Funding in startups

Submitted by D A Santhosh

Importing Libraries

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

```
data = pd.read_csv('/content/investments_VC.csv', encoding = "latin1")
```

```
df = data.copy()
df.head()
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code	reg
0	/organization/waywire	#waywire	http://www.waywire.com	|Entertainment|Politics|Social Media|News|	News	17,50,000	acquired	USA	NY	
1	/organization/tv-communications	&TV Communications	http://enjoyandtv.com	|Games|	Games	40,00,000	operating	USA	CA	Ang
2	/organization/rock-your-paper	'Rock' Your Paper	http://www.rockyourpaper.org	|Publishing|Education|	Publishing	40,000	operating	EST	NaN	Ta
3	/organization/in-touch-network	(In)Touch Network	http://www.InTouchNetwork.com	|Electronics|Guides|Coffee|Restaurants|Music|i...	Electronics	15,00,000	operating	GBR	NaN	Lor
4	/organization/r-ranch-and-mine	-R- Ranch and Mine	NaN	|Tourism|Entertainment|Games|	Tourism	60,000	operating	USA	TX	Da

5 rows × 39 columns

```
df.shape
```

```
(54294, 39)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 54294 entries, 0 to 54293
Data columns (total 39 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   permalink             49438 non-null  object 
 1   name                  49437 non-null  object 
 2   homepage_url          45989 non-null  object 
 3   category_list         45477 non-null  object 
 4    market               45470 non-null  object 
 5    funding_total_usd    49438 non-null  object 
 6   status                48124 non-null  object 
 7   country_code          44165 non-null  object 
 8   state_code            30161 non-null  object 
 9   region                44165 non-null  object 
 10  city                  43322 non-null  object 
 11  funding_rounds        49438 non-null  float64
 12  founded_at            38554 non-null  object 
 13  founded_month         38482 non-null  object 
 14  founded_quarter       38482 non-null  object 
 15  founded_year          38482 non-null  float64
 16  first_funding_at      49438 non-null  object 
 17  last_funding_at       49438 non-null  object 
 18  seed                  49438 non-null  float64
 19  venture               49438 non-null  float64
 20  equity_crowdfunding   49438 non-null  float64
 21  undisclosed           49438 non-null  float64
 22  convertible_note      49438 non-null  float64
 23  debt_financing        49438 non-null  float64
 24  angel                 49438 non-null  float64
 25  grant                 49438 non-null  float64
 26  private_equity        49438 non-null  float64
 27  post_ipo_equity       49438 non-null  float64
 28  post_ipo_debt         49438 non-null  float64
 29  secondary_market      49438 non-null  float64
 30  product_crowdfunding  49438 non-null  float64
 31  round_A               49438 non-null  float64
 32  round_B               49438 non-null  float64
 33  round_C               49438 non-null  float64
 34  round_D               49438 non-null  float64
 35  round_E               49438 non-null  float64
 36  round_F               49438 non-null  float64
 37  round_G               49438 non-null  float64
 38  round_H               49438 non-null  float64
dtypes: float64(23), object(16)
memory usage: 16.2+ MB
```

```
df.describe(include="O").T
```

	count	unique	top	freq
permalink	49438	49436	/organization/treasure-valley-urology-services	2
name	49437	49350	Roost	4
homepage_url	45989	45850	http://spaceport.io	2
category_list	45477	16675	|Software|	3650
market	45470	753	Software	4620
funding_total_usd	49438	14617	-	8531
status	48124	3	operating	41829
country_code	44165	115	USA	28793
state_code	30161	61	CA	9917
region	44165	1089	SF Bay Area	6804
city	43322	4188	San Francisco	2615
founded_at	38554	3369	2012-01-01	2181
founded_month	38482	420	2012-01	2327
founded_quarter	38482	218	2012-Q1	2904
first_funding_at	49438	3914	2012-01-01	468
last_funding_at	49438	3657	2013-01-01	387

```
df.describe(include="d").T
```

	count	mean	std	min	25%	50%	75%	max
funding_rounds	49438.0	1.696205e+00	1.294213e+00	1.0	1.0	1.0	2.0	1.800000e+01
founded_year	38482.0	2.007359e+03	7.579203e+00	1902.0	2006.0	2010.0	2012.0	2.014000e+03
seed	49438.0	2.173215e+05	1.056985e+06	0.0	0.0	0.0	25000.0	1.300000e+08
venture	49438.0	7.501051e+06	2.847112e+07	0.0	0.0	0.0	5000000.0	2.351000e+09
equity_crowdfunding	49438.0	6.163322e+03	1.999048e+05	0.0	0.0	0.0	0.0	2.500000e+07
undisclosed	49438.0	1.302213e+05	2.981404e+06	0.0	0.0	0.0	0.0	2.924328e+08
convertible_note	49438.0	2.336410e+04	1.432046e+06	0.0	0.0	0.0	0.0	3.000000e+08
debt_financing	49438.0	1.888157e+06	1.382046e+08	0.0	0.0	0.0	0.0	3.007950e+10
angel	49438.0	6.541898e+04	6.582908e+05	0.0	0.0	0.0	0.0	6.359026e+07
grant	49438.0	1.628453e+05	5.612088e+06	0.0	0.0	0.0	0.0	7.505000e+08
private_equity	49438.0	2.074286e+06	3.167231e+07	0.0	0.0	0.0	0.0	3.500000e+09
post_ipo_equity	49438.0	6.088736e+05	2.678348e+07	0.0	0.0	0.0	0.0	4.700000e+09
post_ipo_debt	49438.0	4.434360e+05	3.428169e+07	0.0	0.0	0.0	0.0	5.800000e+09
secondary_market	49438.0	3.845592e+04	3.864461e+06	0.0	0.0	0.0	0.0	6.806116e+08
product_crowdfunding	49438.0	7.074227e+03	4.282166e+05	0.0	0.0	0.0	0.0	7.200000e+07
round_A	49438.0	1.243955e+06	5.531974e+06	0.0	0.0	0.0	0.0	3.190000e+08
round_B	49438.0	1.492891e+06	7.472704e+06	0.0	0.0	0.0	0.0	5.420000e+08
round_C	49438.0	1.205356e+06	7.993592e+06	0.0	0.0	0.0	0.0	4.900000e+08
round_D	49438.0	7.375261e+05	9.815218e+06	0.0	0.0	0.0	0.0	1.200000e+09
round_E	49438.0	3.424682e+05	5.406915e+06	0.0	0.0	0.0	0.0	4.000000e+08
round_F	49438.0	1.697692e+05	6.277905e+06	0.0	0.0	0.0	0.0	1.060000e+09
round_G	49438.0	5.767067e+04	5.252312e+06	0.0	0.0	0.0	0.0	1.000000e+09
round_H	49438.0	1.423197e+04	2.716865e+06	0.0	0.0	0.0	0.0	6.000000e+08

```
df.isna().sum()
```

	0
permalink	4856
name	4857
homepage_url	8305
category_list	8817
market	8824
funding_total_usd	4856
status	6170
country_code	10129
state_code	24133
region	10129
city	10972
funding_rounds	4856
founded_at	15740
founded_month	15812
founded_quarter	15812
founded_year	15812
first_funding_at	4856
last_funding_at	4856
seed	4856
venture	4856
equity_crowdfunding	4856
undisclosed	4856
convertible_note	4856
debt_financing	4856
angel	4856
grant	4856
private_equity	4856
post_ipo_equity	4856
post_ipo_debt	4856
secondary_market	4856

product_crowdfunding	4856
round_A	4856
round_B	4856
round_C	4856
round_D	4856
round_E	4856
round_F	4856
round_G	4856
round_H	4856

