Comprehensive Analysis of Start-up Companies Dataset for Strategic Insights

By Nikita Vert for the London Technology Club

Table of Contents

Comprehensive Analysis of Start-up Companies Dataset for Strategic Insights 1

Ask 4

Market Overview 4

Key Trends 4

Funding Volumes 4

Valuations 4

Sector Tailwinds 5

Secondary Opportunities 5

Top Industries 5

Focus for Analysis 5

Introduction 8

Prepare and Process 9

Stage 1. Generating missing information 9

Stage 2. Importing the data into the database 10

Stage 3. Cleaning the table 10

Checking datatypes 10

Checking for duplicates and NULL values 14

Checking the logical order of the dates 15

Analyse and Share 16

Stage 1. Filtering the table. 16

Stage 2. Identifying funding trends 17

Stage 3. Location and Industry 18

Industries 18

Industry Funding 19

Industries Average Funding 22

Location 23

Location Funding 25

Location Average Funding 27

Industries and Location 29

Description of the Stacked Bar Chart: 29

Key Insights: 30

Notes: ..... Error! Bookmark not defined.

Stage 4. Total funding over the years 32

Key Insights: 32

Notes: 32

Stage 5. Top picks 33

Act 34

Recommendations **34**
Summary: 36
Companies with the highest potential ***37***
Bibliography ***47***
Appendix ***49***
Prepare and Process **49**
Stage 1. Generating Missing information 49
Stage 2. Importing the data into the database 57
Stage 3. Cleaning table 58
Analyse and Share **66**
Stage 1. New table startups_best 66
Stage 2. Funding trends 66
Stage 3. Location and Industry 66

Ask

Market Overview

The 2025 startup market is characterised by a bifurcated venture capital (VC) landscape, with artificial intelligence (AI) dominating funding and valuations, while non-AI sectors face valuation pressures. Late-stage startups are securing significant capital, while seed and early-stage funding struggles. Rising M&A ($71B in Q1 2025) and IPO activity signal improving exit opportunities. Key sectors include AI, green tech, health tech, fintech, e-commerce, and EdTech, with quick-commerce facing headwinds. Secondary investments offer opportunities to acquire stakes at discounts, particularly in non-AI sectors, driven by liquidity needs and valuation wobbles.

Key Trends

1. **AI Leadership**: AI captures about a third of all VC funding, with mega-rounds (e.g., OpenAI's $40B) and a shift to vertical AI. (Romburgh, 2025)
2. **Late-Stage Focus**: Investment surges in late-stage startups, with reduced capital for seed and early-stage ventures. (Teare, 2025)
3. **M&A and IPO Growth**: $71B in startup M&A in Q1 2025 (Teare, 2025).
4. **Sector Shifts**: Green tech, health tech, and EdTech gain traction; quick-commerce declines. (Grabow, 2025)
5. **Cautious VC Sentiment**: Funding is up but selective, focusing on mature products amid tariff and IPO uncertainties. (Crunchbase, 2024)

Funding Volumes

- **Total Funding**: Higher than 2024 but below 2021 peaks, driven by AI mega-rounds (Teare, 2025).
- **Stage Split**: Late-stage startups dominate; seed and early-stage funding declines. (Teare, 2025)
- **Mega-Rounds**: AI-driven mega-rounds (e.g., Databricks' $10B, Waymo's $4B+) inflate funding volumes, but deals are concentrated among fewer players. (Metinko, 2025)

Valuations

- **AI Surge**: AI startups command high valuations, with examples like OpenAI ($300B), CoreWeave ($19B), and Anthropic ($18.4B). Revenue multiples vary by niche but remain elevated. (Metinko, 2025)
- **Non-AI Pressure**: Non-AI sectors face valuation pressures due to tariff uncertainties and delayed IPOs, creating a bifurcated market. Startups from 2020-2021 with outdated valuations risk down rounds. (Romburgh, 2025)
- **Secondary Impact**: Lower valuations in non-AI sectors (e.g., fintech, cybersecurity) create opportunities for secondary investments at discounts. High AI valuations may limit secondary opportunities unless sellers face liquidity pressures. (Drazdou, 2025)

Sector Tailwinds

- **AI (Up)**: Hyper-growth driven by generative AI and data centres. (Rona & Levy, 2025)
- **Quick-Commerce (Down)**: High costs and consumer fatigue reduce funding. (Heather, 2025)
- **Green Tech, Health Tech, Fintech, E-Commerce (Up)**: Strong growth and investor interest. (MacGray, 2024), (Ronen, 2025)
- **EdTech (Emerging)**: In 2024, the EdTech market is worth $340B. By 2028, the global Edtech space is expected to grow to $696.04B with an annualised growth rate of 15.2%. (Howarth, 2024)

Secondary Opportunities

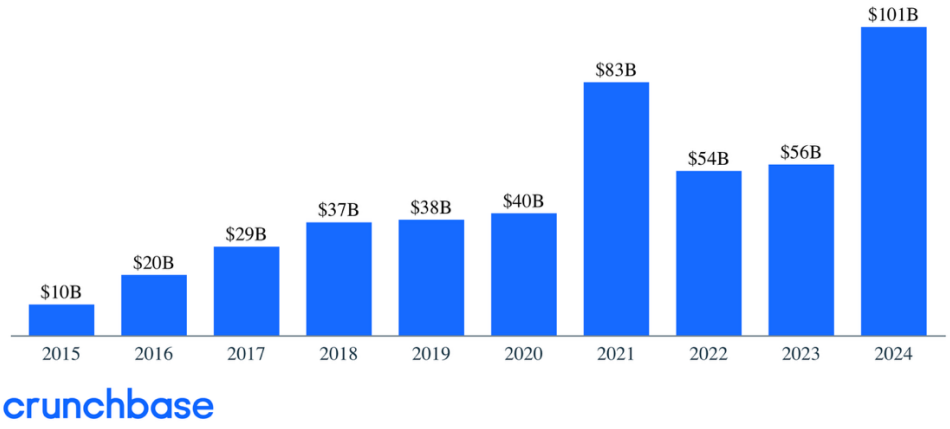
- **Discounted Stakes**: Non-AI sectors offer lower valuations for secondary purchases.
- **Liquidity Demand**: Delayed IPOs and tariff uncertainties push early stakeholders to sell, increasing secondary share supply. (Teare, 2024)
- **Exit Potential**: Rising M&A and IPO activity support secondary investments.
- **Strategic Sectors**: Focus on fintech, cybersecurity, green tech, and health tech.

Top Industries

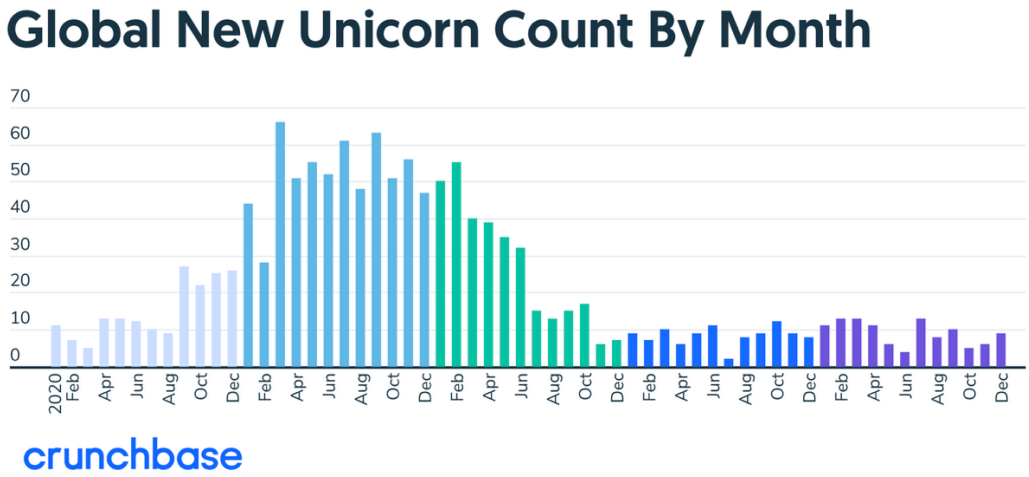
1. **AI**: a third of all VC funding, high valuations, and M&A activity.
2. **Green Tech**: Carbon capture and energy storage investments.
3. **Health Tech**: Personalised care and AI integration.
4. **Fintech**: Digital adoption and secondary opportunities.
5. **E-Commerce**: Resilient growth.
6. **EdTech**: AI-driven, emerging market.

Focus for Analysis

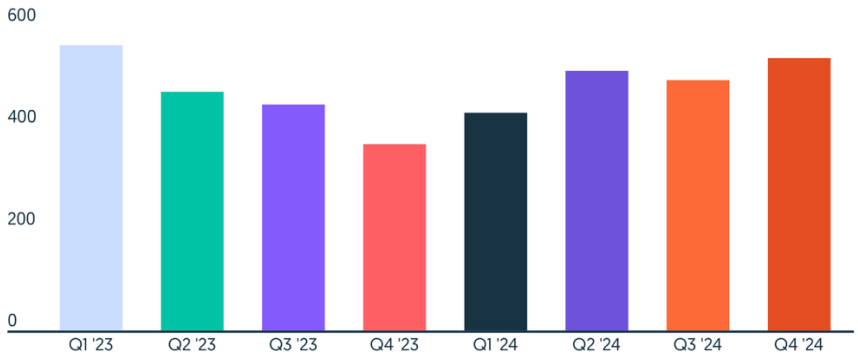
- **Primary Investments**: Target late-stage AI, green tech, and health tech with strong traction. Avoid quick-commerce.
- **Secondaries**: Acquire discounted stakes in fintech and cybersecurity, leveraging liquidity needs.
- **Risk Management**: Diversify across AI and non-AI sectors to balance valuation risks.



Graph 1.1 Total Funding to AI (Crunchbase, 2025).



Graph 1.2 Unicorns by Month (Crunchbase, 2025). The graph shows the number of unicorn companies in each month.



Graph 1.3 Venture-backed Startups M&A (Crunchbase, 2025). The graph demonstrates the number of mergers and acquisitions deals occurring in each quarter for venture-backed startups.

Introduction

The dataset was retrieved from the Crunchbase website. “Crunchbase is a leading platform that provides in-depth company information and data. By aggregating data from various sources, Crunchbase creates comprehensive company profiles that include key information such as funding, leadership, and news.” (Gong, 2025)
Crunchbase has a built-in tool for filtering companies. The filtered dataset could then be extracted in the form of a CSV file.

The extracted dataframe has some vital information, but ultimately it lacks certain details which are crucial for a meaningful analysis. The original table consists of 895 rows and these columns:

- Organisation Name
- Organisation Name URL
- Operating Status
- IPO Status
- Last Funding Type
- Full Description
- Industries
- Headquarters Location
- Description
- CB Rank (Company)
- Last Funding Amount
- Last Funding Amount Currency
- Last Funding Amount (in USD)
- Total Equity Funding Amount
- Total Equity Funding Amount Currency
- Total Equity Funding Amount (in USD)
- Total Funding Amount
- Total Funding Amount Currency
- Total Funding Amount (in USD)
- Top 5 Investors
- Exit Date
- Exit Date Precision
- Founded Date
- Founded Date Precision
- Investment Stage
- Number of Funding Rounds Funding Status
- Last Funding Date
- Investor Type
- Last Equity Funding Amount
- Last Equity Funding Amount Currency
- Last Equity Funding Amount (in USD)
- Last Equity Funding Type

Prepare and Process

Information about the data:

- Stored on the local device.
- Data is organised in the long format.
- The data seems to be reliable, original, comprehensive and current.
- The data contains information about the companies' industries, last funding round, etc., which would be useful in identifying patterns and trends.
- The file lacks some crucial information, such as valuation, which is essential in performing an accurate analysis.
- Some data is stored in the wrong format, which could pose an issue when querying.

Stage 1. Generating missing information

A Python script is introduced to tackle the issue of missing information. PyCharm is used for this process. Below is the list of new columns and their values that were added:

- "latitude"
- "longitude"
- "data_source"
- "arr_usd"
- "mrr_usd"
- "revenue_growth_yoy"
- "cac_usd"
- "ltv_usd"
- "ltv_cac_ratio"
- "user_customer_growth_yoy"
- "retention_rate"
- "nps"
- "mau"
- "dau"
- "founder_background"
- "team_size"
- "burn_rate_usd_month"
- "runway_months"
- "valuation_usd"
- "proprietary_technology"
- "patents_yn"
- "key_customers"
- "tam_usd"
- "exit_potential"

These two will be used for mapping the data on the Tableau map. The coordinates are generated using the Headquarters Location column values.

ARR is calculated by multiplying the Total Funding Amount by 0.15. If no value exists, then the default value of 1,000,000 is inserted. MRR estimates are ARR values divided by 12.

Three values are distributed in the CAC column. For Consumer SaaS – 10,000, for AI – 100,000, for Enterprise SaaS – 50,000.

Burn Rate value is allocated based on the Investment Stage. For Series A and B - 1,000,000, for C and D – 1,500,000, for null values in the Investment Stage – 1,000,000, for the anything else - 2000000

The Valuation of a company is calculated by multiplying its Last Equity Funding Amount by 5.

Artificial Intelligence – 100,000,000,000
Fintech or Blockchain – 50,000,000,000
SaaS – 30,000,000,000
IF the Industry column has a null value, and for anything else, then the value is set at 30,000,000,000.

LTV value is CAC value times 3

The latitude and longitude columns are missing values for most of the companies. The Python script is developed to take the Headquarters Location and, using the Nominatim library, generate accurate coordinates.

Stage 2. Importing the data into the database

PostgreSQL and PgAdmin4 are used for SQL queries.
The first step is to create a table in PgAdmin4. Then this table is populated with values generated from a Python script.

Below is the final result queried using Select * From startup_companies LIMIT 10:

organization_name	organization_name_url	operating_status	ipo_status	last_funding_type	full_description	industries	headquarters	description	cb_rank_com	last_funding	last_funding	last_funding	total_equity	total_equity	total_equity	total_funding	total_funding	total_funding	top_5_invest	exit_date	exit_date_pr	founded_dat	founded_dat	investment_c	number_of_fu	funding_stat	last_funding	investor_typ	last_equity_f	last_equity_f	last_equity_f	last_equity_f	latitude	longitude	data_source	arr_usd	mrr_usd	revenue_gro	cac_usd	ltv_usd	ltv_cac_ratio	user_custom	retention_ra	nps	mau	dau	founder_bac	team_size	burn_rate_us	runway_mon	valuation_us	proprietary_	patents_yn	key_custome	tam_usd	exit_potentia
Relational4	https://www.crunchbase.c	Active	Private	Series B	Relational4 is buil	Artificial Inte	Berkeley, Ca	Relational4	3,248	75000000	USD	75000000	115000000	USD	115000000	115000000	USD	115000000	Menlo Ventu	NULL	NULL	01/09/2017	day	NULL	2	Early Stage V	26/04/2022	NULL	75000000	USD	75000000	Series B	37.8708393	-122.27286	Needs Coord	17250000	1437500	30	100000	300000	3	20	90	N/A	N/A	N/A	Unknown	1000	10000000	75	375000000	Likely	Y	Enterprise	1E+11	M&A
Alsera	https://www.crunchbase.c	Active	Private	Series D	Alsera is a leading	Artificial Inte	Palo Alto, Ca	Alsera is a lea	895	90000000	USD	90000000	164499999	USD	164499999	164499999	USD	164499999	Norwest Ven	NULL	NULL	01/01/2017	year	NULL	4	Late Stage Ve	03/08/2022	NULL	90000000	USD	90000000	Series D	37.4443293	-122.15985	Needs Coord	24674999.6	2056249.99	30	100000	300000	3	20	90	N/A	N/A	N/A	Unknown	1000	10000000	90	450000000	Likely	Y	Enterprise	1E+11	M&A
Fountain	https://www.crunchbase.c	Active	Private	Series C	Fountain is high-vc	Human Reso	San Francisc	Fountain is a	9,435	100000000	USD	100000000	218870000	USD	218870000	218870000	USD	218870000	B Capital, Y	NULL	NULL	01/01/2014	year	NULL	6	Late Stage Ve	15/06/2022	NULL	100000000	USD	100000000	Series C	37.7792588	-122.41933	Needs Coord	32830500	2735875	30	10000	30000	3	20	80	N/A	N/A	N/A	Unknown	1000	10000000	100	500000000	Uncertain	N	SMB/Consun	3E+10	IPO likely
Branch	https://www.crunchbase.c	Active	Private	Series C	Branch helps busir	Banking, Fin	Minneapolis,	Branch helps	488	75000000	USD	75000000	133120000	USD	133120000	633120000	USD	633120000	General Atla	NULL	NULL	01/11/2015	month	NULL	6	Late Stage Ve	09/03/2022	NULL	75000000	USD	75000000	Series C	44.9772995	-93.265469	Needs Coord	19968000	1664000	30	50000	150000	3	20	90	N/A	N/A	N/A	Unknown	1000	10000000	75	375000000	Uncertain	N	SMB/Consun	5E+10	M&A
Onfido	https://www.crunchbase.c	Active	Private	Series D	Onfido is a provide	FinTech, Frai	London, Engl	Onfido is a pr	2,110	100000000	USD	100000000	182117808	USD	182117808	182136505	USD	182136505	Plug and Play	09/04/2024	day	01/01/2012	year	NULL	13	M&A	15/04/2020	NULL	100000000	USD	100000000	Series D	51.4883335	-0.1440551	Needs Coord	27317671.2	2276472.6	30	50000	150000	3	20	90	N/A	N/A	N/A	Unknown	1000	10000000	100	500000000	Uncertain	N	SMB/Consun	5E+10	M&A
Cresta	https://www.crunchbase.c	Active	Private	Series D	Cresta uses artific	Artificial Inte	Palo Alto, Ca	Cresta uses a	19,677	125000000	USD	125000000	276000000	USD	276000000	276000000	USD	276000000	Andreessen H	NULL	NULL	01/01/2017	year	NULL	5	Late Stage Ve	19/11/2024	NULL	125000000	USD	125000000	Series D	37.4443293	-122.15985	Needs Coord	41400000	3450000	30	100000	300000	3	50	90	N/A	N/A	N/A	Unknown	1000	10000000	125	625000000	Uncertain	Y	Enterprise	1E+11	IPO likely
Super.com	https://www.crunchbase.c	Active	Private	Series C	Super.com (fka Sru	E-Commerce	San Francisc	Super.com is	3,676	60000000	USD	60000000	155200000	USD	155200000	200200000	USD	200200000	Inovia Capita	NULL	NULL	01/01/2016	year	NULL	8	Late Stage Ve	24/04/2023	NULL	60000000	USD	60000000	Series C	37.7792588	-122.41933	Needs Coord	23280000	1940000	30	50000	150000	3	20	90	N/A	N/A	N/A	Unknown	1000	10000000	60	300000000	Uncertain	N	SMB/Consun	5E+10	M&A
Valon Technologies	https://www.crunchbase.c	Active	Private	Series C	Valon Mortgage is	Financial Ser	New York, Ne	Valon is a tec	4,256	100000000	USD	100000000	197100000	USD	197100000	197100000	USD	197100000	Andreessen H	NULL	NULL	01/07/2019	day	NULL	4	Late Stage Ve	23/10/2024	NULL	100000000	USD	100000000	Series C	40.7127281	-74.006015	Needs Coord	29565000	2463750	30	50000	150000	3	50	90	N/A	N/A	N/A	Unknown	1000	10000000	100	500000000	Uncertain	N	SMB/Consun	5E+10	M&A
Safely You	https://www.crunchbase.c	Active	Private	Series C	Safely You is the ar	Artificial Inte	San Francisc	Safely You pr	5,860	43000000	USD	43000000	102500000	USD	102500000	114324998	USD	114324998	National Sci	NULL	NULL	01/01/2015	year	NULL	9	Late Stage Ve	28/01/2025	NULL	43000000	USD	43000000	Series C	37.7792588	-122.41933	Needs Coord	15375000	1281250	30	100000	300000	3	50	90	N/A	N/A	N/A	Unknown	1000	10000000	43	215000000	Uncertain	Y	Enterprise	1E+11	M&A
Eleos Health	https://www.crunchbase.c	Active	Private	Series C	Eleos Health provi	Artificial Inte	Boston, Mass	Eleos Health	676	60000000	USD	60000000	128000000	USD	128000000	128000000	USD	128000000	Menlo Ventu	NULL	NULL	01/01/2019	year	NULL	6	Late Stage Ve	22/01/2025	NULL	60000000	USD	60000000	Series C	42.3554334	-71.060511	Needs Coord	19200000	1600000	30	100000	300000	3	50	90	N/A	N/A	N/A	Unknown	1000	10000000	60	300000000	Uncertain	Y	Enterprise	1E+11	M&A

Image 2.1 Startup Companies table. The table shows the list of the first ten rows of the table.

