ANOVA sheet in the attached excel workbook for full normality analysis.

Step2: test the null hypothesis: there is no significant difference between the means of total funds between industries.

SUMMARY					
Groups	*Count*	*Sum*	*Average*	*Variance*	*SD*
Software	152	1058.368833	6.962952851	0.352092	0.593373
Enterprise	72	499.7405849	6.940841457	0.372065	0.609971
Other Industry	356	2462.562544	6.917310518	0.612897	0.782877
Mobile	78	530.7885558	6.804981485	0.653882	0.80863
Advertising	62	418.8258742	6.755256035	0.669758	0.818388
Web	143	926.7435838	6.480724362	0.670021	0.818548
Games/Video	52	348.010681	6.692513096	0.819609	0.905322

The SD values between the largest and smallest is 1.5:1 so we will accept the ANOVA test result.

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	25.08718559	6	4.181197598	7.22724	1.5E-07	2.108549
Within Groups	525.3080306	908	0.578533073			
Total	550.3952162	914				

With an F-value of 7.22, which exceeds the critical F-value of 2.108, and a p-value of 1.5E-07 (smaller than 0.05), we reject the null hypothesis. This indicates a significant mean difference between the groups, suggesting that the logarithmic total fund mean varies significantly across industries.

To identify where significant differences occur between the groups, we conducted a post-hoc analysis by performing pairwise t-tests for all possible group comparisons, resulting in

21 tests. Each p-value was compared to the Bonferroni-corrected alpha level of 0.00238. The comparisons that yielded p-values less than 0.00238 include:

- Software vs. Web with a p-value 1.46268E-05
- Other Industries vs. Web with a p-value 2.2145E-08
- Enterprise vs. Web with a p-value 7.79018E-09

Additionally, Mobile vs. Web approached statistical significance. These results indicate that there are significant differences in the average total funding between startups in the Web industry and those in other industries such as Mobile, Enterprise, and Software. This suggests that startups in the Web industry receive significantly lower levels of funding compared to those in the other industries studied.

MANOVA

For MANOVA analysis we have generated new questions to test hypotheses related to independent polytomous variables and multiple qualitative dependent variables.

Are there significant differences in company metrics of age first last fund and total fund across different states?

Step1: normality distribution

The Age_first_funde doesn't show normality in the original dataset. However, excluding the values where the age>12 years shows an accepted normal distribution. The normal logarithmic fund is normally distributed so we would assume normality per group. We also had to exclude states where the number of records is equal to 1.

Step 2: We need to test the null hypothesis, that there is no significant mean difference between total funds, and the company's age at first and last funding. For this, we will use the MANOVA test where the dependent variables are Log10 of the total fund and Age_first_funding_year and the independent variable is the State code filtering out states with less than 2 companies.

One-way MANOVA						
	stat	*F*	*df1*	*df2*	*p-value*	*part eta-sq*
Pillai Trace	0.184051	2.035503	84	2616	1.41059E-07	0.06135
Wilk's Lambda	0.825675	2.049547	84	2603.511	1.05403E-07	0.062025
Hotelling Trace	0.199539	2.063483	84	2606	7.85003E-08	0.062365
Roy's Lg Root	0.11363					

Figure 15: One-way MANOVA states vs (log10 fund, age first and last funding year)

With the very small p-value from the MANOVA tests, we reject the null hypothesis of no mean differences between the groups. Therefore, we can conclude that the state significantly affects both the average total funding and the age of the company during its first funding round. This also agrees with the single ANOVA tests for each variable.

We further investigated the significance of differences in total funding between state groups to identify where statistical differences occur. We categorized the states (CA, MA, IL, NY, TX, WA) and grouped all other states into a single category, "Other States." We then conducted 21 paired t-tests, adjusting the alpha level using the Bonferroni correction, resulting in a new alpha of 0.0023. The following significant differences were identified:

- CA vs. IL: Significant difference with a p-value of 0.004.
- CA vs. NY: Significant difference with a p-value of 0.00083.
- IL vs. MA: Significant difference with a p-value of 0.000173.
- MA vs. NY: Significant difference with a p-value of 0.000143.
- MA vs. Other States: Significant difference with a p-value of 0.002021.

These results indicate that the observed differences in the mean logarithmic total funding between the states listed are statistically significant and unlikely to have occurred by chance. Specifically:

- Massachusetts (MA) has the highest level of significance in its differences with nearly all other states. This suggests that MA has a significantly higher average funding compared to other states, with the exception of California and New York.
- California (CA) has higher average funding compared to IL and NY, indicating that California's startups receive significantly more funding than those in Illinois and New York.

To understand the factors influencing the static difference, we might need additional data such as the density of startups per state, the type of investment (rounds), and the industry focus.

Do companies in different industries differ significantly in their metrics (age last fund, total fund)?

Step1:Same as the previous MANOVA test, we had to exclude the outliers from the Age_first_funded and remove States with one record.

