Predicting Startup Success

Milestone 2: Data Exploration

Website:

https://wihi1131.github.io/Data-Mining-Project/Data%20Exploration

GitHub:

https://github.com/WiHi1131/Data-Mining-Project

Michael Van Vuuren, Melina Kopischkie, William Hinkley

Abstract

Investors have a lot to consider when investing in a startup. Our aim is to use data mining to create predictive models that simplify this process. Accurate models require a large volume of data. In Milestone 2, we have scraped, cleaned, integrated, and explored the data we need to build our models. We discuss how we scraped, cleaned, integrated our data, then how we used exploratory data analysis to visualize it. In total, we have created one primary dataset and two subsets of that primary dataset. In the next milestone, we will use the appropriate data for our models.

Data Collection

Our data was scraped from two different websites. First, we checked *robots.txt* for permission, then slowly scraped each website to avoid overwhelming their servers. The scraping process went as follows: A request was sent to the webpage to obtain an HTML/JavaScript document, then the document was parsed and searched using Beautiful Soup 4, then the information we needed was extracted using regular expression matchers, and finally the information was stored, then written to a CSV file. To obtain URLs to scrape, we consulted *robots.txt* for sitemaps, XML documents containing URLs. The first website we scraped was OurCrowd, which gave us a base list of URLs. Using this base list of URLs, we scraped CB Insights as well.

The first website was OurCrowd, which for each startup has information like name, website, industries, a summary of what the startup is about, and a sentiment score with a collection of articles. Here is an example of a page that was scraped for a startup called Ro that does online physician screenings: https://www.ourcrowd.com/companies/ro.

The second website, CB Insights, has two tabs, both with pertinent information. The first tab we scraped was the financials tab. This contains information about investors, funding totals, most recent fundings, and for some startups valuations and revenue amounts. Here is an example webpage of the financials of Ro on CB Insights: https://www.cbinsights.com/company/roman-health-ventures/financials. The second tab we scraped was the overview tab, which contained information about founding years, mosaic score changes, a description, and the location for each startup. Here is an example webpage of the overview of Ro on CB Insights: https://www.cbinsights.com/company/roman-health-ventures.

Cleaning

Three datasets: financial, profile, and overview were gathered via webscraping. Cleaning the datasets was a two-step process. First, we imputed and dropped missing values, then we converted each initial data type to its appropriate type.

For imputation, we started by viewing how many null values each column had. If the column was a numerical type and the null values made up less than 25% of the total rows, we would impute the missing values. Our datasets had 5000-7000 rows each, so we imputed the values of a lot of columns. We plotted the distribution of each column and used median if it was moderately-to-strongly skewed to negate outliers and skewness. Otherwise, we would use mean or mode. In most cases the distributions were skewed.

For type conversion, we started by viewing the types of each column, then converted *object* types to more narrow types. For example, categorical types such as funding type were converted into categories to improve space efficiency. Integer types such as investor count were changed to *Int64* types from *float64* types. With date and time information, we formatted these columns for ease of use. We converted these to *datetime* types, and then stored them as formatted *strings* (which are stored as *objects* but now properly formatted).

A step-by-step explanation of the cleaning process can be found in *cleaning.ipynb* under the *cleaning/* directory in our repository: https://github.com/WiHi1131/Data-Mining-Project/.

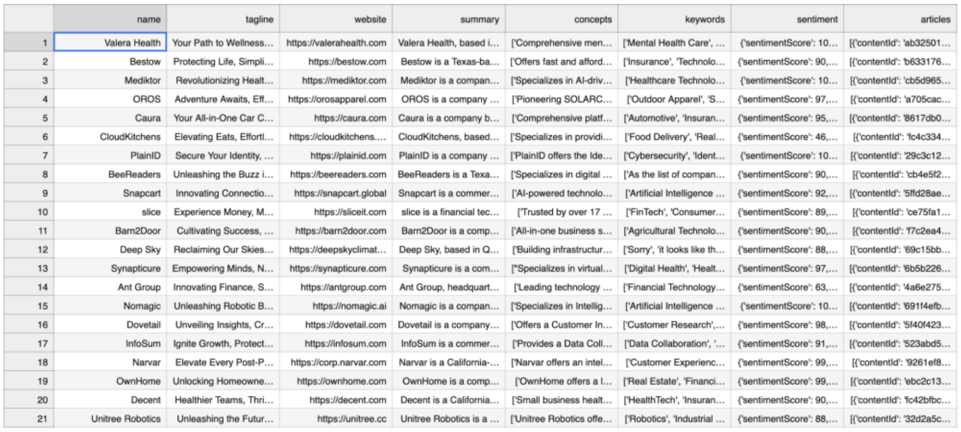
Integration

After scraping and cleaning, we had three datasets. While scraping each website, we collected the homepage URL of the startup being scraped, so the integration process of the cleaned datasets was straightforward. We performed an inner join on the profile dataset with the financial dataset on the column *website*. Then, we performed another inner join using the column *website* on that combined dataset and the overview dataset. In the end, we created a primary dataset called *primary.csv* under the *clean-datasets/* directory. After integrating our cleaned datasets to create a primary dataset, we decided to create two mini-datasets also under *clean-datasets/.* One contains the startups from primary with complete revenue information, and the other contains startups with complete valuation information. Revenue and valuation data were sparse in the primary dataset but may be useful for model building. Finally, we dropped the revenue and valuation columns from the primary dataset.

Datasets before and after cleaning and integration

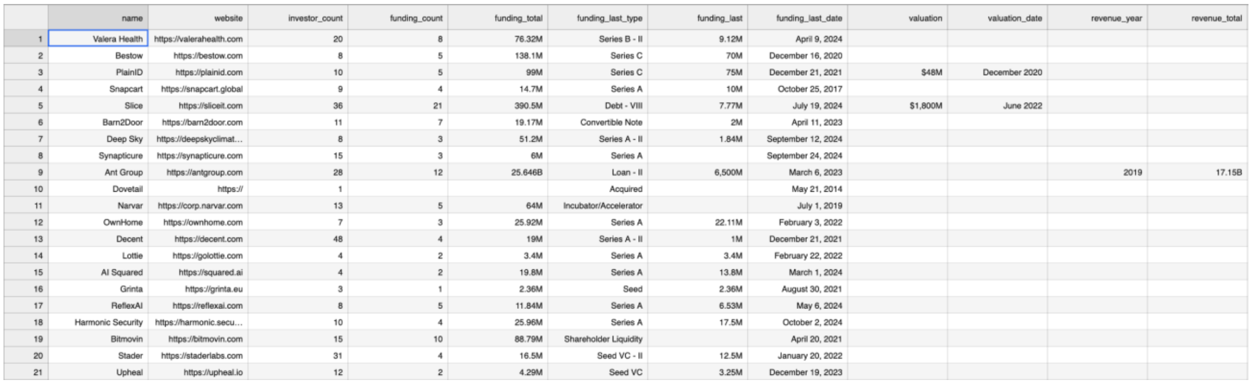
Before:

profiles.csv



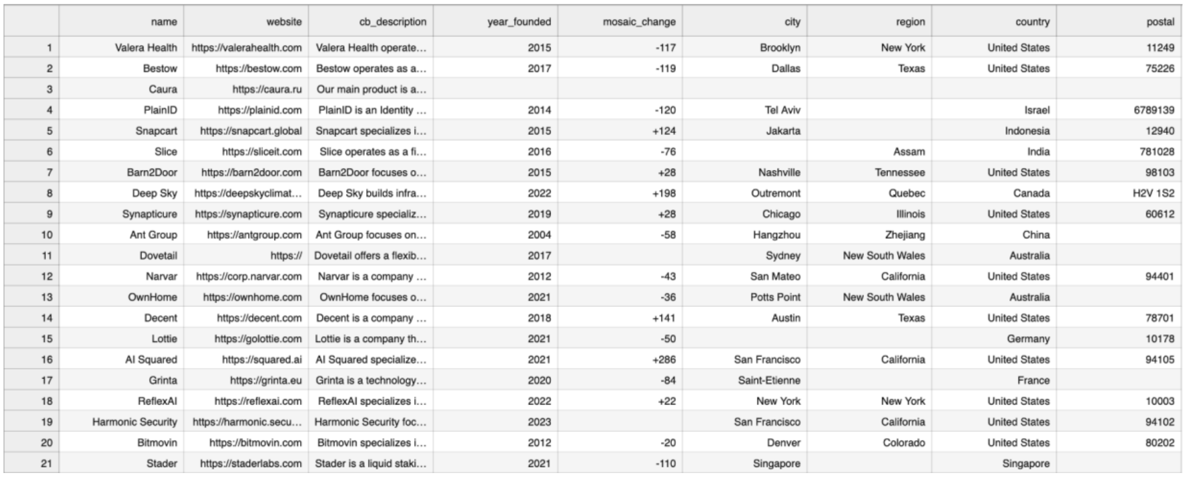
	name	tagline	website	summary	concepts	keywords	sentiment	articles
1	Valera Health	Your Path to Wellness…	https://valerahealth.com	Valera Health, based i…	['Comprehensive men…	['Mental Health Care', …	{'sentimentScore': 10…	[{'contentId': 'ab32501…
2	Bestow	Protecting Life, Simpli…	https://bestow.com	Bestow is a Texas-ba…	['Offers fast and afford…	['Insurance', 'Technolo…	{'sentimentScore': 90,…	[{'contentId': 'b633176…
3	Mediktor	Revolutionizing Healt…	https://mediktor.com	Mediktor is a compan…	['Specializes in AI-driv…	['Healthcare Technolo…	{'sentimentScore': 10…	[{'contentId': 'cb5d965…
4	OROS	Adventure Awaits, Eff…	https://orosapparel.com	OROS is a company …	['Pioneering SOLARC…	['Outdoor Apparel', 'S…	{'sentimentScore': 97,…	[{'contentId': 'a705cac…
5	Caura	Your All-in-One Car C…	https://caura.com	Caura is a company b…	['Comprehensive platf…	['Automotive', 'Insuran…	{'sentimentScore': 95,…	[{'contentId': '8617db0…
6	CloudKitchens	Elevating Eats, Effortl…	https://cloudkitchens.…	CloudKitchens, based…	['Specializes in providi…	['Food Delivery', 'Real…	{'sentimentScore': 46,…	[{'contentId': 'fc4c334…
7	PlainID	Secure Your Identity, …	https://plainid.com	PlainID is a company …	['PlainID offers the Ide…	['Cybersecurity', 'Ident…	{'sentimentScore': 10…	[{'contentId': '29c3c12…
8	BeeReaders	Unleashing the Buzz i…	https://beereaders.com	BeeReaders is a Texa…	['Specializes in digital …	['As the list of compan…	{'sentimentScore': 90,…	[{'contentId': 'cb4e5f2…
9	Snapcart	Innovating Connectio…	https://snapcart.global	Snapcart is a commer…	['AI-powered technolo…	['Artificial Intelligence …	{'sentimentScore': 92,…	[{'contentId': '5ffd28ae…
10	slice	Experience Money, M…	https://sliceit.com	slice is a financial tec…	['Trusted by over 17 …	['FinTech', 'Consumer…	{'sentimentScore': 89,…	[{'contentId': 'ce75fa1…
11	Barn2Door	Cultivating Success, …	https://barn2door.com	Barn2Door is a comp…	['All-in-one business s…	['Agricultural Technolo…	{'sentimentScore': 90,…	[{'contentId': 'f7c2ea4…
12	Deep Sky	Reclaiming Our Skies…	https://deepskyclimat…	Deep Sky, based in Q…	['Building infrastructur…	['Sorry', 'it looks like th…	{'sentimentScore': 88,…	[{'contentId': '69c15bb…
13	Synapticure	Empowering Minds, N…	https://synapticure.com	Synapticure is a com…	["Specializes in virtual…	['Digital Health', 'Healt…	{'sentimentScore': 97,…	[{'contentId': '6b5b226…
14	Ant Group	Innovating Finance, S…	https://antgroup.com	Ant Group, headquart…	['Leading technology …	['Financial Technology…	{'sentimentScore': 63,…	[{'contentId': '4a6e275…
15	Nomagic	Unleashing Robotic B…	https://nomagic.ai	Nomagic is a compan…	['Specializes in Intellig…	['Artificial Intelligence …	{'sentimentScore': 10…	[{'contentId': '691f4efb…
16	Dovetail	Unveiling Insights, Cr…	https://dovetail.com	Dovetail is a company…	['Offers a Customer In…	['Customer Research',…	{'sentimentScore': 98,…	[{'contentId': '5f40f423…
17	InfoSum	Ignite Growth, Protect…	https://infosum.com	InfoSum is a commer…	['Provides a Data Coll…	['Data Collaboration', '…	{'sentimentScore': 91,…	[{'contentId': '523abd5…
18	Narvar	Elevate Every Post-P…	https://corp.narvar.com	Narvar is a California-…	['Narvar offers an intel…	['Customer Experienc…	{'sentimentScore': 99,…	[{'contentId': '9261ef8…
19	OwnHome	Unlocking Homeowne…	https://ownhome.com	OwnHome is a comp…	['OwnHome offers a l…	['Real Estate', 'Financi…	{'sentimentScore': 99,…	[{'contentId': 'ebc2c13…
20	Decent	Healthier Teams, Thri…	https://decent.com	Decent is a California…	['Small business healt…	['HealthTech', 'Insuran…	{'sentimentScore': 90,…	[{'contentId': 'fc42bfbc…
21	Unitree Robotics	Unleashing the Futur…	https://unitree.cc	Unitree Robotics is a …	['Unitree Robotics offe…	['Robotics', 'Industrial …	{'sentimentScore': 88,…	[{'contentId': '32d2a5c…