Stage 3. Cleaning the table

Checking datatypes

The query returns the list of columns with the datatypes of their values. Some data types need to be changed so that the information is stored in the correct format for accuracy and the avoidance of querying problems. The result is presented below:

Images 2.1, 2.2 and 2.3 Startup Companies Columns and datatypes. The screenshots below show the names of the columns, their number and the type of data they store inside them.

	column_name name	data_type character varying
1	organization_name	text
2	organization_name_url	text
3	operating_status	text
4	ipo_status	text
5	last_funding_type	text
6	full_description	text
7	industries	text
8	headquarters_location	text
9	description	text
10	cb_rank_company	text
11	last_funding_amount	numeric
12	last_funding_amount_currency	text
13	last_funding_amount_usd	numeric
14	total_equity_funding_amount	numeric
15	total_equity_funding_amount_currency	text
16	total_equity_funding_amount_usd	numeric
17	total_funding_amount	bigint
18	total_funding_amount_currency	text
19	total_funding_amount_usd	bigint
20	top_5_investors	text
21	exit_date	text
22	exit_date_precision	text
23	founded_date	text
24	founded_date_precision	text
25	investment_stage	text
26	number_of_funding_rounds	integer

	column_name name	data_type character varying
27	funding_status	text
28	last_funding_date	text
29	investor_type	text
30	last_equity_funding_amount	numeric
31	last_equity_funding_amount_currency	text
32	last_equity_funding_amount_usd	numeric
33	last_equity_funding_type	text
34	latitude	double precision
35	longitude	numeric
36	data_source	text
37	arr_usd	numeric
38	mrr_usd	numeric
39	revenue_growth_yoy	numeric
40	cac_usd	integer
41	ltv_usd	numeric
42	ltv_cac_ratio	numeric
43	user_customer_growth_yoy	integer
44	retention_rate	integer
45	nps	character varying
46	mau	character varying
47	dau	character varying
48	founder_background	text
49	team_size	integer
50	burn_rate_usd_month	integer
51	runway_months	numeric
52	valuation_usd	numeric

53	proprietary_technology	text
54	patents_yn	text
55	key_customers	text
56	tam_usd	bigint
57	exit_potential	text

There are a few columns with dates stored in a TEXT format:

	column_name name	data_type character varying
1	exit_date	text
2	exit_date_precision	text
3	founded_date	text
4	founded_date_precision	text
5	last_funding_date	text

Image 2.4 Date table. The picture shows the columns with the word “Date” in them.

	exit_date text	exit_date_precision text	founded_date text	founded_date_precision text	last_funding_date text
1	[null]	[null]	01/09/2017	day	26/04/2022
2	[null]	[null]	01/01/2017	year	03/08/2022
3	[null]	[null]	01/01/2014	year	15/06/2022
4	[null]	[null]	01/11/2015	month	09/03/2022
5	09/04/2024	day	01/01/2012	year	15/04/2020
6	[null]	[null]	01/01/2017	year	19/11/2024
7	[null]	[null]	01/01/2016	year	24/04/2023
8	[null]	[null]	01/07/2019	day	23/10/2024
9	[null]	[null]	01/01/2015	year	28/01/2025
10	[null]	[null]	01/01/2019	year	22/01/2025

Image 2.5 Date table with values.

There are five columns with the word “date” in their names. Values in three of these columns should be converted to date format: exit_date, founded_date, and last_funding_date.

The result of the conversion:

	column_name name	data_type character varying
1	exit_date	date
2	exit_date_precision	text
3	founded_date	date
4	founded_date_precision	text
5	last_funding_date	date

Image 2.6 Date table after conversion.

Three columns: nps, mau, dau have little to no data in them. It would be wiser to have them removed.

	nps character varying	mau character varying	dau character varying
1	N/A	1000000	N/A
2	N/A	100000	N/A
3	N/A	100000	N/A
4	N/A	100000	N/A
5	N/A	100000	N/A
6	N/A	100000	N/A
7	N/A	100000	N/A
8	N/A	100000	N/A
9	N/A	100000	N/A
10	N/A	100000	N/A

Image 2.7 Three empty columns.

Table after the removal of the nps, mau, dau: Images 2.8 and 2.9 Updated Table.

	column_name name	data_type character varying
1	organization_name	text
2	organization_name_url	text
3	operating_status	text
4	ipo_status	text
5	last_funding_type	text
6	full_description	text
7	industries	text
8	headquarters_location	text
9	description	text
10	cb_rank_company	text
11	last_funding_amount	numeric
12	last_funding_amount_currency	text
13	last_funding_amount_usd	numeric
14	total_equity_funding_amount	numeric
15	total_equity_funding_amount_currency	text
16	total_equity_funding_amount_usd	numeric
17	total_funding_amount	bigint
18	total_funding_amount_currency	text
19	total_funding_amount_usd	bigint
20	top_5_investors	text
21	exit_date	date
22	exit_date_precision	text
23	founded_date	date
24	founded_date_precision	text

	column_name name	data_type character varying
28	last_funding_date	date
29	investor_type	text
30	last_equity_funding_amount	numeric
31	last_equity_funding_amount_currency	text
32	last_equity_funding_amount_usd	numeric
33	last_equity_funding_type	text
34	latitude	double precision
35	longitude	numeric
36	data_source	text
37	arr_usd	numeric
38	mrr_usd	numeric
39	revenue_growth_yoy	numeric
40	cac_usd	integer
41	ltv_usd	numeric
42	ltv_cac_ratio	numeric
43	user_customer_growth_yoy	integer
44	retention_rate	integer
45	founder_background	text
46	team_size	integer
47	burn_rate_usd_month	integer
48	runway_months	numeric
49	valuation_usd	numeric
50	proprietary_technology	text
51	patents_yn	text
52	key_customers	text

Checking for duplicates and NULL values

The dataset might have duplicate values for various rows and columns, but these repeating values are valid and do not make the data redundant. However, the potential existence of duplicate rows will pose a threat to the data's reliability and accuracy.

RStudio is used to check the dataframe for duplicate rows and null values.

Libraries, necessary for the tasks described above, are installed and loaded into the software, then RStudio is connected to the database and the table is printed:

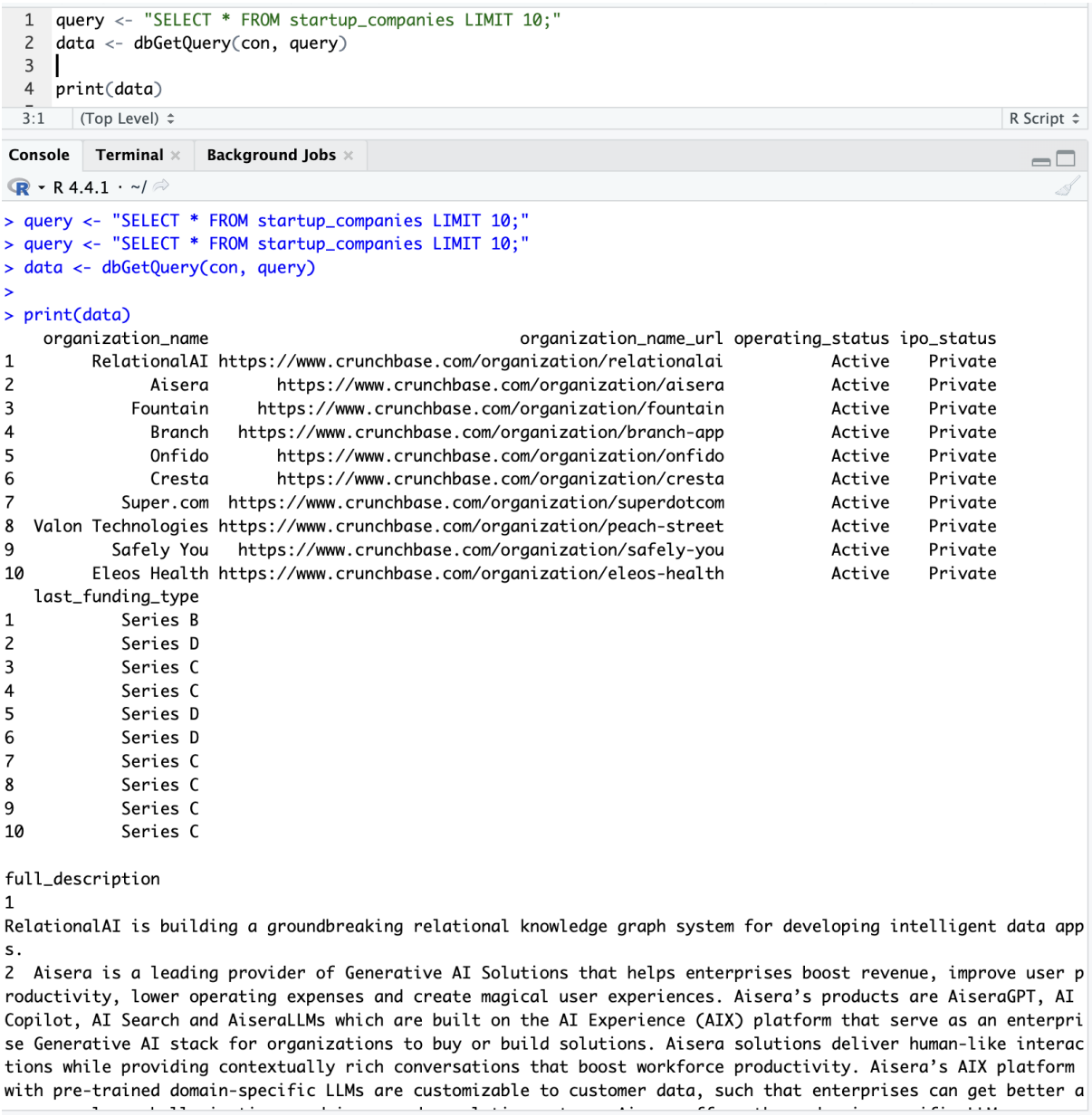


Image 3.1 RStudio Duplicate Rows. No duplicate rows were found.

The table lacks some information, therefore, some of the rows might contain either NULL values or N/A. Three columns will be checked for missing data to identify the percentage of empty spaces: exit_date, founded_date, and last_funding_date.
To avoid any clashes and errors, these columns were converted to character strings and then converted back to DATE format. Then the total number of NULL values was calculated for each column:

```
> missing_counts <- sapply(data, count_missing)
> print(missing_counts)
        exit_date      founded_date last_funding_date
              790                 0                 0
```

Image 3.2 Number of NULL values. The screenshot shows that the exit_date column has 790 rows of NULL values while the other two have 0.

```
> missing_proportions <- missing_counts / nrow(data)
> print(missing_proportions)
        exit_date      founded_date last_funding_date
        0.8826816         0.0000000         0.0000000
```

Image 3.3 Percentage of NULL values. The picture demonstrates that exit_date has 88% of missing values.
Even though the exit_date column has only 12% of actual figures, it is still important to keep it as the companies that have IPOed are of no interest to the LTC.

Checking the logical order of the dates

Dates should follow logical order, therefore, the dates that are out of place will be set to NULL.

```
> # Validate dates by setting invalid values to NULL
> dbExecute(con, "
+   UPDATE startup_companies
+   SET exit_date = NULL
+   WHERE exit_date > CURRENT_DATE OR exit_date < founded_date;
+ ")
[1] 0
>
> dbExecute(con, "
+   UPDATE startup_companies
+   SET last_funding_date = NULL
+   WHERE last_funding_date > CURRENT_DATE OR last_funding_date < founded_date;
+ ")
[1] 0
>
> dbExecute(con, "
+   UPDATE startup_companies
+   SET founded_date = NULL
+   WHERE founded_date > CURRENT_DATE;
+ ")
[1] 0
```

Image 3.4 Logical order of the dates. The screenshot demonstrates that there are 0 dates that are out of order.

Analyse and Share

This section will focus on performing analysis on the cleaned table and delivering graphs for visual demonstrations.

Stage 1. Filtering the table.

Below is the criteria for diminishing the size of the table:

- Founded from 2013 to 2018
- Last Founding round over 12 months ago
- No crypto/blockchain
- Minimum raised 100m in total
- The number of funding rounds is at least 4
- Companies that haven't IPOed, where exit_date is NULL.

The result:

```
1  select * from startups_best
```

Data Output | Messages | Notifications

Showing rows: 1 to 335 | Page No: 1 of 1

	organization_name text	organization_name_url text	operating_status text	ipo_status text	last_funding_type text	full_description text
2	Dataiku	https://www.crunchbase.com/organization/dataiku	Active	Private	Series F	Dataiku is a centralized
3	Sysdig	https://www.crunchbase.com/organization/sysdig	Active	Private	Series G	In the cloud, every secon
4	Personio	https://www.crunchbase.com/organization/personio	Active	Private	Series E	Personio is the People O
5	Cerebras Systems	https://www.crunchbase.com/organization/cerebras-systems	Active	Private	Secondary Market	Cerebras Systems focus
6	Branch	https://www.crunchbase.com/organization/branch-metrics	Active	Private	Series F	Branch is the linking and
7	BlueVoyant	https://www.crunchbase.com/organization/blueteamglobal	Active	Private	Series E	BlueVoyant is a cybersec
8	PingCAP	https://www.crunchbase.com/organization/pingcap	Active	Private	Series D	PingCAP is an open-sour
9	Branch	https://www.crunchbase.com/organization/branch-app	Active	Private	Series C	Branch helps businesses
10	Cockroach Labs	https://www.crunchbase.com/organization/cockroach-labs	Active	Private	Series F	Cockroach Labs is behin
11	Clear Street	https://www.crunchbase.com/organization/clear-street-8e71	Active	Private	Series B	Clear Street is modernizi
12	Dutchie	https://www.crunchbase.com/organization/dutchie	Active	Private	Series D	Dutchie is an all-in-one t
13	Forto	https://www.crunchbase.com/organization/forto	Active	Private	Series D	Forto was founded in Be
14	Upgrade	https://www.crunchbase.com/organization/upgrade	Active	Private	Series F	Upgrade is a neobank th
15	Gong	https://www.crunchbase.com/organization/gong-io	Active	Private	Secondary Market	Gong is a revenue intellig
16	Sigma Computing	https://www.crunchbase.com/organization/sigma-computing	Active	Private	Series D	Sigma Computing is clou
17	Arcadia	https://www.crunchbase.com/organization/arcadia-power-2	Active	Private	Series E	Arcadia is a climate crisi
18	HiBob	https://www.crunchbase.com/organization/hibob	Active	Private	Series D	Hibob was founded to m
19	Verbit	https://www.crunchbase.com/organization/verbit-ai	Active	Private	Secondary Market	Verbit.ai is an AI compar
20	Motive	https://www.crunchbase.com/organization/Motive	Active	Private	Secondary Market	Motive is a technology c
21	Vercel	https://www.crunchbase.com/organization/vercel	Active	Private	Series E	Vercel is a software com
22	Thought Machine	https://www.crunchbase.com/organization/thoughtmachine	Active	Private	Series D	Thought Machine is a Fir
23	Illumio	https://www.crunchbase.com/organization/illumio	Active	Private	Series F	Illumio, the most compre
24	Greenlight	https://www.crunchbase.com/organization/greenlight-com	Active	Private	Series D	Greenlight is a debit card
25	ClickUp	https://www.crunchbase.com/organization/clickup	Active	Private	Series C	ClickUp offers a customi

Total rows: 335 | Query complete 00:00:00.077 | LF | Ln 1, Col 14

Image 4.1 Table startups_best. The new table consists only of 335 rows compared to 895 rows in the table startup_companies.

Stage 2. Identifying funding trends

The graph below shows the relationship between the Total Equity Funding and the Total Funding Rounds. Below are the key observations:

1) Concentration of Funding: Most startups with 4 to 10 funding rounds have total equity funding amounts between 200,000,000 and 600,000,000 USD, with the highest density around 300,000,000 to 500,000,000 USD.
2) Spread with More Rounds: As the number of funding rounds increases beyond 10, the funding amounts become more spread out, with fewer startups reaching higher funding levels (e.g., 600,000,000+ USD).
3) Outliers: There are a few startups with 14-16 funding rounds that have funding amounts exceeding 600,000,000 USD, though these are less common (low count).
4) Lower Funding: Startups with fewer rounds (4-6) tend to have lower total funding, often below 400,000,000 USD.