Step 2: We conducted a MANOVA test to evaluate the null hypothesis that there is no significant difference in the mean total funding and the age of the company's first funding year across different industry categories.

	stat	F	df1	df2	p-value	part eta-sq
Pillai Trace	0.302233	4.901062	64	1762	0	0.151117
Wilk's Lam	0.709688	5.143661	64	1760	0	0.15757
Hotelling T	0.392271	5.387594	64	1758	0	0.163974
Roy's Lg Ro	0.343345					

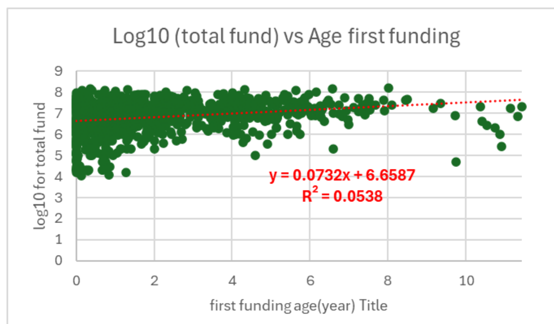
Figure 16: One-way MANOVA Industry vs Log10 total fund, age first and last funding year

The p-value of zero provides strong evidence that our hypothesis is false, indicating a significant difference in means between different industries. Which agrees with the single ANOVA tests. As previously observed, the average funding significantly differs between startups in the Web industry and those in other industries. Further group comparisons are needed to assess how age at last funding varies across these industries.

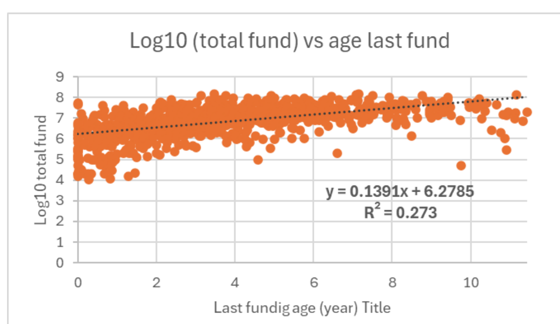
Inferential statistics

Correlation

For this section, we will find the correlation(relationship strength and direction) between the following variables, please note all outliers causing non normality were excluded for this analysis :

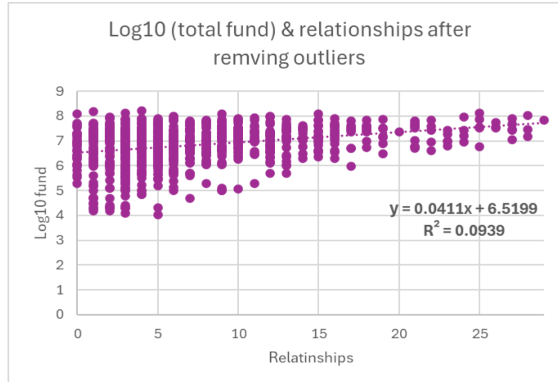


Age First Fund & Log10 of Total Fund: The correlation coefficient (r) between the age of the company at its first funding and log10 of the total fund is 0.252. As the age of the first fund increases the total fund increase slightly. With a p-value 9.98428E-15, the correlation is statistically significant and unlikely to occur by chance.

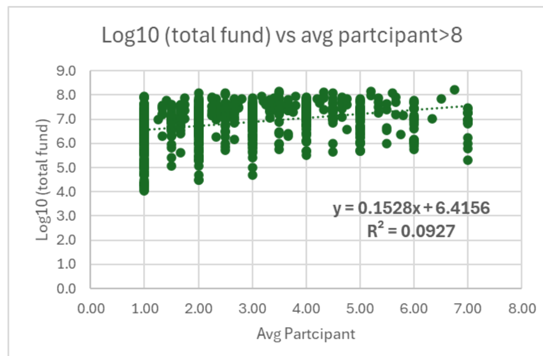


Age Last Fund & Log10 of Total Fund: The correlation coefficient (r) between age of the company at its last funding and log10 of the total fund is 0.55, indicating a moderate positive relationship. This indicates older companies tend to receive higher total funds. With a p-value 1.11465E-73, the correlation is statistically significant and unlikely to occur by chance.

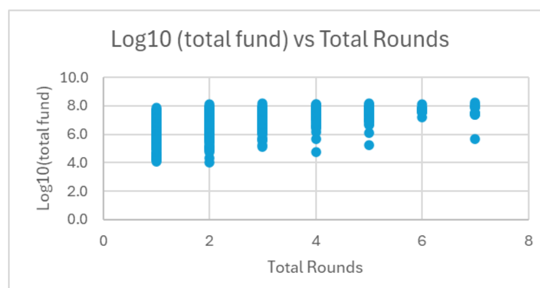
Relationships & Log10 of Total Fund: The correlation coefficient (r) between the number of relationships and log10 of the total fund is 0.306. The more relationships the startup has the total fund slightly increases. With a p-value of 6.43E-21, the correlation is statistically significant and unlikely to occur by chance.



Avg Participant & Log10 of Total Fund: The correlation coefficient (r) between average participants per round and log10 of the total fund is 0.316, indicating a positive relationship. With a p-value of 1.62237E-20, the correlation is statistically significant and unlikely to occur by chance.



Total Fund Rounds & Log10 of Total Fund: The correlation coefficient between total funding rounds and log10 of the total fund is 0.461. More rounds indicate more total funds. With a p-value of 4.71357E-41, the correlation is statistically significant and unlikely to occur by chance.



