Funding In Startups

Problem Statement

The project aims to explore the startup funding landscape by analyzing historical data on various startups, their funding rounds, and funding types across different regions and sectors. The objective is to uncover trends and insights that can guide strategic decision-making for entrepreneurs and investors.

Data Description

permalink - Static hyperlink for the startup on Crunchbase.

name - Name of the startup.

homepage_url - Website address of the startup.

category_list - Categories the startup belongs to.[

market - The market the startup caters to.

funding_total_usd - Total funding received (in USD).

status - Current operating status of the startup (e.g., operating, acquired).

country_code - Country of origin.

state_code - State of origin (if applicable).

region - Region where the startup operates.

city - City of origin.

funding_rounds - Total number of funding rounds the startup has received.

founded_at - Date the startup was founded.

founded_month - Month when the startup was founded.

founded_quarter - Quarter when the startup was founded.

founded_year - Year when the startup was founded.

first_funding_at Date of the first funding round.

last_funding_at - Date of the last funding round.

seed - Seed funding received (in USD).

venture - Venture funding received (in USD).

equity_crowdfunding - Funding received by diluting equity through crowdfunding.

undisclosed - Other undisclosed funding sources.

convertible_note - Funding received from convertible notes.

debt_financing - Funding received through debt financing.

angel - Funding received from angel investors.

grant - Funding received from grants.

private_equity - Funding received from private equity firms.

post_ipo_equity - Equity-based funding received after IPO.

post_ipo_debt - Debt financing received after IPO.

secondary_market - Funding received from secondary market transactions.

product_crowdfunding - Funding received from product-based crowdfunding.

round_A - Funding received in round A.

round_B - Funding received in round B.

round_C - Funding received in round C.

round_D - Funding received in round D.

round_E - Funding received in round E.

round_F - Funding received in round F.

Libraries 📖

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

Loading Data 📀

In [ ]:
```
data = pd.read_csv('/content/drive/MyDrive/Datasets/investments_VC.csv', encoding = "latin1")
```

In [ ]:
```
df = data.copy()
df.head()
```

Out[ ]:

	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	17,50,000	acquired	USA	NY	New York City
1	/organization/tv-communications	&TV Communications	http://enjoyandtv.com	|Games|	Games	40,00,000	operating	USA	CA	Los Angeles
2	/organization/rock-your-paper	'Rock' Your Paper	http://www.rockyourpaper.org	|Publishing|Education|	Publishing	40,000	operating	EST	NaN	Tallinn
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	London
4	/organization/r-ranch-and-mine	-R- Ranch and Mine	NaN	|Tourism|Entertainment|Games|	Tourism	60,000	operating	USA	TX	Dallas

5 rows × 39 columns

Understanding The Data ✨

In [ ]:
```
pd.set_option('display.max_columns', 50)
```

In [ ]:
```
df.sample(1)
```

Out[ ]:

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code	region	city	funding_rounds	fo
17658	/organization/gourmetzoom	GourmetZoom	http://www.GourmetZoom.com	|Internet|	Internet	-	operating	USA	NY	New York City	New York	1.0	2

In [ ]:
```
df.shape
```

Out[ ]:
```
(54294, 39)
```

In [ ]:
```
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
```

In [ ]:

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

Out[ ]:

	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

In [ ]:

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

Out[ ]:

	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

In [ ]:

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

Out[ ]:

	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

dtype: int64

In [ ]:

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

Out[ ]:

	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

Column Names

In [ ]:

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

Out[ ]:

```
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')
```

Handling Null Values ⌨️

In [ ]:

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

Out[ ]:

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code
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
2	/organization/rock-your-paper	'Rock' Your Paper	http://www.rockyourpaper.org	|Publishing|Education|	Publishing	40,000	operating	EST	NaN
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
4	/organization/r-ranch-and-mine	-R- Ranch and Mine	NaN	|Tourism|Entertainment|Games|	Tourism	60,000	operating	USA	TX
...	...	...	...	...	...	...	...	...	...
49433	/organization/zzish	Zzish	http://www.zzish.com	|Analytics|Gamification|Developer APIs|iOS|And...	Education	3,20,000	operating	GBR	NaN
49434	/organization/zznode-science-and-technology-co...	ZZNode Science and Technology	http://www.zznode.com	|Enterprise Software|	Enterprise Software	15,87,301	operating	CHN	NaN
49435	/organization/zzzzapp-com	Zzzzapp Wireless ltd.	http://www.zzzzapp.com	|Web Development|Advertising|Wireless|Mobile|	Web Development	97,398	operating	HRV	NaN
49436	/organization/a-list-games	[a]list games	http://www.alistgames.com	|Games|	Games	93,00,000	operating	NaN	NaN
49437	/organization/x	[x+1]	http://www.xplusone.com/	|Enterprise Software|	Enterprise Software	4,50,00,000	operating	USA	NY

49438 rows × 39 columns

In [ ]:
```
df.isna().all(axis=1).sum()
```

Out[ ]:
```
0
```

In [ ]:
```
df.isna().sum()
```

Out[ ]:

	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

dtype: int64

Permalink - handling duplicates

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

Out[ ]:

	count
permalink	
/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

dtype: int64

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

Out[ ]:

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code	region	city	funding_rounds	founded_at	foun
44033	/organization/treasure-valley-urology-services	Treasure Valley Urology Services	NaN	|Biotechnology|	Biotechnology	3,32,194	operating	USA	TX	Austin	Austin	4.0	2004-01-01	
44034	/organization/treasure-valley-urology-services	Treasure Valley Urology Services	NaN	NaN	NaN	3,32,194	operating	USA	TX	Austin	Austin	1.0	2004-01-01	

In [ ]:
```
df = df.drop(44034)
```

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

Out[ ]:

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code	region	city	funding_rounds	founded_at	foun
44033	/organization/treasure-valley-urology-services	Treasure Valley Urology Services	NaN	|Biotechnology|	Biotechnology	3,32,194	operating	USA	TX	Austin	Austin	4.0	2004-01-01	

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

Out[ ]:

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code	region	city	funding_rounds	founded_at	found
33939	/organization/prysm	Prysm	http://www.prysm.com/	NaN	NaN	29,30,80,123	operating	NaN	NaN	NaN	NaN	1.0	NaN	
33940	/organization/prysm	Prysm	http://www.prysm.com	|Displays|Hardware + Software|	Displays	29,30,80,123	operating	USA	CA	SF Bay Area	San Jose	3.0	2005-01-01	

In [ ]:
```
df = df.drop(33939)
```

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

Out[ ]:

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code	region	city	funding_rounds	founded_at	founde
33940	/organization/prysm	Prysm	http://www.prysm.com	|Displays|Hardware + Software|	Displays	29,30,80,123	operating	USA	CA	SF Bay Area	San Jose	3.0	2005-01-01	

Observation: Duplicate rows are removed

Name column - Handling Null

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

Out[ ]:

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code	region	city	funding_rounds	founded_at	founded_month	fo
28221	/organization/tell-it-in	NaN	http://tellitin10.com	|Startups|	Startups	25,000	closed	NaN	NaN	NaN	NaN	1.0	2011-10-01	2011-10	

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

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

Out[ ]:
```
0
```

Observation: Null value is replaced in name column

In [ ]:
```
df.columns
```

Out[ ]:
```
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')
```

homepage_url column

In [ ]:
```
#filling missing URLs with "Unknown"
df['homepage_url'].fillna('Unknown', inplace=True)
```

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

Out[ ]:
```
0
```

category_list column

In [ ]:
```
#filling missing URLs with "Unknown"
df['category_list'].fillna('Unknown', inplace=True)
```

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

Out[ ]:
```
0
```

market column

In [ ]:
```
#filling missing URLs with "Unknown"
df['market'].fillna('Unknown', inplace=True)
```

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

Out[ ]:
```
0
```

Funding_total_usd column

In [ ]:
```
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
```

Out[ ]:
```
dtype('float64')
```

In [ ]:
```
df.sample(1)
```

Out[ ]:

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code	region	city	funding_rounds	founded_
3027	/organization/aquaback-technologies	Aquaback Technologies	http://aquaback.com	|Clean Technology|	Clean Technology	735000.0	operating	USA	MA	Boston	Tewksbury	1.0	Na

Observation: Removed "-" and replaced with 0

In [ ]:
```
df['funding_total_usd'].isna().sum()
```

Out[ ]:
```
0
```

status column

In [ ]:
```
#filling missing URLs with "Unknown"
df['status'].fillna('Unknown', inplace=True)
df['status'].isna().sum()
```

Out[ ]:
```
0
```

In [ ]:
```
df.isna().sum()
```

Out[ ]:

	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

dtype: int64

country_code, state_code, region, city columns

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

founded_at, founded_month, founded_quarter, founded_year columns

In [ ]:
```
#drop rows where these values are null
df.dropna(subset=['founded_at', 'founded_month', 'founded_quarter', 'founded_year'], inplace=True)
```

In [ ]:
```
df.isna().sum()
```

Out[ ]:

	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

Observation : Since These columns are date-related, and missing dates could affect the analysis. We will drop rows where these values are null, as imputation might affect the accuracy of time-based analysis.

In [ ]:

```
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
```

In [ ]:
```
df.sample(2)
```

Out[ ]:

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code	region	city	funding_rounds	founded_a
29481	/organization/north-shore-innoventures	North Shore InnoVentures	http://www.nsiv.org	|Biotechnology|	Biotechnology	311500.0	operating	USA	MA	Boston	Beverly	2.0	2008-01-0
532	/organization/abbey-house-media	Abbey House Media	Unknown	Unknown	Unknown	600000.0	operating	USA	TX	Austin	Austin	1.0	2006-01-0

Handling Data types of the columns🛠️

In [ ]:
```
## 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')
```

In [ ]:
```
# Converting founded_year to int
df['founded_year'] = df['founded_year'].astype(int)
```

In [ ]:
```
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
```

In [ ]:

```python
# 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
```

Out[ ]:

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code	region	cit
1492	/organization/agflow	AgFlow	http://www.agflow.com	|Software|	Software	0.0	operating	CHE	Unknown	Geneva	Genev
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	Firenz
14524	/organization/exploco	Exploco	http://www.exploco.com	|Adventure Travel|	Adventure Travel	0.0	operating	AUS	Unknown	Perth	Pert
29695	/organization/nubank	Nubank	https://www.nubank.com.br/	|Consumer Internet|Financial Services|	Financial Services	16300000.0	operating	BRA	Unknown	Sao Paulo	Sã Paul
31865	/organization/peoplegoal	PeopleGoal	http://www.peoplegoal.com	|Enterprise Software|	Enterprise Software	0.0	operating	Unknown	Unknown	Unknown	Unknow
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	Austi

In [ ]:

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

In [ ]:

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

Out[ ]:

	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

The data is cleaned and the data types of the columns are checked.