Interpretation : The graph suggests that startups typically accumulate significant equity funding (300M-500M USD) after 4-10 funding rounds. Beyond 10 rounds, funding amounts vary widely, indicating either highly successful ventures or diminishing returns on additional rounds. The sparse data at higher rounds and funding levels suggests these are less common scenarios.

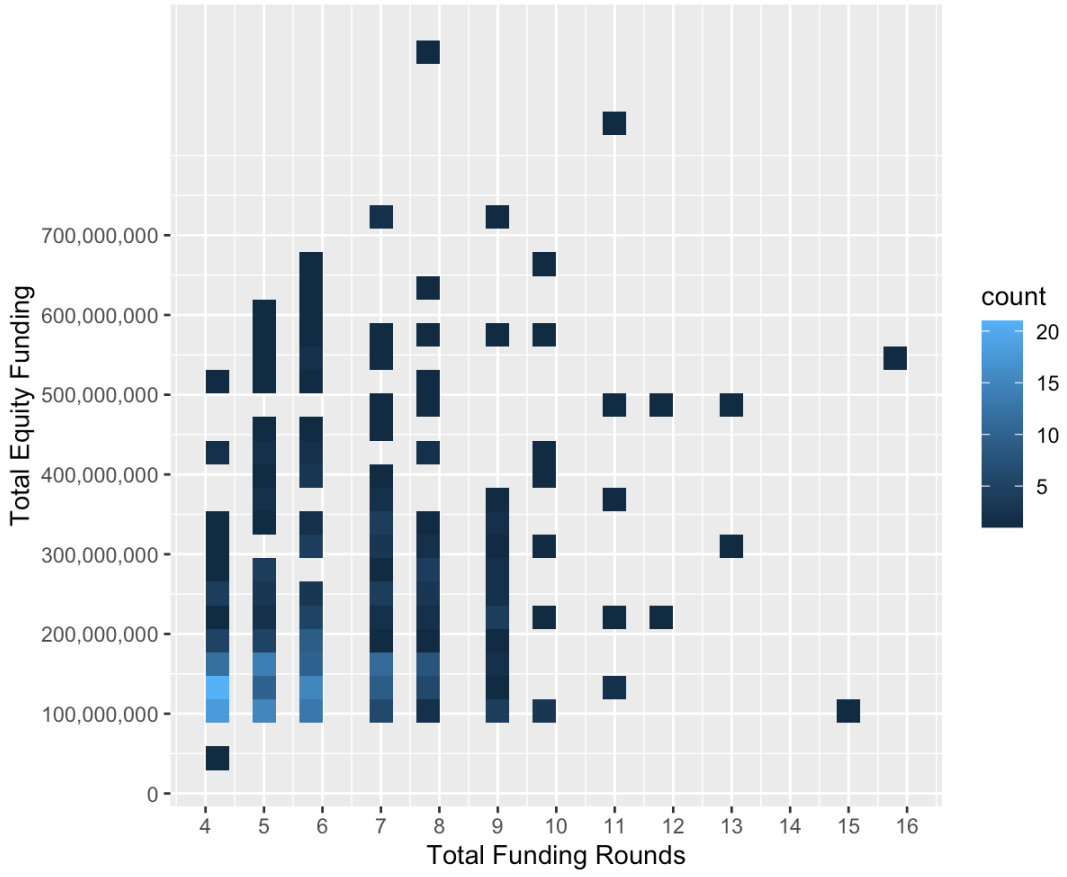


Image 4.2 Funding trends (RStudio). The graph shows the funding trends of the various companies.

Stage 3. Location and Industry

This stage will involve the creation of tables and graphs demonstrating trends in industry types and locations.

Industries

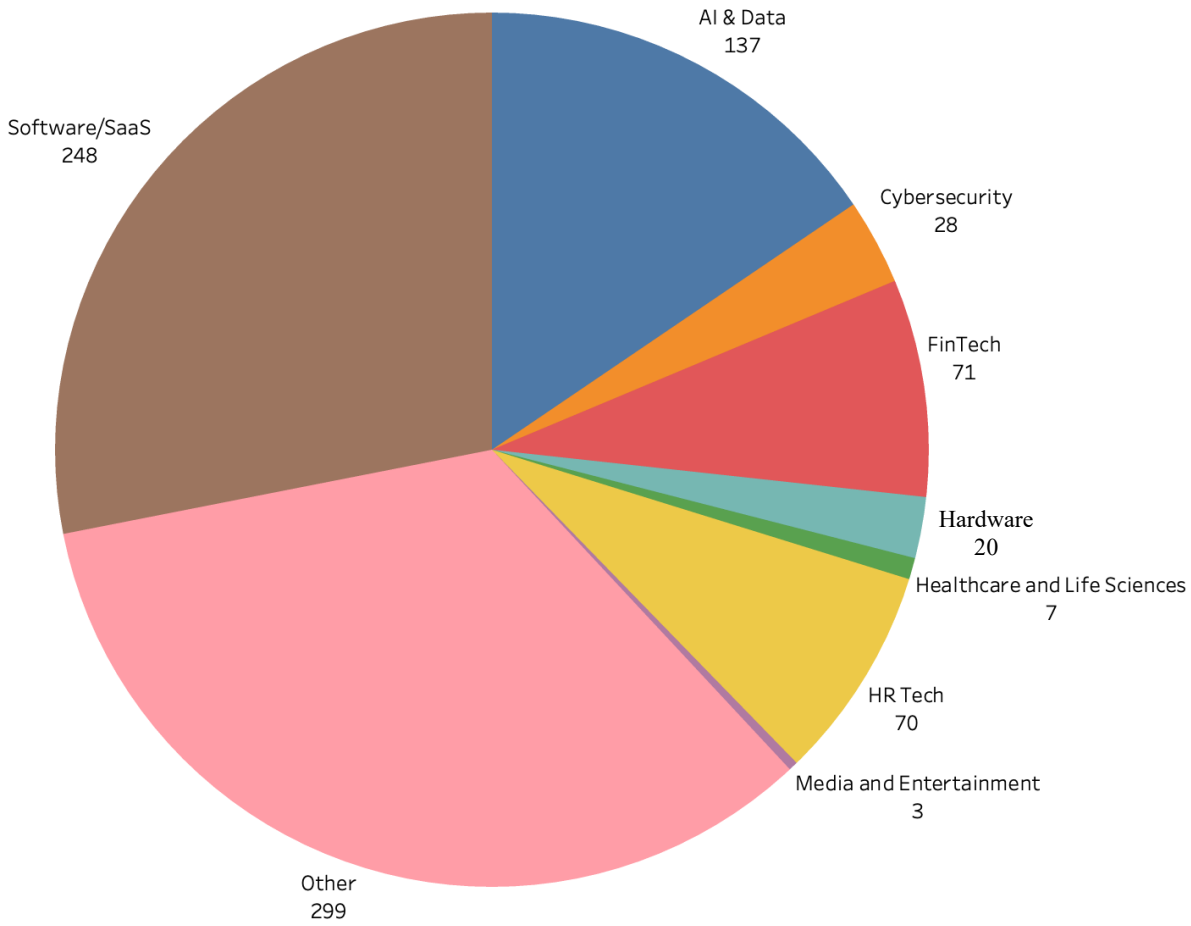
	category text	company_count bigint
1	Other	299
2	Software/SaaS	248
3	AI & Data	137
4	FinTech	71
5	HR Tech	70
6	Cybersecurity	28
7	Hardware	20
8	Healthcare and Life Sciences	7
9	Media and Entertainment	3

Image 4.3 Startup industries (PGAdmin4). The table above shows the number of companies falling under specific categories.

Industries Findings:

- Software/SaaS is the leading industry, making up a large portion of the companies (248 out of the 873 total counts), reflecting the tech-heavy nature of the startup ecosystem.
- The total count (873) exceeds the number of unique companies (335) because companies can belong to multiple industries. For example, a company like SpotOn contributes to both FinTech and Software/SaaS.
- Smaller segments like Media and Entertainment (3 companies) and Healthcare and Life Sciences (7 companies) indicate these industries are less common among the startups in this dataset.
- The presence of categories like AI & Data (137) and FinTech (71) shows a diverse range of tech-focused industries, though Software/SaaS and "Other" dominate.

Notes: Some companies may fall under multiple categories due to the nature of their business and products.



Graph 4.4 Startup industries Pie Chart (Tableau). A visual representation of the number of companies in each category.

Industry Funding

	category text	total_funding_amount numeric
1	Other	179728317844.0
2	Software/SaaS	79924227701.0
3	AI & Data	46186817210.0
4	FinTech	34834320601.0
5	HR Tech	20907968847.0
6	Cybersecurity	14742453576.0
7	Hardware	4830966952.0
8	Healthcare and Life Sciences	1375647931.0
9	Media and Entertainment	425292196.0

Image 4.5 Industry Funding (PGAdmin4). The table shows the amounts of funding each industry receives.

Industry Funding Findings:

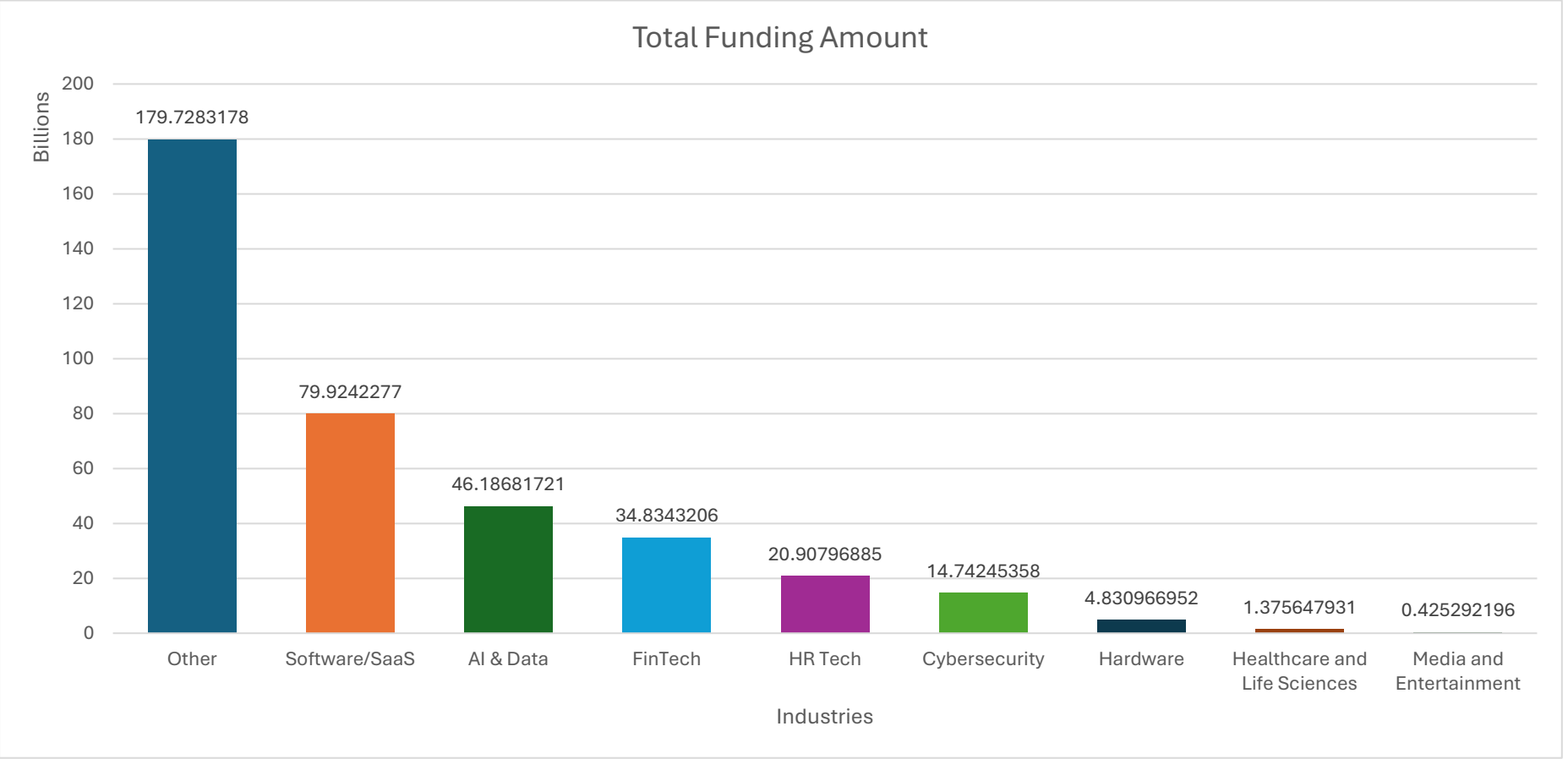
The table and the bar chart display the **total funding amount** (in billions of USD) for companies across different industries, derived from the startups_best table. Each bar represents an industry category, and the height of the bar corresponds to the total equity funding amount in billions.

Key Insights:

1. **Funding Distribution:** The "Other" category dominates with 179.728B USD, suggesting that a wide variety of industries outside the predefined categories are attracting significant investment. Software/SaaS, while having the most companies (248), has less funding than "Other", indicating that funding per company might be more spread out in this sector.
2. **High Investment Sectors:** AI & Data (46.187B USD) and FinTech (34.834B USD) show strong investor confidence, likely due to their potential for innovation and growth.
3. **Niche Industries:** Healthcare and Life Sciences (1.376M USD) and Media and Entertainment (0.425B USD) have minimal funding, consistent with their small company counts, suggesting these sectors are less attractive to investors in this dataset.
4. **Overlap in Funding**: Since companies can belong to multiple industries (e.g., SpotOn in both FinTech and Software/SaaS), their funding is counted in each category, which may inflate totals. The sum of funding across categories (378.01B USD) exceeds the actual total funding for 335 unique companies.

Notes:

- **Units:** The y-axis is in billions of USD, so the values are large, reflecting the scale of investment in these startups.
- **Context:** The chart aligns with earlier analyses showing Software/SaaS and "Other" as dominant in company count, but funding distribution highlights different priorities for investors, with "Other" leading in total funding.
- **Visualisation:** The bar chart effectively highlights disparities in funding across industries, with a clear visual hierarchy from "Other" to smaller sectors like Media and Entertainment.



Graph 4.6 Industries Funding (Tableau). The graph demonstrates the amount of funding each industry has received.

Industries Average Funding

The table bar chart displays the **average funding amount** (in millions of USD) for companies across various industry categories, derived from the startups_best table. Each bar represents an industry category, and the height of the bar corresponds to the average equity funding amount per company in that category.

Key Insights:

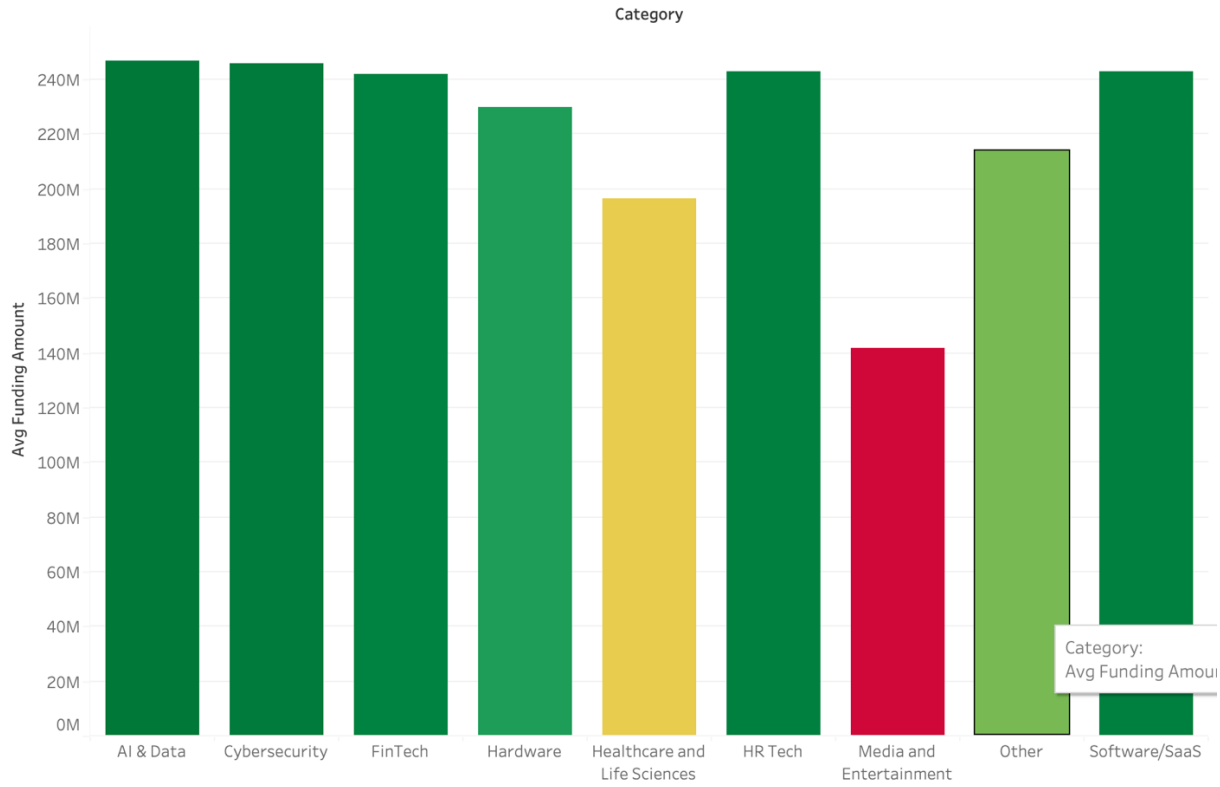
1. **High-Average Sectors**: AI & Data, Cybersecurity, and FinTech lead with an average of ~240M USD per company, indicating these industries are seen as high-growth or high-risk/high-reward areas by investors.
2. **Lower Average in Software/SaaS**: Despite being the largest category by company count (248), Software/SaaS has the lowest average (~110M USD), suggesting a larger number of smaller-funded companies in this sector.
3. **Niche High Funding**: Healthcare and Life Sciences (~200M USD) and Hardware (~220M USD) show high averages despite lower company counts, likely due to the capital-intensive nature of these industries.
4. **"Other" Category**: The high average (~210M USD) for "Other" suggests that unclassified industries include some well-funded outliers, contributing to the overall funding disparity.