financials.csv



	name	website	investor_count	funding_count	funding_total	funding_last_type	funding_last	funding_last_date	valuation	valuation_date	revenue_year	revenue_total
1	Valera Health	https://valerahealth.com	20	8	76.32M	Series B - II	9.12M	April 9, 2024				
2	Bestow	https://bestow.com	8	5	138.1M	Series C	70M	December 16, 2020				
3	PlainID	https://plainid.com	10	5	99M	Series C	75M	December 21, 2021	$48M	December 2020		
4	Snapcart	https://snapcart.global	9	4	14.7M	Series A	10M	October 25, 2017				
5	Slice	https://sliceit.com	36	21	390.5M	Debt - VIII	7.77M	July 19, 2024	$1,800M	June 2022		
6	Barn2Door	https://barn2door.com	11	7	19.17M	Convertible Note	2M	April 11, 2023				
7	Deep Sky	https://deepskyclimat…	8	3	51.2M	Series A - II	1.84M	September 12, 2024				
8	Synapticure	https://synapticure.com	15	3	6M	Series A		September 24, 2024				
9	Ant Group	https://antgroup.com	28	12	25.646B	Loan - II	6,500M	March 6, 2023			2019	17.15B
10	Dovetail	https://	1			Acquired		May 21, 2014				
11	Narvar	https://corp.narvar.com	13	5	64M	Incubator/Accelerator		July 1, 2019				
12	OwnHome	https://ownhome.com	7	3	25.92M	Series A	22.11M	February 3, 2022				
13	Decent	https://decent.com	48	4	19M	Series A - II	1M	December 21, 2021				
14	Lottie	https://golottie.com	4	2	3.4M	Series A	3.4M	February 22, 2022				
15	AI Squared	https://squared.ai	4	2	19.8M	Series A	13.8M	March 1, 2024				
16	Grinta	https://grinta.eu	3	1	2.36M	Seed	2.36M	August 30, 2021				
17	ReflexAI	https://reflexai.com	8	5	11.84M	Series A	6.53M	May 6, 2024				
18	Harmonic Security	https://harmonic.secu…	10	4	25.96M	Series A	17.5M	October 2, 2024				
19	Bitmovin	https://bitmovin.com	15	10	88.79M	Shareholder Liquidity		April 20, 2021				
20	Stader	https://staderlabs.com	31	4	16.5M	Seed VC - II	12.5M	January 20, 2022				
21	Upheal	https://upheal.io	12	2	4.29M	Seed VC	3.25M	December 19, 2023				

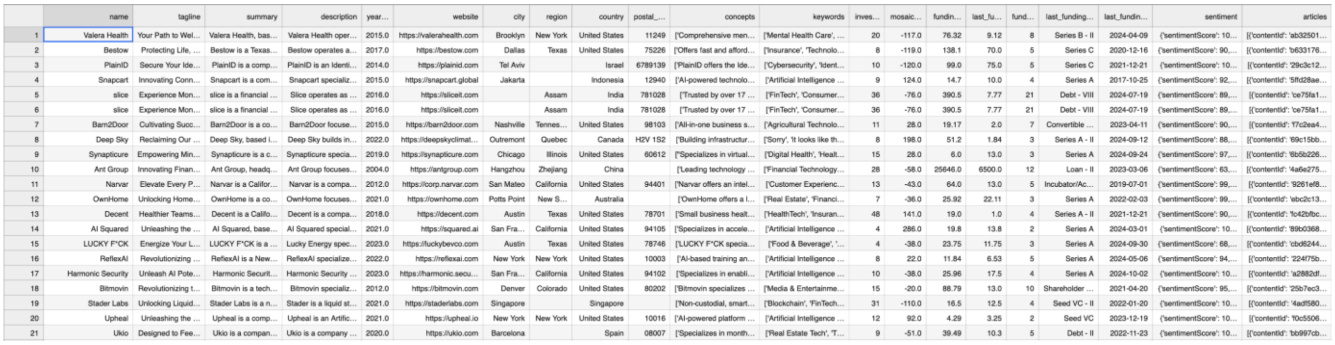
overviews.csv



	name	website	cb_description	year_founded	mosaic_change	city	region	country	postal
1	Valera Health	https://valerahealth.com	Valera Health operate…	2015	-117	Brooklyn	New York	United States	11249
2	Bestow	https://bestow.com	Bestow operates as a…	2017	-119	Dallas	Texas	United States	75226
3	Caura	https://caura.ru	Our main product is a…						
4	PlainID	https://plainid.com	PlainID is an Identity …	2014	-120	Tel Aviv		Israel	6789139
5	Snapcart	https://snapcart.global	Snapcart specializes i…	2015	+124	Jakarta		Indonesia	12940
6	Slice	https://sliceit.com	Slice operates as a fi…	2016	-76		Assam	India	781028
7	Barn2Door	https://barn2door.com	Barn2Door focuses o…	2015	+28	Nashville	Tennessee	United States	98103
8	Deep Sky	https://deepskyclimat…	Deep Sky builds infra…	2022	+198	Outremont	Quebec	Canada	H2V 1S2
9	Synapticure	https://synapticure.com	Synapticure specializ…	2019	+28	Chicago	Illinois	United States	60612
10	Ant Group	https://antgroup.com	Ant Group focuses on…	2004	-58	Hangzhou	Zhejiang	China	
11	Dovetail	https://	Dovetail offers a flexib…	2017		Sydney	New South Wales	Australia	
12	Narvar	https://corp.narvar.com	Narvar is a company …	2012	-43	San Mateo	California	United States	94401
13	OwnHome	https://ownhome.com	OwnHome focuses o…	2021	-36	Potts Point	New South Wales	Australia	
14	Decent	https://decent.com	Decent is a company …	2018	+141	Austin	Texas	United States	78701
15	Lottie	https://golottie.com	Lottie is a company th…	2021	-50			Germany	10178
16	AI Squared	https://squared.ai	AI Squared specialize…	2021	+286	San Francisco	California	United States	94105
17	Grinta	https://grinta.eu	Grinta is a technology…	2020	-84	Saint-Etienne		France	
18	ReflexAI	https://reflexai.com	ReflexAI specializes i…	2022	+22	New York	New York	United States	10003
19	Harmonic Security	https://harmonic.secu…	Harmonic Security foc…	2023		San Francisco	California	United States	94102
20	Bitmovin	https://bitmovin.com	Bitmovin specializes i…	2012	-20	Denver	Colorado	United States	80202
21	Stader	https://staderlabs.com	Stader is a liquid staki…	2021	-110	Singapore		Singapore	

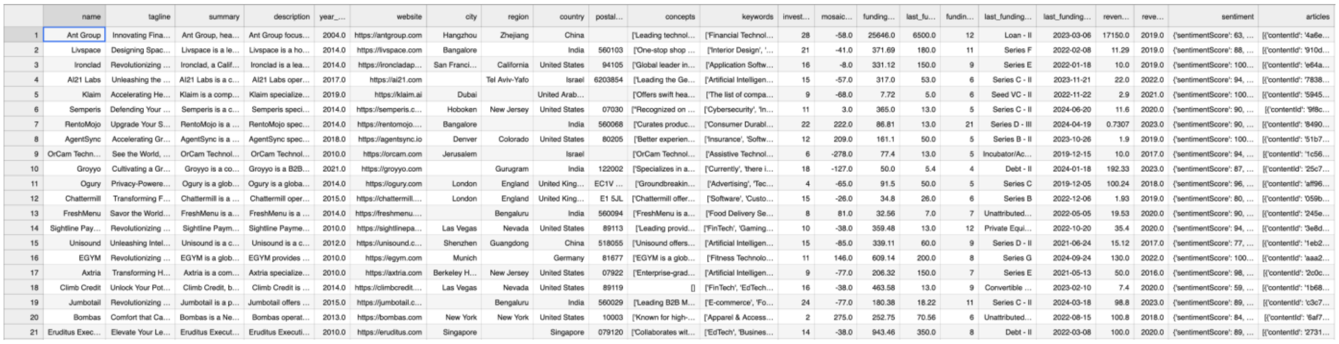
After:

primary.csv



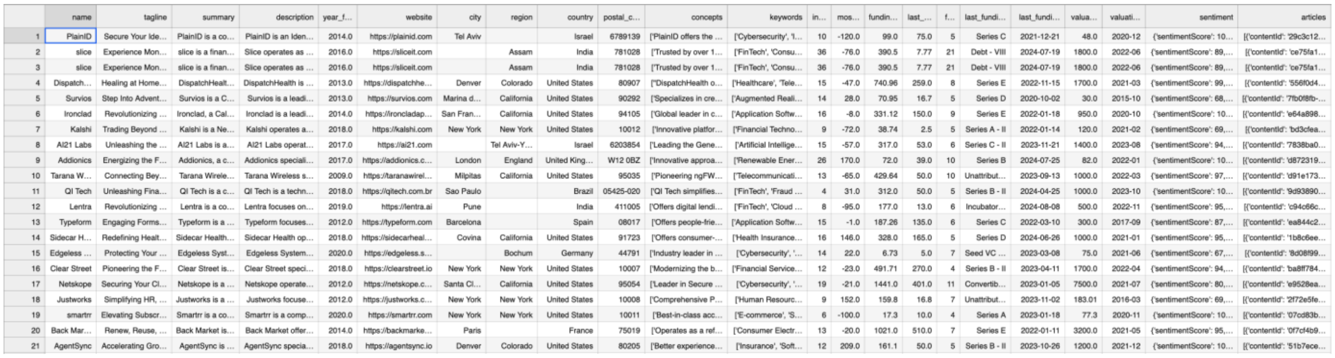
	name	tagline	summary	description	year...	website	city	region	country	postal_...	concepts	keywords	inves...	mosaic...	fundin...	last_fu...	fund...	last_funding...	last_fundin...	sentiment	articles
1	Valera Health	Your Path to Wel...	Valera Health, bas...	Valera Health oper...	2015.0	https://valerahealth.com	Brooklyn	New York	United States	11249	['Comprehensive men...	['Mental Health Care', ...	20	-117.0	76.32	9.12	8	Series B - II	2024-04-09	{'sentimentScore': 10...	[{'contentId': 'ab32501...
2	Bestow	Protecting Life, ...	Bestow is a Texas...	Bestow operates a...	2017.0	https://bestow.com	Dallas	Texas	United States	75226	['Offers fast and afford...	['Insurance', 'Technolo...	8	-119.0	138.1	70.0	5	Series C	2020-12-16	{'sentimentScore': 90,...	[{'contentId': 'b633176...
3	PlainID	Secure Your Ide...	PlainID is a comp...	PlainID is an Identi...	2014.0	https://plainid.com	Tel Aviv		Israel	6789139	['PlainID offers the Ide...	['Cybersecurity', 'Ident...	10	-120.0	99.0	75.0	5	Series C	2021-12-21	{'sentimentScore': 10...	[{'contentId': '29c3c12...
4	Snapcart	Innovating Conn...	Snapcart is a com...	Snapcart specializ...	2015.0	https://snapcart.global	Jakarta		Indonesia	12940	['AI-powered technolo...	['Artificial Intelligence ...	9	124.0	14.7	10.0	4	Series A	2017-10-25	{'sentimentScore': 92,...	[{'contentId': '5ffd28ae...
5	slice	Experience Mon...	slice is a financial ...	Slice operates as ...	2016.0	https://sliceit.com		Assam	India	781028	['Trusted by over 17 ...	['FinTech', 'Consumer...	36	-76.0	390.5	7.77	21	Debt - VIII	2024-07-19	{'sentimentScore': 89,...	[{'contentId': 'ce75fa1...
6	slice	Experience Mon...	slice is a financial ...	Slice operates as ...	2016.0	https://sliceit.com		Assam	India	781028	['Trusted by over 17 ...	['FinTech', 'Consumer...	36	-76.0	390.5	7.77	21	Debt - VIII	2024-07-19	{'sentimentScore': 89,...	[{'contentId': 'ce75fa1...
7	Barn2Door	Cultivating Succ...	Barn2Door is a co...	Barn2Door focuse...	2015.0	https://barn2door.com	Nashville	Tennes...	United States	98103	['All-in-one business s...	['Agricultural Technolo...	11	28.0	19.17	2.0	7	Convertible ...	2023-04-11	{'sentimentScore': 90,...	[{'contentId': 'f7c2ea4...
8	Deep Sky	Reclaiming Our ...	Deep Sky, based i...	Deep Sky builds in...	2022.0	https://deepskyclimat...	Outremont	Quebec	Canada	H2V 1S2	['Building infrastructur...	['Sorry', 'It looks like th...	8	198.0	51.2	1.84	3	Series A - II	2024-09-12	{'sentimentScore': 88,...	[{'contentId': '69c15bb...
9	Synapticure	Empowering Min...	Synapticure is a c...	Synapticure specia...	2019.0	https://synapticure.com	Chicago	Illinois	United States	60612	["Specializes in virtual...	['Digital Health', 'Healt...	15	28.0	6.0	13.0	3	Series A	2024-09-24	{'sentimentScore': 97,...	[{'contentId': '6b5b226...
10	Ant Group	Innovating Finan...	Ant Group, headq...	Ant Group focuses...	2004.0	https://antgroup.com	Hangzhou	Zhejiang	China		['Leading technology ...	['Financial Technology...	28	-58.0	25646.0	6500.0	12	Loan - II	2023-03-06	{'sentimentScore': 63,...	[{'contentId': '4a6e275...
11	Narvar	Elevate Every P...	Narvar is a Califor...	Narvar is a compa...	2012.0	https://corp.narvar.com	San Mateo	California	United States	94401	['Narvar offers an intel...	['Customer Experienc...	13	-43.0	64.0	13.0	5	Incubator/Ac...	2019-07-01	{'sentimentScore': 99,...	[{'contentId': '9261ef8...
12	OwnHome	Unlocking Home...	OwnHome is a co...	OwnHome focuses...	2021.0	https://ownhome.com	Potts Point	New S...	Australia		['OwnHome offers a l...	['Real Estate', 'Financi...	7	-36.0	25.92	22.11	3	Series A	2022-02-03	{'sentimentScore': 99,...	[{'contentId': 'ebc2c13...
13	Decent	Healthier Teams...	Decent is a Califo...	Decent is a compa...	2018.0	https://decent.com	Austin	Texas	United States	78701	['Small business healt...	['HealthTech', 'Insuran...	48	141.0	19.0	1.0	4	Series A - II	2021-12-21	{'sentimentScore': 90,...	[{'contentId': 'fc42bfbc...
14	AI Squared	Unleashing the ...	AI Squared, base...	AI Squared special...	2021.0	https://squared.ai	San Fra...	California	United States	94105	['Specializes in accele...	['Artificial Intelligence ...	4	286.0	19.8	13.8	2	Series A	2024-03-01	{'sentimentScore': 10...	[{'contentId': '89b0368...
15	LUCKY F*CK	Energize Your L...	LUCKY F*CK is a ...	Lucky Energy spec...	2023.0	https://luckybevco.com	Austin	Texas	United States	78746	['LUCKY F*CK specia...	['Food & Beverage', '...	4	-38.0	23.75	11.75	3	Series A	2024-09-30	{'sentimentScore': 68,...	[{'contentId': 'cbd6244...
16	ReflexAI	Revolutionizing ...	ReflexAI is a New...	ReflexAI specialize...	2022.0	https://reflexai.com	New York	New York	United States	10003	['AI-based training an...	['Artificial Intelligence ...	8	22.0	11.84	6.53	5	Series A	2024-05-06	{'sentimentScore': 94,...	[{'contentId': '224f75b...
17	Harmonic Security	Unleash AI Pote...	Harmonic Securit...	Harmonic Security ...	2023.0	https://harmonic.secu...	San Fra...	California	United States	94102	['Specializes in enabli...	['Artificial Intelligence ...	10	-38.0	25.96	17.5	4	Series A	2024-10-02	{'sentimentScore': 10...	[{'contentId': 'a2882df...
18	Bitmovin	Revolutionizing t...	Bitmovin is a tech...	Bitmovin specializ...	2012.0	https://bitmovin.com	Denver	Colorado	United States	80202	['Bitmovin specializes ...	['Media & Entertainme...	15	-20.0	88.79	13.0	10	Shareholder ...	2021-04-20	{'sentimentScore': 95,...	[{'contentId': '25b7ec3...
19	Stader Labs	Unlocking Liquid...	Stader Labs is a n...	Stader is a liquid st...	2021.0	https://staderlabs.com	Singapore		Singapore		['Non-custodial, smart...	['Blockchain', 'FinTech...	31	-110.0	16.5	12.5	4	Seed VC - II	2022-01-20	{'sentimentScore': 10...	[{'contentId': '4adf580...
20	Upheal	Unleashing the ...	Upheal is a comp...	Upheal is an Artific...	2021.0	https://upheal.io	New York	New York	United States	10016	['AI-powered platform ...	['Artificial Intelligence ...	12	92.0	4.29	3.25	2	Seed VC	2023-12-19	{'sentimentScore': 10...	[{'contentId': 'f0c5506...
21	Ukio	Designed to Fee...	Ukio is a compan...	Ukio is a company ...	2020.0	https://ukio.com	Barcelona		Spain	08007	['Specializes in month...	['Real Estate Tech', 'T...	9	-51.0	39.49	10.3	5	Debt - II	2022-11-23	{'sentimentScore': 10...	[{'contentId': 'bb997cb...

primary-with-revenue.csv



	name	tagline	summary	description	year_...	website	city	region	country	postal...	concepts	keywords	invest...	mosaic...	funding...	last_fu...	fundin...	last_funding...	last_funding...	reven...	reve...	sentiment	articles
1	Ant Group	Innovating Fina...	Ant Group, hea...	Ant Group focus...	2004.0	https://antgroup.com	Hangzhou	Zhejiang	China		['Leading technol...	['Financial Technol...	28	-58.0	25646.0	6500.0	12	Loan - II	2023-03-06	17150.0	2019.0	{'sentimentScore': 63, ...	[{'contentId': '4a6e...
2	Livspace	Designing Spac...	Livspace is a le...	Livspace is a ho...	2014.0	https://livspace.com	Bangalore		India	560103	['One-stop shop ...	['Interior Design', '...	21	-41.0	371.69	180.0	11	Series F	2022-02-08	11.29	2019.0	{'sentimentScore': 88, ...	[{'contentId': '910d...
3	Ironclad	Revolutionizing ...	Ironclad, a Calif...	Ironclad is a lea...	2014.0	https://ironcladap...	San Franci...	California	United States	94105	['Global leader in...	['Application Softw...	16	-8.0	331.12	150.0	9	Series E	2022-01-18	10.0	2019.0	{'sentimentScore': 100...	[{'contentId': 'e64a...
4	AI21 Labs	Unleashing the ...	AI21 Labs is a c...	AI21 Labs oper...	2017.0	https://ai21.com		Tel Aviv-Yafo	Israel	6203854	['Leading the Ge...	['Artificial Intelligen...	15	-57.0	317.0	53.0	6	Series C - II	2023-11-21	22.0	2022.0	{'sentimentScore': 94, ...	[{'contentId': '7838...
5	Klaim	Accelerating He...	Klaim is a comp...	Klaim specialize...	2019.0	https://klaim.ai	Dubai		United Arab...		['Offers swift hea...	['The list of compa...	9	-68.0	7.72	5.0	6	Seed VC - II	2022-11-22	2.9	2021.0	{'sentimentScore': 100...	[{'contentId': '5945...
6	Semperis	Defending Your ...	Semperis is a c...	Semperis speci...	2014.0	https://semperis.c...	Hoboken	New Jersey	United States	07030	["Recognized on ...	['Cybersecurity', 'In...	11	3.0	365.0	13.0	5	Series C - II	2024-06-20	11.6	2020.0	{'sentimentScore': 90, ...	[{'contentId': '9f8c...
7	RentoMojo	Upgrade Your S...	RentoMojo is a ...	RentoMojo spec...	2014.0	https://rentomojo....	Bangalore		India	560068	['Curates produc...	['Consumer Durabl...	22	222.0	86.81	13.0	21	Series D - III	2024-04-19	0.7307	2023.0	{'sentimentScore': 90, ...	[{'contentId': '8490...
8	AgentSync	Accelerating Gr...	AgentSync is a ...	AgentSync spec...	2018.0	https://agentsync.io	Denver	Colorado	United States	80205	['Better experien...	['Insurance', 'Softw...	12	209.0	161.1	50.0	5	Series B - II	2023-10-26	1.9	2019.0	{'sentimentScore': 100...	[{'contentId': '51b7...
9	OrCam Techn...	See the World, ...	OrCam Technol...	OrCam Technol...	2010.0	https://orcam.com	Jerusalem		Israel		['OrCam Technol...	['Assistive Technol...	6	-278.0	77.4	13.0	5	Incubator/Ac...	2019-12-15	10.0	2017.0	{'sentimentScore': 94, ...	[{'contentId': '1c56...
10	Groyyo	Cultivating a Gr...	Groyyo is a co...	Groyyo is a B2B...	2021.0	https://groyyo.com		Gurugram	India	122002	['Specializes in a...	['Currently', 'there i...	18	-127.0	50.0	5.4	4	Debt - II	2024-01-18	192.33	2023.0	{'sentimentScore': 87, ...	[{'contentId': '25c7...
11	Ogury	Privacy-Powere...	Ogury is a glob...	Ogury is a globa...	2014.0	https://ogury.com	London	England	United King...	EC1V ...	['Groundbreakin...	['Advertising', 'Tec...	4	-65.0	91.5	50.0	5	Series C	2019-12-05	100.24	2018.0	{'sentimentScore': 96, ...	[{'contentId': 'aff96...
12	Chattermill	Transforming F...	Chattermill is a ...	Chattermill oper...	2015.0	https://chattermill....	London	England	United King...	E1 5JL	['Chattermill offer...	['Software', 'Custo...	15	-26.0	34.8	26.0	6	Series B	2022-12-06	1.93	2019.0	{'sentimentScore': 80, ...	[{'contentId': '059b...
13	FreshMenu	Savor the World...	FreshMenu is a...	FreshMenu is a ...	2014.0	https://freshmenu....		Bengaluru	India	560094	['FreshMenu is a...	['Food Delivery Se...	8	81.0	32.56	7.0	7	Unattributed...	2022-05-05	19.53	2020.0	{'sentimentScore': 90, ...	[{'contentId': '245e...
14	Sightline Pay...	Revolutionizing ...	Sightline Paym...	Sightline Payme...	2010.0	https://sightlinepa...	Las Vegas	Nevada	United States	89113	['Leading provid...	['FinTech', 'Gaming...	10	-38.0	359.48	13.0	12	Private Equi...	2022-10-20	35.4	2020.0	{'sentimentScore': 94, ...	[{'contentId': '3e8d...
15	Unisound	Unleashing Intel...	Unisound is a c...	Unisound is a c...	2012.0	https://unisound.c...	Shenzhen	Guangdong	China	518055	['Unisound offers...	['Artificial Intelligen...	15	-85.0	339.11	60.0	9	Series D - II	2021-06-24	15.12	2017.0	{'sentimentScore': 77, ...	[{'contentId': '1eb2...
16	EGYM	Revolutionizing ...	EGYM is a glob...	EGYM provides ...	2010.0	https://egym.com	Munich		Germany	81677	['EGYM is a glob...	['Fitness Technolo...	11	146.0	609.14	200.0	8	Series G	2024-09-24	130.0	2022.0	{'sentimentScore': 100...	[{'contentId': 'aaa2...
17	Axtria	Transforming H...	Axtria is a com...	Axtria specialize...	2010.0	https://axtria.com	Berkeley H...	New Jersey	United States	07922	['Enterprise-grad...	['Artificial Intelligen...	9	-77.0	206.32	150.0	7	Series E	2021-05-13	50.0	2016.0	{'sentimentScore': 98, ...	[{'contentId': '2c0c...
18	Climb Credit	Unlock Your Pot...	Climb Credit, b...	Climb Credit is ...	2014.0	https://climbcredit....	Las Vegas	Nevada	United States	89119	[]	['FinTech', 'EdTech...	16	-38.0	463.58	13.0	9	Convertible ...	2023-02-10	7.4	2020.0	{'sentimentScore': 59, ...	[{'contentId': '1b68...
19	Jumbotail	Revolutionizing ...	Jumbotail is a p...	Jumbotail offers ...	2015.0	https://jumbotail.c...		Bengaluru	India	560029	['Leading B2B M...	['E-commerce', 'Fo...	24	-77.0	180.38	18.22	11	Series C - II	2024-03-18	98.8	2023.0	{'sentimentScore': 89, ...	[{'contentId': 'c3c7...
20	Bombas	Comfort that Ca...	Bombas is a Ne...	Bombas operat...	2013.0	https://bombas.com	New York	New York	United States	10003	['Known for high-...	['Apparel & Access...	2	275.0	252.75	70.56	6	Unattributed...	2022-08-15	100.8	2018.0	{'sentimentScore': 84, ...	[{'contentId': '6af7...
21	Eruditus Exec...	Elevate Your Le...	Eruditus Execut...	Eruditus Executi...	2010.0	https://eruditus.com	Singapore		Singapore	079120	['Collaborates wit...	['EdTech', 'Busines...	14	-38.0	943.46	350.0	8	Debt - II	2022-03-08	100.0	2020.0	{'sentimentScore': 89, ...	[{'contentId': '2731...