```
# percentage of null values
np.round((df.isna().sum()/df.shape[0]*100),2).reset_index().sort_values(by=0, ascending=False)
```

	index	0
8	state_code	44.45
13	founded_month	29.12
15	founded_year	29.12
14	founded_quarter	29.12
12	founded_at	28.99
10	city	20.21
7	country_code	18.66
9	region	18.66
4	market	16.25
3	category_list	16.24
2	homepage_url	15.30
6	status	11.36
1	name	8.95
28	post_ipo_debt	8.94
29	secondary_market	8.94
30	product_crowdfunding	8.94
31	round_A	8.94
32	round_B	8.94
0	permalink	8.94
33	round_C	8.94
34	round_D	8.94
35	round_E	8.94
26	private_equity	8.94
36	round_F	8.94
37	round_G	8.94
27	post_ipo_equity	8.94
19	venture	8.94
25	grant	8.94
24	angel	8.94
23	debt_financing	8.94

22	convertible_note	8.94
21	undisclosed	8.94
20	equity_crowdfunding	8.94
18	seed	8.94
17	last_funding_at	8.94
16	first_funding_at	8.94
11	funding_rounds	8.94
5	funding_total_usd	8.94
38	round_H	8.94

Columns

```
df.columns = df.columns.str.strip()
df.columns
```

```
Index(['permalink', 'name', 'homepage_url', 'category_list', 'market',
       'funding_total_usd', 'status', 'country_code', 'state_code', 'region',
       'city', 'funding_rounds', 'founded_at', 'founded_month',
       'founded_quarter', 'founded_year', 'first_funding_at',
       'last_funding_at', 'seed', 'venture', 'equity_crowdfunding',
       'undisclosed', 'convertible_note', 'debt_financing', 'angel', 'grant',
       'private_equity', 'post_ipo_equity', 'post_ipo_debt',
       'secondary_market', 'product_crowdfunding', 'round_A', 'round_B',
       'round_C', 'round_D', 'round_E', 'round_F', 'round_G', 'round_H'],
      dtype='object')
```

Null Values

```
#dropping rows where all values are nan
df = df.dropna(how="all")
df
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_coc
0	/organization/waywire	#waywire	http://www.waywire.com	|Entertainment|Politics|Social Media|News|	News	17,50,000	acquired	USA	N
1	/organization/tv-communications	&TV Communications	http://enjoyandtv.com	|Games|	Games	40,00,000	operating	USA	C
2	/organization/rock-your-paper	'Rock' Your Paper	http://www.rockyourpaper.org	|Publishing|Education|	Publishing	40,000	operating	EST	Na
3	/organization/in-touch-network	(In)Touch Network	http://www.InTouchNetwork.com	|Electronics|Guides|Coffee|Restaurants|Music|i...	Electronics	15,00,000	operating	GBR	Na
4	/organization/r-ranch-and-mine	-R- Ranch and Mine	NaN	|Tourism|Entertainment|Games|	Tourism	60,000	operating	USA	T
...	...	...	...	...	...	...	...	...	
49433	/organization/zzish	Zzish	http://www.zzish.com	|Analytics|Gamification|Developer APIs|iOS|And...	Education	3,20,000	operating	GBR	Na
49434	/organization/zznode-science-and-technology-co...	ZZNode Science and Technology	http://www.zznode.com	|Enterprise Software|	Enterprise Software	15,87,301	operating	CHN	Na
49435	/organization/zzzzapp-com	Zzzzapp Wireless ltd.	http://www.zzzzapp.com	|Web Development|Advertising|Wireless|Mobile|	Web Development	97,398	operating	HRV	Na
49436	/organization/a-list-games	[a]list games	http://www.alistgames.com	|Games|	Games	93,00,000	operating	NaN	Na
49437	/organization/x	[x+1]	http://www.xplusone.com/	|Enterprise Software|	Enterprise Software	4,50,00,000	operating	USA	N

49438 rows × 39 columns

```
df.isna().all(axis=1).sum()
```

```
0
```

```
df.isna().sum()
```

	0
permalink	0
name	1
homepage_url	3449
category_list	3961
market	3968
funding_total_usd	0
status	1314
country_code	5273
state_code	19277
region	5273
city	6116
funding_rounds	0
founded_at	10884
founded_month	10956
founded_quarter	10956
founded_year	10956
first_funding_at	0
last_funding_at	0
seed	0
venture	0
equity_crowdfunding	0
undisclosed	0
convertible_note	0
debt_financing	0
angel	0
grant	0
private_equity	0
post_ipo_equity	0
post_ipo_debt	0
secondary_market	0

product_crowdfunding	0
round_A	0
round_B	0
round_C	0
round_D	0
round_E	0
round_F	0
round_G	0
round_H	0

Duplicates

```
df["permalink"].value_counts()
```

permalink	count
/organization/treasure-valley-urology-services	2
/organization/prysm	2
/organization/waywire	1
/organization/polybona	1
/organization/pollfish	1
...	...
/organization/game-ventures	1
/organization/game9z	1
/organization/gameaccount-network	1
/organization/gameanalytics	1
/organization/x	1

49436 rows × 1 columns

```
df[df["permalink"] == "/organization/treasure-valley-urology-services"]
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code	region	...	secondary_market	product_crow
44033	/organization/treasure-valley-urology-services	Treasure Valley Urology Services	NaN	|Biotechnology|	Biotechnology	3,32,194	operating	USA	TX	Austin	...	0.0	
44034	/organization/treasure-valley-urology-services	Treasure Valley Urology Services	NaN	NaN	NaN	3,32,194	operating	USA	TX	Austin	...	0.0	

2 rows × 39 columns

```
df = df.drop(44034)
```

```
df[df["permalink"] == "/organization/treasure-valley-urology-services"]
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code	region	...	secondary_market	product_crow
44033	/organization/treasure-valley-urology-services	Treasure Valley Urology Services	NaN	|Biotechnology|	Biotechnology	3,32,194	operating	USA	TX	Austin	...	0.0	

1 rows × 39 columns

```
df[df["permalink"] == "/organization/prysm"]
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code	region	...	secondary_market	product_crow
33939	/organization/prysm	Prysm	http://www.prysm.com/	NaN	NaN	29,30,80,123	operating	NaN	NaN	NaN	...	0.0	
33940	/organization/prysm	Prysm	http://www.prysm.com	|Displays|Hardware + Software|	Displays	29,30,80,123	operating	USA	CA	SF Bay Area	...	0.0	

2 rows × 39 columns

```
df = df.drop(33939)
```

```
df[df["permalink"] == "/organization/prysm"]
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code	region	...	secondary_market	product_crowd
33940	/organization/prysm	Prysm	http://www.prysm.com	|Displays|Hardware + Software|	Displays	29,30,80,123	operating	USA	CA	SF Bay Area	...	0.0	

1 rows × 39 columns

Hence duplicates are removed

Null values

```
df[df["name"].isna()]
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code	region	...	secondary_market	product_crowdfunding	r
28221	/organization/tell-it-in	NaN	http://tellitin10.com	|Startups|	Startups	25,000	closed	NaN	NaN	NaN	...	0.0	0.0	

1 rows × 39 columns

```
val = df[df["permalink"] == "/organization/tell-it-in"]["permalink"].str.split("/",expand = True)[2]
df["name"].fillna(val, inplace = True)
```

```
df["name"].isna().sum()
```

```
0
```

Replacing null values in the name column

```
df.columns
```

```
Index(['permalink', 'name', 'homepage_url', 'category_list', 'market',
       'funding_total_usd', 'status', 'country_code', 'state_code', 'region',
       'city', 'funding_rounds', 'founded_at', 'founded_month',
       'founded_quarter', 'founded_year', 'first_funding_at',
       'last_funding_at', 'seed', 'venture', 'equity_crowdfunding',
       'undisclosed', 'convertible_note', 'debt_financing', 'angel', 'grant',
       'private_equity', 'post_ipo_equity', 'post_ipo_debt',
       'secondary_market', 'product_crowdfunding', 'round_A', 'round_B',
       'round_C', 'round_D', 'round_E', 'round_F', 'round_G', 'round_H'],
      dtype='object')
```

Filling Missing URL