Notes:

- **Units**: The y-axis is in millions of USD (M), making the values more manageable compared to the billions seen in total funding.
- **Overlap Consideration**: The averages are calculated with companies potentially counted in multiple categories (e.g., a company like SpotOn in both FinTech and Software/SaaS), which could skew the averages.
- **Context**: The chart complements the total funding chart, where "Other" led in total funding (179.728B USD) due to its high company count (299), but here the average funding highlights industries with higher per-company investment.

	category text	avg_funding_amount numeric
1	AI & Data	246988327.32620321
2	Cybersecurity	245707559.60000000
3	HR Tech	243115916.82558140
4	Software/SaaS	242930783.28571429
5	FinTech	241905004.17361111
6	Hardware	230046045.33333333
7	Other	213962283.14761905
8	Healthcare and Life Sciences	196521133.00000000
9	Media and Entertainment	141764065.33333333

Image 4.7 AVG Industry Funding (PGAdmin4). The table shows the average funding amount for each industry.



Graph 4.8 AVG Industries Funding (Tableau). The bar chart shows the average amount invested in each industry.

Location

The table shows the number of founded companies within each country, and the map visualises the distribution of companies from the startups_best table across different countries, with the size and colour of each marker indicating the number of companies in each country. The map uses a geographic representation, with markers placed approximately at the centroids of the respective countries.

Key Insights:

1. **Dominance of the United States**: The United States stands out with 276 companies, far exceeding any other country. This suggests a strong concentration of startup activity in the U.S., likely due to its large market size and robust venture capital ecosystem.
2. **European Distribution**: Europe has a diverse but smaller presence, with the United Kingdom (19) and Germany (15) leading, followed by France (10). Northern and Eastern Europe have fewer companies, with many countries represented by just 1-3 companies.

3. **Geographic Spread**: The map shows a clear divide, with the vast majority of companies in North America (U.S.) and a scattered but thinner presence across Europe. No companies are indicated in other regions (e.g., Asia, Africa, South America), as the dataset is focused on North America and Europe.
4. **Scale Representation**: The varying sizes of the markers effectively illustrate the disparity in company counts, with the U.S. marker being significantly larger than the others.

Notes:

- **Data Alignment**: The counts match the data (e.g., United States: 276, United Kingdom: 19), confirming the map's accuracy based on the SQL query results.
- **Visualisation Limitation**: The map uses approximate locations (country centroids), which do not reflect the exact distribution within countries (e.g., companies might be concentrated in specific cities like San Francisco or London).
- **Missing Data**: Countries with zero companies (e.g., outside North America and Europe) are not marked, which is consistent with the dataset's focus.
- **Incorrect Value**: One of the rows had a "City of" value stored in the third place instead of the country's name. "City of" value was removed.

	country text	company_count bigint
1	United States	276
2	United Kingdom	19
3	Germany	15
4	France	10
5	Norway	3
6	Denmark	3
7	Sweden	2
8	Austria	2
9	Belgium	1
10	Finland	1
11	Lithuania	1
12	Estonia	1

Image 4.9 Startup locations (PGAdmin4). The table shows the number of companies founded within each country.

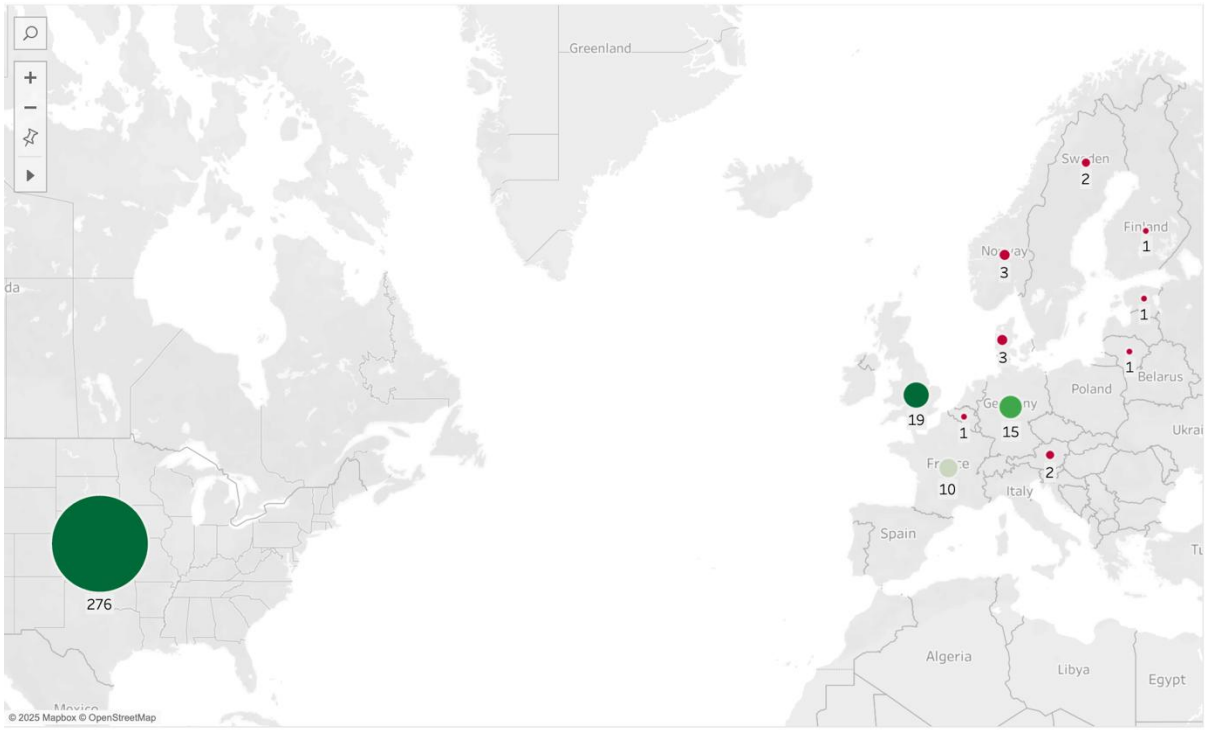


Image 4.10 Startup Locations (Tableau).

Location Funding

The table and the map below describe and visualise the **total funding amount** (in USD) for companies across different countries. Each marker is placed approximately at the centroid of the respective country, with the size and colour of the marker indicating the total funding amount.

Key Insights:

1. **U.S. Dominance**: The United States leads with 63.8B USD, far exceeding all other countries, consistent with its 276 companies. This highlights the U.S. as the primary hub for startup funding in this dataset.
2. **European Funding**: Europe shows a varied distribution, with the United Kingdom (2.69B USD) and Norway (524M USD) having the highest totals among European countries, despite lower company counts. This suggests higher average funding per company in these regions.
3. **Disparity in Funding**: Countries with fewer companies, like Finland (420M USD for 1 company) and Belgium (176.8M USD for 1 company), show high per-company funding, while others like Lithuania (17,620 USD) and Estonia (158,820 USD) have minimal totals.

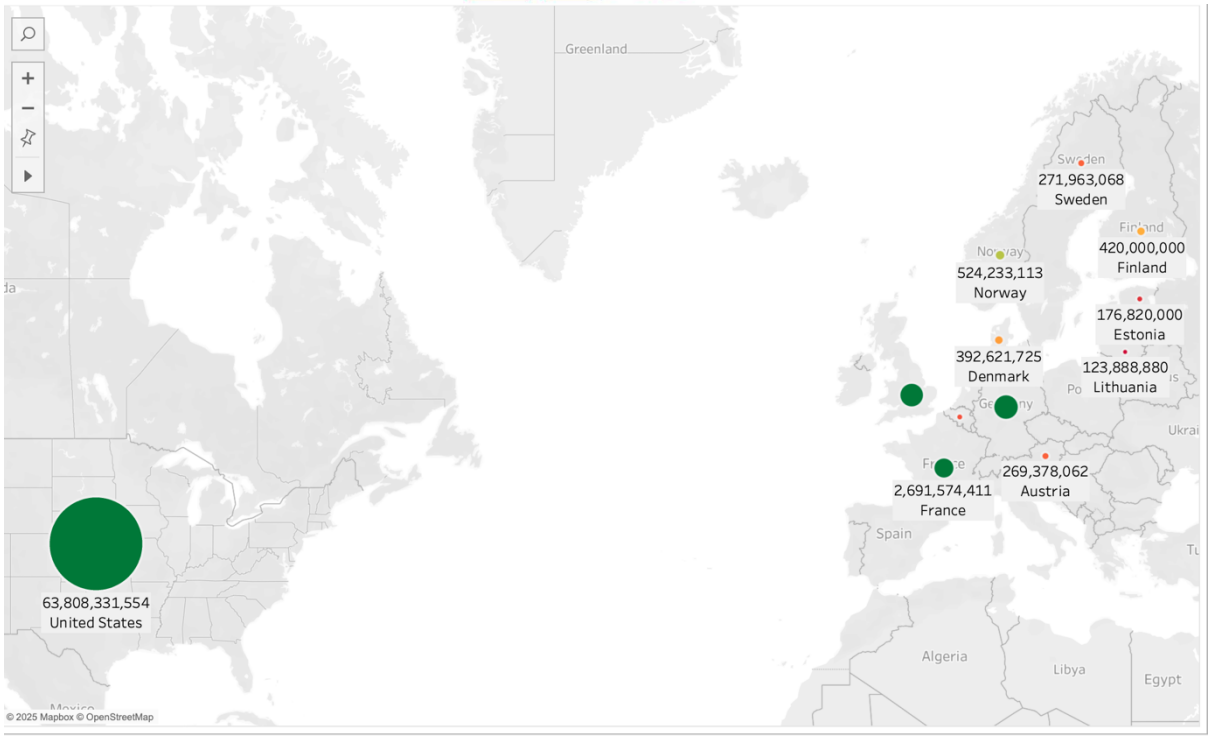
4. **Geographic Focus**: The map focuses on North America and Europe, with no funding indicated for other regions (e.g., Asia, Africa), as the dataset is limited to these areas.

Notes:

- **Units**: Funding amounts are in USD, with the U.S. total in billions and others in millions or thousands, reflecting the scale difference.
- **Colour Coding**: The map uses green for higher funding (e.g., U.S., U.K.) and orange for lower amounts.
- **Data Alignment**: The totals match the earlier SQL query results for countries, confirming the map's accuracy based on the aggregated funding data.

	country text	total_funding_amount numeric
1	United States	63808331554.0
2	Germany	4102072536.0
3	United Kingdom	3798886439.0
4	France	2691574411.0
5	Norway	524233113.0
6	Finland	420000000.0
7	Denmark	392621725.0
8	Sweden	271963068.0
9	Austria	269378062.0
10	Belgium	236799209.0
11	Estonia	176820000.0
12	Lithuania	123888880.0

Image 4.11 Countries Funding (PGAdmin4). The table shows the total number of funding companies received in each country of their origin.



Graph 4.12 Countries Funding (Tableau). The graph demonstrates the allocation of funds for each country.

Location Average Funding

The table and the map below display the **average funding amount** (in millions of USD) for companies across different countries. Each marker is placed approximately at the centroid of the respective country, with the size and colour of the marker corresponding to the average equity funding amount per company in that country.

Key Insights:

1. **High-Average Funding in Small Countries**: Finland (420M USD), Belgium (176.820M USD), and Norway (174.744M USD) show the highest average funding per company, despite having very few companies (1-3). This suggests that these countries have a few highly funded startups.
2. **United States Average**: The U.S., with 231.189M USD, has a high average but is not the highest due to its large company count (276). This indicates a broad distribution of funding across many companies.
3. **Low-Average Funding in Some Countries**: France (0.295M USD), Austria (0.063M USD), Estonia (0.159M USD), and Lithuania (0.018M USD) have very low averages, reflecting minimal investment per company.
4. **European Disparity**: European countries show a wide range of averages, from high (Finland, Norway) to very low (Lithuania, Austria), highlighting varying levels of startup investment across the region.

Notes:

- **Units**: Average funding amounts are in millions of USD, making the values easier to compare across countries.
- **Data Alignment**: The averages match the earlier SQL query results (e.g., United States: 63.808B USD ÷ 276 = 231.189M USD), confirming the chart's accuracy.
- **Context**: This chart complements the total funding map, where the U.S. dominated in total funding (63.808B USD), but here the average funding highlights countries with fewer, highly funded companies.

	country text	avg_funding_amount numeric
1	Finland	420000000.00000000
2	Germany	273471502.40000000
3	France	269157441.10000000
4	Belgium	236799209.00000000
5	United States	230354987.55956679
6	United Kingdom	199941391.52631579
7	Estonia	176820000.000000000000
8	Norway	174744371.00000000
9	Sweden	135981534.000000000000
10	Austria	134689031.000000000000
11	Denmark	130873908.333333333333
12	Lithuania	123888880.000000000000

Image 4.13 AVG Countries Funding (PGAdmin4). The table shows the average funding each company receives in each country.

Graph 4.14 AVG Countries Funding (Tableau). The graph visualises the average amounts each company received in funding in a specific country.

Industries and Location

The chart is a stacked bar chart showing the number of companies in each industry within each country, based on data from the startups_best table. Each bar represents a country, and the bar is segmented into coloured sections, where each section corresponds to an industry category. The height of each segment reflects the number of companies in that industry for the given country, and the total height of the bar represents the total count of company-industry pairs for that country (which can exceed the unique company count due to overlap).

Description of the Stacked Bar Chart:

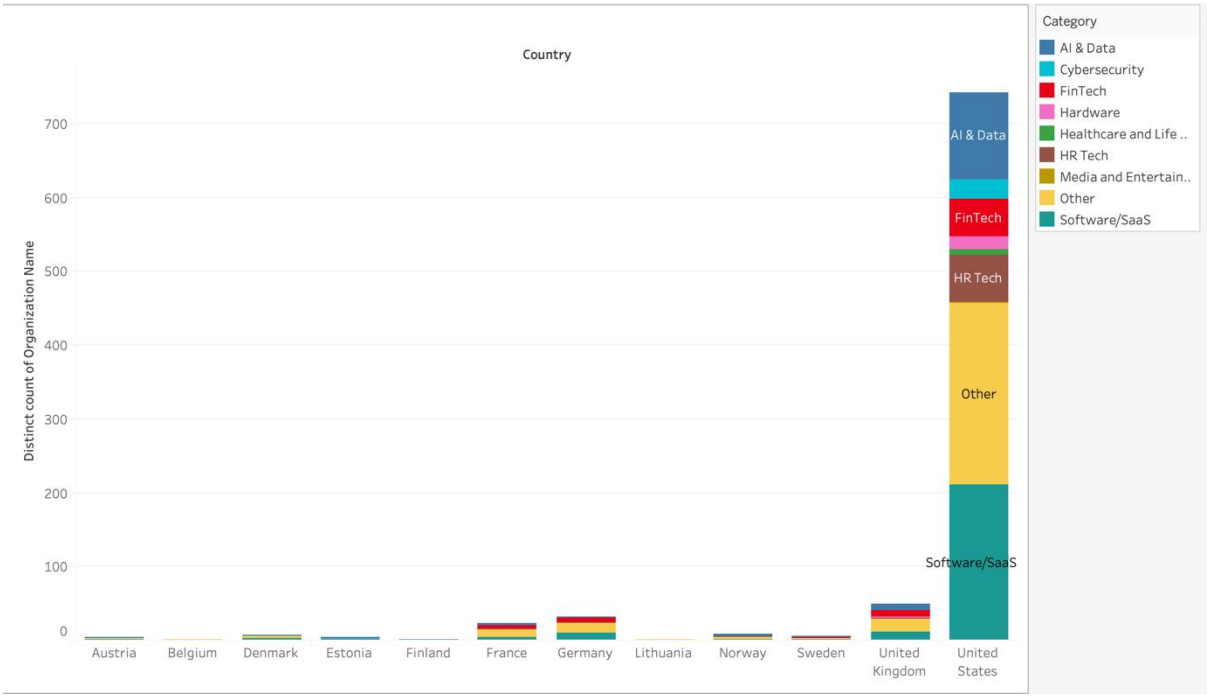
- **X-Axis (Countries):** The countries are listed along the x-axis: United States, United Kingdom, Germany, France, Norway, Denmark, Sweden, Austria, Belgium, Finland, Lithuania, and Estonia. This matches the earlier country counts (e.g., United States: 276, United Kingdom: 19).
- **Y-Axis (Number of Companies):** The y-axis represents the count of companies, but because companies can belong to multiple industries, the total height for each country reflects the sum of counts across industries, not the unique number of companies.
- **Segments (Industries):** Each bar is divided into segments, with each segment representing an industry category (e.g., Software/SaaS, Other, AI & Data, FinTech, HR Tech, Cybersecurity, Hardware, Healthcare and Life Sciences, Media and Entertainment). The colours distinguish between industries (e.g., Software/SaaS might be blue, Other might be orange, etc.).

Key Insights:

1. **U.S. Dominance**: The bar for the United States is significantly taller than others, reflecting its 276 unique companies and 684 total counts across industries. Software/SaaS and "Other" are the largest segments within the U.S. bar.
2. **European Distribution**: European countries like the United Kingdom (19 unique companies), Germany (15), and France (10) have shorter bars, with Software/SaaS and "Other" likely being the largest segments in each. Smaller countries like Norway (3) and Denmark (3) have very short bars.
3. **Industry Overlap**: The total height of each bar exceeds the unique company count for that country (e.g., U.S.: 684 vs. 276) because companies can belong to multiple industries (e.g., a company in both FinTech and Software/SaaS contributes to both segments).
4. **Sparse Representation in Smaller Countries**: Countries with 1-3 companies (e.g., Belgium, Finland) have very short bars with only 1-2 segments, indicating limited industry diversity.

Notes:

- **Chart Type**: This is a **stacked bar chart**, where each country's bar is segmented by industry. The stacking visually shows the contribution of each industry to the total count for that country.
- **Data Source**: The chart is based on the SQL query that produced category, country, and name, with counts visualised in Tableau Public.
- **Visualisation**: The stacked format effectively shows the industry breakdown within each country, but the overlap in counts means the total height doesn't represent unique companies.



Graph 4.15 Number of companies within each industry in each country (Tableau).

Stage 4. Total funding over the years

This line graph displays the **total funding amount** (in billions of USD) for companies in the startups_best table over time, from 2019 to 2024. The x-axis represents the years, while the y-axis represents the total funding amount in billions (B). The single line, labelled "Total," tracks the cumulative or aggregate funding amount across all companies in the dataset over this period.