primary-with-valuation.csv

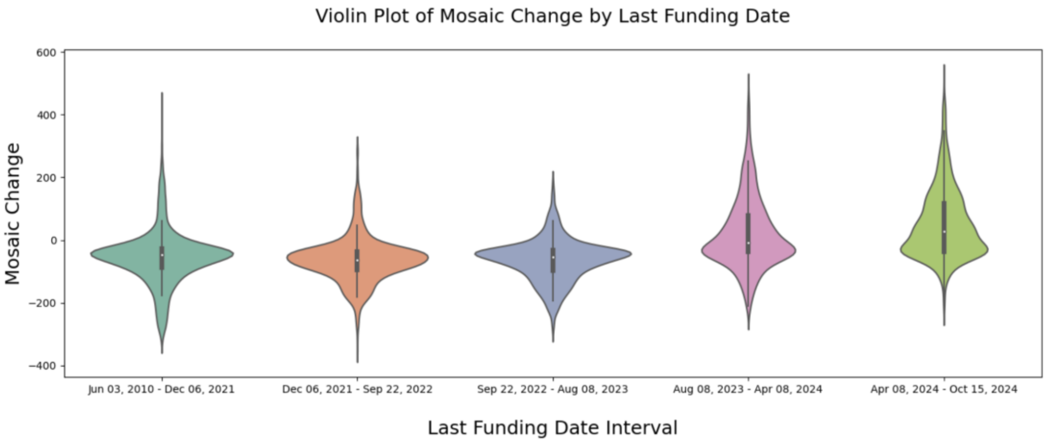


	name	tagline	summary	description	year_f...	website	city	region	country	postal_c...	concepts	keywords	in...	mos...	fundin...	last_...	f...	last_fundi...	last_fundi...	valua...	valuati...	sentiment	articles
1	PlainID	Secure Your Ide...	PlainID is a co...	PlainID is an Iden...	2014.0	https://plainid.com	Tel Aviv		Israel	6789139	['PlainID offers the ...	['Cybersecurity', 'I...	10	-120.0	99.0	75.0	5	Series C	2021-12-21	48.0	2020-12	{'sentimentScore': 10...	[{'contentId': '29c3c12...
2	slice	Experience Mon...	slice is a finan...	Slice operates as ...	2016.0	https://sliceit.com		Assam	India	781028	['Trusted by over 1...	['FinTech', 'Consu...	36	-76.0	390.5	7.77	21	Debt - VIII	2024-07-19	1800.0	2022-06	{'sentimentScore': 89,...	[{'contentId': 'ce75fa1...
3	slice	Experience Mon...	slice is a finan...	Slice operates as ...	2016.0	https://sliceit.com		Assam	India	781028	['Trusted by over 1...	['FinTech', 'Consu...	36	-76.0	390.5	7.77	21	Debt - VIII	2024-07-19	1800.0	2022-06	{'sentimentScore': 89,...	[{'contentId': 'ce75fa1...
4	Dispatch...	Healing at Home...	DispatchHealt...	DispatchHealth is ...	2013.0	https://dispatchhe...	Denver	Colorado	United States	80907	["DispatchHealth o...	['Healthcare', 'Tele...	15	-47.0	740.96	259.0	8	Series E	2022-11-15	1700.0	2021-03	{'sentimentScore': 99,...	[{'contentId': '556f0d4...
5	Survios	Step Into Advent...	Survios is a C...	Survios is a leadi...	2013.0	https://survios.com	Marina d...	California	United States	90292	['Specializes in cre...	['Augmented Reali...	14	28.0	70.95	16.7	5	Series D	2020-10-02	30.0	2015-10	{'sentimentScore': 68,...	[{'contentId': '7fb0f8fb-...
6	Ironclad	Revolutionizing ...	Ironclad, a Cal...	Ironclad is a leadi...	2014.0	https://ironcladap...	San Fran...	California	United States	94105	['Global leader in c...	['Application Softw...	16	-8.0	331.12	150.0	9	Series E	2022-01-18	950.0	2020-10	{'sentimentScore': 10...	[{'contentId': 'e64a898...
7	Kalshi	Trading Beyond ...	Kalshi is a Ne...	Kalshi operates a...	2018.0	https://kalshi.com	New York	New York	United States	10012	['Innovative platfor...	['Financial Techno...	9	-72.0	38.74	2.5	5	Series A - II	2022-01-14	120.0	2021-02	{'sentimentScore': 69,...	[{'contentId': 'bd3cfea...
8	AI21 Labs	Unleashing the ...	AI21 Labs is a...	AI21 Labs operat...	2017.0	https://ai21.com		Tel Aviv-Y...	Israel	6203854	['Leading the Gene...	['Artificial Intellige...	15	-57.0	317.0	53.0	6	Series C - II	2023-11-21	1400.0	2023-08	{'sentimentScore': 94,...	[{'contentId': '7838ba0...
9	Addionics	Energizing the F...	Addionics, a c...	Addionics speciali...	2017.0	https://addionics.c...	London	England	United King...	W12 0BZ	['Innovative approa...	['Renewable Ener...	26	170.0	72.0	39.0	10	Series B	2024-07-25	82.0	2022-01	{'sentimentScore': 10...	[{'contentId': 'd872319...
10	Tarana W...	Connecting Bey...	Tarana Wirele...	Tarana Wireless s...	2009.0	https://taranawirel...	Milpitas	California	United States	95035	['Pioneering ngFW...	['Telecommunicati...	13	-65.0	429.64	50.0	10	Unattribut...	2023-09-13	1000.0	2022-03	{'sentimentScore': 97,...	[{'contentId': 'd91e173...
11	QI Tech	Unleashing Fina...	QI Tech is a c...	QI Tech is a techn...	2018.0	https://qitech.com.br	Sao Paulo		Brazil	05425-020	['QI Tech simplifies...	['FinTech', 'Fraud ...	4	31.0	312.0	50.0	5	Series B - II	2024-04-25	1000.0	2023-10	{'sentimentScore': 10...	[{'contentId': '9d93890...
12	Lentra	Revolutionizing ...	Lentra is a co...	Lentra focuses on...	2019.0	https://lentra.ai	Pune		India	411005	['Offers digital lendi...	['FinTech', 'Cloud ...	8	-95.0	177.0	13.0	6	Incubator...	2024-08-08	500.0	2022-11	{'sentimentScore': 95,...	[{'contentId': 'c94c66c...
13	Typeform	Engaging Forms...	Typeform is a ...	Typeform focuses...	2012.0	https://typeform.com	Barcelona		Spain	08017	['Offers people-frie...	['Application Softw...	15	-1.0	187.26	135.0	6	Series C	2022-03-10	300.0	2017-09	{'sentimentScore': 87,...	[{'contentId': 'ea844c2...
14	Sidecar H...	Redefining Healt...	Sidecar Health...	Sidecar Health op...	2018.0	https://sidecarheal...	Covina	California	United States	91723	['Offers consumer-...	['Health Insurance...	16	146.0	328.0	165.0	5	Series D	2024-06-26	1000.0	2021-01	{'sentimentScore': 95,...	[{'contentId': '1b8c6ee...
15	Edgeless ...	Protecting Your ...	Edgeless Syst...	Edgeless System...	2020.0	https://edgeless.s...		Bochum	Germany	44791	['Industry leader in ...	['Cybersecurity', '...	14	22.0	6.73	5.0	7	Seed VC ...	2023-03-08	75.0	2021-06	{'sentimentScore': 67,...	[{'contentId': '8d08f99...
16	Clear Street	Pioneering the F...	Clear Street is...	Clear Street speci...	2018.0	https://clearstreet.io	New York	New York	United States	10007	['Modernizing the b...	['Financial Service...	12	-23.0	491.71	270.0	4	Series B - II	2023-04-11	1700.0	2022-04	{'sentimentScore': 94,...	[{'contentId': 'ba8ff784...
17	Netskope	Securing Your Cl...	Netskope is a ...	Netskope operate...	2012.0	https://netskope.c...	Santa Cl...	California	United States	95054	['Leader in Secure ...	['Cybersecurity', '...	19	-21.0	1441.0	401.0	11	Convertib...	2023-01-05	7500.0	2021-07	{'sentimentScore': 80,...	[{'contentId': 'e9528ea...
18	Justworks	Simplifying HR, ...	Justworks is a ...	Justworks focuse...	2012.0	https://justworks.c...	New York	New York	United States	10008	['Comprehensive P...	['Human Resourc...	9	152.0	159.8	16.8	7	Unattribut...	2023-11-02	183.01	2016-03	{'sentimentScore': 69,...	[{'contentId': '2f72e5fe...
19	smartrr	Elevating Subscr...	Smartrr is a co...	Smartrr is a comp...	2020.0	https://smartrr.com	New York	New York	United States	10011	['Best-in-class acc...	['E-commerce', 'S...	6	-100.0	17.3	10.0	4	Series A	2023-01-18	77.3	2020-11	{'sentimentScore': 10...	[{'contentId': '07cd83b...
20	Back Mar...	Renew, Reuse, ...	Back Market is...	Back Market offer...	2014.0	https://backmarke...	Paris		France	75019	['Operates as a ref...	['Consumer Electr...	13	-20.0	1021.0	510.0	7	Series E	2022-01-11	3200.0	2021-05	{'sentimentScore': 95,...	[{'contentId': '0f7cf4b9...
21	AgentSync	Accelerating Gro...	AgentSync is ...	AgentSync specia...	2018.0	https://agentsync.io	Denver	Colorado	United States	80205	['Better experience...	['Insurance', 'Soft...	12	209.0	161.1	50.0	5	Series B - II	2023-10-26	1200.0	2021-12	{'sentimentScore': 10...	[{'contentId': '51b7ece...