```
df['homepage_url'].fillna('Unknown', inplace=True)
```

```
df['homepage_url'].isna().sum()
```

```
0
```

```
df['category_list'].fillna('Unknown', inplace=True)
```

```
df['category_list'].isna().sum()
```

```
0
```

```
df['market'].fillna('Unknown', inplace=True)
```

```
df['market'].isna().sum()
```

```
0
```

Removing "-" and replacing by '0'

```
df["funding_total_usd"] = df["funding_total_usd"].str.strip()
df["funding_total_usd"] = df["funding_total_usd"].str.replace(",","")
df["funding_total_usd"] = df["funding_total_usd"].replace("-","0")
df["funding_total_usd"] = df["funding_total_usd"].astype(float)
df["funding_total_usd"].dtype
```

```
dtype('float64')
```

```
df.sample(1)
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code	region	...	secondary_market	pro
33114	/organization/postcron	Postcron	http://postcron.com	|Twitter Applications|Productivity Software|We...	Twitter Applications	143083.0	operating	ARG	NaN	Cordoba, ARG	...	0.0	

1 rows × 39 columns

```
df['status'].fillna('Unknown', inplace=True)
df['status'].isna().sum()
```

```
0
```

```
df.isna().sum()
```

	0
permalink	0
name	0
homepage_url	0
category_list	0
market	0
funding_total_usd	0
status	0
country_code	5272
state_code	19276
region	5272
city	6115
funding_rounds	0
founded_at	10883
founded_month	10955
founded_quarter	10955
founded_year	10955
first_funding_at	0
last_funding_at	0
seed	0
venture	0
equity_crowdfunding	0
undisclosed	0
convertible_note	0
debt_financing	0
angel	0
grant	0
private_equity	0
post_ipo_equity	0
post_ipo_debt	0
secondary_market	0

product_crowdfunding	0
round_A	0
round_B	0
round_C	0
round_D	0
round_E	0
round_F	0
round_G	0
round_H	0

Country code / state code / region / city wise column

```
for col in ['country_code', 'state_code', 'region', 'city']:
    df[col].fillna('Unknown', inplace=True)
```

Dropping Found tear , month , at and quarter

```
df.dropna(subset=['founded_at', 'founded_month', 'founded_quarter', 'founded_year'], inplace=True)
```

```
df.isna().sum()
```

	0
permalink	0
name	0
homepage_url	0
category_list	0
market	0
funding_total_usd	0
status	0
country_code	0
state_code	0
region	0
city	0
funding_rounds	0
founded_at	0
founded_month	0
founded_quarter	0
founded_year	0
first_funding_at	0
last_funding_at	0
seed	0
venture	0
equity_crowdfunding	0
undisclosed	0
convertible_note	0
debt_financing	0
angel	0
grant	0
private_equity	0
post_ipo_equity	0
post_ipo_debt	0
secondary_market	0

product_crowdfunding	0
round_A	0
round_B	0
round_C	0
round_D	0
round_E	0
round_F	0
round_G	0
round_H	0

Since these columns are related to dates, missing values could impact the analysis. Therefore, we will remove rows with null values, as filling in the missing dates might compromise the accuracy of time-based analysis.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 38481 entries, 0 to 49437
Data columns (total 39 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   permalink             38481 non-null  object 
 1   name                  38481 non-null  object 
 2   homepage_url          38481 non-null  object 
 3   category_list         38481 non-null  object 
 4   market                38481 non-null  object 
 5   funding_total_usd     38481 non-null  float64
 6   status                38481 non-null  object 
 7   country_code          38481 non-null  object 
 8   state_code            38481 non-null  object 
 9   region                38481 non-null  object 
 10  city                  38481 non-null  object 
 11  funding_rounds        38481 non-null  float64
 12  founded_at            38481 non-null  object 
 13  founded_month         38481 non-null  object 
 14  founded_quarter       38481 non-null  object 
 15  founded_year          38481 non-null  float64
 16  first_funding_at      38481 non-null  object 
 17  last_funding_at       38481 non-null  object 
 18  seed                  38481 non-null  float64
 19  venture               38481 non-null  float64
 20  equity_crowdfunding   38481 non-null  float64
 21  undisclosed           38481 non-null  float64
 22  convertible_note      38481 non-null  float64
```

```
 23  debt_financing        38481 non-null  float64
 24  angel                 38481 non-null  float64
 25  grant                 38481 non-null  float64
 26  private_equity        38481 non-null  float64
 27  post_ipo_equity       38481 non-null  float64
 28  post_ipo_debt         38481 non-null  float64
 29  secondary_market      38481 non-null  float64
 30  product_crowdfunding  38481 non-null  float64
 31  round_A               38481 non-null  float64
 32  round_B               38481 non-null  float64
 33  round_C               38481 non-null  float64
 34  round_D               38481 non-null  float64
 35  round_E               38481 non-null  float64
 36  round_F               38481 non-null  float64
 37  round_G               38481 non-null  float64
 38  round_H               38481 non-null  float64
dtypes: float64(24), object(15)
memory usage: 11.7+ MB
```

```
df.sample(2)
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code	region	...	secondary_market	proc
4423	/organization/barcoding	Barcoding	http://www.barcoding.com	Unknown	Unknown	0.0	operating	USA	MD	Baltimore	...	0.0	
28977	/organization/neuaer	NewAer	http://www.newaer.com	|Mobile|	Mobile	0.0	operating	USA	CA	Los Angeles	...	0.0	

2 rows × 39 columns

Converting date-related columns to datetime

```
date_columns = ['founded_at', 'founded_month', 'founded_quarter', 'first_funding_at', 'last_funding_at']
df[date_columns] = df[date_columns].apply(pd.to_datetime, errors='coerce')
```

```
df['founded_year'] = df['founded_year'].astype(int)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 38481 entries, 0 to 49437
Data columns (total 39 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   permalink             38481 non-null  object
 1   name                  38481 non-null  object
 2   homepage_url          38481 non-null  object
 3   category_list         38481 non-null  object
```

```
 4   market                38481 non-null  object
 5   funding_total_usd     38481 non-null  float64
 6   status                38481 non-null  object
 7   country_code          38481 non-null  object
 8   state_code            38481 non-null  object
 9   region                38481 non-null  object
 10  city                  38481 non-null  object
 11  funding_rounds        38481 non-null  float64
 12  founded_at            38481 non-null  datetime64[ns]
 13  founded_month         38481 non-null  datetime64[ns]
 14  founded_quarter       38481 non-null  datetime64[ns]
 15  founded_year          38481 non-null  int64
 16  first_funding_at      38475 non-null  datetime64[ns]
 17  last_funding_at       38479 non-null  datetime64[ns]
 18  seed                  38481 non-null  float64
 19  venture               38481 non-null  float64
 20  equity_crowdfunding   38481 non-null  float64
 21  undisclosed           38481 non-null  float64
 22  convertible_note      38481 non-null  float64
 23  debt_financing        38481 non-null  float64
 24  angel                 38481 non-null  float64
 25  grant                 38481 non-null  float64
 26  private_equity        38481 non-null  float64
 27  post_ipo_equity       38481 non-null  float64
 28  post_ipo_debt         38481 non-null  float64
 29  secondary_market      38481 non-null  float64
 30  product_crowdfunding  38481 non-null  float64
 31  round_A               38481 non-null  float64
 32  round_B               38481 non-null  float64
 33  round_C               38481 non-null  float64
 34  round_D               38481 non-null  float64
 35  round_E               38481 non-null  float64
 36  round_F               38481 non-null  float64
 37  round_G               38481 non-null  float64
 38  round_H               38481 non-null  float64
dtypes: datetime64[ns](5), float64(23), int64(1), object(10)
memory usage: 11.7+ MB
```

```
# Select the columns with dtype 'datetime64[ns]'
datetime_columns = df.select_dtypes(include=['datetime64[ns]']).columns

# Check for NaT values in the datetime columns
# Create a boolean mask where NaT exists
nat_mask = df[datetime_columns].isna().any(axis=1)

# Filter the DataFrame to show only rows with NaT values
rows_with_nat = df[nat_mask]
rows_with_nat
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code	region	...
1492	/organization/agflow	AgFlow	http://www.agflow.com	|Software|	Software	0.0	operating	CHE	Unknown	Geneva	...
6661	/organization/buru-buru	Buru Buru	http://www.buru-buru.com	|Startups|Internet|Retail|Design|Art|E-Commerce|	Startups	0.0	operating	ITA	Unknown	Firenze	...
14524	/organization/exploco	Exploco	http://www.exploco.com	|Adventure Travel|	Adventure Travel	0.0	operating	AUS	Unknown	Perth	...
29695	/organization/nubank	Nubank	https://www.nubank.com.br/	|Consumer Internet|Financial Services|	Financial Services	16300000.0	operating	BRA	Unknown	Sao Paulo	...
31865	/organization/peoplegoal	PeopleGoal	http://www.peoplegoal.com	|Enterprise Software|	Enterprise Software	0.0	operating	Unknown	Unknown	Unknown	...
37313	/organization/securenet-payment-systems	SecureNet Payment Systems	http://www.securenet.com	|Trading|Mobile Payments|Payments|E-Commerce|	Payments	18000000.0	acquired	USA	TX	Austin	...

6 rows × 39 columns