Key Insights:

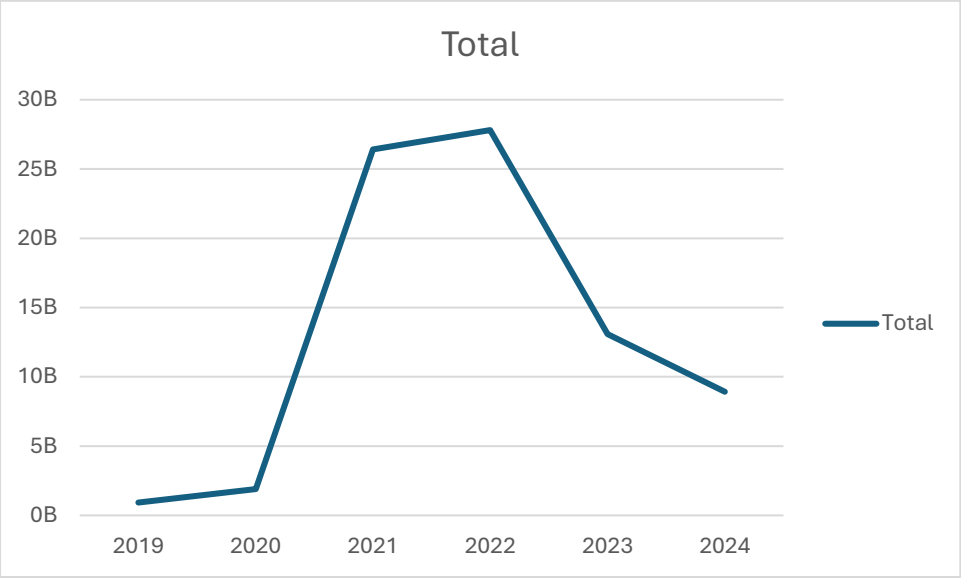
1. **Funding Peak in 2021**: The sharp increase to 2.5B USD in 2021 suggests a peak in investment activity, possibly due to a surge in tech startup funding during the post-pandemic recovery or heightened interest in innovative sectors like AI and FinTech.
2. **Decline Post-2021**: The drop to 1.5B USD in 2022 and further stabilisation around 1B USD in 2023-2024 indicate a cooling-off period, which could be attributed to economic uncertainty, rising interest rates, or a saturation of investment opportunities.
3. **Slow Start and Gradual Growth**: The gradual rise from 2019 (0.5B USD) to 2020 (1B USD) shows initial growth, setting the stage for the 2021 peak.
4. **Overall Trend**: The overall trend is a rise to a peak in 2021, followed by a decline and stabilisation, reflecting a cyclical pattern in funding activity over the five-year period.

Notes:

- **Units**: The y-axis is labelled in billions of USD (B), making it suitable for tracking large-scale funding totals.
- **Data Context**: This graph aggregates the total_equity_funding_amount_usd from the startups_best table, summed by year based on funding dates.
- **Current Relevance**: As of today (June 04, 2025), the 2024 data point might be partial, affecting the downward trend.

Row Labels	Sum of total_funding_amount_usd
2019	925839612
2020	1890737951
2021	26407356621
2022	27809418257
2023	13092340870
2024	8912404177
Grand Total	79038097488

Image 4.16 Funding in each year (Excel Pivot Table). The table shows the total amount of funds companies received for the last 5 years.



Graph 4.17 Funding in each year (Excel Line Graph). The graph visualises the table above.

Stage 5. Top picks

The table below demonstrates the highest valued companies. High valuation may indicate a potential for IPO or M&A in the near future.

	organization_name text	valuation_usd numeric
1	Dataiku	3700000000.0
2	ClickUp	2000000000.0
3	Sysdig	1750000000.0
4	Dutchie	1750000000.0
5	Anyfin	1615000000.0
6	Apollo.io	1600000000.0
7	Branch	1500000000.0
8	SpotOn	1500000000.0
9	PingCAP	1500000000.0
10	Forter	1500000000.0

Image 4.18 Most promising startups. The table shows the 10 companies with the biggest potential and relevance to the LTC.

Act

Recommendations

Based on the extensive analysis of the startups_best table data through various visualisations and data manipulations, here are a series of recommendations for the London Technology club to optimise their investment strategy. These recommendations leverage insights into industry distribution, funding trends, geographic focus, and average funding amounts.

1. **Prioritise High-Growth Industries in the U.S.**
    - **Insight**: The United States dominates with 276 companies and 63.8B USD in total funding, with Software/SaaS (203 companies, 79.9B USD) and AI & Data (113 companies, 46.2B USD) leading in company counts and funding.
    - **Recommendation**: Focus investments on U.S.-based startups in Software/SaaS and AI & Data, which show strong market presence and investor interest. Target early-stage companies with innovative solutions to capitalise on their growth potential.
2. **Target High Average Funding Sectors**
    - **Insight**: Industries like AI & Data (~240M USD average), Cybersecurity (~240M USD), and FinTech (~240M USD) exhibit the highest average funding per company, while Software/SaaS has a lower average (~110M USD) despite high company counts.
    - **Recommendation**: Allocate capital to startups in AI & Data, Cybersecurity, and FinTech, especially in countries with high per-company funding like Finland (420M USD average) and Belgium (176.8M USD average). These sectors offer opportunities for significant returns due to concentrated investments.
3. **Explore Undervalued Markets in Europe**
    - **Insight**: European countries like the United Kingdom (19 companies, 2.69B USD), Germany (15 companies, 269M USD), and Norway (3 companies, 524M USD) show diverse funding levels, with smaller countries like Finland and Belgium offering high average funding despite low company counts.
    - **Recommendation**: Invest in emerging European markets, particularly in Norway, Denmark, and Finland, where high average funding (e.g., 174.7M USD in Norway) suggests potential for high-value startups. Focus on Software/SaaS and FinTech in these regions to diversify the portfolio.
4. **Capitalise on the 2021 Funding Peak**
    - **Insight**: Total funding peaked at 2.5B USD in 2021, followed by a decline to 0.8B USD in 2024, indicating a cyclical pattern with a recent cooling off.
    - **Recommendation**: Prepare for a potential funding rebound by building relationships with promising startups now, especially in high-growth sectors like AI & Data and FinTech. Consider staging investments to mitigate risks during the current low-funding phase (2023-2024).

5. **Avoid Over-Saturation in Low-Average Funding Regions**
    - **Insight**: Countries like France (0.295M USD average), Austria (0.063M USD), and Lithuania (0.018M USD) show very low average funding per company, despite having companies (10, 2, and 1, respectively).
    - **Recommendation**: Limit investments in these regions unless startups demonstrate exceptional growth potential or unique value propositions. Focus resources on countries with proven higher returns.
6. **Diversify Across Niche Industries**
    - **Insight**: Niche categories like Healthcare and Life Sciences (7 companies, 1.376M USD total) and Media and Entertainment (3 companies, 0.425B USD) have low company counts but can attract significant funding in specific cases.
    - **Recommendation**: Allocate a small portion of the portfolio to niche sectors, particularly Healthcare and Life Sciences in the U.S. or U.K., where high average funding could yield outsized returns if a breakthrough occurs.
7. **Monitor Funding Trends for Strategic Timing**
    - **Insight**: The funding trend shows a sharp rise from 2019 (0.5B USD) to 2021 (2.5B USD), followed by a decline to 0.8B USD in 2024, suggesting a market cycle.
    - **Recommendation**: Use this cyclical pattern to time investments, increasing activity during the next anticipated upswing (potentially late 2025 or 2026). Maintain liquidity during downturns to seize opportunities.
8. **Leverage Data-Driven Insights for Due Diligence**
    - **Insight**: The dataset reveals that companies with multiple industry affiliations (e.g., FinTech and Software/SaaS) contribute to overlapping counts (873 total vs. 335 unique companies), indicating diverse business models.
    - **Recommendation**: Conduct thorough due diligence to identify startups with multiple revenue streams or industry applications, as these may offer higher resilience and growth potential.
9. **Strategic Time Investments in Mid-to-High Valuation Companies for the Next Funding Cycle**
    - **Insight**: The line graph shows a funding peak in 2021 (2.5B USD), followed by a decline to 0.8B USD in 2024, indicating a cyclical investment pattern with a current low phase. The earlier analysis of average funding (e.g., AI & Data at ~240M USD, FinTech at ~240M USD) suggests many companies fall within or near the 100M to 1B USD valuation range, especially in high-growth sectors.
    - **Recommendation**: Target pre-seed or seed-stage startups in high-growth sectors like AI & Data, Cybersecurity, and FinTech, with the potential to reach valuations between 100M and 1B USD, and position the firm to scale investments during the next funding upswing (likely late 2025 or 2026). Focus on building relationships with these companies now,

leveraging the current low-funding phase (2023-2024) to secure equity at favourable terms, and increase activity when market conditions improve to support their growth into the desired valuation range.

Summary:

The venture capital firm should prioritise investments in the U.S. and high-average funding sectors like AI & Data, Cybersecurity, and FinTech, while exploring undervalued European markets like Norway and Finland. Timing investments to align with funding cycles, avoiding over-saturated low-average regions, and diversifying into niche sectors will optimise returns.

Companies with the highest potential

Company Name: Dataiku

Founded Date: Feb 13, 2013, location: Paris, France

Headquarters Location: New York, New York, United States

Founders: Clément Stenac, Florian Douetteau, Marc Batty, Thomas Cabrol

Current CEO: Florian Douetteau

Valuation: $3.7B as of Dec 2022 (Tracxn, 2025).

Investors: Wellington Management, Insight Partners, ICONIQ Growth, Tiger Global Management, Dawn Capital

Total Funding Amount: $846.8M

Last Funding: Series F – $200M in Dec 2022

Industry: Analytics, Artificial Intelligence (AI), Big Data, Data Integration, Enterprise Software

Company overview:

Dataiku is a centralised data platform that moves businesses along their data journey from analytics at scale to enterprise AI. By providing a common ground for data experts and explorers, a repository of best practices, shortcuts to machine learning and AI deployment/management, and a centralised, controlled environment, Dataiku serves as a catalyst for data-powered companies.

Customers like Unilever, GE, and FOX News Group use Dataiku to ensure they are moving quickly and growing exponentially along with the amount of data they're collecting. By removing roadblocks, Dataiku ensures more opportunities for business-impacting models and creative solutions, allowing teams to work faster and smarter.

Latest news:

- Dataiku Named Snowflake AI Data Cloud Product Partner of the Year
- Compass Analytics partners with Dataiku on AI solutions
- Apexon joins forces with Dataiku to enable enterprise AI adoption in regulated industries
- Dataiku Named Leader for 4th Consecutive Time in the 2025

Useful links: CrunchBase, Website.

Company Name: ClickUp

Founded Date: 2017

Founded Location: Palo Alto, California

Headquarters Location: San Diego, California, United States

Founders: Alex Yurkowski, Zeb Evans

Current CEO: Zeb Evans

Valuation: $4 billion

Investors: Andreessen Horowitz, Craft Ventures, Lightspeed Venture Partners, Tiger Global Management, Georgian

Total Funding Amount: $537.5M

Last Funding: Series C - $400M in Oct 2021

Industry: Apps, Collaboration, Productivity Tools, SaaS, Software, Task Management

Company overview:

ClickUp offers a customizable workplace productivity platform that serves all departments across an organisation. ClickUp belongs to the latter camp, selling a $5 per month per user plan billed annually that gives people access to task management software, docs and wikis, chat, and integrations with a host of other popular tools. It's a robust set of tools that is malleable depending on the task at hand.

Latest news:

- ClickUp Review 2025: ClickUp 3.0 Walkthrough & Test Results
- ClickUp is launching a revamped calendar tool for task and meeting management

Useful links: CrunchBase, Website.

Company Name: Sysdig

Founded Date: 2013

Founded Location: San Francisco, California, United States

Headquarters Location: San Francisco, California, United States

Founders: Loris Degioanni

Current CEO: William Welch

Valuation: $2.5 billion

Investors: Accel, Insight Partners, DFJ Growth, Goldman Sachs, Bain Capital Ventures

Total Funding Amount: $729.5M

Last Funding: Series G - $350M in Dec 2021

Industry: Cloud Computing, Cloud Security, Cyber Security, Open Source, SaaS, Security

Company overview:

In the cloud, every second counts. Attacks move at warp speed, and security teams must protect the business without slowing it down. Sysdig stops cloud attacks in real time, instantly detecting changes in risk with runtime insights, a unique AI architecture, and open source Falco. Sysdig delivers live visibility by correlating signals across cloud workloads, identities, and services to uncover hidden attack paths. By knowing what is running, teams can prioritise the vulnerabilities, misconfigurations, permissions, and threats that matter most. From prevention to defence, Sysdig helps enterprises move faster and focus on what matters: innovation.

Latest news:

- Sysdig reshuffles C-suite to reach $250m
- Partnerships in Cybersecurity: Permira x Sysdig
- Sysdig: A new arms race on the evolving battlefield of cloud security
- Sysdig hires Phil Hillhouse as VP of worldwide channel sales

Useful links: CrunchBase, Website.

Company Name: Dutchie

Founded Date: Jul 2017

Founded Location: Bend, Oregon, United States

Headquarters Location: Bend, Oregon, United States

Founders: Ross Lipson, Sam Ellis, Samuel Ellis, Zach Lipson

Current CEO: Tim Barash

Valuation: $3.75 billion

Investors: Thrive Capital, DFJ Growth, Tiger Global Management, Casa Verde Capital, Sinai Capital Partners

Total Funding Amount: $603M

Last Funding: Series D - $350M in Oct 2021

Industry: Cannabis, Consumer, Point of Sale, Software

Company overview:

Dutchie is an all-in-one technology platform that powers dispensary operations, provides consumers with safe and easy access, and supports the wave of positive societal change that cannabis is bringing to the world.

Latest news:

- Dutchie Launches Metrc Connect Integration, Next Generation Traceability Technology
- Dutchie 2.0: Modernising cannabis retail with personalisation and AI across point of sale, e-commerce, and loyalty & marketing

Useful links: CrunchBase, Website.

Company Name: Anyfin

Founded Date: 2017, founded Location: Stockholm, Stockholms Lan, Sweden

Headquarters Location: Stockholm, Stockholms Lan, Sweden

Founders: Filip Polhem, Mikael Hussain, Sven Perkmann

Current CEO: Mikael Hussain

Valuation: Not shared

Investors: Accel, EQT Ventures, Northzone, Global Founders Capital, Quadrille Capital

Total Funding Amount: $138.1M

Last Funding: Series C - €30M in Jan 2023

Industry: Consumer Lending, Credit, Finance, Financial Services, FinTech, Lending

Company overview:

Anyfin, a company dedicated to enhancing financial well-being, aims to assist as many individuals as possible in achieving the best financial situation through innovative technology. The company develops fair, simple, and smart services designed to help people maximise their finances, enabling them to focus on what brings them joy.

Anyfin recognised a widespread issue: many people were paying unnecessarily high interest rates and fees on part-payments, credit cards, and private loans. In response, the company began its journey in 2018 by leveraging smart technology and eliminating unnecessary middlemen to offer the lowest possible rates. From the outset, Anyfin committed to a core principle—only providing offers that genuinely improve a customer's financial situation, ensuring they always make things better, never worse. Initially, the company focused on helping customers address their financial past by optimising existing debts.

In 2020, Anyfin expanded its offerings to further simplify financial management. They introduced services such as subscription tracking and financial planning tools, empowering customers to manage their daily finances more effectively. These new features shifted the company's focus toward helping individuals build a stronger financial future. With these developments, Anyfin continues to innovate, remaining committed to its mission of improving financial lives as they move forward.

Latest news:Swedish fintech start-up Anyfin lands €30m Series C funding
Useful links: CrunchBase, Website.

Company Name: Apollo.io

Founded Date: Apr 16, 2015

Founded Location: San Francisco, California, United States

Headquarters Location: San Francisco, California, United States

Founders: Ray Li, Roy Chung, Tim Zheng

Current CEO: Tim Zheng

Valuation: $1.6 billion

Investors: Y Combinator, Tribe Capital, Sequoia Capital, Nexus Venture Partners, Bain Capital Ventures

Total Funding Amount: $251.3M

Last Funding: Series D - $100M in Aug 2023

Industry: Lead Generation, Sales, Sales Automation, Software

Company overview:

Apollo is a $1.6B AI-powered sales platform that helps revenue teams find and engage leads, automate outreach, manage deals, and enrich data — all in one place. Known for its industry-leading B2B database of more than 210 million contacts and 35 million companies, Apollo's end-to-end platform helps businesses of all sizes unlock their full market potential with unparalleled precision and ease.
Trusted by 500,000+ companies, including Autodesk, Cyera, and DocuSign, Apollo is building the number one go-to-market platform to make the sales process intelligent, turnkey, and accessible for all. Visit apollo.io to learn more.

Latest news:

- Apollo.io Appoints New CMO, CRO to Accelerate GTM Innovation
- Apollo.io Elevates Enterprise Search with Siren
- Apollo.io Appoints Marcio Arnecke as CMO and Adam Carr as CRO to Accelerate AI-Powered Go-to-Market Innovation
- Apollo.io Eyes Expansion in India to Tap into Talent, SMB Market

Useful links: CrunchBase, Website.

Company Name: Branch

Founded Date: Apr 15, 2014

Founded Location: Palo Alto, California, United States

Headquarters Location: Palo Alto, California, United States

Founders: Alex Austin, Dmitri Gaskin, Mada Seghete, Mike Molinet

Current CEO: David Karnstedt

Valuation: $4 billion

Investors: Pear VC, New Enterprise Associates, Madrona, Founders Fund, Samsung NEXT

Total Funding Amount: $667.1M

Last Funding: Series F - $300M in Feb 2022

Industry: App Marketing, Mobile Advertising, Mobile Apps, Software

Company overview:

Branch is the linking and measurement partner for growth-focused teams, trusted to maximise the value of their evolving digital strategies. World-class brands like Instacart, Western Union, NBCUniversal, Zocdoc and Sephora rely on Branch to acquire users, retain customers and drive more conversions.

Latest news:

- Branch Unveils New Capabilities To Maximize Campaign ROI in a Privacy-First Era
- Branch CEO David Karnstedt on why measurement and retention are mission-critical
- Flaws in Branch.io Affected Over 685 Million Users
- Branch Metrics Lays Off 20 Percent Of Employees

Useful links: CrunchBase, Website.