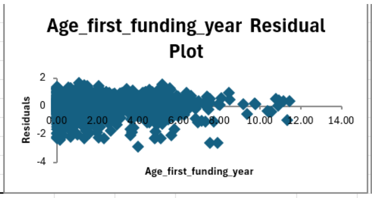
From our correlation analysis, we found that all the quantitative variables—such as relationships, average participants, age at first funding, and age at last funding—show significant correlations with the logarithmic total funding. Among these variables, **Age_ Last_Fund** has the highest correlation with the logarithmic total funding and is the most statistically significant.

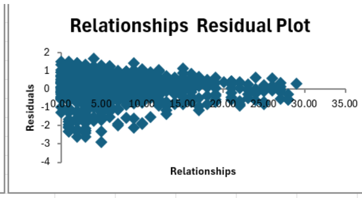
Regression analysis

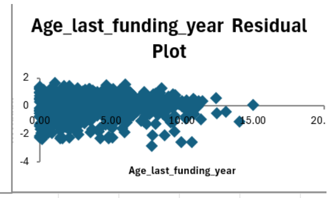
We conducted a linear regression analysis to examine the impact of three continuous variables—age at first funding, age at last funding, and total relationships—on the logarithm of total fund size (log10 of total fund).

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.34072
R Square	0.11609
Adjusted R Square	0.11311
Standard Error	0.72805
Observations	895







ANOVA

	df	SS	MS	F	gnificance F
Regression	3	62.0281	20.676	39.0069	1.1E-23
Residual	891	472.284	0.53006		
Total	894	534.312			

	Coefficient	andard Err	t Stat	P-value	ower 95%	Jpper 95%	ower 95.0%	pper 95.0%
Intercept	6.3702	0.05141	123.913	0	6.26931	6.4711	6.26931	6.4711
Age_first_funding_year	-0.0139	0.01742	-0.797	0.42569	-0.0481	0.0203	-0.0481	0.0203
Age_last_funding_year	0.05517	0.01378	4.00346	6.8E-05	0.02812	0.08221	0.02812	0.08221
Relationships	0.03688	0.00453	8.13971	1.3E-15	0.02799	0.04577	0.02799	0.04577

Our findings indicate that:

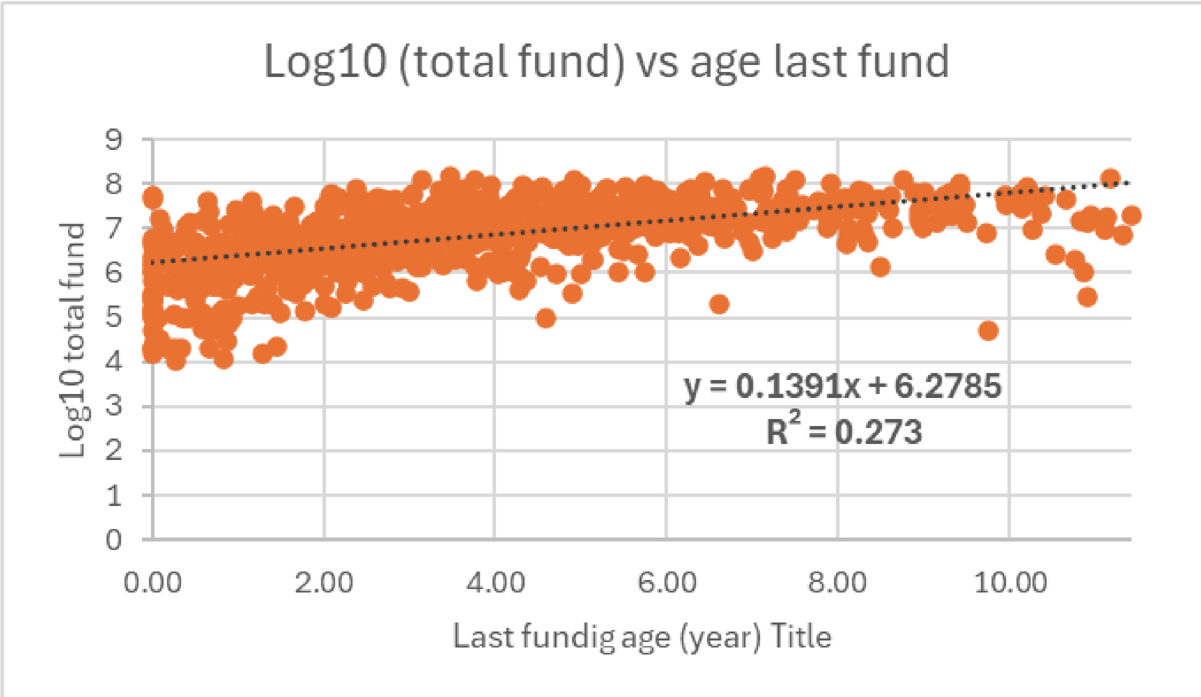
1. **Age at First Funding**: The p-value associated with age at first funding suggests that there is no statistically significant relationship with total fund size ($p > 0.05$). Consequently, age at first funding does not emerge as a significant predictor of total fund size in this analysis.

2. **Combined Effect of Variables**: When all three variables were included together in the regression model, the R^2 value decreased. This suggests that the combined model did not improve the prediction accuracy compared to using age at last funding alone.

Regression Age last fund predictor

Based on these observations, we proceed with a focused predictor linear regression analysis involving age at last funding and total fund size. Removing the outliers of the age larger than 11.43 at last fund.