```
df.dropna(inplace = True)
```

```
df.isna().sum()
```

	0
permalink	0
name	0
homepage_url	0
category_list	0
market	0
funding_total_usd	0
status	0
country_code	0
state_code	0
region	0
city	0
funding_rounds	0
founded_at	0
founded_month	0
founded_quarter	0
founded_year	0
first_funding_at	0
last_funding_at	0
seed	0
venture	0
equity_crowdfunding	0
undisclosed	0
convertible_note	0
debt_financing	0
angel	0
grant	0
private_equity	0
post_ipo_equity	0
post_ipo_debt	0
secondary_market	0

product_crowdfunding	0
round_A	0
round_B	0
round_C	0
round_D	0
round_E	0
round_F	0
round_G	0
round_H	0

dtype: int64

Cleaned data

```
df.shape
```

```
(38475, 39)
```

Saving the cleaned data

```
#dff = pd.read_csv('/content/cleaned_startup_funding_data.csv')
```

```
df.head()
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code	region
0	/organization/waywire	#waywire	http://www.waywire.com	|Entertainment|Politics|Social Media|News|	News	1750000.0	acquired	USA	NY	New York City
2	/organization/rock-your-paper	'Rock' Your Paper	http://www.rockyourpaper.org	|Publishing|Education|	Publishing	40000.0	operating	EST	Unknown	Tallinn
3	/organization/in-touch-network	(In)Touch Network	http://www.InTouchNetwork.com	|Electronics|Guides|Coffee|Restaurants|Music|i...	Electronics	1500000.0	operating	GBR	Unknown	London
4	/organization/r-ranch-and-mine	-R- Ranch and Mine	Unknown	|Tourism|Entertainment|Games|	Tourism	60000.0	operating	USA	TX	Dallas
5	/organization/club-domains	.Club Domains	http://nic.club/	|Software|	Software	7000000.0	Unknown	USA	FL	Ft. Lauderdale

5 rows × 39 columns

Funding Overview

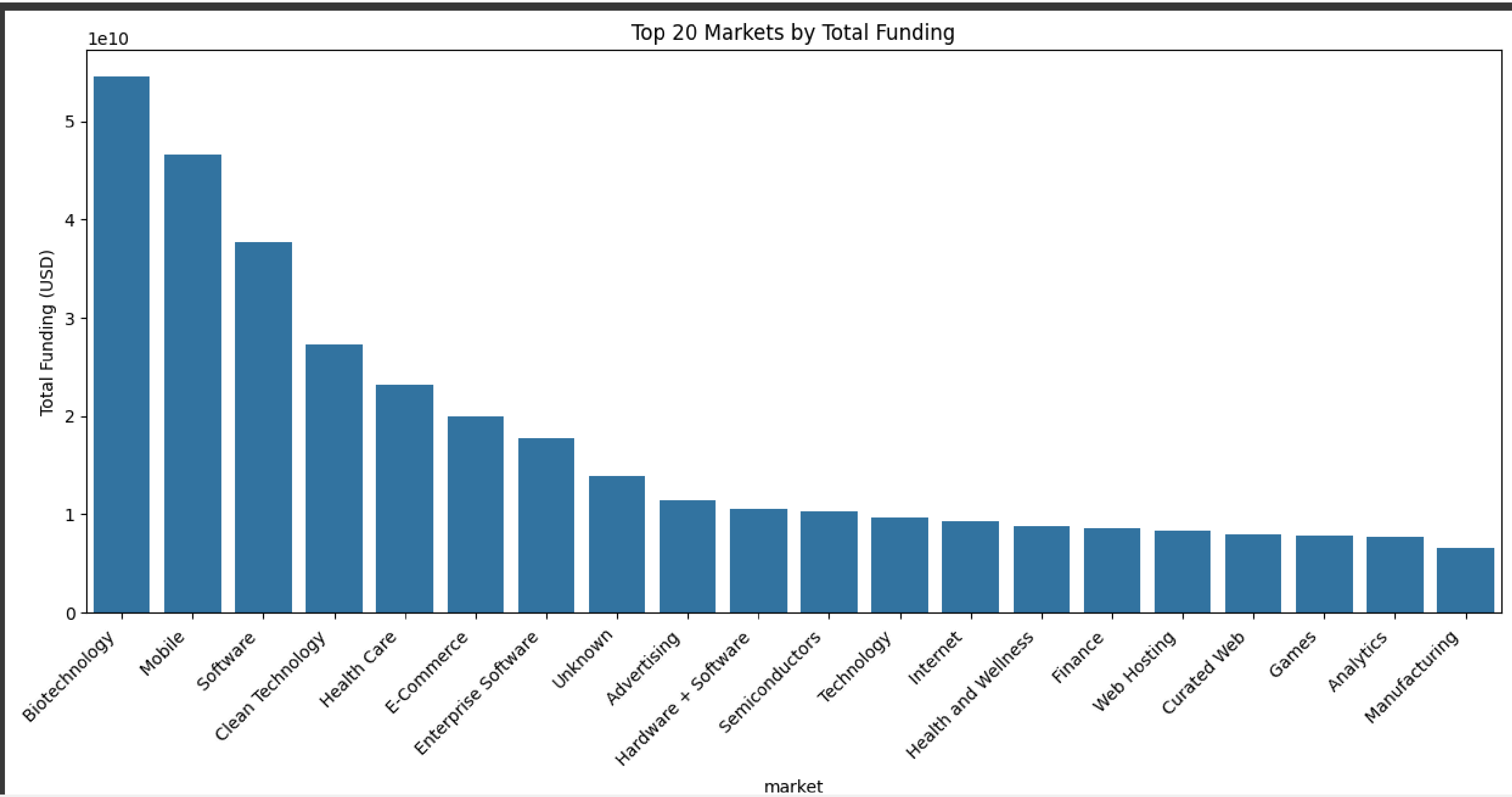
```
print(f"Total number of startups: {len(df)}")
print(f"Total funding: ${df['funding_total_usd'].sum():,.0f}")
print(f"Average funding per startup: ${df['funding_total_usd'].mean():,.0f}")
print(f"Median funding per startup: ${df['funding_total_usd'].median():,.0f}")
```

```
Total number of startups: 38475
Total funding: $534,119,397,445
Average funding per startup: $13,882,246
Median funding per startup: $1,000,000
```

Distribution across markets

```
market_funding = df.groupby('market')['funding_total_usd'].agg(['sum', 'mean', 'count']).sort_values('sum', ascending=False).head(20)

plt.figure(figsize=(15, 6))
sns.barplot(x=market_funding.index, y=market_funding['sum'])
plt.title('Top 20 Markets by Total Funding')
plt.xticks(rotation=45, ha='right')
plt.ylabel('Total Funding (USD)')
plt.show()
```



Biotechnology takes the top spot with approximately $50 billion USD in funding, with the Mobile and Software industries coming in next.