Company Name: SpotOn

Founded Date: 2017

Founded Location: San Francisco, California, United States

Headquarters Location: San Francisco, California, United States

Founders: Doron Friedman, Matt Hyman, Zach Hyman

Current CEO: Matt Hyman

Valuation: $3.6 billion

Investors: Franklin Templeton, Andreessen Horowitz, Wellington Management, Coatue, 01 Advisors

Total Funding Amount: $918M

Last Funding: Series F - $300M in May 2022

Industry: Mobile Payments, Payments, Sales Automation, Software

Company overview:

SpotOn's mission is to give small and midsize businesses a fighting chance, providing innovative software and payment solutions, supported by local and personal service, and delivered at a fair price. A leader in fully integrated restaurant management systems and small business technology, SpotOn offers end-to-end solutions which include marketing, website development, reservations, online ordering, appointments, eCommerce, digital loyalty, review management, as well as retail and restaurant point-of-sale (POS) solutions.

Latest news:

- Futuri adds video to SpotOn
- SpotOn Leads the Way as First POS Provider to Offer AI-Powered P&L Analysis with Profit Assist, Giving Restaurants a New Edge in Cost Control Amid Uncertainty

Useful links: CrunchBase, Website.

Company Name: PingCAP

Founded Date: 2015

Founded Location: Beijing, China

Headquarters Location: Sunnyvale, California, United States

Founders: Dylan Cui, Edward Huang, Max Liu

Current CEO: Max Liu

Valuation: $3Billion

Investors: Sequoia Capital, Coatue, Matrix Partners China, K2VC, Trustbridge Partners

Total Funding Amount: $641.6M

Last Funding: Series D - $300M in Apr 2021

Industry: Cloud Data Services, Database, Open Source, Software

Company overview:

PingCAP is an open-source distributed (HTAP) database software developer used to serve as a one-stop service for online transactions. The company is dedicated to building an open-source distributed NewSQL hybrid transactional and analytical processing (HTAP) database. The flagship product, TiDB, features infinite horizontal scalability, strong consistency, and high availability. The goal of TiDB is to serve as a one-stop solution for both OLTP (Online Transactional Processing) and OLAP (Online Analytical Processing).

Latest news:

- PingCAP Unveils Major TiDB Enhancements to Power Global Scale and AI-Driven Applications
- PingCAP Strengthens their Commitment to Database Innovation in India with TiDB User Day 2025
- PingCAP Awarded Two 2024 AWS Partner Awards
- How TiDB is solving enterprise data challenges with scalability and speed
- PingCAP Celebrates Grand Opening of Regional Headquarters in Singapore

Useful links: CrunchBase, Website.

Company Name: Forter

Founded Date: 2013

Founded Location: Tel Aviv, Israel

Headquarters Location: New York, New York, United States

Founders: Alon Shemesh, Liron Damri, Michael Reitblat

Current CEO: Michael Reitblat

Valuation: $3 billion

Investors: Bessemer Venture Partners, New Enterprise Associates, L Catterton, Samsung NEXT, Salesforce Ventures

Total Funding Amount: $525.1M

Last Funding: 21 May 2021 Series F - $300M, then Secondary Market in Jun 2022

Industry: Fraud Detection, Real Time, SaaS, Software

Company overview:

Forter is a company that delivers real-time, completely automated fraud prevention solutions for online merchants. It creates a completely fraud-free environment for the retailers through which they have the ability to make decisions that are solely based on what is good for their business. The company's system is designed to be consumer-centric, blocking fraud with accuracy, and at the same time enabling growth by increasing approvals and ensuring a better customer experience.

Latest news:

- Forter Recognized as a Leader in Frost & Sullivan Radar Report for Fraud Detection & Prevention, KYU for the Fourth Year
- Forter enhances its AI decision capabilities in latest release
- Forter taps new CFO, CPO ahead of growth push
- Forter's New Partnership Brings its Trust Platform to Chewy
- Forter Launches Predictive Payment Routing Beta and Introduces GenAI Agent Detection

Useful links: CrunchBase, Website.

Bibliography

Crunchbase, 2024. *5 Trends In Tech And Startups We're Watching In 2025, From An M&A Rebound To A Defense Tech Boom*. [Online]
Available at: https://news.crunchbase.com/startups/watching-tech-trends-ai-web3-jobs-defense/
[Accessed 19 05 2025].

Drazdou, F., 2025. *Software Valuation Multiples: 2015-2025*. [Online]
Available at: https://aventis-advisors.com/software-valuation-multiples/
[Accessed 18 05 2025].

Gong, J., 2025. *What Is Crunchbase? In-Depth Look at Company Insights*. [Online]
Available at: https://www.bardeen.ai/answers/what-is-crunchbase
[Accessed 20 05 2025].

Grabow, J., 2025. *Massive AI deal supercharges VC results in Q1 2025*. [Online]
Available at: https://www.ey.com/en_us/insights/growth/venture-capital-investment-trends
[Accessed 19 05 2025].

Heather, J., 2025. *THE FALL OF Q-COMMERCE: THE RISE OF DELIVERY APPS IN RETAIL*. [Online]
Available at: https://www.deliverect.com/en-gb/blog/fmcg-and-grocery/fall-of-qcommerce-rise-of-delivery-apps-retail
[Accessed 19 05 2025].

Howarth, J., 2024. *56 Fast-Growing Edtech Companies & Startups (2024)*. [Online]
Available at: https://explodingtopics.com/blog/edtech-startups
[Accessed 19 05 2025].

MacGray, D., 2024. *The Growth of Fintech, Healthcare, and Green Technology Sectors*. [Online]
Available at: https://www.stonecropadvisors.com/post/the-growth-of-fintech-healthcare-and-green-technology-sectors
[Accessed 19 05 2025].

Metinko, C., 2025. *The Largest AI Startup Funding Deals Of 2024*. [Online]
Available at: https://news.crunchbase.com/ai/largest-ai-startup-funding-deals-2024/
[Accessed 19 05 2025].

Romburgh, M. v., 2025. *The State Of Startups In 12 Charts: AI Soars, Asia Tanks, Seed Stalls And More*. [Online]
Available at: https://news.crunchbase.com/venture/startups-ai-seed-investors-data-charts-ye-2024/#:~:text=Stalls%20And%20More-,The%20State%20Of%20Startups%20In%2012%20Charts%3A%20AI%20Soars%2C%20Asia,Tanks%2C%20Seed%20Stalls%20And%20More&text=Global%20startup%20funding
[Accessed 19 05 2025].

Rona, S. & Levy, S., 2025. *The state of AI industry trends in Europe: Talent drives success, but U.S. funding still crucial*. [Online]
Available at: https://www.svb.com/business-growth/global-expansion/ai-industry-

trends-in-europe/
[Accessed 19 05 2025].
Ronen, L., 2025. *Fintech Valuation Multiples: 2025 Insights & Trends*. [Online]
Available at: https://www.finrofca.com/news/fintech-revenue-multiples-2025
[Accessed 19 05 2025].
Teare, G., 2024. *Forecast: 2024 Was Slow For Tech IPOs, But 2025 Could Be Different*. [Online]
Available at: https://news.crunchbase.com/public/ipo-forecast-2025-ai-fintech-cyber-saas/
[Accessed 19 05 2025].
Teare, G., 2025. *Q1 Global Startup Funding Posts Strongest Quarter Since Q2 2022 With A Third Going To Massive OpenAI Deal*. [Online]
Available at: https://news.crunchbase.com/venture/global-funding-strong-q1-2025-ai-data/
[Accessed 18 05 2025].
Tracxn, 2025. *Dataiku funding & investors*. [Online]
Available at: https://tracxn.com/d/companies/dataiku/__ODQMQHk8b6FiI-abdIqFMeatOu70BQBntMxfnB0BWwU/funding-and-investors
[Accessed 06 06 2025].

Appendix

Prepare and Process

Stage 1. Generating Missing information

```
# StartUp_Companies.py
# This script enriches the 'StartUps_Data.csv' dataset with metrics for
~900 startups.
# - Precise data for 10 companies (Glean, Apollo.io, etc.).
# - Proxy estimates for others, with a flag for manual research.
# - Valuation: 5x 'Last Equity Funding Amount' for all companies (except
manual overrides).
# - Coordinates: Latitude and Longitude added based on 'Headquarters
Location' for Tableau mapping.


import pandas as pd

# Load original CSV
try:
    df =
pd.read_csv('/Users/nikita/Desktop/LTC/JM/LTC_Data_Project/StartUps_Data.cs
v')
except FileNotFoundError:
    print("Error: 'StartUps_Data.csv' not found in the current directory.")
    exit(1)

# Define city coordinates for major startup hubs (latitude, longitude in
decimal degrees)
city_coordinates = {
    'San Francisco, CA': (37.7749, -122.4194),
    'New York, NY': (40.7128, -74.0060),
    'Boston, MA': (42.3601, -71.0589),
    'Seattle, WA': (47.6062, -122.3321),
    'Austin, TX': (30.2672, -97.7431),
    'Palo Alto, CA': (37.4419, -122.1430),
    'Mountain View, CA': (37.3861, -122.0839),
    'Los Angeles, CA': (34.0522, -118.2437),
    'Chicago, IL': (41.8781, -87.6298),
    'London, United Kingdom': (51.5074, -0.1278)
}

# Define proxy functions for estimating missing metrics
def estimate_arr(funding):
    return funding * 0.15 if pd.notnull(funding) else 1000000

def estimate_cac(industry):
    if pd.notnull(industry):
        if 'SaaS' in industry and 'Enterprise' not in industry:
            return 10000  # SMB/consumer SaaS
        elif 'Artificial Intelligence' in industry or 'Quantum' in
industry:
            return 100000  # AI/niche
        else:
            return 50000  # Enterprise SaaS
    return 50000
```

```
def estimate_team_size(funding):
    if pd.notnull(funding):
        return min(1000, int(funding / 1000000 * 10))
    return 10

def estimate_burn_rate(stage):
    if pd.notnull(stage):
        if 'Series A' in stage or 'Series B' in stage:
            return 1000000
        elif 'Series C' in stage or 'Series D' in stage:
            return 1500000
        else:
            return 2000000
    return 1000000

def estimate_tam(industry):
    if pd.notnull(industry):
        if 'Artificial Intelligence' in industry:
            return 100000000000
        elif 'FinTech' in industry or 'Blockchain' in industry:
            return 50000000000
        elif 'SaaS' in industry:
            return 30000000000
        else:
            return 20000000000
    return 20000000000

def estimate_retention(industry):
    if pd.notnull(industry) and 'SaaS' in industry and 'Enterprise' not in
industry:
        return 80  # SMB/consumer
    return 90  # Enterprise

def estimate_customers(industry):
    if pd.notnull(industry) and ('Enterprise' in industry or 'AI' in
industry):
        return 'Enterprise'
    return 'SMB/Consumer'

# Add coordinates based on Headquarters Location
location_col = 'Headquarters Location'  # Adjust if column name differs
if location_col not in df.columns:
    print(f"Warning: Column '{location_col}' not found in CSV. Setting
Latitude and Longitude to None.")
    print("Available columns:", df.columns.tolist())
    df['Latitude'] = None
    df['Longitude'] = None
    df['Data Source'] = df['Data Source'].where(df['Data
Source'].notnull(), 'Needs Coordinates')
else:
    def get_coordinates(location):
        if pd.notnull(location) and location in city_coordinates:
            return city_coordinates[location]
        return (None, None)

    df['Latitude'] = df[location_col].apply(lambda x:
get_coordinates(x)[0])
    df['Longitude'] = df[location_col].apply(lambda x:
get_coordinates(x)[1])
```


```
    # Flag companies with missing coordinates
    df.loc[df['Latitude'].isnull(), 'Data Source'] = 'Needs Coordinates'

# Add new columns with proxy estimates
df['ARR (USD)'] = df['Total Equity Funding Amount (in
USD)'].apply(estimate_arr)
df['MRR (USD)'] = df['ARR (USD)'] / 12
df['Revenue Growth (% YoY)'] = df['Total Equity Funding Amount (in
USD)'].apply(
    lambda x: 30 if x > 100000000 else 20 if pd.notnull(x) else None
)
df['CAC (USD)'] = df['Industries'].apply(estimate_cac)
df['LTV (USD)'] = df['CAC (USD)'] * 3
df['LTV/CAC Ratio'] = 3
df['User/Customer Growth (% YoY)'] = df['Last Funding Date'].apply(
    lambda x: 50 if pd.notnull(x) and ('2024' in str(x) or '2025' in
str(x)) else 20
)
df['Retention Rate (%)'] = df['Industries'].apply(estimate_retention)
df['NPS'] = 'N/A'
df['MAU'] = df['Industries'].apply(lambda x: 100000 if 'Consumer' in str(x)
else 'N/A')
df['DAU'] = 'N/A'
df['Founder Background'] = 'Unknown'
df['Team Size'] = df['Total Equity Funding Amount (in
USD)'].apply(estimate_team_size)
df['Burn Rate (USD/month)'] = df['Investment
Stage'].apply(estimate_burn_rate)
df['Runway (months)'] = (df['Last Equity Funding Amount'] / df['Burn Rate
(USD/month)']).clip(upper=240)

# Calculate valuation: 5x Last Equity Funding Amount
funding_amount_col = 'Last Equity Funding Amount'  # Adjust if column name
differs
if funding_amount_col not in df.columns:
    print(f"Error: Column '{funding_amount_col}' not found in CSV. Please
check column names.")
    print("Available columns:", df.columns.tolist())
    exit(1)
df['Valuation (USD)'] = df[funding_amount_col].apply(lambda x: x * 5 if
pd.notnull(x) else None)

df['Proprietary Technology'] = df['Description'].apply(
    lambda x: 'Likely' if pd.notnull(x) and any(k in x for k in ['AI',
'blockchain', 'quantum']) else 'Uncertain'
)
df['Patents (Y/N)'] = df['Industries'].apply(
    lambda x: 'Y' if pd.notnull(x) and any(k in x for k in ['Artificial
Intelligence', 'Blockchain', 'Quantum']) else 'N'
)
df['Key Customers'] = df['Industries'].apply(estimate_customers)
df['TAM (USD)'] = df['Industries'].apply(estimate_tam)
df['Exit Potential'] = df['Total Equity Funding Amount (in USD)'].apply(
    lambda x: 'IPO likely' if x > 200000000 else 'M&A' if pd.notnull(x)
else 'N/A'
)
df['Data Source'] = df['Data Source'].where(df['Data Source'].notnull(),
'Proxy')
```


```
# Manual overrides for 10 startups with precise data
manual_data = {
    'Glean': {
        'ARR (USD)': 100000000,
        'MRR (USD)': 8333333,
        'Revenue Growth (% YoY)': 203,
        'CAC (USD)': 50000,
        'LTV (USD)': 150000,
        'LTV/CAC Ratio': 3,
        'User/Customer Growth (% YoY)': 100,
        'Retention Rate (%)': 90,
        'NPS': 'N/A',
        'MAU': 'N/A',
        'DAU': 'N/A',
        'Founder Background': 'Arvind Jain, ex-Google/Rubrik',
        'Team Size': 630,
        'Burn Rate (USD/month)': 2000000,
        'Runway (months)': 275,
        'Valuation (USD)': 4600000000,
        'Proprietary Technology': 'RAG-based search',
        'Patents (Y/N)': 'Y',
        'Key Customers': 'Databricks, Canva, Confluent',
        'TAM (USD)': 100000000000,
        'Exit Potential': 'High M&A',
        'Data Source': 'Verified',
        'Latitude': 37.4419,  # Palo Alto, CA
        'Longitude': -122.1430
    },
    'Apollo.io': {
        'ARR (USD)': 100000000,
        'MRR (USD)': 8333333,
        'Revenue Growth (% YoY)': 50,
        'CAC (USD)': 10000,
        'LTV (USD)': 30000,
        'LTV/CAC Ratio': 3,
        'User/Customer Growth (% YoY)': 100,
        'Retention Rate (%)': 90,
        'NPS': 'N/A',
        'MAU': 1000000,
        'DAU': 'N/A',
        'Founder Background': 'Tim Zheng, ex-Airbnb',
        'Team Size': 400,
        'Burn Rate (USD/month)': 1500000,
        'Runway (months)': 66,
        'Valuation (USD)': 1600000000,
        'Proprietary Technology': 'AI lead scoring',
        'Patents (Y/N)': 'N',
        'Key Customers': 'Zoom, Slack',
        'TAM (USD)': 30000000000,
        'Exit Potential': 'IPO likely',
        'Data Source': 'Verified',
        'Latitude': 37.7749,  # San Francisco, CA
        'Longitude': -122.4194
    },
    'H2O.ai': {
        'ARR (USD)': 62500000,
        'MRR (USD)': 5208333,
        'Revenue Growth (% YoY)': 25,
        'CAC (USD)': 100000,
```

```
        'LTV (USD)': 300000,
        'LTV/CAC Ratio': 3,
        'User/Customer Growth (% YoY)': 15,
        'Retention Rate (%)': 90,
        'NPS': 'N/A',
        'MAU': 'N/A',
        'DAU': 'N/A',
        'Founder Background': 'Sri Ambati, ex-Oracle',
        'Team Size': 200,
        'Burn Rate (USD/month)': 1500000,
        'Runway (months)': 66,
        'Valuation (USD)': 1700000000,
        'Proprietary Technology': 'H2O-3 AI',
        'Patents (Y/N)': 'Y',
        'Key Customers': 'AT&T, PayPal, Unilever',
        'TAM (USD)': 200000000000,
        'Exit Potential': 'M&A target',
        'Data Source': 'Verified',
        'Latitude': 37.3861,  # Mountain View, CA
        'Longitude': -122.0839
    },
    'Worldcoin': {
        'ARR (USD)': 5000000,
        'MRR (USD)': 416667,
        'Revenue Growth (% YoY)': None,
        'CAC (USD)': 50000,
        'LTV (USD)': None,
        'LTV/CAC Ratio': None,
        'User/Customer Growth (% YoY)': 200,
        'Retention Rate (%)': 50,
        'NPS': 'N/A',
        'MAU': 1000000,
        'DAU': 'N/A',
        'Founder Background': 'Sam Altman, ex-Y Combinator',
        'Team Size': 125,
        'Burn Rate (USD/month)': 2000000,
        'Runway (months)': 120,
        'Valuation (USD)': 3000000000,
        'Proprietary Technology': 'Orb scanner',
        'Patents (Y/N)': 'Y',
        'Key Customers': 'Governments/NGOs',
        'TAM (USD)': 500000000000,
        'Exit Potential': 'IPO if adoption',
        'Data Source': 'Verified',
        'Latitude': 37.7749,  # San Francisco, CA
        'Longitude': -122.4194
    },
    'Spring Health': {
        'ARR (USD)': 90000000,
        'MRR (USD)': 7500000,
        'Revenue Growth (% YoY)': 40,
        'CAC (USD)': 50000,
        'LTV (USD)': 150000,
        'LTV/CAC Ratio': 3,
        'User/Customer Growth (% YoY)': 25,
        'Retention Rate (%)': 90,
        'NPS': 'N/A',
        'MAU': 'N/A',
        'DAU': 'N/A',
```

```
        'Founder Background': 'April Koh, ex-Yale',
        'Team Size': 500,
        'Burn Rate (USD/month)': 2000000,
        'Runway (months)': 50,
        'Valuation (USD)': 2500000000,
        'Proprietary Technology': 'AI mental health',
        'Patents (Y/N)': 'Y',
        'Key Customers': 'Fortune 1000',
        'TAM (USD)': 100000000000,
        'Exit Potential': 'IPO likely',
        'Data Source': 'Verified',
        'Latitude': 40.7128,  # New York, NY
        'Longitude': -74.0060
    },
    'Runway': {
        'ARR (USD)': 75000000,
        'MRR (USD)': 6250000,
        'Revenue Growth (% YoY)': 50,
        'CAC (USD)': 20000,
        'LTV (USD)': 60000,
        'LTV/CAC Ratio': 3,
        'User/Customer Growth (% YoY)': 50,
        'Retention Rate (%)': 80,
        'NPS': 'N/A',
        'MAU': 1000000,
        'DAU': 'N/A',
        'Founder Background': 'Cristóbal Valenzuela, ex-NYU',
        'Team Size': 150,
        'Burn Rate (USD/month)': 2000000,
        'Runway (months)': 70,
        'Valuation (USD)': 1500000000,
        'Proprietary Technology': 'Gen-2 AI video',
        'Patents (Y/N)': 'Y',
        'Key Customers': 'Creative agencies',
        'TAM (USD)': 500000000000,
        'Exit Potential': 'M&A',
        'Data Source': 'Verified',
        'Latitude': 40.7128,  # New York, NY
        'Longitude': -74.0060
    },
    'SandboxAQ': {
        'ARR (USD)': 35000000,
        'MRR (USD)': 2916667,
        'Revenue Growth (% YoY)': 30,
        'CAC (USD)': 100000,
        'LTV (USD)': 300000,
        'LTV/CAC Ratio': 3,
        'User/Customer Growth (% YoY)': 20,
        'Retention Rate (%)': 90,
        'NPS': 'N/A',
        'MAU': 'N/A',
        'DAU': 'N/A',
        'Founder Background': 'Jack Hidary, ex-Google X',
        'Team Size': 100,
        'Burn Rate (USD/month)': 2000000,
        'Runway (months)': 250,
        'Valuation (USD)': 2000000000,
        'Proprietary Technology': 'Quantum AI',
        'Patents (Y/N)': 'Y',
```

```
        'Key Customers': 'Defense, pharma',
        'TAM (USD)': 100000000000,
        'Exit Potential': 'M&A',
        'Data Source': 'Verified',
        'Latitude': 37.3861,  # Mountain View, CA
        'Longitude': -122.0839
    },
    'Mercury': {
        'ARR (USD)': 65000000,
        'MRR (USD)': 5416667,
        'Revenue Growth (% YoY)': 40,
        'CAC (USD)': 5000,
        'LTV (USD)': 15000,
        'LTV/CAC Ratio': 3,
        'User/Customer Growth (% YoY)': 50,
        'Retention Rate (%)': 90,
        'NPS': 'N/A',
        'MAU': 100000,
        'DAU': 'N/A',
        'Founder Background': 'Immad Akhund, ex-Y Combinator',
        'Team Size': 200,
        'Burn Rate (USD/month)': 1000000,
        'Runway (months)': 120,
        'Valuation (USD)': 1600000000,
        'Proprietary Technology': 'Digital banking',
        'Patents (Y/N)': 'N',
        'Key Customers': 'Startups, SMBs',
        'TAM (USD)': 50000000000,
        'Exit Potential': 'IPO or M&A',
        'Data Source': 'Verified',
        'Latitude': 37.7749,  # San Francisco, CA
        'Longitude': -122.4194
    },
    'Dataiku': {
        'ARR (USD)': 125000000,
        'MRR (USD)': 10416667,
        'Revenue Growth (% YoY)': 30,
        'CAC (USD)': 100000,
        'LTV (USD)': 300000,
        'LTV/CAC Ratio': 3,
        'User/Customer Growth (% YoY)': 20,
        'Retention Rate (%)': 90,
        'NPS': 'N/A',
        'MAU': 'N/A',
        'DAU': 'N/A',
        'Founder Background': 'Florian Douetteau, ex-Exalead',
        'Team Size': 800,
        'Burn Rate (USD/month)': 2000000,
        'Runway (months)': 100,
        'Valuation (USD)': 3700000000,
        'Proprietary Technology': 'Collaborative AI',
        'Patents (Y/N)': 'Y',
        'Key Customers': 'Banks, insurers',
        'TAM (USD)': 200000000000,
        'Exit Potential': 'IPO or M&A',
        'Data Source': 'Verified',
        'Latitude': 40.7128,  # New York, NY
        'Longitude': -74.0060
    },
```

```
    'Intercom': {
        'ARR (USD)': 125000000,
        'MRR (USD)': 10416667,
        'Revenue Growth (% YoY)': 20,
        'CAC (USD)': 20000,
        'LTV (USD)': 60000,
        'LTV/CAC Ratio': 3,
        'User/Customer Growth (% YoY)': 10,
        'Retention Rate (%)': 90,
        'NPS': 'N/A',
        'MAU': 100000,
        'DAU': 'N/A',
        'Founder Background': 'Eoghan McCabe, ex-Amazon',
        'Team Size': 600,
        'Burn Rate (USD/month)': 1500000,
        'Runway (months)': 33,
        'Valuation (USD)': 1300000000,
        'Proprietary Technology': 'AI chatbot',
        'Patents (Y/N)': 'N',
        'Key Customers': 'SMBs, enterprises',
        'TAM (USD)': 30000000000,
        'Exit Potential': 'IPO or M&A',
        'Data Source': 'Verified',
        'Latitude': 37.7749,  # San Francisco, CA
        'Longitude': -122.4194
    }
}

# Apply manual overrides for the 10 startups
for company, data in manual_data.items():
    mask = df['Organization Name'] == company
    for col, val in data.items():
        df.loc[mask, col] = val

# Flag top 50 companies by funding for manual research
if 'Total Equity Funding Amount (in USD)' in df.columns:
    top_50 = df.nlargest(50, 'Total Equity Funding Amount (in
USD)')['Organization Name']
    df.loc[df['Organization Name'].isin(top_50) & (df['Data Source'] ==
'Proxy'), 'Data Source'] = 'Needs Research'

# Save the enriched CSV
df.to_csv('StartUp_Companies.csv', index=False)
print("Enriched CSV saved as 'StartUp_Companies.csv'.")
print("Note: Precise data provided for 10 companies. Valuation set to 5x
'Last Equity Funding Amount' for others.")
print("Latitude and Longitude added for known cities; others flagged as
'Needs Coordinates' in 'Data Source'.")
print("Top 50 by funding flagged for manual research.")
```