The regression equation **y(log10 total fund=0.1391x(age last fund)+6.2785**, indicates that for every unit increase in age at last funding (in years), the logarithm of total fund size (log10) is predicted to increase by 0.1391 units. The R^2 value of 0.273 demonstrates that age at last funding alone explains approximately 27.3% of the variance in total fund size.



Total Fund=10^(6.2785+0.1391x age last fund+6.2785)

The predicted Total Fund increases by a factor of approximately 10^(0.1391), which equals 1.3491. Therefore, each additional year in Age Last Fund is associated with a predicted 34.9% increase in Total Fund, according to the regression model. However, the threshold 10^(6.2785) (approximately $1,900,000) indicates that the prediction starts becoming meaningful for companies whose Total Fund at the starting point is greater than $1,900,000.

In summary, this model predicts that companies can raise more money as they get older, only if they have enough runaways, either previous funds or revenue generation that help the startup survive until the next round.

Multi-regression using numerical and dichotomous variables

We try to create a regression model that has better ability to explain the total funds raised by a startup. So we included the dichotomous variables in our model such as is_top 500, has_angel, has_VC, and other numerical ones such as the average participants, relationships, and total funding rounds.

Multicollinearity variables:

Age at last funding and **age at first funding** are highly correlated (r=0.7), indicating multicollinearity. When both variables were included in the analysis, age at first funding lost its statistical significance. To address this, we chose to use **age last fund** as the predictor, as it shows the strongest positive correlation with logarithmic total funding.

For all the predictor variables outliers have been excluded from this model including Relationships larger than 30, rounds larger than 8, Ages larger than11.43, and participants larger than 7.5.

Regression Statistics	
Multiple R	0.806209
R Square	0.649972
Adjusted R Square	0.647103
Standard Error	0.464017
Observations	862

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	7	341.443	48.77757	226.543908	8.4E-190
Residual	854	183.8763	0.215312		
Total	861	525.3193			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	5.892465	0.049757	118.4243	0	5.794805	5.990126	5.794805	5.990126
Age_last_funding_year	0.063167	0.007511	8.409642	1.7144E-16	0.048424	0.07791	0.048424	0.07791
Avg_participants	0.060108	0.011736	5.121692	3.7443E-07	0.037073	0.083143	0.037073	0.083143
Total_funding_rounds	0.183468	0.015501	11.83558	4.9205E-30	0.153043	0.213894	0.153043	0.213894
has_angel	-0.83419	0.041354	-20.1718	2.6451E-74	-0.91535	-0.75302	-0.91535	-0.75302
has_VC	-0.10213	0.038166	-2.67582	0.00759738	-0.17704	-0.02722	-0.17704	-0.02722
is_top500	0.299289	0.045526	6.573959	8.5211E-11	0.209932	0.388645	0.209932	0.388645
Relationships	0.015625	0.003024	5.166596	2.9691E-07	0.009689	0.021561	0.009689	0.021561

Our predictor equation:

Y(Total Logarithmic fund)*=5.89+0.06(Age_Last_Fund)+0.06(Avg_Partcipant)+.183(Total Funding Rounds)+0.0156(Relationships)+0.299(TOP 500) – 0.834(Angle Investor) – 0.102(VC).*

Or

Total Fund*= 10^(5.89+0.06(Age_Last_Fund)+0.06(Avg_Partcipant)+.183(Total Funding Rounds)+0.0156(Relationships)+0.299(TOP 500) – 0.834(Angle Investor) – 0.102(VC)).*

Where:

- Age_Last_Fund: Age of the company at the time of the last funding.
- Avg_Participants: Average number of participants.
- Total_Funding_Rounds: Total number of funding rounds.

- Relationships: Number of relationships.
- TOP_500: A binary variable (1 if the company is in the top 500, 0 otherwise).
- Angel_Investor: A binary variable (1 if the company has an angel investor, 0 otherwise).
- VC: A binary variable (1 if the company has a venture capitalist, 0 otherwise).

Interpretation

Positive Coefficients:

- Age_Last_Fund (0.06): For each additional year of the company's age, the logarithmic total fund is expected to increase by 6%.
- Avg_Participants (0.06): For each additional participant, the logarithmic total fund is expected to increase by 6%.
- Relationships: For each additional relationship, the logarithmic total fund is expected to increase by 1.6%.
- Total_Funding_Rounds (0.185): Each additional funding round is associated with an increase of 18% in the logarithmic total fund.
- TOP_500 (0.2975): If the company is in the top 500, it is expected to raise approximately 30% more in logarithmic total funds.

Negative Coefficients:

- **Angel_Investor (-0.845)**: Companies with an angel investors only are expected to have a logarithmic total funds that are approximately 84.5% lower than startups which raised later round of investment.
- **VC (-0.104)**: Companies with venture capitalists are expected to have a logarithmic total fund that is 10.4% less than startups with higher rounds of fund.

Insights:

With R^2 0.647, meaning model can explain/ predict 64% of the total fund using the 7 variables that we have. The regression model indicates that while some predictors have a measurable effect like top 500 and having angel investors only, others may have minimal impact or be influenced by factors not included in the dataset. Potential additional factors could include team size, milestones achieved, investor types, and the founding team's background. Further analysis could help clarify the effects of angel and VC investors on the total fund. For example, questions to explore include whether angel or VC investors impact the total fund differently at the early stages of a company (e.g., during Series A funding) compared to later stages or if their influence continues throughout the company's lifecycle.