Sectors like Clean Technology, Health Care, and E-commerce also secure notable investments.

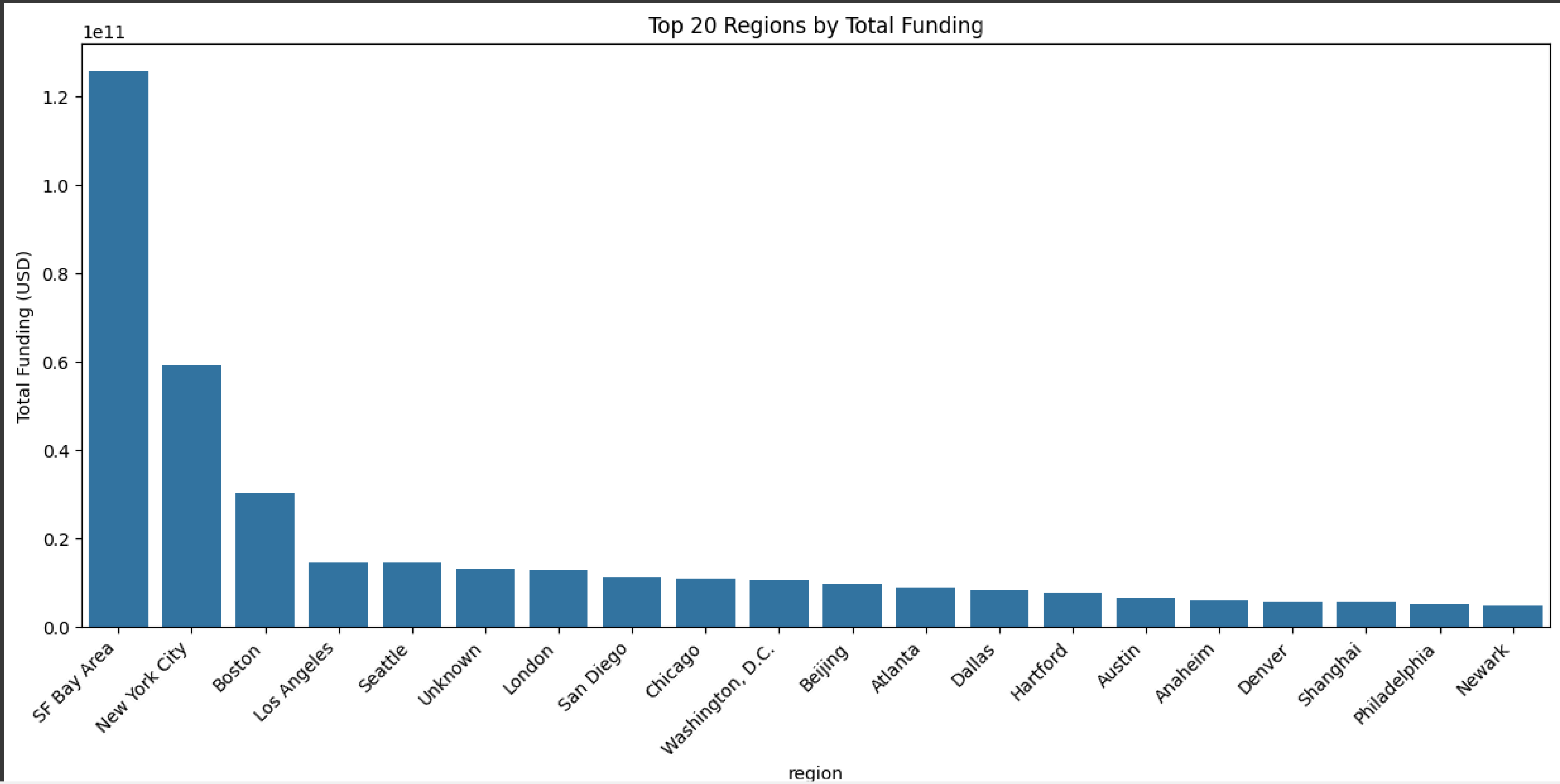
On the other hand, industries such as Analytics, Manufacturing, and Games rank lower in terms of funding among the top 20 markets.

Distribution across regions

```
region_funding = df.groupby('region')['funding_total_usd'].agg(['sum', 'mean', 'count']).sort_values('sum', ascending=False).head(20)

plt.figure(figsize=(15, 6))
```

```
sns.barplot(x=region_funding.index, y=region_funding['sum'])
plt.title('Top 20 Regions by Total Funding')
plt.xticks(rotation=45, ha='right')
plt.ylabel('Total Funding (USD)')
plt.show()
```



SF Bay Area Leading the Way: The San Francisco Bay Area holds a dominant position with more than $120 billion in funding, reaffirming its status as a global center for technology and startups.

New York City's Notable Standing: NYC follows closely, securing over $60 billion in funding, showcasing its influence in both the finance and expanding tech industries.

Global Cities Highlighted: While major US cities like Boston, Los Angeles, and Seattle rank prominently, international centers such as London, Beijing, and Shanghai also feature, highlighting the worldwide reach of startup ecosystems.

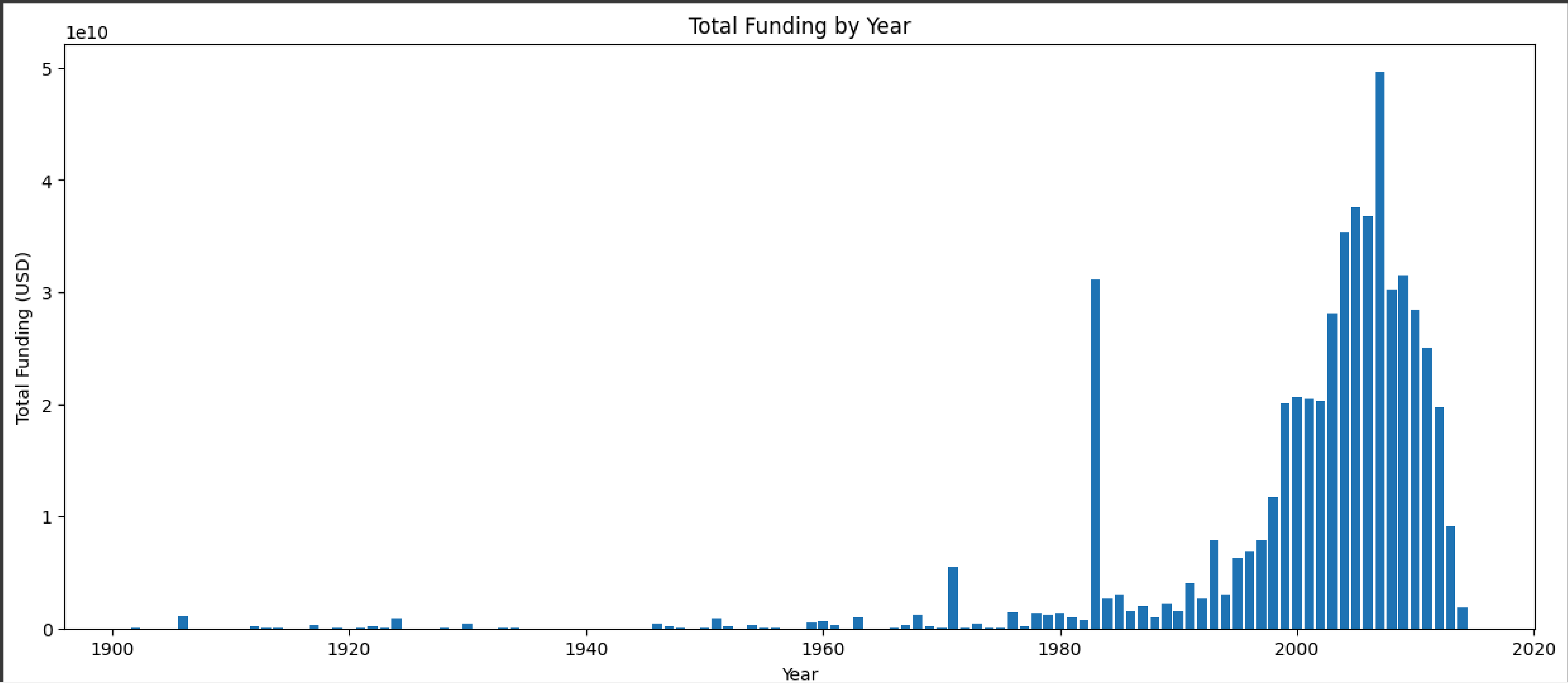
Funding Disparities: There is a sharp decline in funding beyond the top regions, indicating a strong concentration in just a few key locations.