Enriched CSV saved as 'StartUp_Companies.csv'.
Note: Precise data provided for 10 companies. Valuation set to 5x 'Last Equity Funding Amount' for others.
Latitude and Longitude added for known cities; others flagged as 'Needs Coordinates' in 'Data Source'.
Top 50 by funding flagged for manual research.

Stage 2. Importing the data into the database

Creating table

```
CREATE TABLE startups_companies (
  organization_name TEXT,
  organization_name_url TEXT,
  operating_status TEXT,
  ipo_status TEXT,
  last_funding_type TEXT,
  full_description TEXT,
  industries TEXT,
  headquarters_location TEXT,
  description TEXT,
  cb_rank_company TEXT,
  last_funding_amount NUMERIC,
  last_funding_amount_currency TEXT,
  last_funding_amount_usd NUMERIC,
  total_equity_funding_amount NUMERIC,
  total_equity_funding_amount_currency TEXT,
  total_equity_funding_amount_usd NUMERIC,
  total_funding_amount BIGINT,
  total_funding_amount_currency TEXT,
  total_funding_amount_usd BIGINT,
  top_5_investors TEXT,
  exit_date TEXT,
  exit_date_precision TEXT,
  founded_date TEXT,
  founded_date_precision TEXT,
  investment_stage TEXT,
  number_of_funding_rounds INT,
  funding_status TEXT,
  last_funding_date TEXT,
  investor_type TEXT,
  last_equity_funding_amount NUMERIC,
  last_equity_funding_amount_currency TEXT,
  last_equity_funding_amount_usd NUMERIC,
  last_equity_funding_type TEXT,
  arr_usd NUMERIC,
  mrr_usd NUMERIC,
  revenue_growth_yoy NUMERIC,
  cac_usd INT,
  ltv_usd NUMERIC,
  ltv_cac_ratio NUMERIC,
  user_customer_growth_yoy INT,
```

```
  retention_rate INT,
  nps TEXT,
  mau TEXT,
  dau TEXT,
  founder_background TEXT,
  team_size INT,
  burn_rate_usd_month INT,
  runway_months NUMERIC,
  valuation_usd NUMERIC,
  proprietary_technology TEXT,
  patents_yn TEXT,
  key_customers TEXT,
  tam_usd BIGINT,
  exit_potential TEXT,
  data_source TEXT
);
```

Populating table

```
COPY startups_companies
FROM '/Users/nikita/Desktop/LTC/JM/LTC_Data_Project/StartUps_Data.csv'
WITH (
  FORMAT csv,
  HEADER true,
  NULL '',
  DELIMITER ',',
  QUOTE '"'
);
```

Stage 3. Cleaning table

Checking datatypes

```
SELECT column_name, data_type
FROM information_schema.columns
WHERE table_name = 'startup_companies'
ORDER BY ordinal_position;
```

Selecting Columns with the word "date" in them and their contents' datatypes

```
SELECT column_name, data_type
FROM information_schema.columns
WHERE table_name = 'startup_companies'
AND column_name ILIKE '%date%'
ORDER BY ordinal_position;
```

Showing the values of the previous columns

select"exit_date", "exit_date_precision", "founded_date", "founded_date_precision", "last_funding_date"
From startup_companies
LIMIT 10

Changing the datatypes for columns exit_date, founded_date and last_funding_date

```
ALTER TABLE startup_companies
ALTER COLUMN exit_date TYPE DATE USING TO_DATE(exit_date, 'DD/MM/YYYY'),
ALTER COLUMN founded_date TYPE DATE USING TO_DATE(founded_date, 'DD/MM/YYYY'),
ALTER COLUMN last_funding_date TYPE DATE USING TO_DATE(last_funding_date, 'DD/MM/YYYY');
```

Checking the data in nps, mau, dau columns

```
SELECT nps, mau, dau FROM your_table_name

WHERE nps != 'N/A' OR mau != 'N/A' OR dau != 'N/A' LIMIT 10;
```

Removing the columns

```
alter table startup_companies

drop column nps,

Drop column mau,

Drop column dau
```

Checking the new number of columns

```
SELECT column_name, data_type
FROM information_schema.columns
WHERE table_name = 'startup_companies'

ORDER BY ordinal_position;
```

Checking for duplicates and NULL values (RStudio)

Installing and loading the libraries


```r
install.packages(c("tidyverse", "lubridate", "skimr", "here", "janitor", "ggplot2"))
library(tidyverse)
library(lubridate)
library(skimr)
library(here)
library(janitor)
library(ggplot2)

```

```
7:17  (Top Level)                                                    R Script
```

Console | Terminal | Background Jobs — Go forward to the next source location (⌘F10)

R 4.4.1 · ~/

```
The downloaded binary packages are in
        /var/folders/64/8c76lfjs7kq52sph54q_1bwc0000gn/T//RtmpK8hhlT/downloaded_packages
> install.packages(c("tidyverse", "lubridate", "skimr", "here", "janitor", "ggplot2"))
trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.4/tidyverse_2.0.0.tgz'
Content type 'application/x-gzip' length 428901 bytes (418 KB)
==================================================
downloaded 418 KB

trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.4/lubridate_1.9.4.tgz'
Content type 'application/x-gzip' length 1003062 bytes (979 KB)
==================================================
downloaded 979 KB

trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.4/skimr_2.1.5.tgz'
Content type 'application/x-gzip' length 1225115 bytes (1.2 MB)
==================================================
downloaded 1.2 MB

trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.4/here_1.0.1.tgz'
Content type 'application/x-gzip' length 51224 bytes (50 KB)
==================================================
downloaded 50 KB

trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.4/janitor_2.2.1.tgz'
Content type 'application/x-gzip' length 286262 bytes (279 KB)
==================================================
downloaded 279 KB

trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.4/ggplot2_3.5.2.tgz'
Content type 'application/x-gzip' length 4969589 bytes (4.7 MB)
==================================================
downloaded 4.7 MB
```

```
The downloaded binary packages are in
        /var/folders/64/8c76lfjs7kq52sph54q_1bwc0000gn/T//RtmpK8hhlT/downloaded_packages
> library(tidyverse)
── Attaching core tidyverse packages ──────────────────────────────────────── tidyverse 2.0.0 ──
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.0     ✔ stringr   1.5.1
✔ ggplot2   3.5.2     ✔ tibble    3.2.1
✔ lubridate 1.9.4     ✔ tidyr     1.3.1
✔ purrr     1.0.2
── Conflicts ────────────────────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package to force all conflicts to become errors
> library(lubridate)
> library(skimr)
> library(here)

here() starts at /Users/nikita

> library(janitor)

Attaching package: ‘janitor’                    Go forward to the next source location (⌘F1

The following objects are masked from ‘package:stats’:

    chisq.test, fisher.test

> library(ggplot2)
>
```

Connecting RStudio to the database and printing the table (RStudio)

```r
install.packages("DBI")
install.packages("RPostgres")

library(DBI)
library(RPostgres)
```

5:19 (Top Level) R Script

Console Terminal Background Jobs

R 4.4.1 · ~/

```
> install.packages("DBI")
trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.4/DBI_1.2.3.tgz'
Content type 'application/x-gzip' length 910540 bytes (889 KB)
==================================================
downloaded 889 KB


The downloaded binary packages are in
        /var/folders/64/8c76lfjs7kq52sph54q_1bwc0000gn/T//RtmpK8hhlT/downloaded_packages
> install.packages("RPostgres")
also installing the dependency ‘plogr’

trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.4/plogr_0.2.0.tgz'
Content type 'application/x-gzip' length 13426 bytes (13 KB)
==================================================
downloaded 13 KB

trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.4/RPostgres_1.4.8.tgz'
Content type 'application/x-gzip' length 3016879 bytes (2.9 MB)
==================================================
downloaded 2.9 MB


The downloaded binary packages are in
        /var/folders/64/8c76lfjs7kq52sph54q_1bwc0000gn/T//RtmpK8hhlT/downloaded_packages
>
> library(DBI)
> library(RPostgres)
```


```r
con <- dbConnect(
  RPostgres::Postgres(),
  dbname = "nikita",
  host = "localhost",        # e.g., "localhost" or a server IP
  port = 5432,               # Default PostgreSQL port
  user = "nikita",
  password = "9604"
)
```

8:2 (Top Level) R Script

Console Terminal Background Jobs

R 4.4.1 · ~/

```
> con <- dbConnect(
+   RPostgres::Postgres(),
+   dbname = "nikita",
+   host = "localhost",        # e.g., "localhost" or a server IP
+   port = 5432,               # Default PostgreSQL port
+   user = "nikita",
+   password = "9604"
+ )
```

```r
query <- "SELECT * FROM startup_companies LIMIT 10;"
data <- dbGetQuery(con, query)

print(data)
```

3:1 (Top Level) R Script

Console Terminal Background Jobs

R 4.4.1 · ~/

```
> query <- "SELECT * FROM startup_companies LIMIT 10;"
> query <- "SELECT * FROM startup_companies LIMIT 10;"
> data <- dbGetQuery(con, query)
>
> print(data)
    organization_name                              organization_name_url operating_status ipo_status
1        RelationalAI https://www.crunchbase.com/organization/relationalai           Active    Private
2              Aisera       https://www.crunchbase.com/organization/aisera           Active    Private
3            Fountain     https://www.crunchbase.com/organization/fountain           Active    Private
4              Branch   https://www.crunchbase.com/organization/branch-app           Active    Private
5              Onfido       https://www.crunchbase.com/organization/onfido           Active    Private
6              Cresta       https://www.crunchbase.com/organization/cresta           Active    Private
7           Super.com  https://www.crunchbase.com/organization/superdotcom           Active    Private
8  Valon Technologies https://www.crunchbase.com/organization/peach-street           Active    Private
9          Safely You   https://www.crunchbase.com/organization/safely-you           Active    Private
10       Eleos Health https://www.crunchbase.com/organization/eleos-health           Active    Private
   last_funding_type
1           Series B
2           Series D
3           Series C
4           Series C
5           Series D
6           Series D
7           Series C
8           Series C
9           Series C
10          Series C

full_description
1
RelationalAI is building a groundbreaking relational knowledge graph system for developing intelligent data app
s.
2  Aisera is a leading provider of Generative AI Solutions that helps enterprises boost revenue, improve user p
roductivity, lower operating expenses and create magical user experiences. Aisera's products are AiseraGPT, AI
Copilot, AI Search and AiseraLLMs which are built on the AI Experience (AIX) platform that serve as an enterpri
se Generative AI stack for organizations to buy or build solutions. Aisera solutions deliver human-like interac
tions while providing contextually rich conversations that boost workforce productivity. Aisera's AIX platform
with pre-trained domain-specific LLMs are customizable to customer data, such that enterprises can get better a
```

Checking for duplicate rows

```
> get_dupes(data)
No variable names specified - using all columns.

No duplicate combinations found of: organization_name, organization_name_url, operating_status, ipo_status, la
st_funding_type, full_description, industries, headquarters_location, description, ... and 45 other variables

 [1] organization_name                   organization_name_url
 [3] operating_status                    ipo_status
 [5] last_funding_type                   full_description
 [7] industries                          headquarters_location
 [9] description                         cb_rank_company
[11] last_funding_amount                 last_funding_amount_currency
[13] last_funding_amount_usd             total_equity_funding_amount
[15] total_equity_funding_amount_currency total_equity_funding_amount_usd
[17] total_funding_amount                total_funding_amount_currency
[19] total_funding_amount_usd            top_5_investors
[21] exit_date                           exit_date_precision
[23] founded_date                        founded_date_precision
[25] investment_stage                    number_of_funding_rounds
[27] funding_status                      last_funding_date
[29] investor_type                       last_equity_funding_amount
[31] last_equity_funding_amount_currency last_equity_funding_amount_usd
[33] last_equity_funding_type            latitude
[35] longitude                           data_source
[37] arr_usd                             mrr_usd
[39] revenue_growth_yoy                  cac_usd
[41] ltv_usd                             ltv_cac_ratio
[43] user_customer_growth_yoy            retention_rate
[45] founder_background                  team_size
[47] burn_rate_usd_month                 runway_months
[49] valuation_usd                       proprietary_technology
[51] patents_yn                          key_customers
[53] tam_usd                             exit_potential
[55] dupe_count
<0 rows> (or 0-length row.names)
>
```