In [ ]:
```
df.shape
```

Out[ ]:
```
(38475, 39)
```

Saving the cleaned dataset 📷

In [ ]:
```
#dff = pd.read_csv('/content/cleaned_startup_funding_data.csv')
```

In [ ]:
```
df.head()
```

Out[ ]:

	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

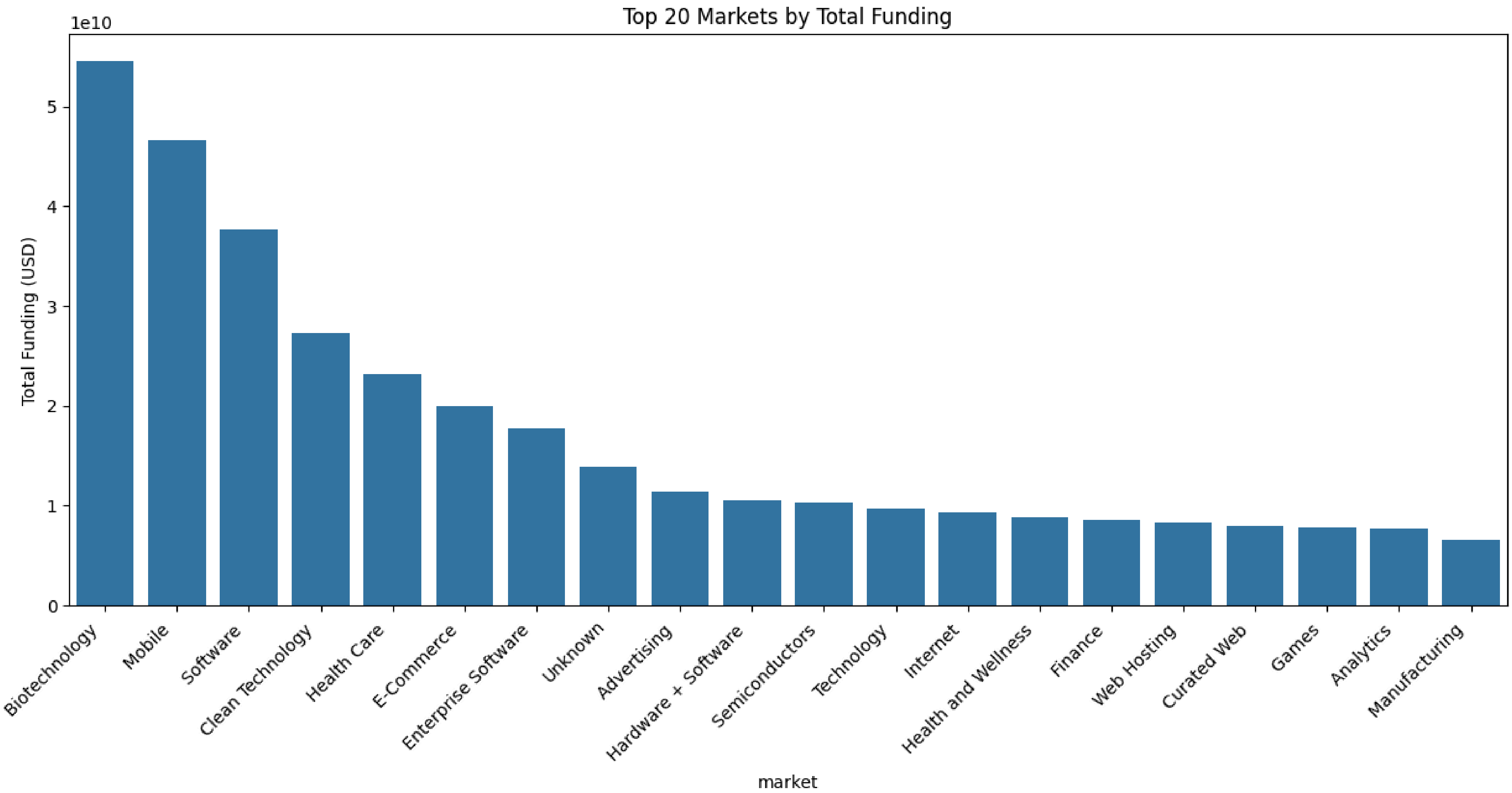
Overview of Funding 🦉 💸

```
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()
```



Observation:

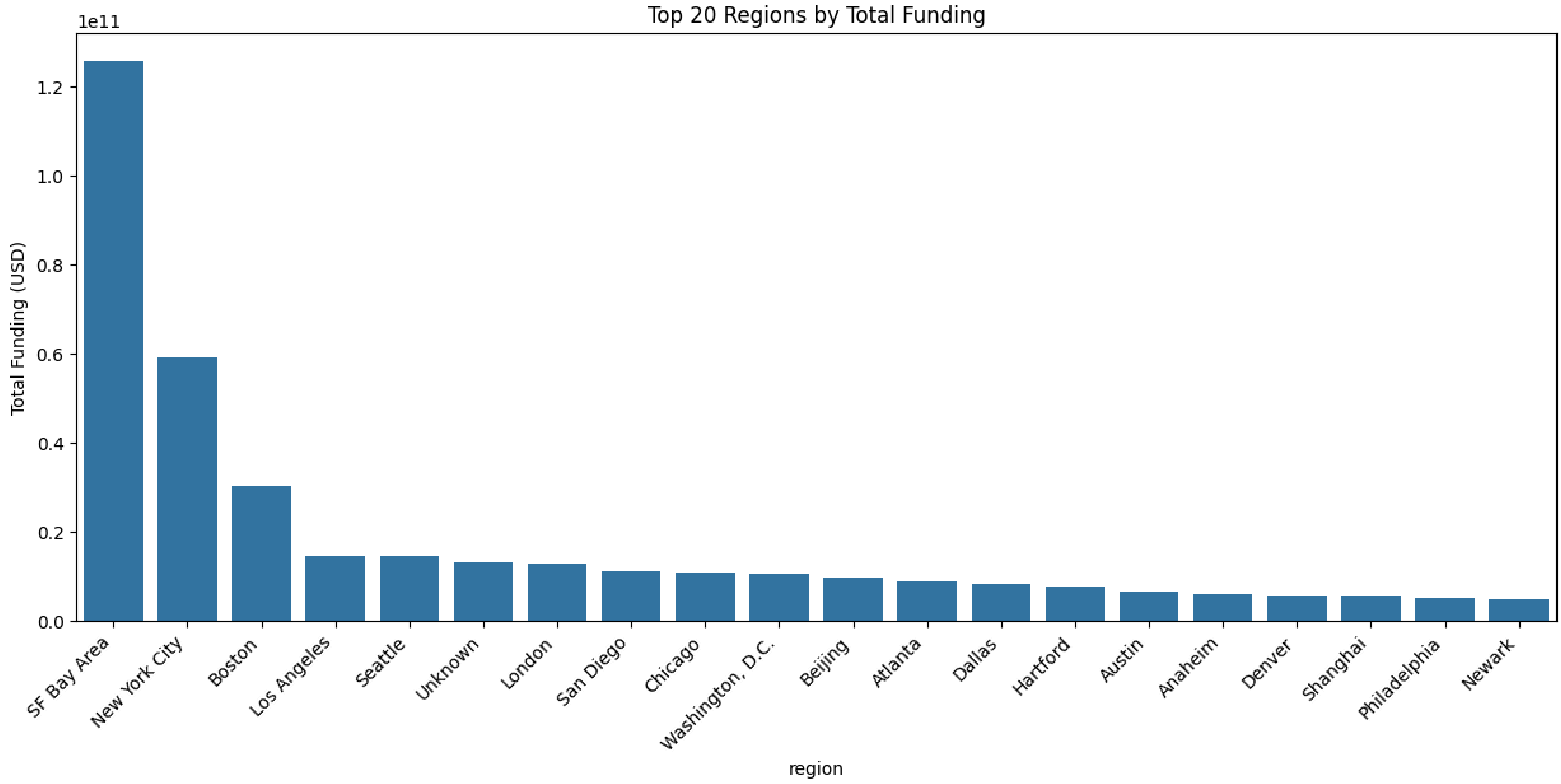
1. Biotechnology tops the list with nearly $50 billion USD, followed by Mobile and Software markets.

2. Industries like Clean Technology, Health Care, and E-commerce also receive substantial funding.

3. Analytics, Manufacturing, and Games are among the lower-funded markets within the top 20

```
# 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()
```



Observation

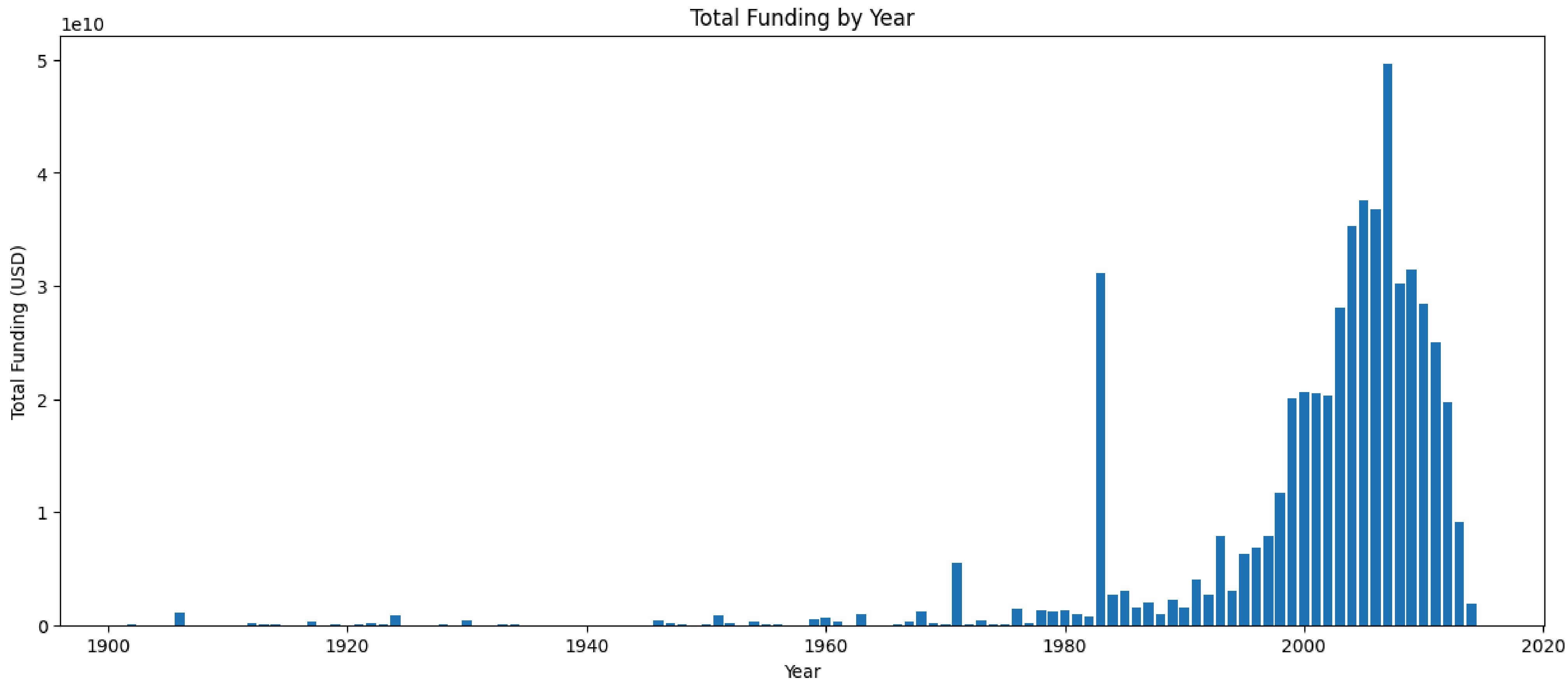
- SF Bay Area Dominance: The San Francisco Bay Area leads significantly in total funding, surpassing $120 billion, underscoring its position as a global tech and startup hub.
- New York City's Strong Presence: NYC follows with over $60 billion in funding, emphasizing its role in finance and growing tech sectors.
- Global Cities: Major US cities like Boston, Los Angeles, and Seattle rank high, but international hubs like London, Beijing, and Shanghai also appear, reflecting the global nature of startup ecosystems.
- Funding Gaps: There's a steep drop in funding beyond the top regions, highlighting concentration in a few key areas.