Yearly Funding

```
df['founded_year'] = pd.to_datetime(df['founded_at']).dt.year

yearly_funding = df.groupby('founded_year')['funding_total_usd'].sum().reset_index()

plt.figure(figsize=(15, 6))
plt.bar(yearly_funding['founded_year'], yearly_funding['funding_total_usd'])
plt.title('Total Funding by Year')
plt.xlabel('Year')
plt.ylabel('Total Funding (USD)')
plt.show()
```



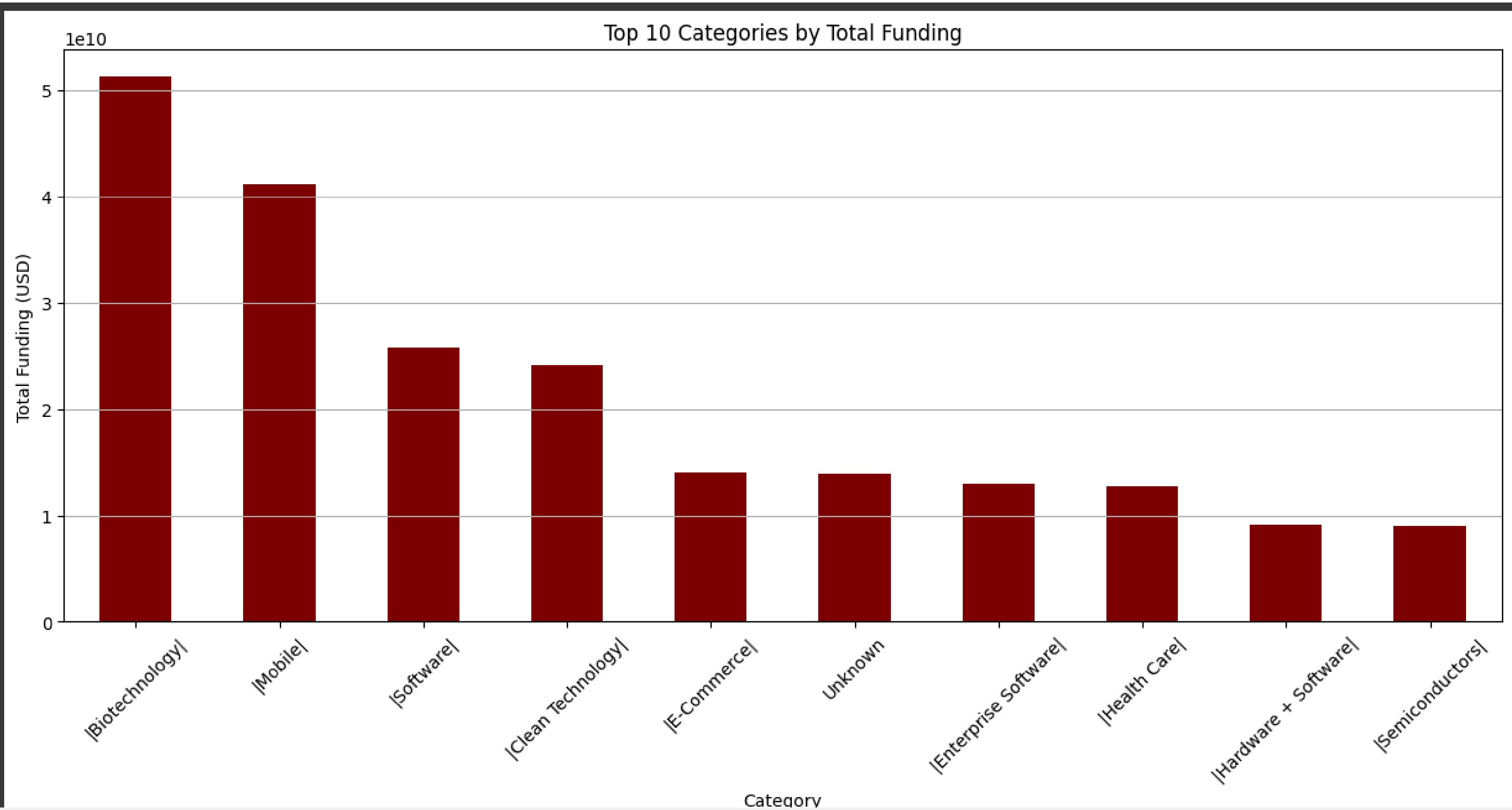
Limited funding activity before 1980: The amount of funding was minimal, suggesting that venture capital or structured startup funding was not widely practiced.

Notable surge in funding from the late 1990s to early 2000s: This corresponds with the dot-com boom, during which numerous tech companies secured substantial investments.

Peak during the early 2000s: Funding reached its highest point during this time, likely driven by significant investments in technology and innovation.

Analyzing funding distribution across different categories

```
# Analyzing funding distribution across different categories
category_funding = df.groupby('category_list')['funding_total_usd'].sum().sort_values(ascending=False)
plt.figure(figsize=(15, 6))
category_funding.head(10).plot(kind='bar', color='maroon')
plt.title('Top 10 Categories by Total Funding')
plt.xlabel('Category')
plt.ylabel('Total Funding (USD)')
plt.xticks(rotation=45)
plt.grid(axis='y')
plt.show()
```