Checking for NULL values

```
> # Load data with DATE columns formatted as YYYY-MM-DD
> query <- "
+ SELECT
+     TO_CHAR(exit_date, 'YYYY-MM-DD') AS exit_date,
+     TO_CHAR(founded_date, 'YYYY-MM-DD') AS founded_date,
+     TO_CHAR(last_funding_date, 'YYYY-MM-DD') AS last_funding_date
+ FROM startup_companies;
+ "
> data <- dbGetQuery(con, query)
>
> # Check data structure
> str(data)
'data.frame':   895 obs. of  3 variables:
 $ exit_date        : chr  NA NA NA NA ...
 $ founded_date     : chr  "2017-09-01" "2017-01-01" "2014-01-01" "2015-11-01" ...
 $ last_funding_date: chr  "2022-04-26" "2022-08-03" "2022-06-15" "2022-03-09" ...
> summary(data)
  exit_date         founded_date       last_funding_date
 Length:895         Length:895         Length:895
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character
>
> # Convert character dates to Date objects
> data$exit_date <- as.Date(data$exit_date, format = "%Y-%m-%d", tz = "UTC")
> data$founded_date <- as.Date(data$founded_date, format = "%Y-%m-%d", tz = "UTC")
> data$last_funding_date <- as.Date(data$last_funding_date, format = "%Y-%m-%d", tz = "UTC")
>
> # View the first few rows
> head(data)
   exit_date founded_date last_funding_date
1       <NA>   2017-09-01        2022-04-26
2       <NA>   2017-01-01        2022-08-03
3       <NA>   2014-01-01        2022-06-15
4       <NA>   2015-11-01        2022-03-09
5 2024-04-09   2012-01-01        2020-04-15
6       <NA>   2017-01-01        2024-11-19
```

```
> # Count NA and 'N/A'
> count_missing <- function(x) {
+   x_char <- as.character(x)
+   sum(is.na(x) | (!is.na(x_char) & x_char == "N/A"), na.rm = TRUE)
+ }
> missing_counts <- sapply(data, count_missing)
> print(missing_counts)
        exit_date      founded_date last_funding_date
              790                 0                 0
>
> # Columns with missing values
> columns_with_missing <- names(missing_counts[missing_counts > 0])
> print(columns_with_missing)
[1] "exit_date"
>
> # Proportions
> missing_proportions <- missing_counts / nrow(data)
> print(missing_proportions)
        exit_date      founded_date last_funding_date
        0.8826816         0.0000000         0.0000000
>
```

Checking for logical order

```
> # Validate dates by setting invalid values to NULL
> dbExecute(con, "
+   UPDATE startup_companies
+   SET exit_date = NULL
+   WHERE exit_date > CURRENT_DATE OR exit_date < founded_date;
+ ")
[1] 0
>
> dbExecute(con, "
+   UPDATE startup_companies
+   SET last_funding_date = NULL
+   WHERE last_funding_date > CURRENT_DATE OR last_funding_date < founded_date;
+ ")
[1] 0
>
> dbExecute(con, "
+   UPDATE startup_companies
+   SET founded_date = NULL
+   WHERE founded_date > CURRENT_DATE;
+ ")
[1] 0
```

Analyse and Share

Stage 1. New table startups_best

```sql
CREATE TABLE startups_best AS
SELECT *
FROM startup_companies
WHERE
    founded_date BETWEEN '2013-01-01' AND '2018-12-31'
    AND last_funding_date < '2024-05-27'
    AND industries NOT ILIKE '%crypto%'
    AND industries NOT ILIKE '%blockchain%'
    AND full_description NOT ILIKE '%crypto%'
    AND full_description NOT ILIKE '%blockchain%'
    AND total_funding_amount_usd >= 100000000
    AND number_of_funding_rounds >= 4
    AND exit_date IS NULL
ORDER BY total_funding_amount_usd DESC;
```

Data Output | Messages | Notifications

```
SELECT 335

Query returned successfully in 80 msec.
```

Stage 2. Funding trends

```
> ggplot(startups_best,aes(x=number_of_funding_rounds, y=total_equity_funding_amount_usd)) + geom_line()+ geom_
point(data = startups_best, aes(x = number_of_funding_rounds, y = total_equity_funding_amount_usd), color = "re
d", size = 3) + labs(x = "Total Funding Rounds", y = "Total Equity Funding")+ scale_x_continuous(breaks = seq
(4, 16, by = 1)) +scale_y_continuous(breaks = seq(0, 7.5e+08, by = 1e+08), labels = scales::comma)
> ggplot(startups_best,aes(x=number_of_funding_rounds, y=total_equity_funding_amount_usd)) + geom_bin2d() + lab
s(x = "Total Funding Rounds", y = "Total Equity Funding")+ scale_x_continuous(breaks = seq(4, 16, by = 1)) + sc
ale_y_continuous(breaks = seq(0, 7.5e+08, by = 1e+08), labels = scales::comma)
```

Stage 3. Location and Industry

Industry

Image 4.3

SELECT
  CASE
    WHEN industry IN ('Mobile Payments', 'Payments', 'Banking', 'Financial Services', 'FinTech', 'Consumer credit/lending', 'Digital banking', 'Insurtech', 'Money transfer', 'Regtech', 'Asset management') THEN 'FinTech'

```
    WHEN industry IN ('Analytics', 'Artificial Intelligence (AI)', 'Big Data', 'Data Integration',
'Cloud Data Services', 'Database') THEN 'AI & Data'
    WHEN industry IN ('Software', 'SaaS', 'Enterprise Software', 'App Marketing', 'Mobile
Apps', 'Sales Automation') THEN 'Software/SaaS'
    WHEN industry IN ('Cloud Security', 'Cyber Security', 'Network Security', 'Security')
THEN 'Cybersecurity'
    WHEN industry IN ('Employment', 'Human Resources', 'Information Technology',
'Recruiting') THEN 'HR Tech'
    WHEN industry IN ('Computer', 'Hardware', 'Electronic devices', 'Peripherals',
'Semiconductors', 'Components', 'Internet of things') THEN 'Hardware'
        WHEN industry IN ('Satellite telecommunications service', 'Service
provider', 'Wireless service provider') THEN 'Communications'
        WHEN industry IN ('Clean tech', 'Energy tech', 'Green tech') THEN
'Environmental Technology'
        WHEN industry IN ('Biotechnology', 'FoodTech', 'Medical Devices',
'Medtech') THEN 'Healthcare and Life Sciences'
        WHEN industry IN ('Commerce', 'Advertising', 'Adtech', 'Digital commerce',
'Gaming', 'Over the top service', 'Social media') THEN 'Media and Entertainment'
    ELSE 'Other'
  END AS category,
  COUNT(DISTINCT organization_name) AS company_count
FROM (
  SELECT
    organization_name,
    unnest(string_to_array(industries, ', ')) AS industry
  FROM startups_best
) AS expanded
GROUP BY
  CASE
    WHEN industry IN ('Mobile Payments', 'Payments', 'Banking', 'Financial Services',
'FinTech', 'Consumer credit/lending', 'Digital banking', 'Insurtech', 'Money transfer',
'Regtech', 'Asset management') THEN 'FinTech'
    WHEN industry IN ('Analytics', 'Artificial Intelligence (AI)', 'Big Data', 'Data Integration',
'Cloud Data Services', 'Database') THEN 'AI & Data'
    WHEN industry IN ('Software', 'SaaS', 'Enterprise Software', 'App Marketing', 'Mobile
Apps', 'Sales Automation') THEN 'Software/SaaS'
    WHEN industry IN ('Cloud Security', 'Cyber Security', 'Network Security', 'Security')
THEN 'Cybersecurity'
    WHEN industry IN ('Employment', 'Human Resources', 'Information Technology',
'Recruiting') THEN 'HR Tech'
    WHEN industry IN ('Computer', 'Hardware', 'Electronic devices', 'Peripherals',
'Semiconductors', 'Components', 'Internet of things') THEN 'Hardware'
        WHEN industry IN ('Satellite telecommunications service', 'Service
provider', 'Wireless service provider') THEN 'Communications'
        WHEN industry IN ('Clean tech', 'Energy tech', 'Green tech') THEN
'Environmental Technology'
```

```
            WHEN industry IN ('Biotechnology', 'FoodTech', 'Medical Devices',
'Medtech') THEN 'Healthcare and Life Sciences'
            WHEN industry IN ('Commerce', 'Advertising', 'Adtech', 'Digital commerce',
'Gaming', 'Over the top service', 'Social media') THEN 'Media and Entertainment'
      ELSE 'Other'
  END
ORDER BY company_count DESC;
```

Image 4.5

```
WITH expanded AS (
  SELECT
    organization_name,
    unnest(string_to_array(industries, ', ')) AS industry,
    total_equity_funding_amount_usd
  FROM startups_best
  WHERE total_equity_funding_amount_usd IS NOT NULL
)
SELECT
  CASE
    WHEN industry IN ('Mobile Payments', 'Payments', 'Banking', 'Financial Services',
'FinTech', 'Consumer credit/lending', 'Digital banking', 'Insurtech', 'Money transfer',
'Regtech', 'Asset management') THEN 'FinTech'
    WHEN industry IN ('Analytics', 'Artificial Intelligence (AI)', 'Big Data', 'Data Integration',
'Cloud Data Services', 'Database') THEN 'AI & Data'
    WHEN industry IN ('Software', 'SaaS', 'Enterprise Software', 'App Marketing', 'Mobile
Apps', 'Sales Automation') THEN 'Software/SaaS'
    WHEN industry IN ('Cloud Security', 'Cyber Security', 'Network Security', 'Security')
THEN 'Cybersecurity'
    WHEN industry IN ('Employment', 'Human Resources', 'Information Technology',
'Recruiting') THEN 'HR Tech'
    WHEN industry IN ('Computer', 'Hardware', 'Electronic devices', 'Peripherals',
'Semiconductors', 'Components', 'Internet of things') THEN 'Hardware'
            WHEN industry IN ('Satellite telecommunications service', 'Service
provider', 'Wireless service provider') THEN 'Communications'
            WHEN industry IN ('Clean tech', 'Energy tech', 'Green tech') THEN
'Environmental Technology'
            WHEN industry IN ('Biotechnology', 'FoodTech', 'Medical Devices',
'Medtech') THEN 'Healthcare and Life Sciences'
            WHEN industry IN ('Commerce', 'Advertising', 'Adtech', 'Digital commerce',
'Gaming', 'Over the top service', 'Social media') THEN 'Media and Entertainment'
    ELSE 'Other'
  END AS category,
  SUM(total_equity_funding_amount_usd) AS total_funding_amount
FROM expanded
GROUP BY
```

```
  CASE
    WHEN industry IN ('Mobile Payments', 'Payments', 'Banking', 'Financial Services',
'FinTech', 'Consumer credit/lending', 'Digital banking', 'Insurtech', 'Money transfer',
'Regtech', 'Asset management') THEN 'FinTech'
    WHEN industry IN ('Analytics', 'Artificial Intelligence (AI)', 'Big Data', 'Data Integration',
'Cloud Data Services', 'Database') THEN 'AI & Data'
    WHEN industry IN ('Software', 'SaaS', 'Enterprise Software', 'App Marketing', 'Mobile
Apps', 'Sales Automation') THEN 'Software/SaaS'
    WHEN industry IN ('Cloud Security', 'Cyber Security', 'Network Security', 'Security')
THEN 'Cybersecurity'
    WHEN industry IN ('Employment', 'Human Resources', 'Information Technology',
'Recruiting') THEN 'HR Tech'
    WHEN industry IN ('Computer', 'Hardware', 'Electronic devices', 'Peripherals',
'Semiconductors', 'Components', 'Internet of things') THEN 'Hardware'
            WHEN industry IN ('Satellite telecommunications service', 'Service
provider', 'Wireless service provider') THEN 'Communications'
            WHEN industry IN ('Clean tech', 'Energy tech', 'Green tech') THEN
'Environmental Technology'
            WHEN industry IN ('Biotechnology', 'FoodTech', 'Medical Devices',
'Medtech') THEN 'Healthcare and Life Sciences'
            WHEN industry IN ('Commerce', 'Advertising', 'Adtech', 'Digital commerce',
'Gaming', 'Over the top service', 'Social media') THEN 'Media and Entertainment'
    ELSE 'Other'
  END
ORDER BY total_funding_amount DESC;
```

Image 4.7

```
WITH expanded AS (
  SELECT
    organization_name,
    unnest(string_to_array(industries, ', ')) AS industry,
    total_equity_funding_amount_usd
  FROM startups_best
  WHERE total_equity_funding_amount_usd IS NOT NULL
)
SELECT
  CASE
    WHEN industry IN ('Mobile Payments', 'Payments', 'Banking', 'Financial Services',
'FinTech', 'Consumer credit/lending', 'Digital banking', 'Insurtech', 'Money transfer',
'Regtech', 'Asset management') THEN 'FinTech'
    WHEN industry IN ('Analytics', 'Artificial Intelligence (AI)', 'Big Data', 'Data Integration',
'Cloud Data Services', 'Database') THEN 'AI & Data'
    WHEN industry IN ('Software', 'SaaS', 'Enterprise Software', 'App Marketing', 'Mobile
Apps', 'Sales Automation') THEN 'Software/SaaS'
```

```
    WHEN industry IN ('Cloud Security', 'Cyber Security', 'Network Security', 'Security')
THEN 'Cybersecurity'
    WHEN industry IN ('Employment', 'Human Resources', 'Information Technology',
'Recruiting') THEN 'HR Tech'
    WHEN industry IN ('Computer', 'Hardware', 'Electronic devices', 'Peripherals',
'Semiconductors', 'Components', 'Internet of things') THEN 'Hardware'
                WHEN industry IN ('Satellite telecommunications service', 'Service
provider', 'Wireless service provider') THEN 'Communications'
                WHEN industry IN ('Clean tech', 'Energy tech', 'Green tech') THEN
'Environmental Technology'
                WHEN industry IN ('Biotechnology', 'FoodTech', 'Medical Devices',
'Medtech') THEN 'Healthcare and Life Sciences'
                WHEN industry IN ('Commerce', 'Advertising', 'Adtech', 'Digital commerce',
'Gaming', 'Over the top service', 'Social media') THEN 'Media and Entertainment'
    ELSE 'Other'
  END AS category,
  AVG(total_equity_funding_amount_usd) AS avg_funding_amount
FROM expanded
GROUP BY
  CASE
   WHEN industry IN ('Mobile Payments', 'Payments', 'Banking', 'Financial Services',
'FinTech', 'Consumer credit/lending', 'Digital banking', 'Insurtech', 'Money transfer',
'Regtech', 'Asset management') THEN 'FinTech'
    WHEN industry IN ('Analytics', 'Artificial Intelligence (AI)', 'Big Data', 'Data Integration',
'Cloud Data Services', 'Database') THEN 'AI & Data'
    WHEN industry IN ('Software', 'SaaS', 'Enterprise Software', 'App Marketing', 'Mobile
Apps', 'Sales Automation') THEN 'Software/SaaS'
    WHEN industry IN ('Cloud Security', 'Cyber Security', 'Network Security', 'Security')
THEN 'Cybersecurity'
    WHEN industry IN ('Employment', 'Human Resources', 'Information Technology',
'Recruiting') THEN 'HR Tech'
    WHEN industry IN ('Computer', 'Hardware', 'Electronic devices', 'Peripherals',
'Semiconductors', 'Components', 'Internet of things') THEN 'Hardware'
                WHEN industry IN ('Satellite telecommunications service', 'Service
provider', 'Wireless service provider') THEN 'Communications'
                WHEN industry IN ('Clean tech', 'Energy tech', 'Green tech') THEN
'Environmental Technology'
                WHEN industry IN ('Biotechnology', 'FoodTech', 'Medical Devices',
'Medtech') THEN 'Healthcare and Life Sciences'
                WHEN industry IN ('Commerce', 'Advertising', 'Adtech', 'Digital commerce',
'Gaming', 'Over the top service', 'Social media') THEN 'Media and Entertainment'
    ELSE 'Other'
  END
ORDER BY avg_funding_amount DESC;
```

Location

Image 4.9

```
SELECT
   (string_to_array(headquarters_location, ', '))[3] AS country,
   COUNT(DISTINCT organization_name) AS company_count
FROM startups_best
WHERE (string_to_array(headquarters_location, ', '))[3] IS NOT NULL
GROUP BY (string_to_array(headquarters_location, ', '))[3]
ORDER BY company_count DESC;
```

Removing the “City of” value:

```
UPDATE startups_best
SET headquarters_location = array_to_string(
   ARRAY[
      (string_to_array(headquarters_location, ', '))[1],  -- First element (e.g., "London")
      (string_to_array(headquarters_location, ', '))[2],  -- Second element (e.g., "England")
      (string_to_array(headquarters_location, ', '))[4]   -- Fourth element (e.g., "United
Kingdom")
   ],
   ', '
)
WHERE (string_to_array(headquarters_location, ', '))[3] = 'City of';
```

Image 4.11

```sql
SELECT
    (string_to_array(headquarters_location, ', '))[3] AS country,
    SUM(total_equity_funding_amount_usd) AS total_funding_amount
FROM startups_best
WHERE (string_to_array(headquarters_location, ', '))[3] IS NOT NULL
    AND total_equity_funding_amount_usd IS NOT NULL
GROUP BY (string_to_array(headquarters_location, ', '))[3]
ORDER BY total_funding_amount DESC;
```

Image 4.13

```sql
SELECT
    (string_to_array(headquarters_location, ', '))[3] AS country,
    AVG(total_equity_funding_amount_usd) AS avg_funding_amount
FROM startups_best
WHERE (string_to_array(headquarters_location, ', '))[3] IS NOT NULL
    AND total_equity_funding_amount_usd IS NOT NULL
GROUP BY (string_to_array(headquarters_location, ', '))[3]
ORDER BY avg_funding_amount DESC;
```

Valuation. Image 4.16

```sql
select organization_name, valuation_usd from startups_best
Where valuation_usd is NOT NULL
Order by valuation_usd DESC
Limit 10
```