In [ ]:

```
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()
```



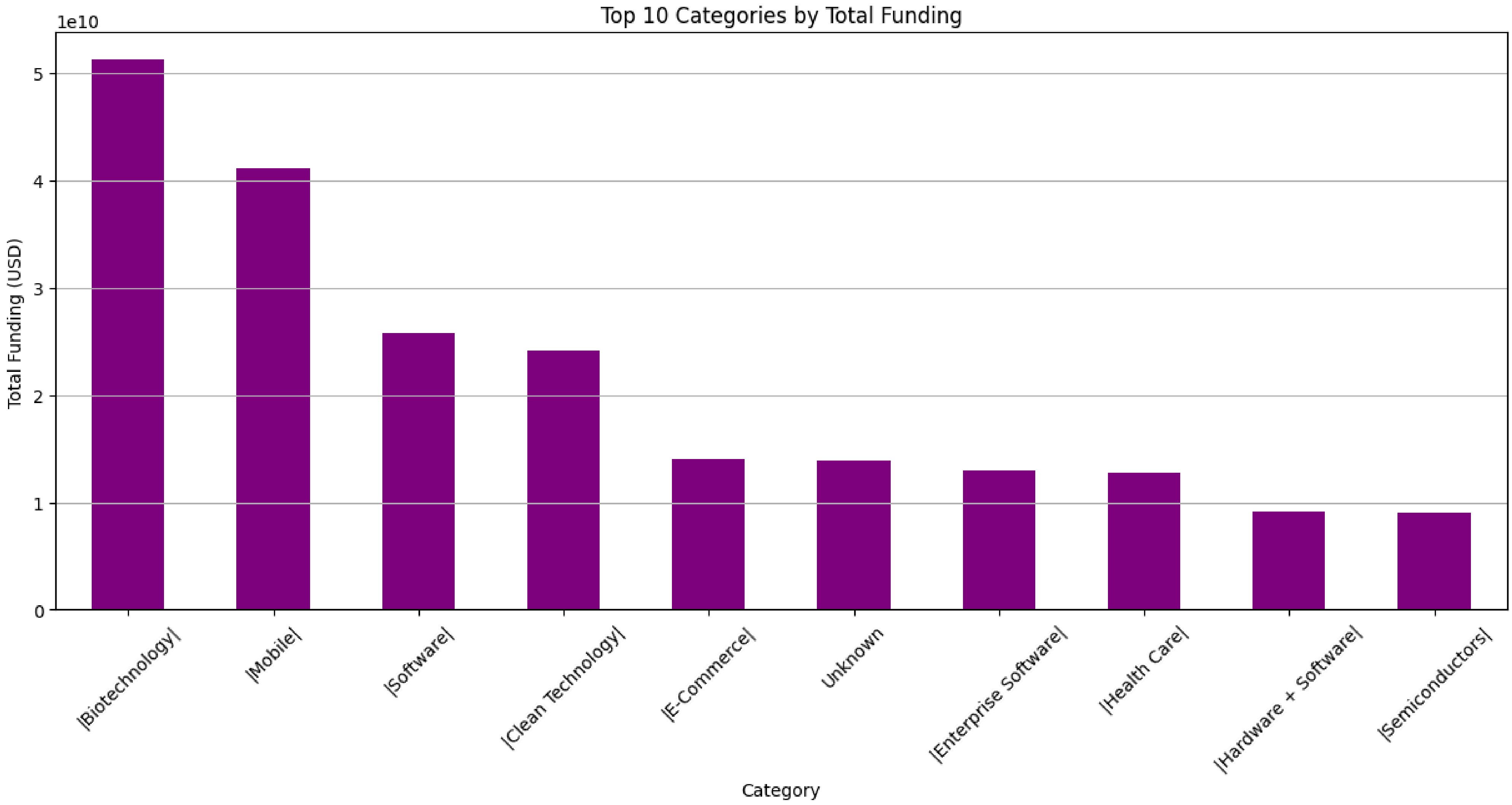
Observation

- Minimal funding activity before 1980: Funding amounts are almost negligible, indicating that venture capital or formalized funding for startups wasn't common.
- Significant spike in funding in the late 1990s to early 2000s: This aligns with the dot-com boom, where many tech companies attracted large investments.
- Peak in early 2000s: Funding hit its highest during this period, possibly reflecting large investments in technology and innovation.

In [ ]:

```
# 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='purple')
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()
```



Observation

- Biotechnology leads with over $50 billion in funding.

Mobile follows at around $45 billion.