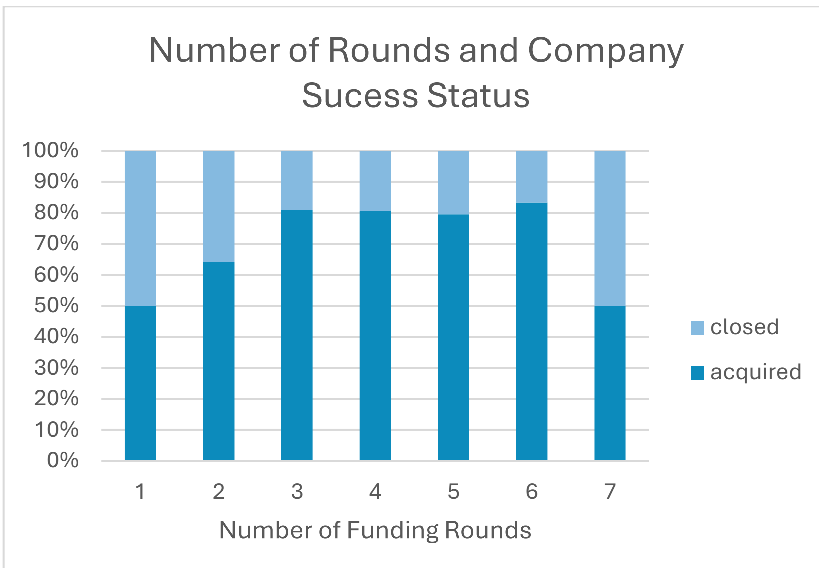
Overall Study Results

We used both descriptive (Odds Ratio, Chi-Square, ANOVA, MANOVA) and inferential (Correlation & Linear Regression) statistics to analyze the dataset and address our FINER research questions. Below is a summary of our findings for each question:

How the raised fund influence the success of the startup company and what factors have influenced the total fund?

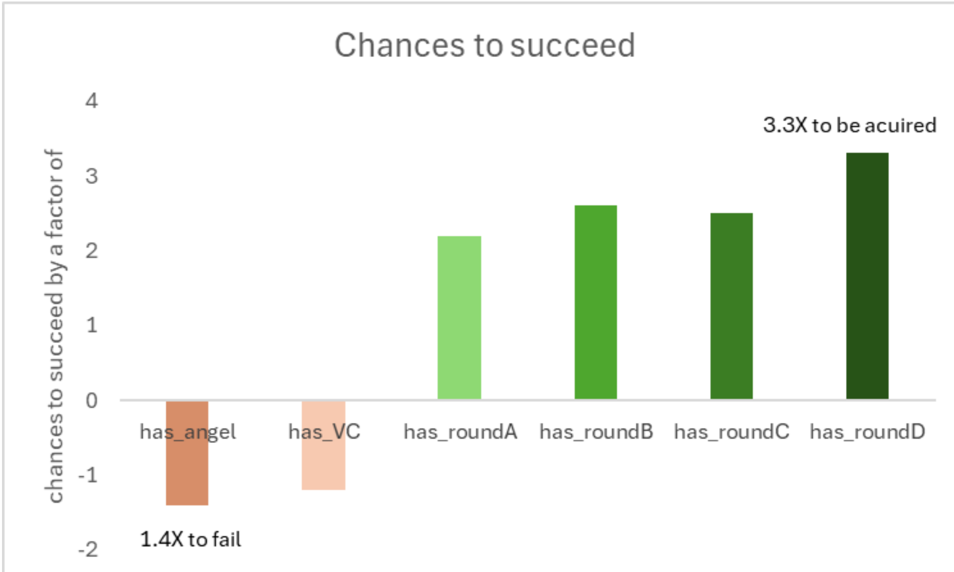
Using a t-test, we concluded that acquired companies tend to have higher average fundraising amounts, demonstrating a significant statistical difference based on company status. Consequently, we aim to investigate the factors influencing the increase in the average total funds that a startup can raise.

How did the total funds raised and the total number of rounds correlate with the startup's success?



We found a positive correlation of $r=0.4$ between the total rounds raised and the logarithmic total fund, suggesting that more funding rounds are associated with higher total funds, which aligns with intuitive expectations. Combining this with our previous finding of a significant difference between total funds and startup success, we can infer that additional rounds may lead to increased funding and potentially higher success rates. However, our linear regression model indicates that each additional funding round is associated with only an 18% increase in the logarithmic total fund, which may not represent a substantial impact. Other data set that includes the fund per round can be helpful to investigating the impact of the rounds on the total funds and the success of the company.