Visualizations

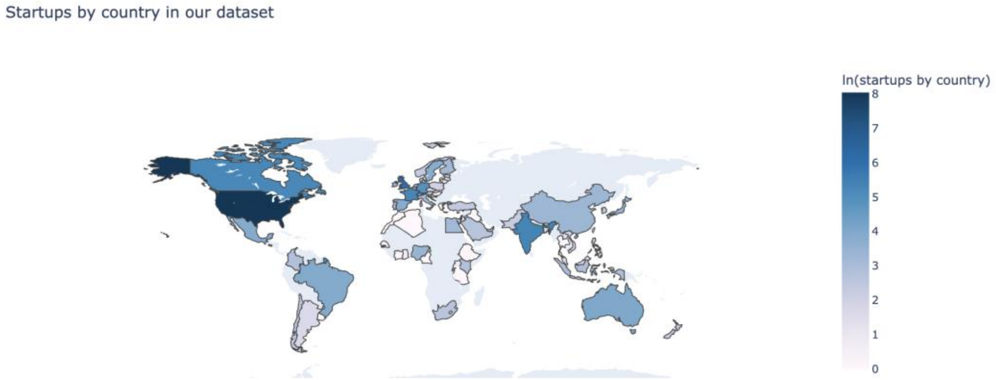
1. Mosaic change vs date of most recent funding

The mosaic score for a company is an internal metric designed by CB Insights for evaluating the long-term success of startups. Without a paid account, only the mosaic score change was available to scrape for each startup. This plot puts startups into bins based on their most recent funding date, then creates a violin plot for each bin to show the distribution of its mosaic score changes. From this plot, we can see that startups that have recently received funding tend to have more positive mosaic score changes. This intuitively makes sense: if a startup has not been funded for a long time, it is less likely to be a long-term success.



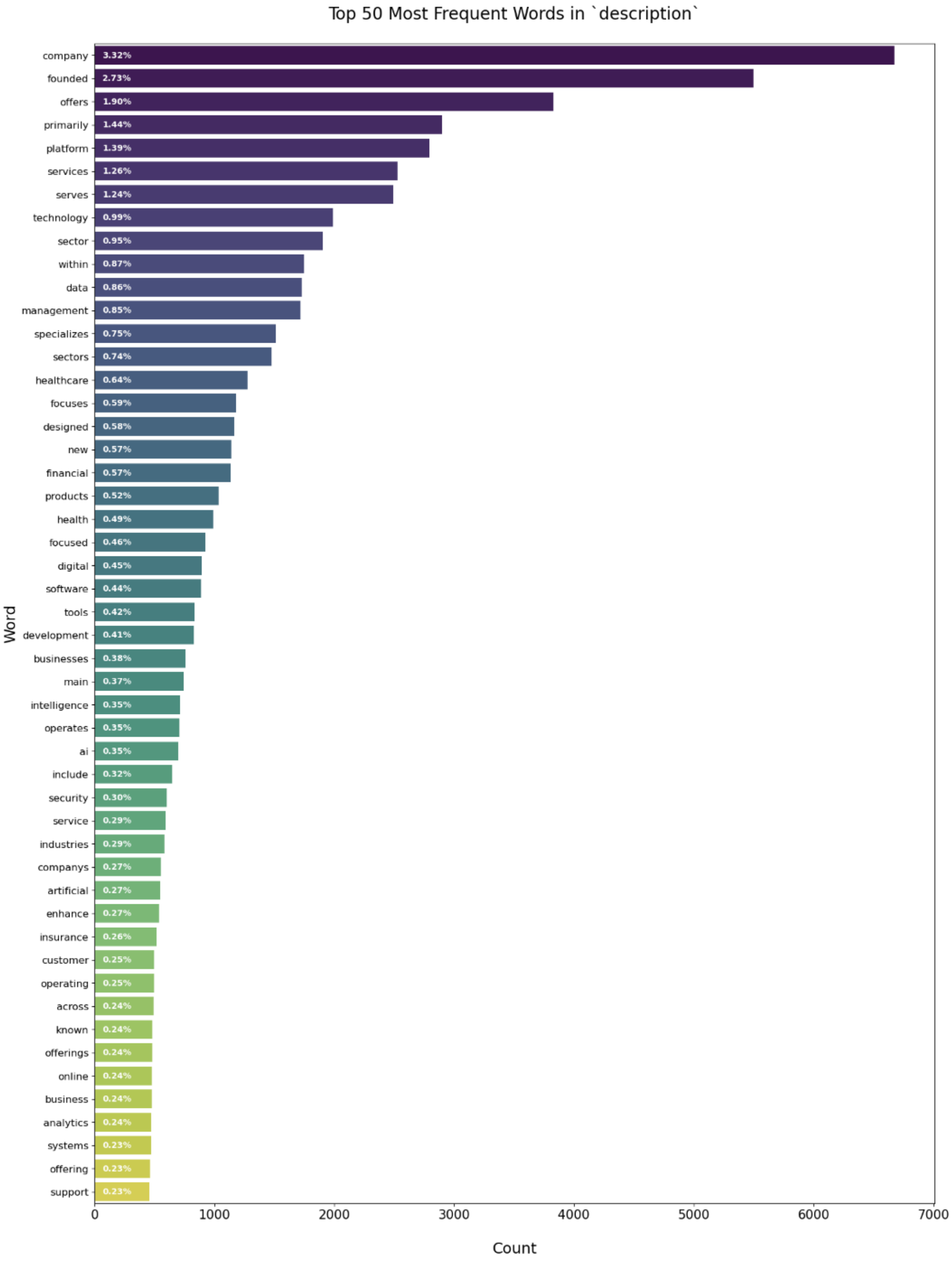
2. Number of startups by country

We have the country information for each of our startups, so we can plot a map showing how many startups are in each country in our dataset. From this plot, we get an idea of where the startups we are analyzing are located, and we can also decide if there are countries we want to drop or countries we want to gather more data for.



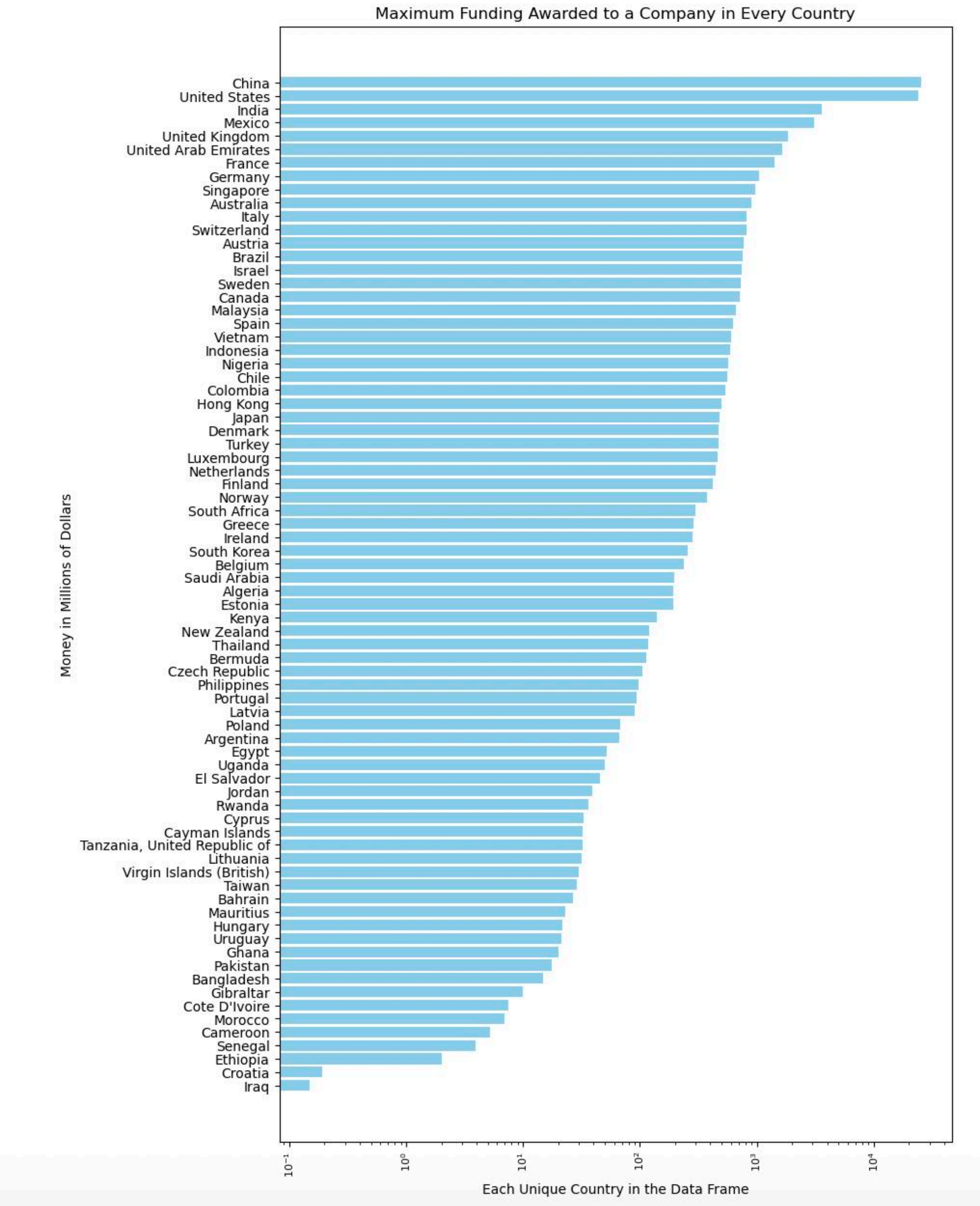
3. Most common words in startup descriptions

Each of the startups in our dataset has a description written by CB Insights. Using these descriptions, we can extract the most common words and create a frequency plot. Some of the descriptions contain city names, so we are getting common city names and removing them from the descriptions before processing the words. We are also moving stopping words like 'the', 'and', and 'but'. From this frequency plot, we can get an idea of what types of companies we are looking at, and what the most popular areas are.



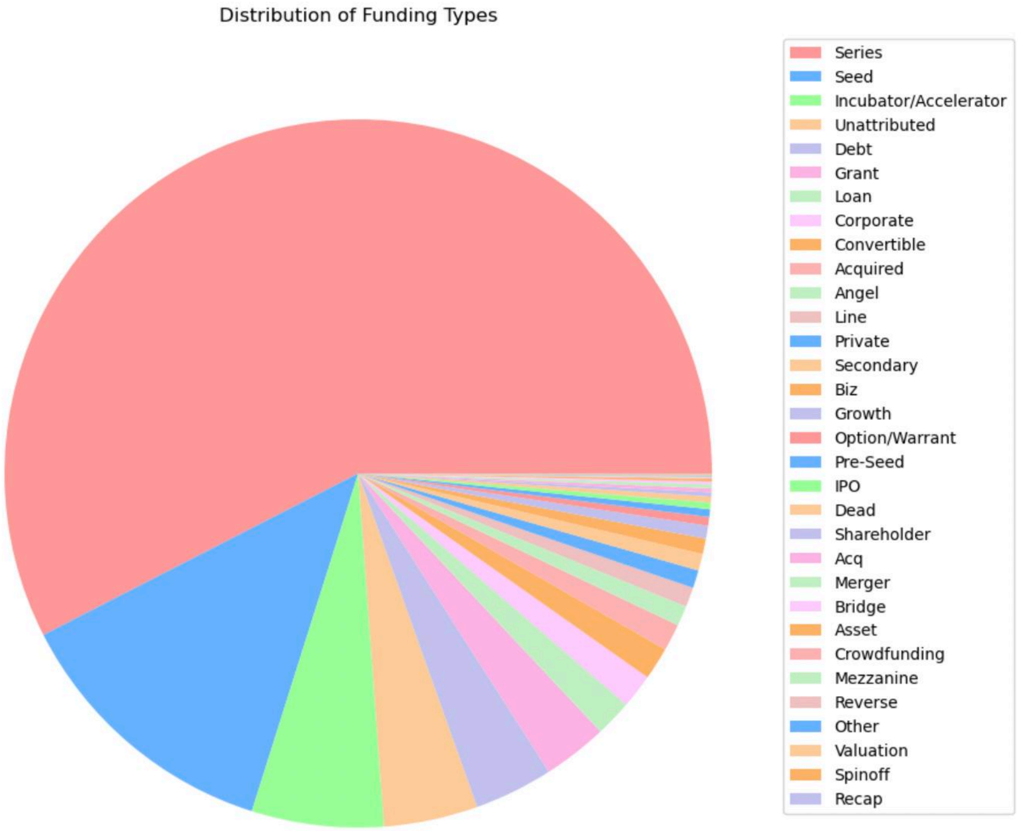
4. Maximum funding awarded to a company in every country

This graph illustrates the maximum amount of funding an individual company receives in each country. This will be vital to our analysis because we can predict if a company will receive more funding based on the country, they founded their company in.