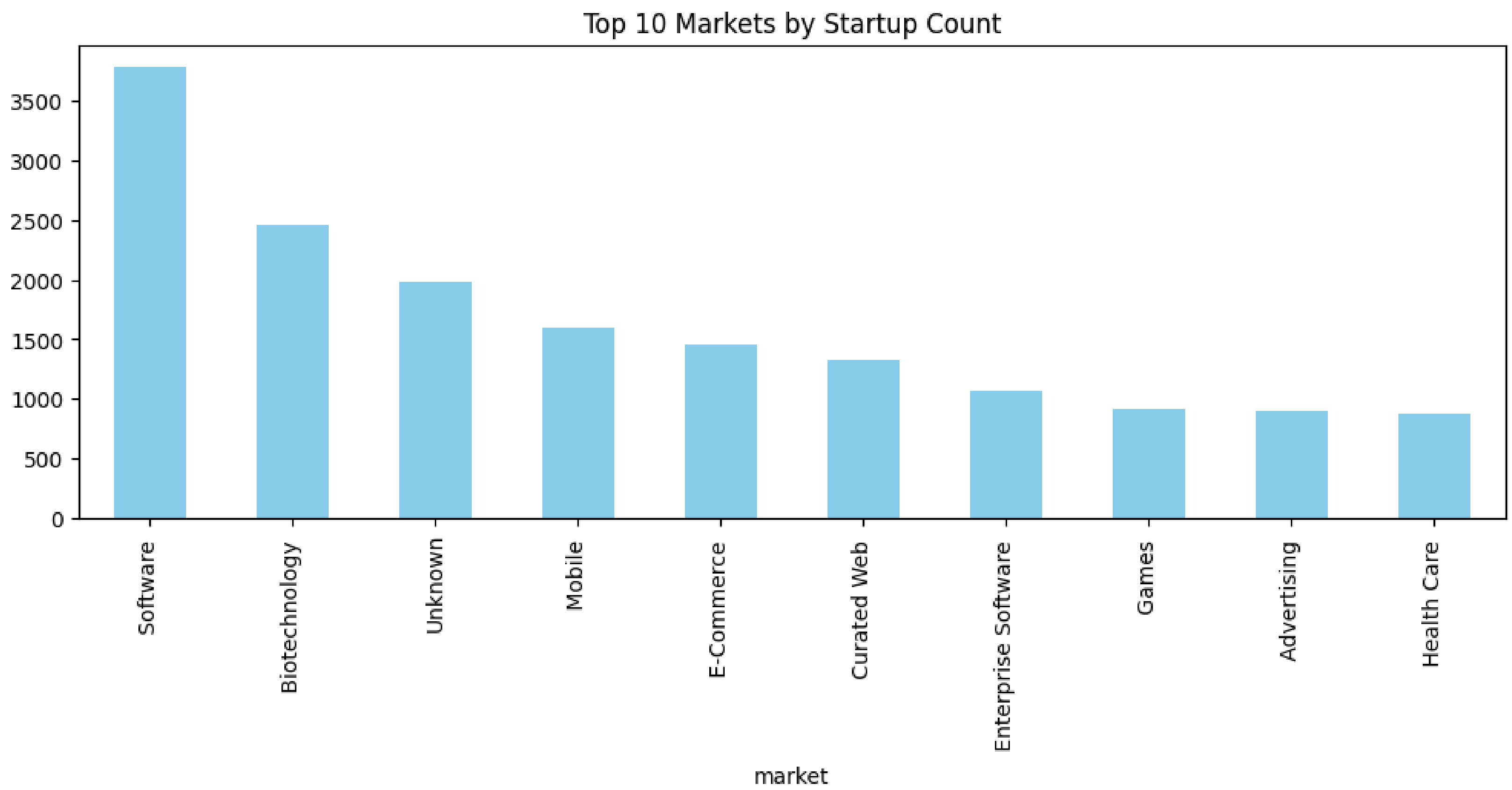
- Software and Clean Technology are mid-range with about $25 billion each.

E-Commerce and Unknown are funded at around $15 billion.

- Enterprise Software, Health Care, and Hardware + Software are near $10 billion.

Semiconductors has the lowest funding, under $10 billion.

In [ ]:
```
    # Bar chart for funding by market
plt.figure(figsize=(12, 4))
df['market'].value_counts().head(10).plot(kind='bar', color='skyblue')
plt.title('Top 10 Markets by Startup Count')
plt.show()
```