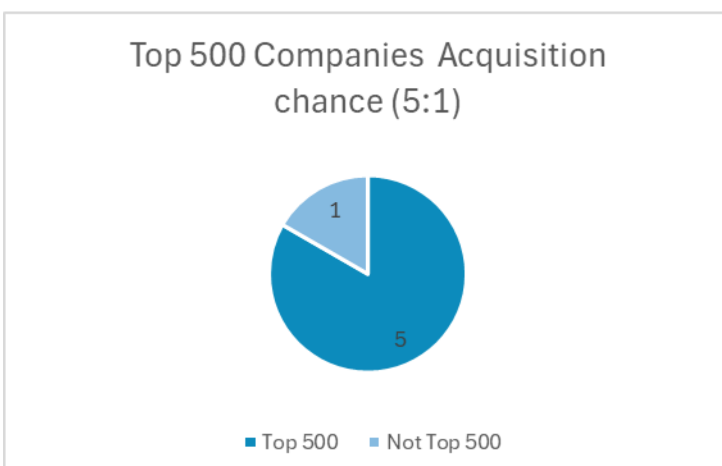
What fund round had the most influence on the company’s success?



The analysis suggest a potential positive association between raising higher rounds of investment (A, B, C, and D) and the success of the company in being acquired. For instance, startups that raised round D are **3.3 times more** likely to be acquired successfully.

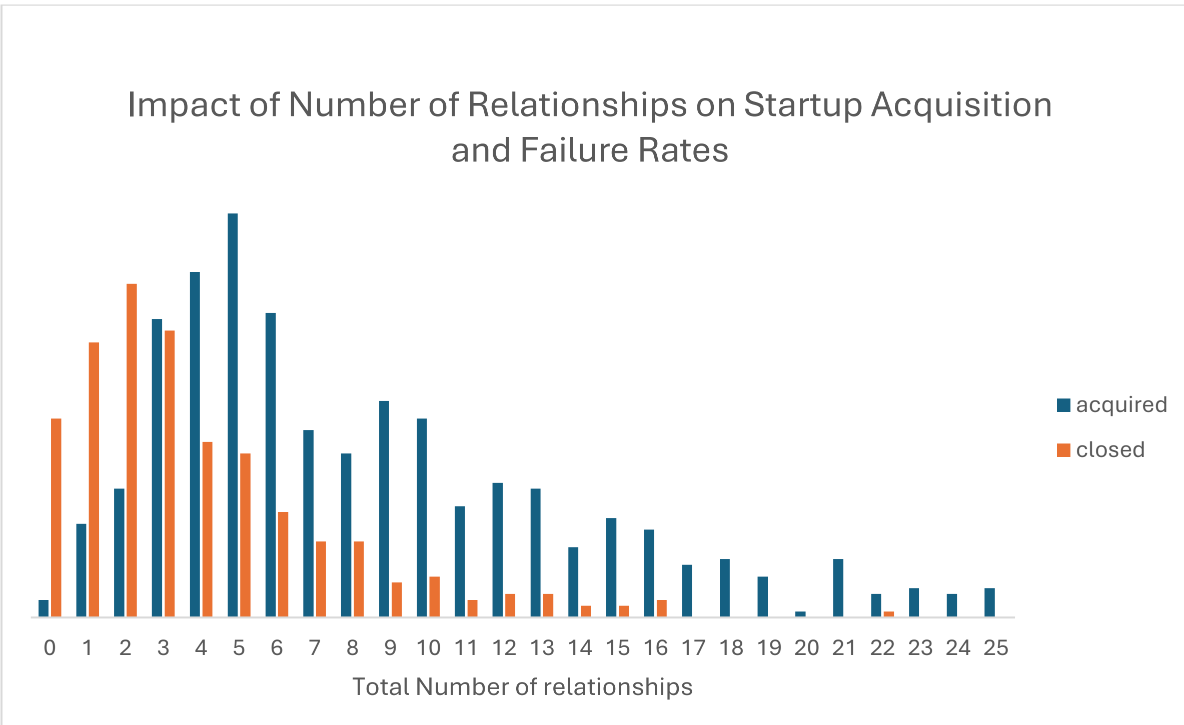
Moreover, The analysis indicates a negative association between angel and VC rounds and the company's success, which is not a favorable interpretation suggesting that companies that raised angel rounds are 1.4 times more likely to fail, whereas those that raised angel and/or VC rounds exclusively are 1.2 times more likely to fail.

How did being recognized in media lists like the top 500 startups affect the total fund?



The odds ratio calculations suggest that Startups which were listed as one of the top 500 are **5 times more** likely to be successfully acquired. Moreover, our regression model suggests that companies in top 500 list are expected to raise approximately 30% more in logarithmic total funds.