5. Distribution of funding types

This pie chart illustrates the last funding type an individual company received. This information will be important for our analysis because we can predict the success of a company based on the type of funding that they receive.

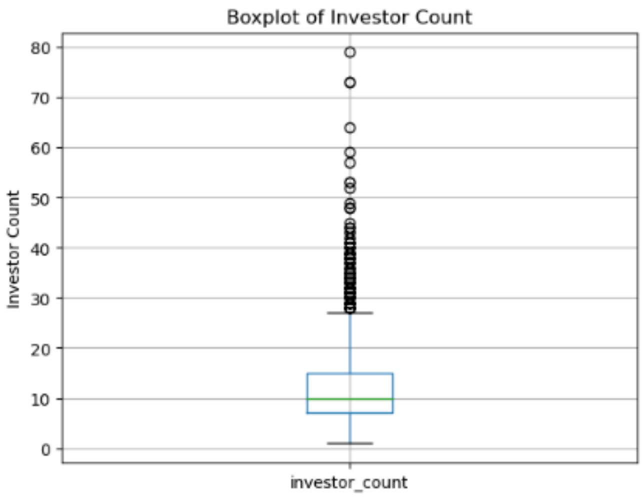


6. Boxplots

Boxplots were created for each relevant numerical variable to get a visual indication of central tendency and spread of each attribute.

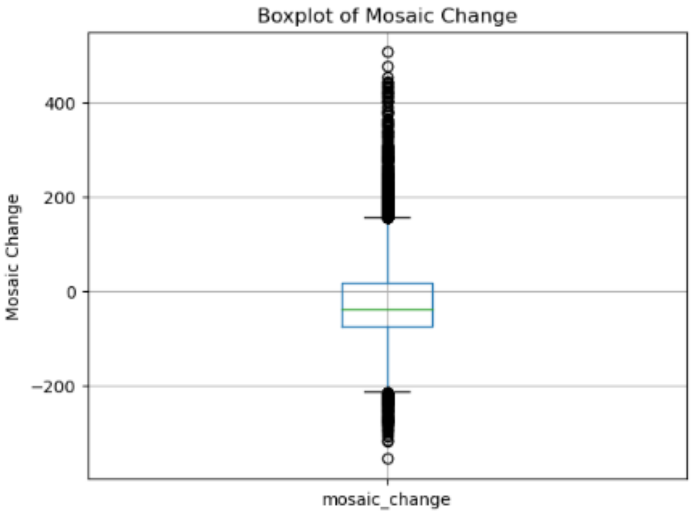
Investor Count:

This boxplot shows a median of 10 investors, with the middle 50% of the data lying close to the median, a low whisker of 1 and a high whisker a bit higher than 25. The data appears somewhat symmetrically spread with a bit of a right skew towards higher numbers of investors, and a significant number of high outliers.



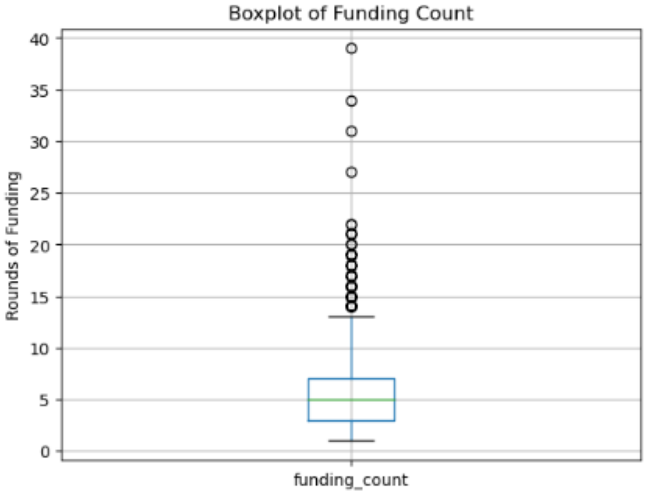
Mosaic Change:

This boxplot shows a smaller negative mean, and a very symmetrical looking spread, with a very small right skew. There is a significant number of both high and low outliers. This boxplot shows a heavy right skew, but a symmetrical middle 50% of the data, with a median of 5 (close to the mean) of the number of rounds of funding. There are high outliers but less outliers than in previous boxplots.



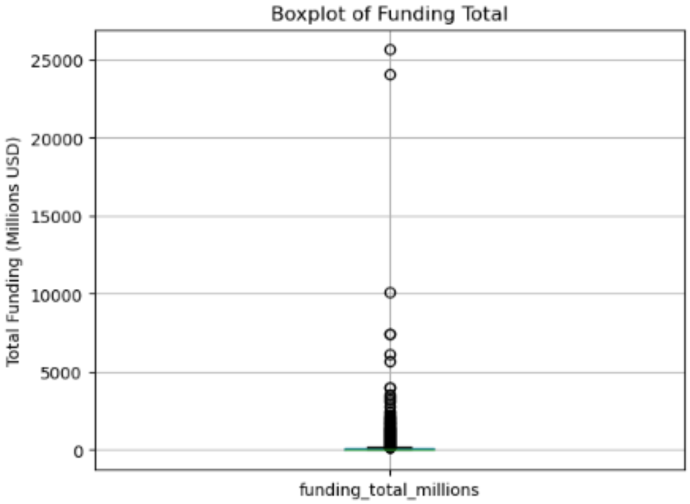
Funding Count:

This boxplot shows a heavy right skew, but a symmetrical middle 50% of the data, with a median of 5 (close to the mean) of the number of rounds of funding. There are high outliers but less outliers than in previous boxplots.

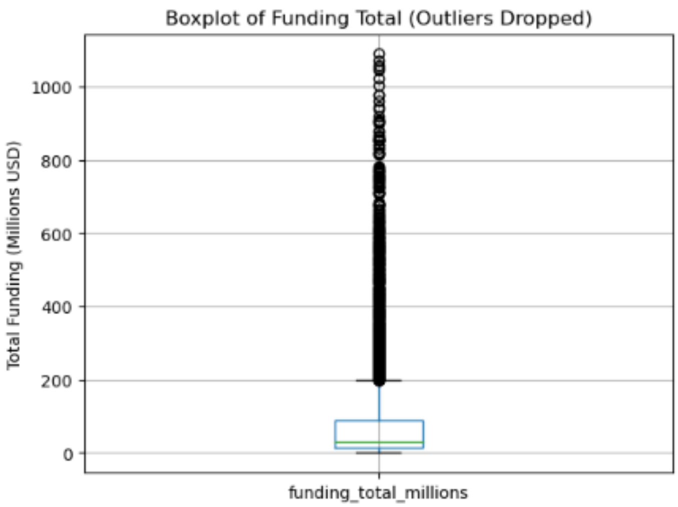


Funding Total:

This boxplot is undiscernible because of the significant high outliers. Removal of the top 50 outliers was conducted, and the data replotted:

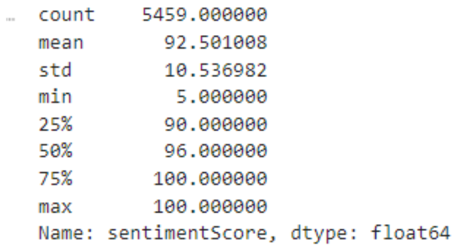


This boxplot is now discernible due to removing outliers. We see a very small median with a heavy right skew, and still a significant number of very high outliers. The lowest whisker is close to 0 and not below.

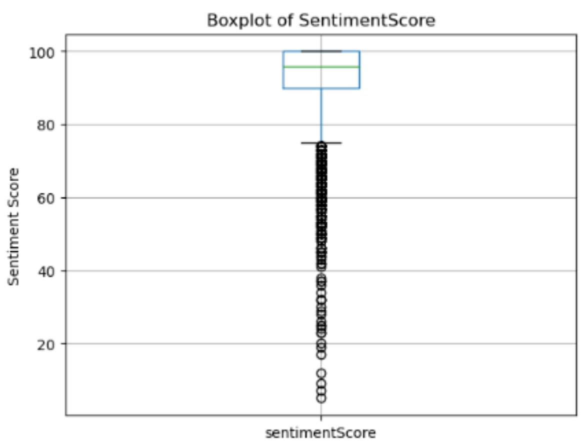


Sentiment Score:

After sentiment scores were extracted, summary statistics were generated into the table below and this boxplot was created. We see a very high mean of around 92, indicating that most sentiment scores are very high. This is likely due to bias in selecting startups that were able to gain high amounts of funding (with a mean of 114 million USD).



```
count    5459.000000
mean       92.501008
std        10.536982
min         5.000000
25%        90.000000
50%        96.000000
75%       100.000000
max       100.000000
Name: sentimentScore, dtype: float64
```



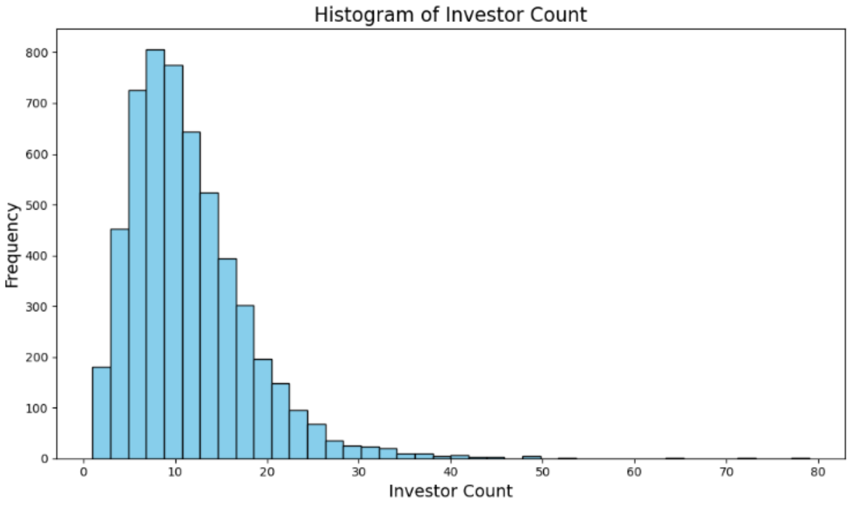
This boxplot shows most of the data being very close to 100. There is a strong left skew to this data and plentiful low outliers.

7. Histograms

Histograms were created for each relevant numerical variable to get a visual sense of the distribution of each attribute.

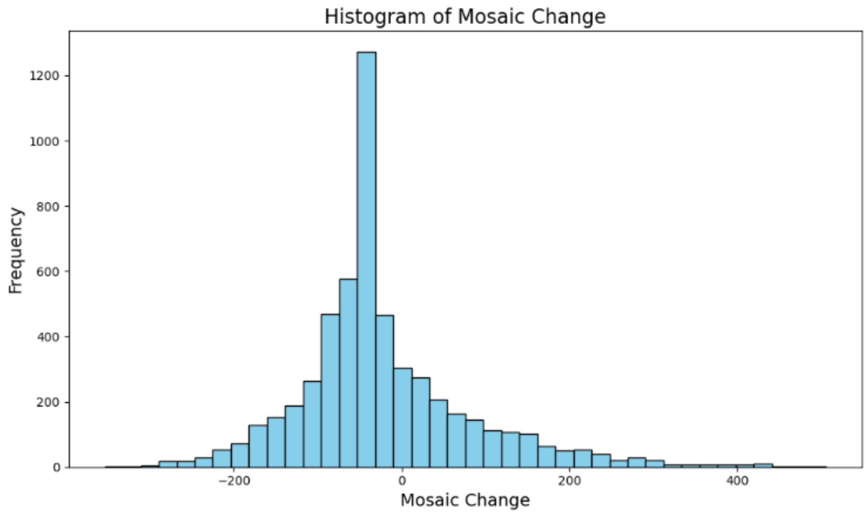
Investor Count:

This histogram looks mostly normally distributed around the mean/median, with a long right-handed tail.