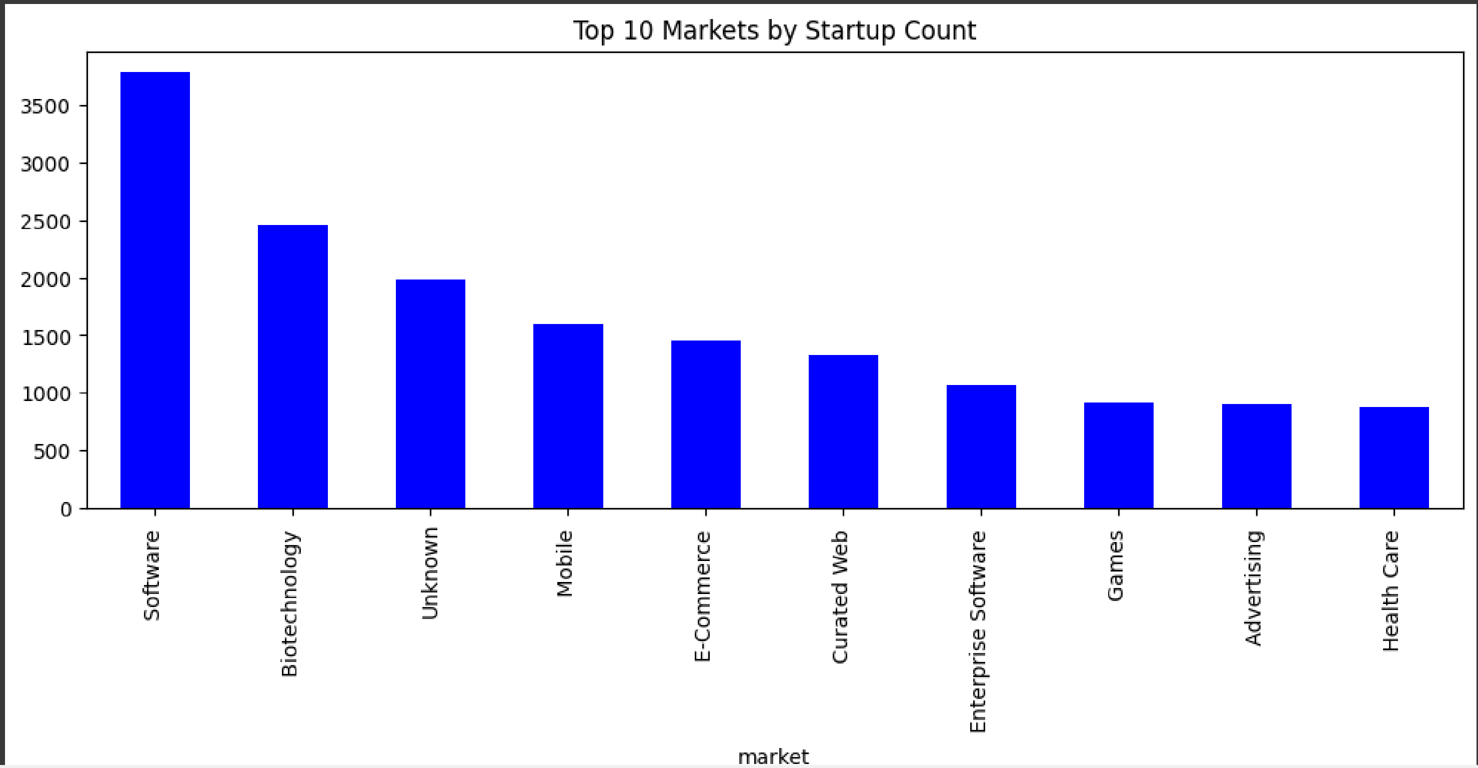
Biotechnology takes the lead with more than 50 billion dollars in funding, followed by Mobile at approximately $45 billion.

Software and Clean Technology fall in the mid-range, each securing around 25 billion dollars in funding. E-Commerce and the "Unknown" category receive about 15 billion dollars each.

Enterprise Software, Health Care, and Hardware + Software attract close to 10 billion dollars, while Semiconductors have the lowest funding, coming in below 10 billion dollars.

Bar chart for funding by market

```
plt.figure(figsize=(12, 4))
df['market'].value_counts().head(10).plot(kind='bar', color='blue')
plt.title('Top 10 Markets by Startup Count')
plt.show()
```



Software leads with the largest number of startups, exceeding 3,500. Biotechnology is in second place, followed by an Unknown category.

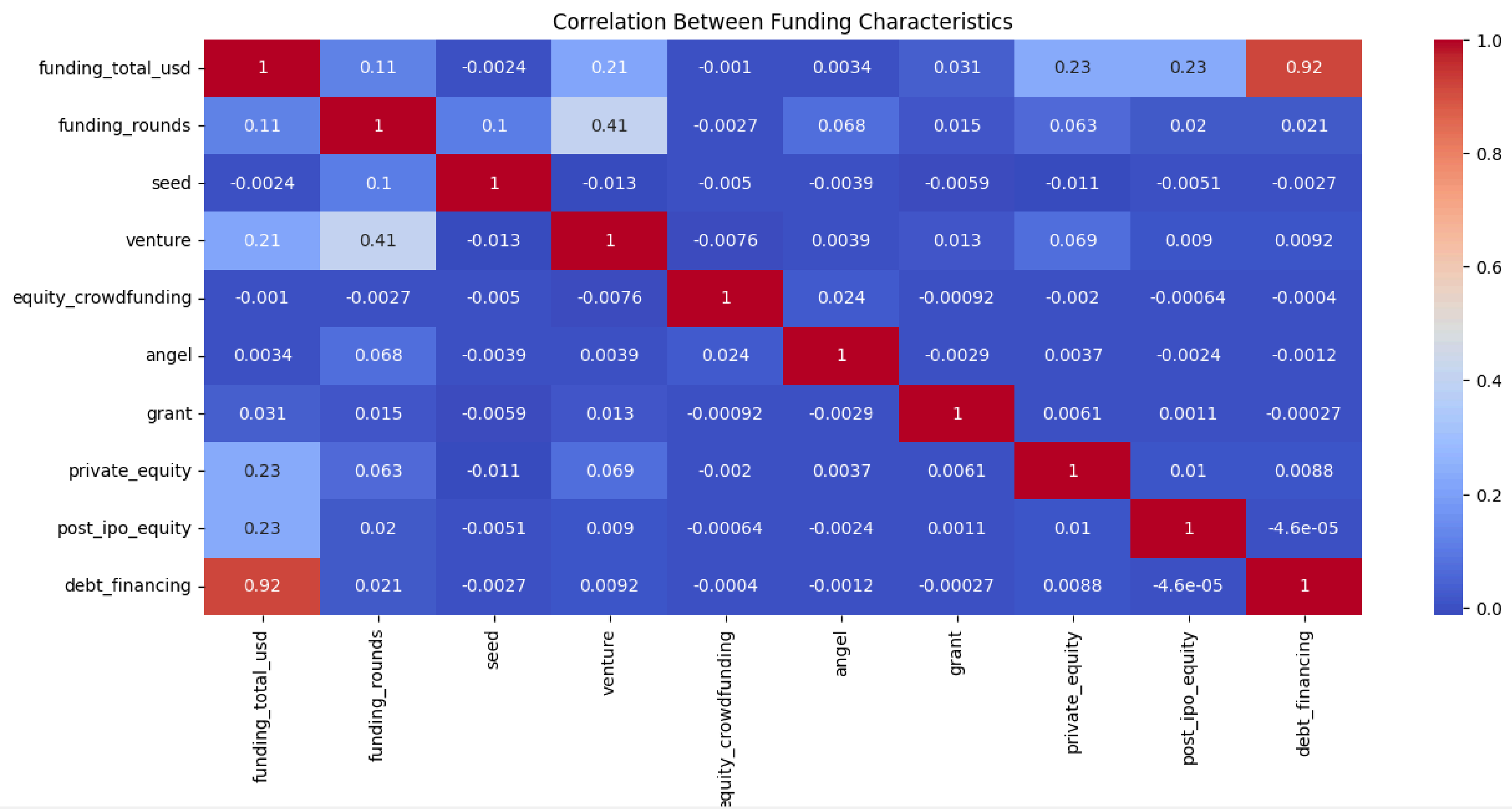
Mobile and E-Commerce have a moderate number of startups.

Curated Web, Enterprise Software, Games, and Advertising show a steady presence, while Health Care has the fewest startups among the top 10 sectors.