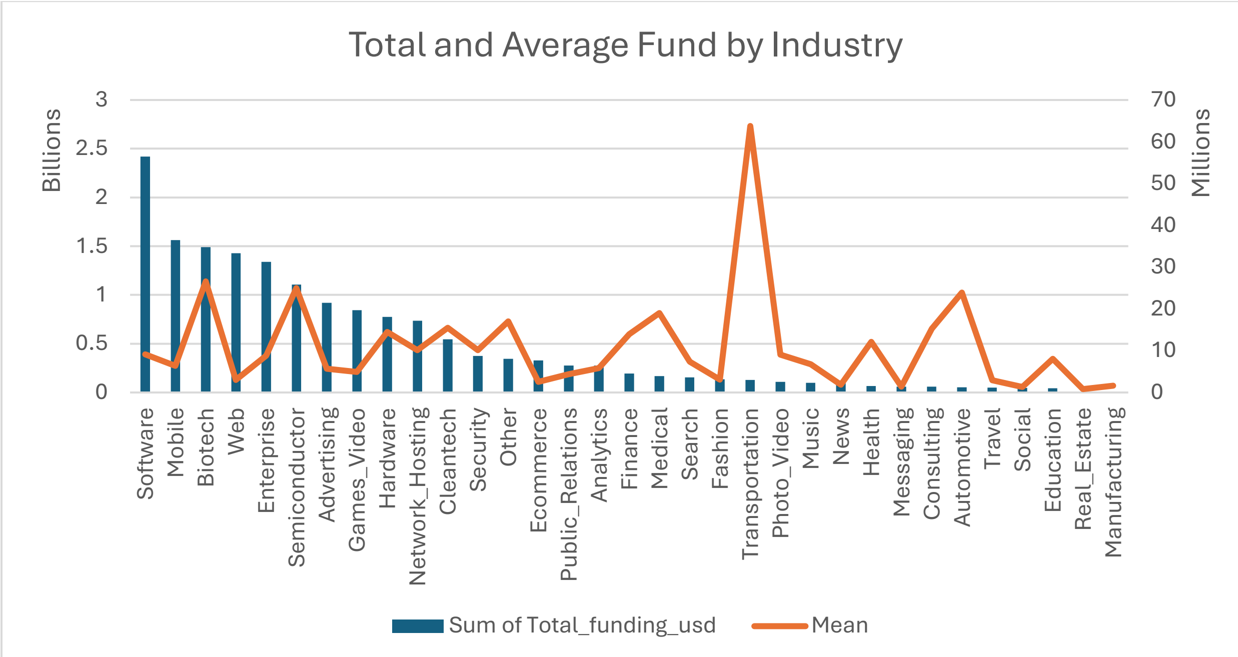
How did the number of relationships affect the total fund? Any correlations with the startup's success?



The positive t-value indicates that successful companies have on average, a higher number of relationships compared to closed companies. A weak correlation coefficient of 0.29 positive correlation between the total relationship and the logarithmic total fund. The total relationship has a very small coefficient to predict the logarithmic total fund of 1.6% increase for each additional relationship. This finding was surprising and additional investigation may be done to understand the relationship variable, types of relationships, number of relationships per round, or cluster analysis to see the actual effect.

How did the startup industry influence a startup's success and failure rates?

What industries had the highest fund rates?

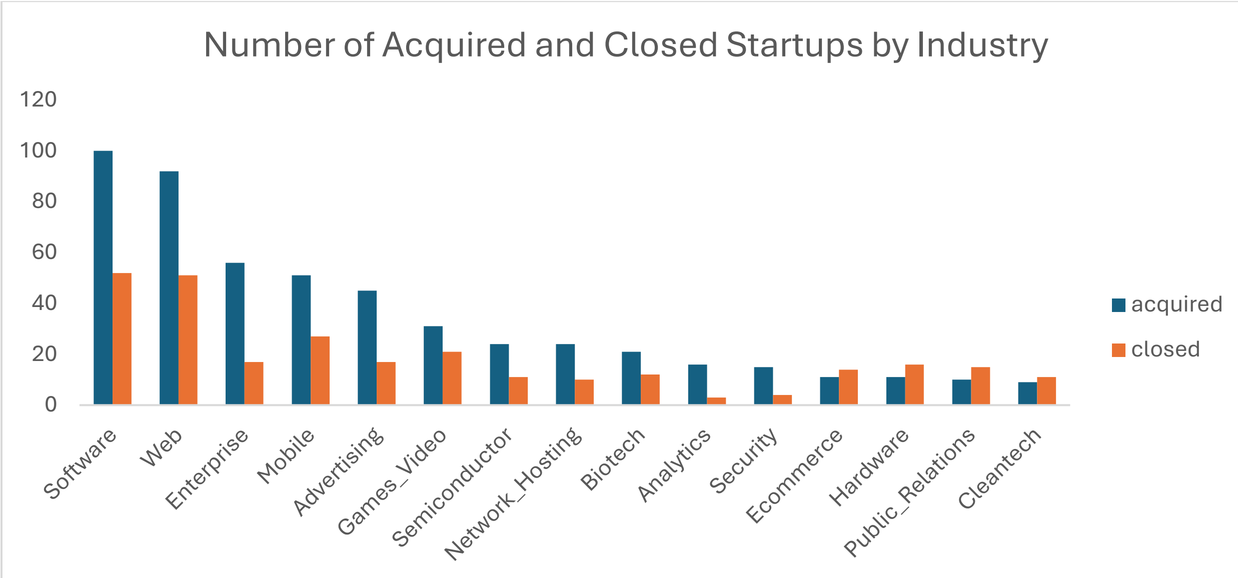


We found a *significant statistical difference of p-value of 1.5E-07* in the average total funding between startups in various industries. The highest average total investment amounts were observed in the Software, Web, Mobile, Enterprise, and Advertising sectors, indicating that these industries receive substantial funding.

However, the average total fund per company was higher in the Transportation, Biotech, Semiconductor, Automotive, and Medical industries.

Comparing the average fund for all industries we found that the statistical difference occurs due to startups in web industry receiving significantly less average fun compared to almost all other industries other than advertising, and gaming industries.

What industries had the highest success rate?



The chi-square value of 29.15 indicates a statistically significant difference in startup success rates across industries, with a p-value of 0.009954 at a 95% confidence level.