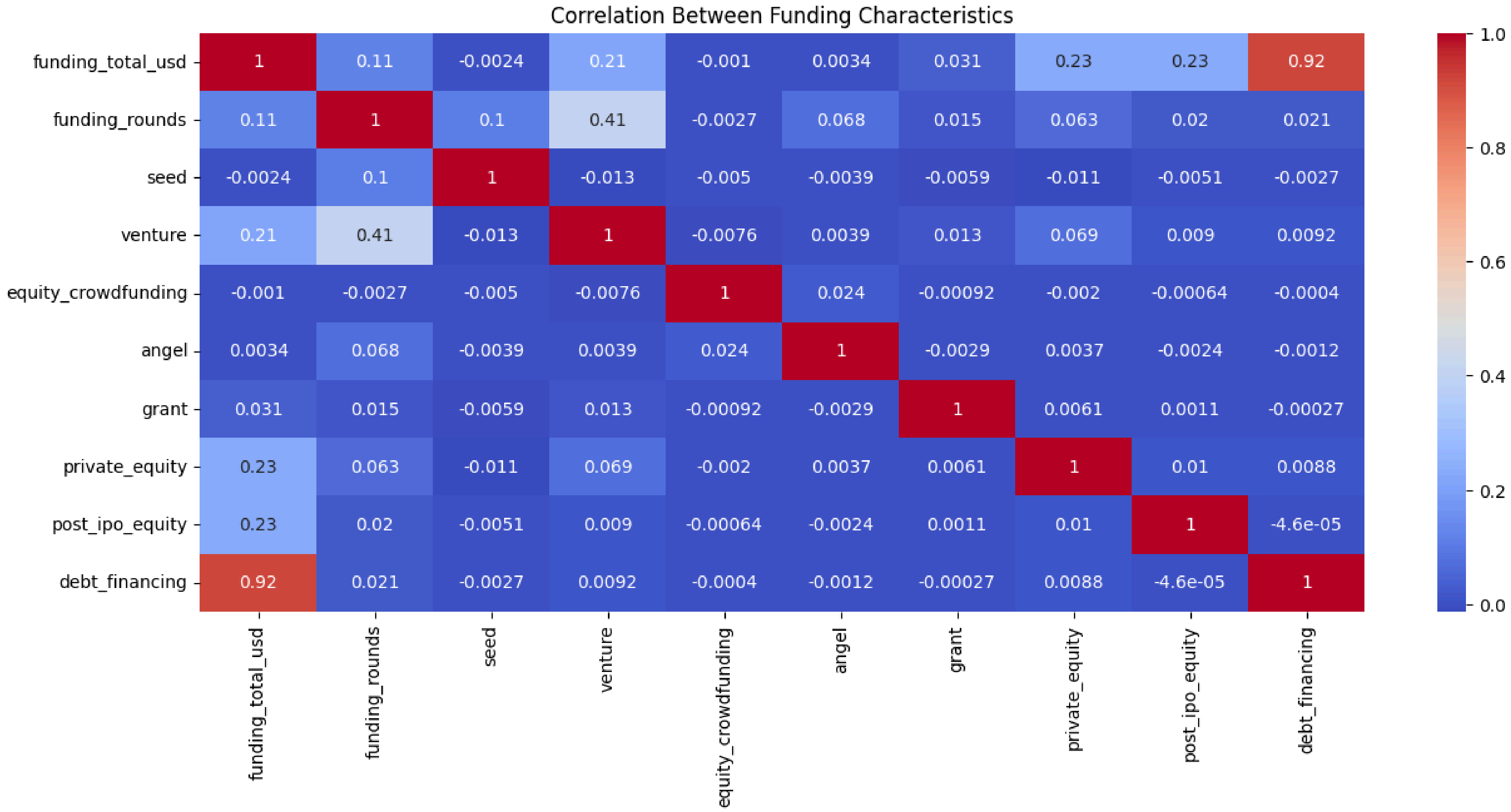
Observation

- Software dominates with the highest number of startups (over 3,500).
- Biotechnology ranks second, followed by an Unknown category.
- Mobile and E-Commerce are mid-level in startup count.
- Curated Web, Enterprise Software, Games, and Advertising have moderate presence
- Health Care has the lowest startup count among the top 10.

In [ ]:
```
# 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()
```



In [ ]:

```
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:                Mon, 07 Oct 2024   Prob (F-statistic):               0.00
Time:                        18:24:16   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.
```

Observation

R-squared (0.847):

- An R-squared value of 0.847 means that 84.7% of the variance in total funding (USD) is explained by Debt Financing alone.
- This is a very high value, indicating a strong linear relationship between debt financing and total funding. It suggests that companies with higher total funding are significantly relying on debt financing.

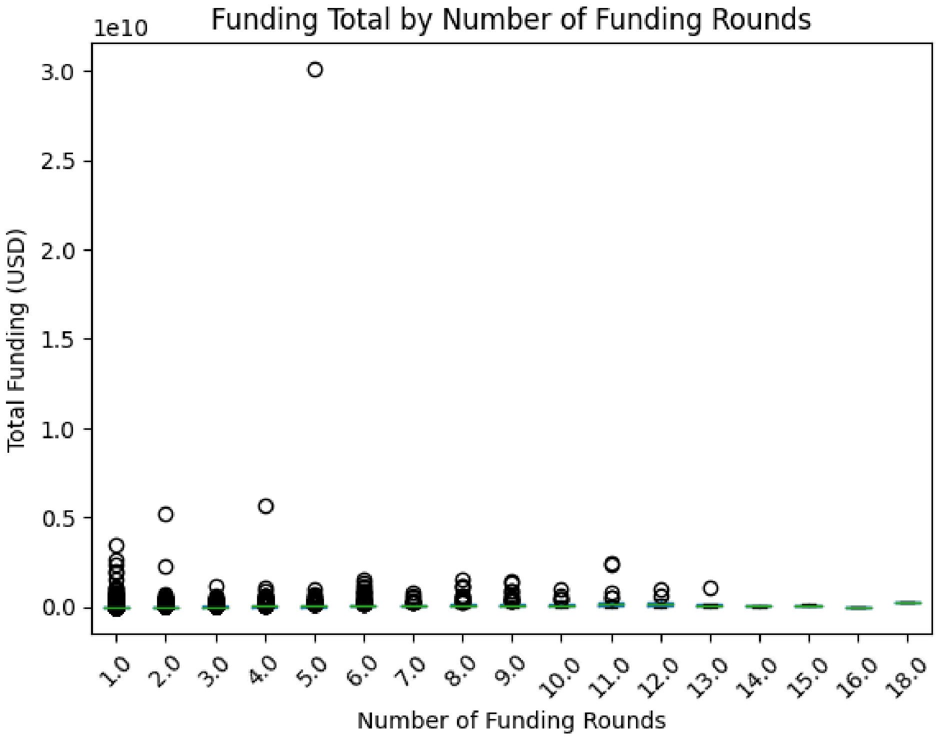
P-value (F-statistic = 0.00):

- The F-statistic p-value of 0.00 is less than the 0.05 threshold, meaning the relationship is statistically significant.
- This confirms that the relationship between Debt Financing and Total Funding is not due to random chance.