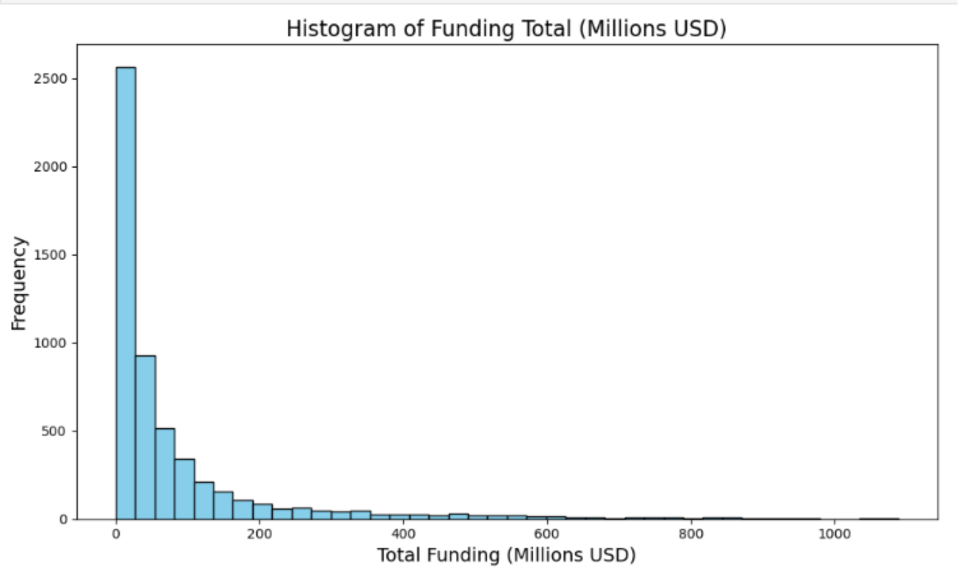
Mosaic Change:

This histogram appears symmetrical and has a very large peak around the mean – there is a slight but noticeable right skew.



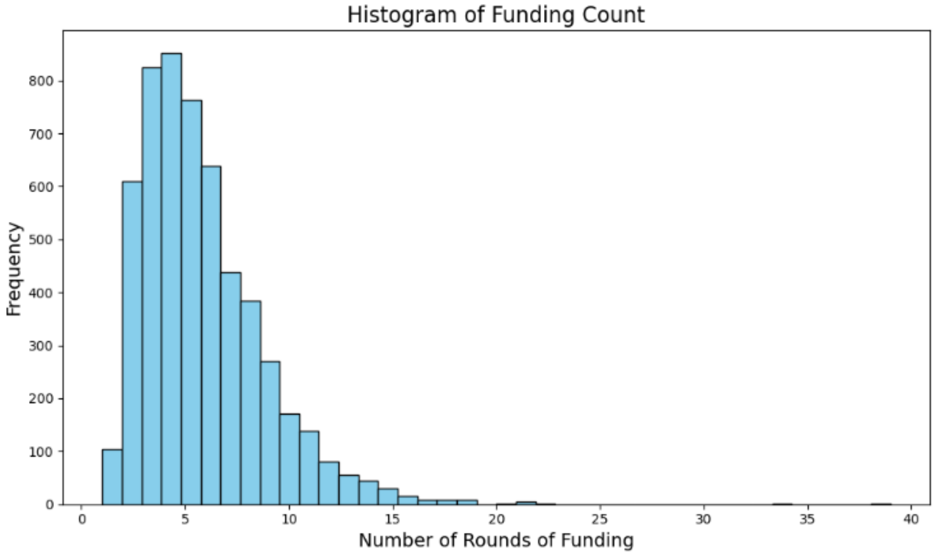
Funding Total:

This histogram is heavily right skewed. Most of the data is located close to 0. We know the mean is above 100, but this is likely heavily influenced by outliers, making the median a more robust measurement at 31 million.



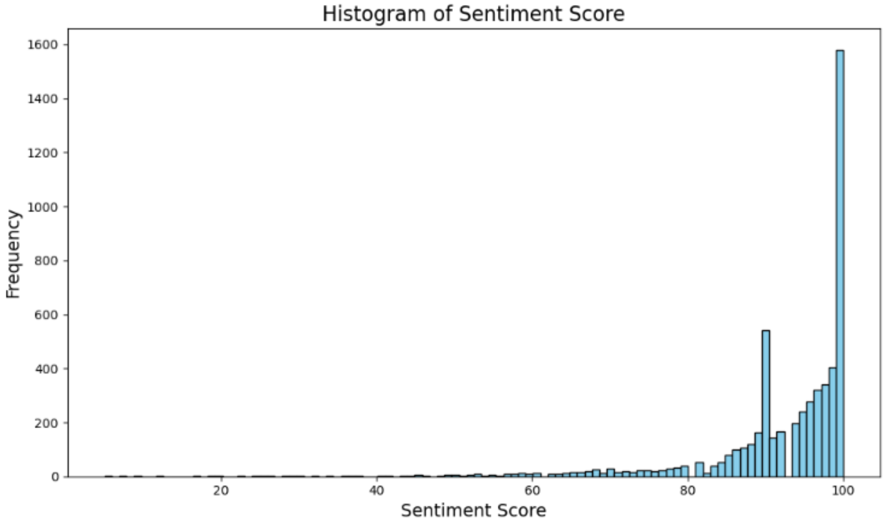
Funding Count:

This histogram looks slightly normal, being unimodal, but also containing a long right-tail.



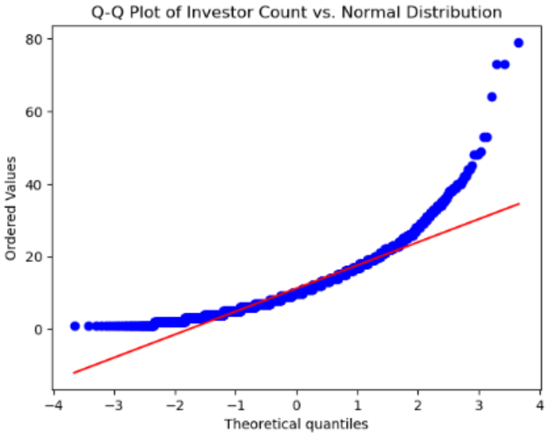
Sentiment score:

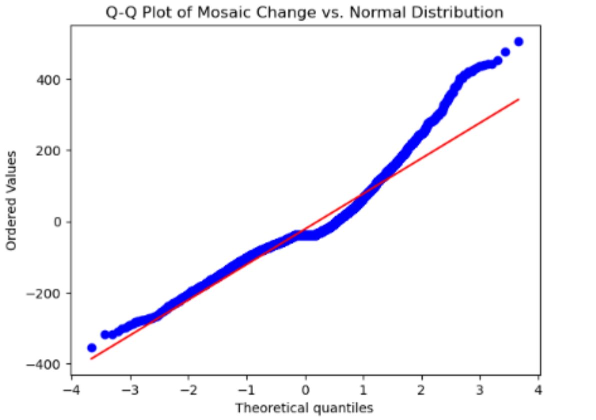
This histogram shows that most sentiment scores are 100, with a very long left-handed tail. The mean is close to the median even with so many outliers, indicating how many values are directly at 100. It may be worth further investigation to find out how sentiment score is determined and if it has much meaning to our analysis.

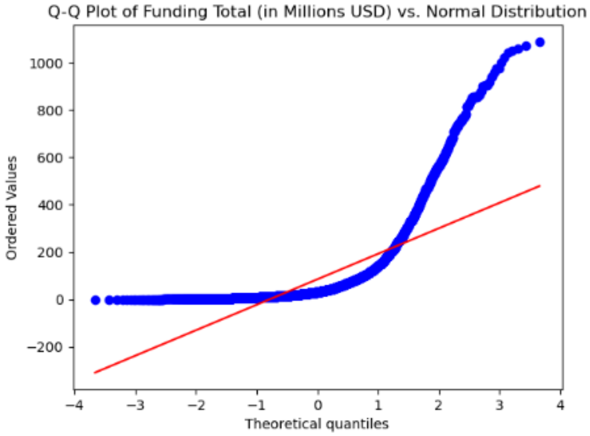


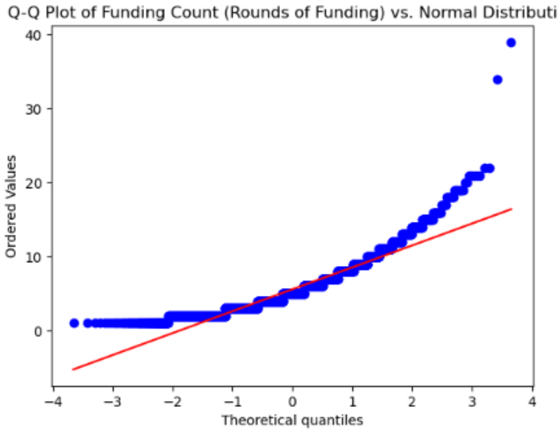
8. QQ-plots

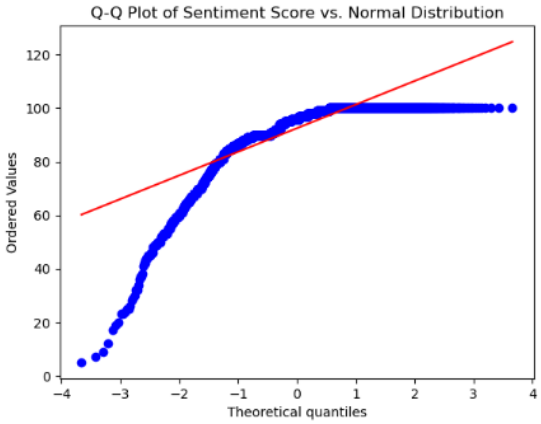
QQ-plots were created for each relevant numerical variable to visually determine how normally distributed they were. We can visually measure how close the distribution of each numerical variable is to the theoretical Normal distribution by how closely they approximate the straight red line in each plot. Each variable shows a marked difference between its distribution and a normal distribution. The Mosaic score is the most normally distributed variable, as it is the closest to the straight line. Both Funding Total and Sentiment Score show the most significant deviations, as we have seen already in above plots - both of these variables had peaks at one end of their value ranges or the other. Investor Count and Funding Count showed distributions in between these two aforementioned plots, not being completely normally distributed by any means, but closer than Funding Total and/or Sentiment Score. The deviation from a Gaussian distribution for each variable indicates that normalization may be worthwhile for all numerical variables.