Creating a correlation matrix

```
correlation_cols = ['funding_total_usd', 'funding_rounds', 'seed', 'venture', 'equity_crowdfunding',
                    'angel', 'grant', 'private_equity', 'post_ipo_equity', 'debt_financing']
corr_matrix = df[correlation_cols].corr()
```

```
plt.figure(figsize=(15, 6))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Between Funding Characteristics')
plt.show()
```




```
import statsmodels.api as sm
# Step 1: Define dependent and independent variables
X = df['debt_financing']  # Independent variable (Debt Financing)
y = df['funding_total_usd']  # Dependent variable (Total Funding)

# Step 2: Add a constant to the independent variable (for the intercept)
X = sm.add_constant(X)
```

```
# Step 3: Fit the regression model
model = sm.OLS(y, X).fit()

# Step 4: Print the summary of the regression analysis
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      funding_total_usd   R-squared:                       0.847
Model:                            OLS   Adj. R-squared:                  0.847
Method:                 Least Squares   F-statistic:                 2.134e+05
Date:                Wed, 09 Oct 2024   Prob (F-statistic):               0.00
Time:                        12:00:07   Log-Likelihood:            -7.4743e+05
No. Observations:               38475   AIC:                         1.495e+06
Df Residuals:                   38473   BIC:                         1.495e+06
Df Model:                           1
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const            1.19e+07   3.37e+05     35.298      0.000    1.12e+07    1.26e+07
debt_financing     1.0037      0.002    461.979      0.000       0.999       1.008
==============================================================================
Omnibus:                   113282.423   Durbin-Watson:                   1.991
Prob(Omnibus):                  0.000   Jarque-Bera (JB):      12633204989.011
Skew:                          41.545   Prob(JB):                         0.00
Kurtosis:                    2808.968   Cond. No.                     1.55e+08
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.55e+08. This might indicate that there are
strong multicollinearity or other numerical problems.
```

R-squared (0.847):

An R-squared value of 0.847 indicates that 84.7% of the variability in total funding (USD) can be attributed to Debt Financing alone.

This is considered a very high value, suggesting a strong linear correlation between debt financing and total funding. It implies that companies with greater total funding tend to heavily depend on debt financing.

P-value (F-statistic = 0.00):

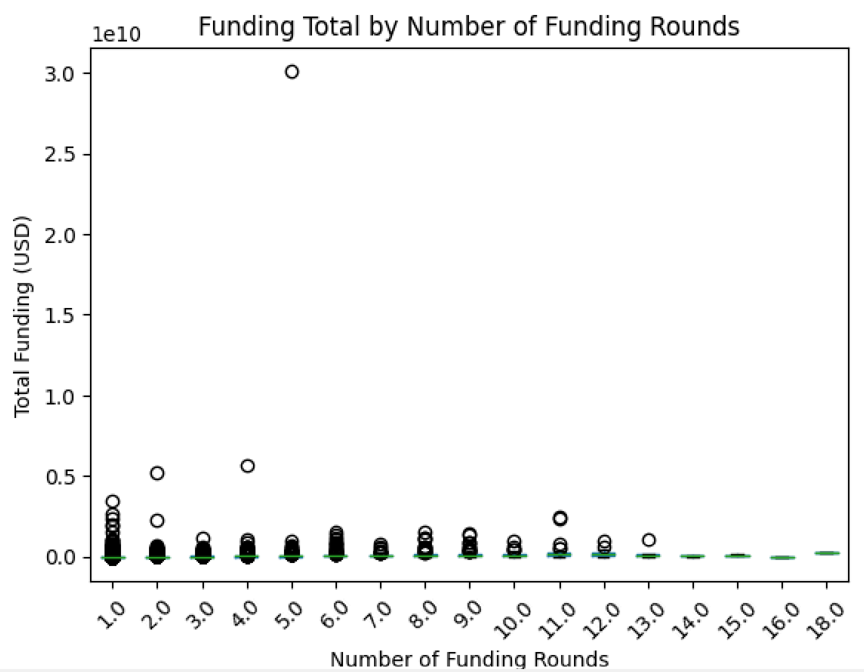
The p-value associated with the F-statistic is 0.00, which is below the 0.05 threshold, signifying that the relationship is statistically significant.

This indicates that the connection between Debt Financing and Total Funding is unlikely to have occurred by random chance.

Analyze funding success based on funding rounds

```
plt.figure(figsize=(20, 2))
df.boxplot(column='funding_total_usd', by='funding_rounds', grid=False)
plt.title('Funding Total by Number of Funding Rounds')
plt.suptitle('')
plt.xlabel('Number of Funding Rounds')
plt.ylabel('Total Funding (USD)')
plt.xticks(rotation=45)
plt.show()
```

<Figure size 2000x200 with 0 Axes>



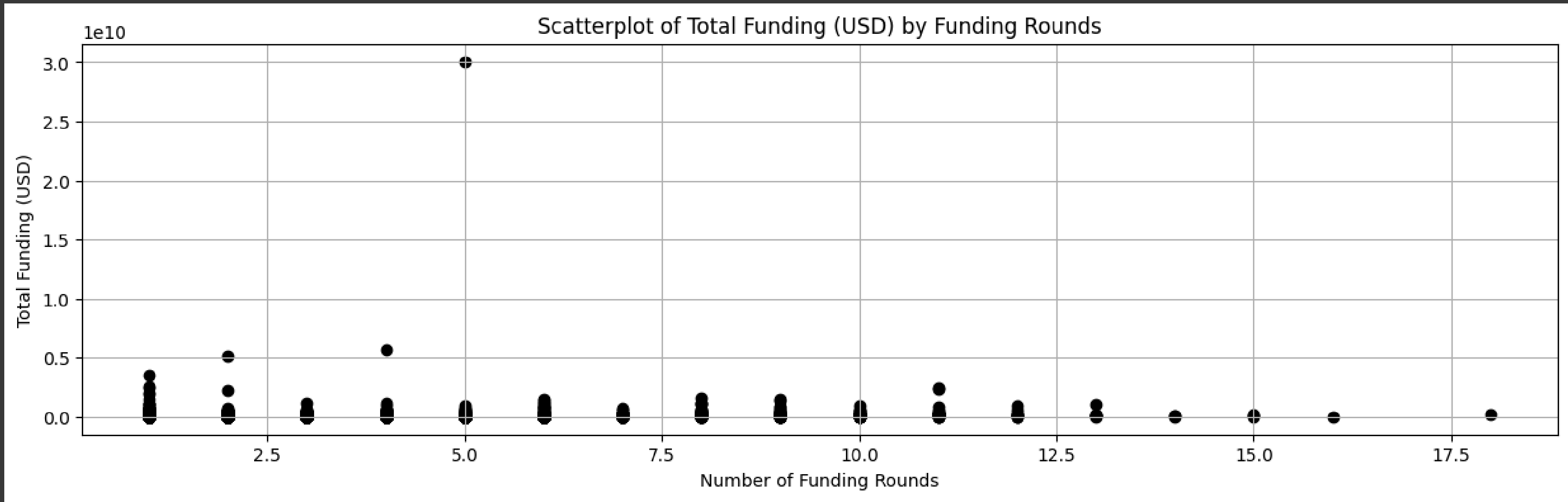
Companies typically obtain moderate levels of funding through several rounds; however, a select few manage to secure exceptionally high amounts early in their funding journey.

Analyzing the outliers—particularly those with fewer funding rounds yet remarkably higher funding—can yield valuable insights into the factors that led to their significant success.

Scatterplot for 'funding_total_usd' vs 'funding_rounds'

```
plt.figure(figsize=(15, 4))
plt.scatter(df['funding_rounds'], df['funding_total_usd'], color='black')
plt.title('Scatterplot of Total Funding (USD) by Funding Rounds')
```

```
plt.xlabel('Number of Funding Rounds')
plt.ylabel('Total Funding (USD)')
plt.grid(True)
plt.show()
```



The scatterplot illustrates that the majority of companies secure funding in a limited number of rounds (ranging from 1 to 7), with total funding typically below $1 billion.

However, there are a few notable outliers that have raised significantly larger amounts, including one company that approached $30 billion over 5 funding rounds.

There is no evident linear correlation between the number of funding rounds and total funding, as the data exhibits considerable variability.

Boxplot for 'funding_total_usd'

```
plt.figure(figsize=(10, 4))
plt.boxplot(df['funding_total_usd'], vert=False, patch_artist=True, boxprops=dict(facecolor="skyblue"))
plt.title('Boxplot of Total Funding (USD)')
plt.xlabel('Total Funding (USD)')
plt.show()
```