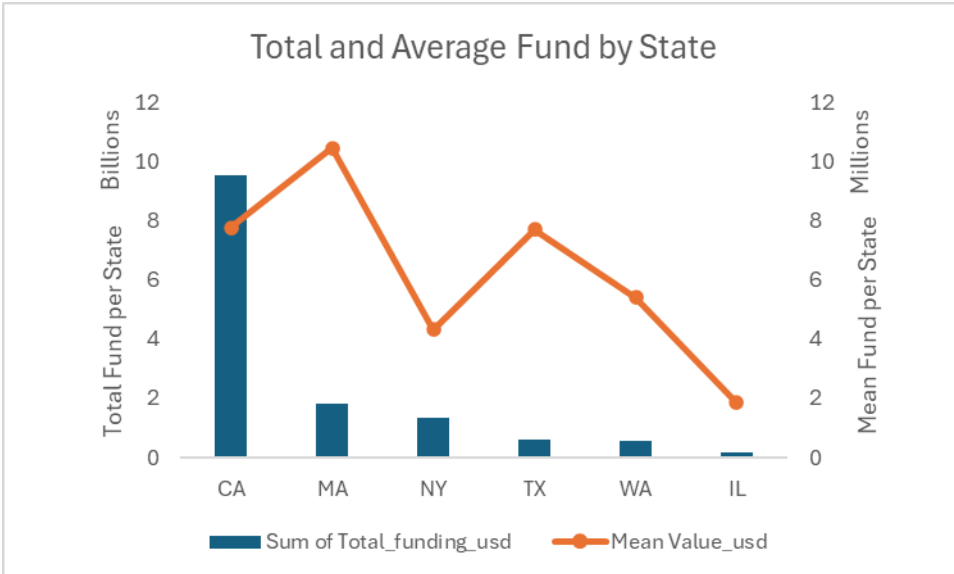
The frequency table reveals that the **Software**, **Web**, and **Mobile** industries have the highest numbers of both acquisitions and closures. Notably, the ratio of acquisitions to closures in these industries is almost 2:1, suggesting that success rates (acquisitions) are nearly double the failure rates (closures) within these sectors.

We may need a larger dataset or national data on the number of startups per industry to ensure our sample is representative.

Does a and b correlate together?

The analysis indicates that while higher investments in certain sectors appear to correlate with higher success rates for companies, this relationship is more reflective of industry trends rather than individual company performance. Specifically, industries with the highest average funding per company do not necessarily exhibit the highest success rates. This discrepancy may be attributed to other factors not covered in our analysis, such as the inherent complexity of the business and ideas. For example, medical companies often face more complexity compared to software companies.

How did the location of the startups correlate with the startup's success? and how the startup's location influenced the amount of funds the startups raised?



The chi-square value of 23.92 indicates a statistically significant difference in startup success rates by state, with a p-value of 0.0044 at a 95% confidence level.

Our dataset shows that the highest number of acquired startups are in California (CA), followed by New York (NY), Massachusetts (MA), and Texas (TX). This may reflect the influence of Silicon Valley or suggest that our sample may not fully represent startups across all states.

Additionally, there is a significant relationship between location and total funding, with a p-value of 7.44E-05. Massachusetts startups receive the highest average funding compared to other states, excluding NY and CA. Furthermore, California startups receive significantly more funding than those in Illinois (IL) and New York (NY).

Further Work & Additional data

Our analysis successfully addressed our FINER questions and identified key relationships, but some unexpected findings emerged regarding the influence of relationships, angel investors, and venture capital on startup funding and success. Notably, our predictive model explained only 64% of the variance in total funds, indicating that our predictions could be refined.

To enhance our analysis, we should incorporate additional data, such as:

- **Detailed Investment Round Information:** More specifics about each funding round and the sequential rounds effect.
- **Founding Team Characteristics:** Data on the education, experience, and age of founding team members.
- **Startup populations by state and industry:** this would allow us to explore whether specific industries in particular states show significant differences in funding and success rates, improving our understanding of these interactions.

Updating our dataset from sources like Crunchbase or similar platforms could provide deeper insights.

Conclusion

With a 95% confidence interval and a p-value of 4.23E-21 our study confirms that acquired companies significantly tend to have higher average funding compared to closed companies. Specifically, companies that raised a Series D round are 3.3 times more likely to be acquired, while those with angel investment are 1.4 times more likely to fail. Recognition as a top 500 startup markedly increases the likelihood of acquisition by a factor of 5 and contributes an increase of 30% to the total predicted funding. Additionally, location and industry play a critical role in startup success and funding, with Massachusetts and California startups showing higher average funding and acquisition rates. Conversely, startups in the Web industry receive significantly less average funding compared to other sectors. Age at the time of the last funding round shows a strong correlation, with each additional year predicting a 34.9% increase in total funding, provided the company has sufficient runway. However, our model accounts for only 64% of the variation in total funding, suggesting that incorporating additional data could significantly enhance prediction accuracy and provide deeper insights into factors affecting startup funding.

All relevant data, analyses, and results can be reviewed in the attached Excel workbook.

References

KC, M. (2020, September 20). *Startup Success Prediction*. Retrieved from Kaggle: https://www.kaggle.com/datasets/manishkc06/startup-success-prediction/data

Panthena, R. (2018). *Startup-Success-Prediction*. Retrieved from Github: https://github.com/RamkishanPanthena/Startup-Success-Prediction