Q-Q Plot of Investor Count vs. Normal Distribution
Ordered Values
Theoretical quantiles
80
60
40
20
0
−4
−3
−2
−1
0
1
2
3
4


Q-Q Plot of Mosaic Change vs. Normal Distribution
Ordered Values
Theoretical quantiles
400
200
0
−200
−400
−4
−3
−2
−1
0
1
2
3
4

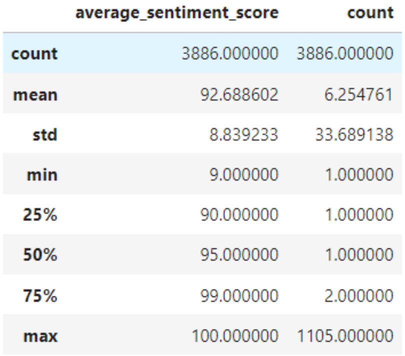

Q-Q Plot of Funding Total (in Millions USD) vs. Normal Distribution
Ordered Values
Theoretical quantiles
1000
800
600
400
200
0
−200
−4
−3
−2
−1
0
1
2
3
4


Q-Q Plot of Funding Count (Rounds of Funding) vs. Normal Distributi
Ordered Values
Theoretical quantiles
40
30
20
10
0
−4
−3
−2
−1
0
1
2
3
4


Q-Q Plot of Sentiment Score vs. Normal Distribution
Ordered Values
Theoretical quantiles
120
100
80
60
40
20
0
−4
−3
−2
−1
0
1
2
3
4

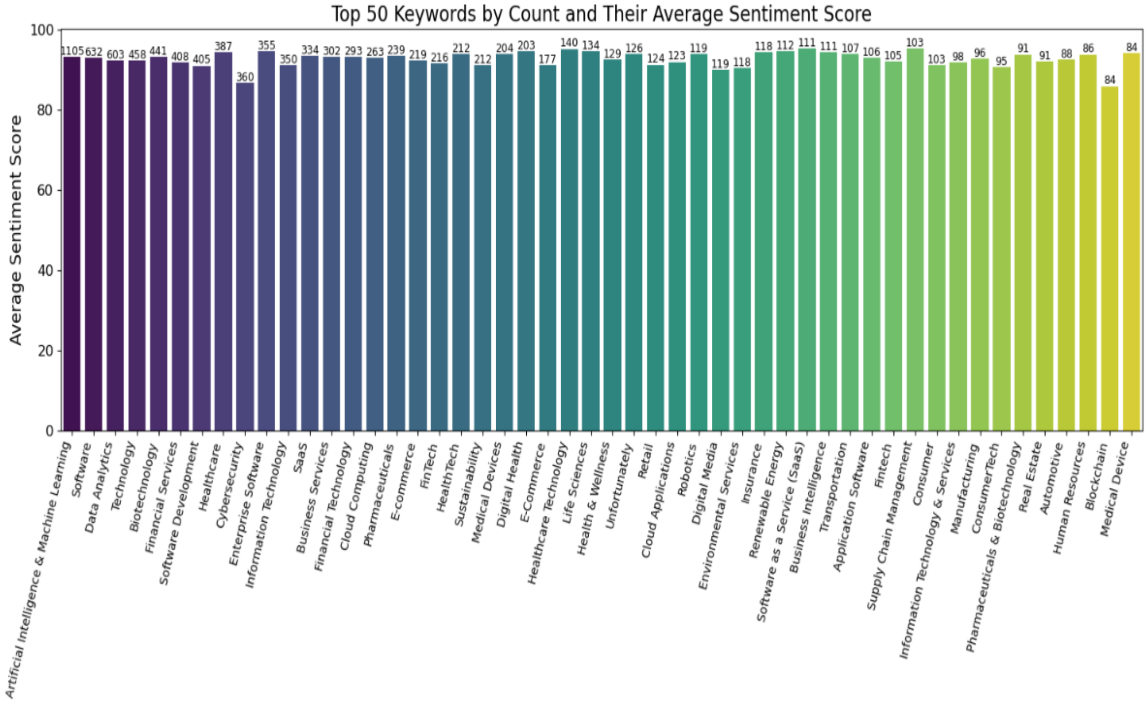
9. Keywords vs sentiment score

To understand more about how keywords may have influenced sentiment analysis, we created a separate dataframe that aggregated the average sentiment value score for each industry keyword present in the dataset. We iteratively updated the average sentiment score for each keyword as it was encountered in the dataset, and also counted the instances of each keyword present. Below is a table describing summary statistics for this new dataframe:



	average_sentiment_score	count
count	3886.000000	3886.000000
mean	92.688602	6.254761
std	8.839233	33.689138
min	9.000000	1.000000
25%	90.000000	1.000000
50%	95.000000	1.000000
75%	99.000000	2.000000
max	100.000000	1105.000000

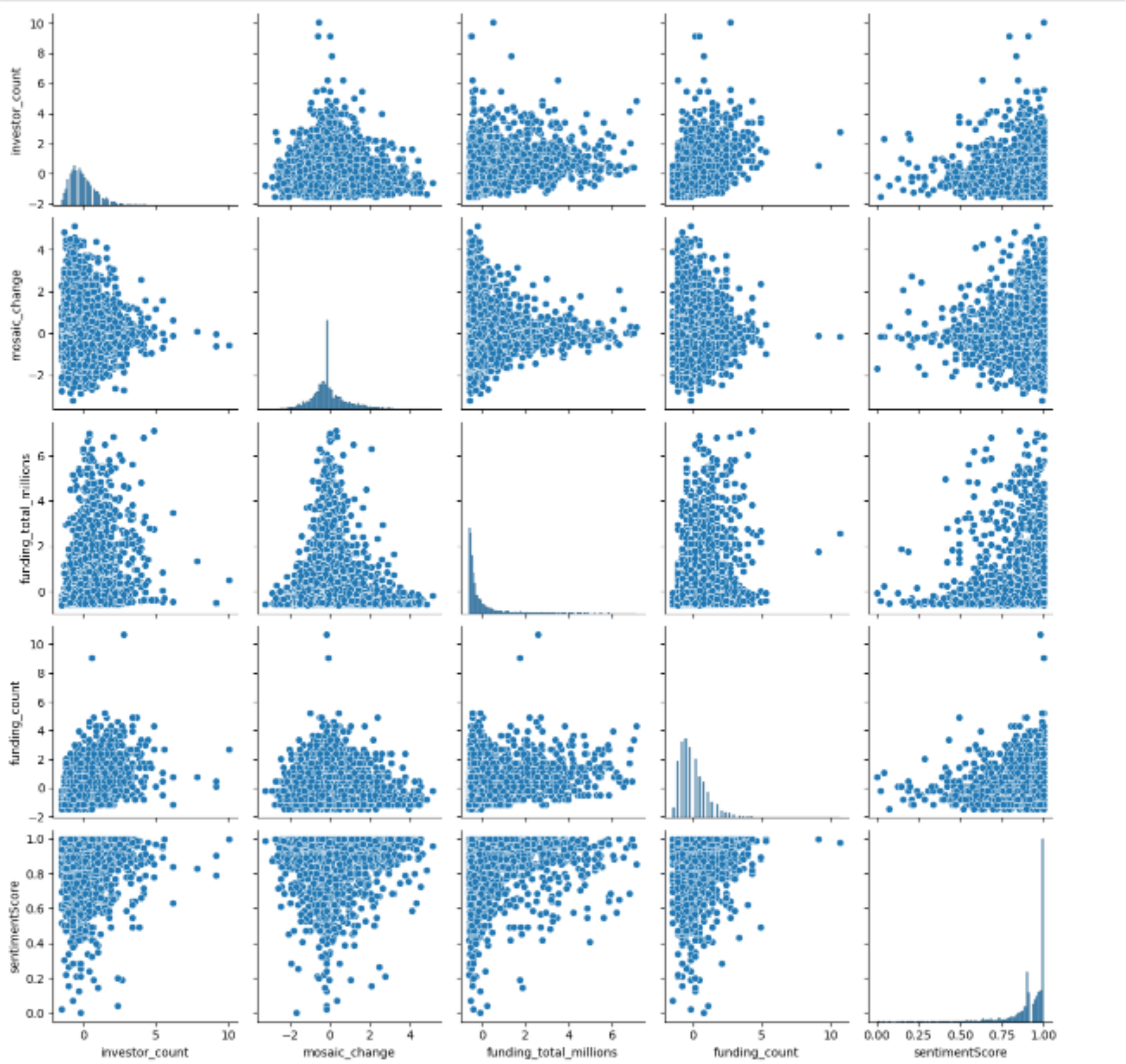
Again, we see a median and mean very close together, but with values very highly skewed. Many of these values don't have much meaning since they have a very low sample size (the count attribute) of 1. We decided to visualize the average sentiment score of the top 50 most numerous words from the dataset (more would take up too much space):



This graph shows the average sentiment score for each industry keyword and the number of each keyword present in the dataset. We do not see hardly any variation here. This is likely a good indication that our dataset is fairly biased, either in the way that sentiment scores are calculated, or in the fact that every startup in the dataset is very successful, having been able to secure 114 million dollars of funding on average, so it is likely that these industry buzzwords associated with the industry of each startup all are attached to very positive sentiments.

10. Pairplot

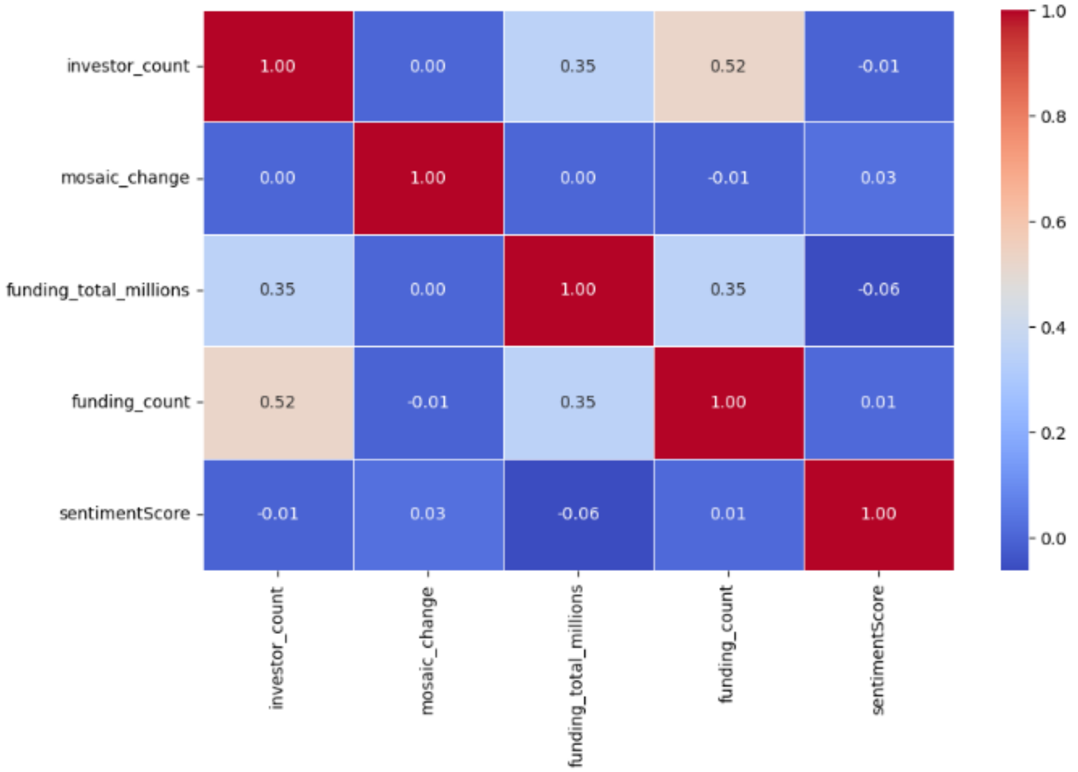
We can get a sense of whether any of our numerical variables are correlated with each other by creating a pairplot of scatterplots of each numerical variable plotted against every other one. This can help us determine which variables might be important for any kind of linear regression or other types of machine learning models:



We can see some potentially positive relationships, especially between the funding count and the investor count. This intuitively makes sense, as you would expect more rounds of funding as the number of investors increased. Overall, though, the data looks very scattered with very little indication of linearity between any of the numerical variables. We would perhaps expect that more investors and more total funding, for example, would have a stronger positive relationship. We see this somewhat, but it does not appear to be a very strong linear relationship. This does not bode well for future model-building. Let's look at a heatmap showing correlation coefficients between each numerical variable.

11. Heat map

This heatmap shows color-coded correlation coefficients between each numerical variable, with a stronger correlation being the reddest, and the weakest being the bluest. We immediately notice how most of the heatmap is very blue, with almost no correlation between most variables (with extremely small coefficients, a few even being 0). We see a moderate positive correlation between funding count and investor count, and a weak positive correlation between total funding and funding count and total funding and investor count. But most variables appear to be as uncorrelated as possible. Again, this will prove to be a problem for further model-building and analysis, and we may need to explore other sources of data to create effective models.


investor_count
mosaic_change
funding_total_millions
funding_count
sentimentScore
investor_count
mosaic_change
funding_total_millions
funding_count
sentimentScore