In [ ]:

```
# 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()
```

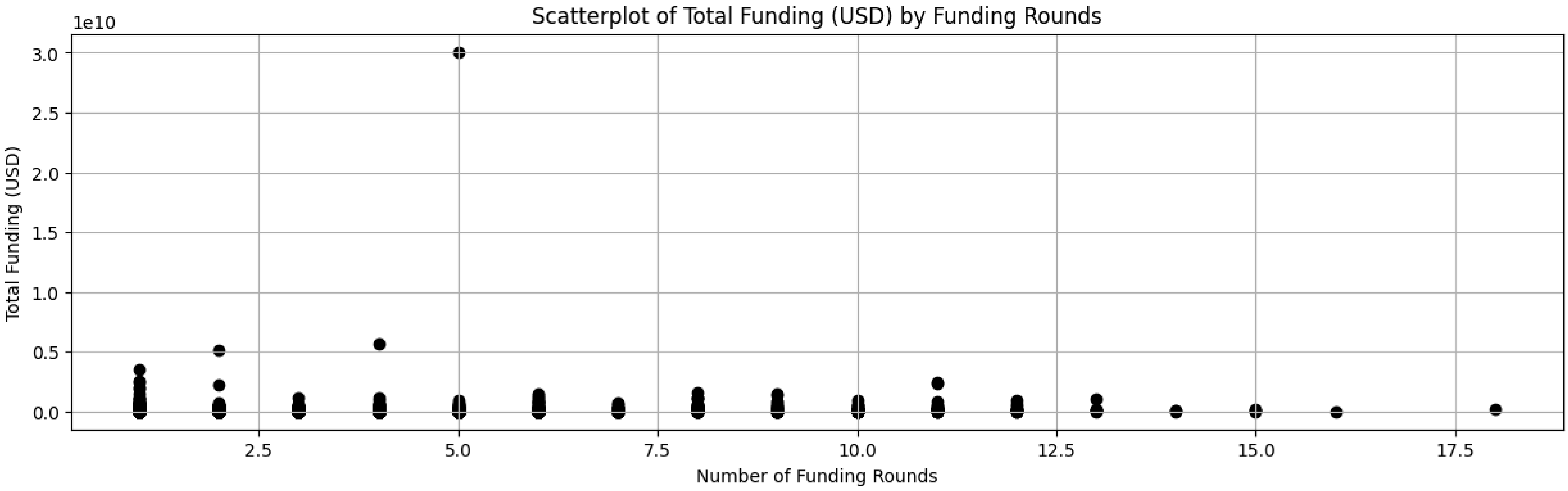
```
<Figure size 2000x200 with 0 Axes>
```



Observation

- Companies generally receive moderate funding across multiple rounds, but a small group of companies can secure extremely high funding early in the process.
- Investigating the outliers (especially those with fewer funding rounds but significantly higher funding) can provide insights into what factors contributed to such high success.

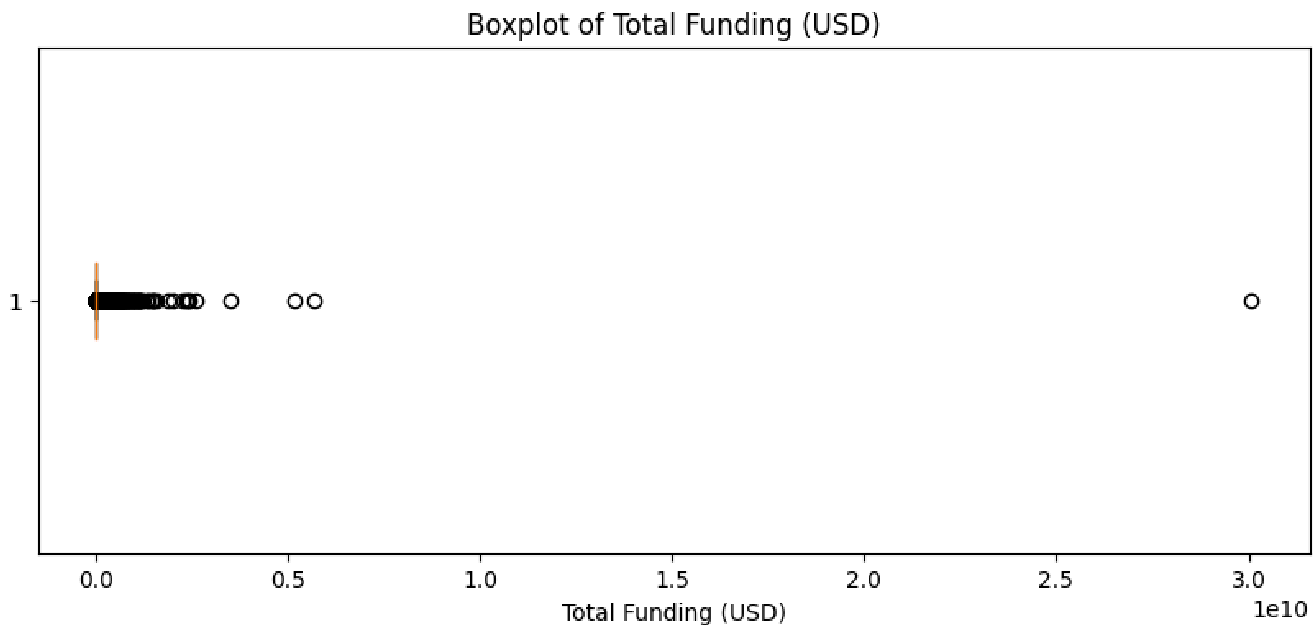
In [ ]:
```python
# Scatterplot for 'funding_total_usd' vs 'funding_rounds'
plt.figure(figsize=(15, 4))
plt.scatter(df['funding_rounds'], df['funding_total_usd'], color='black')
plt.title('Scatterplot of Total Funding (USD) by Funding Rounds')
plt.xlabel('Number of Funding Rounds')
plt.ylabel('Total Funding (USD)')
plt.grid(True)
plt.show()
```



Observation

- The scatterplot shows that most companies raise funds in fewer rounds (1-7), with total funding generally under $1 billion.
- A few outliers raised much more, including one near $30 billion in 5 funding rounds.
- There's no clear linear relationship between funding rounds and total funding, with significant variability in the data.

In [ ]:
```python
# 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()
```



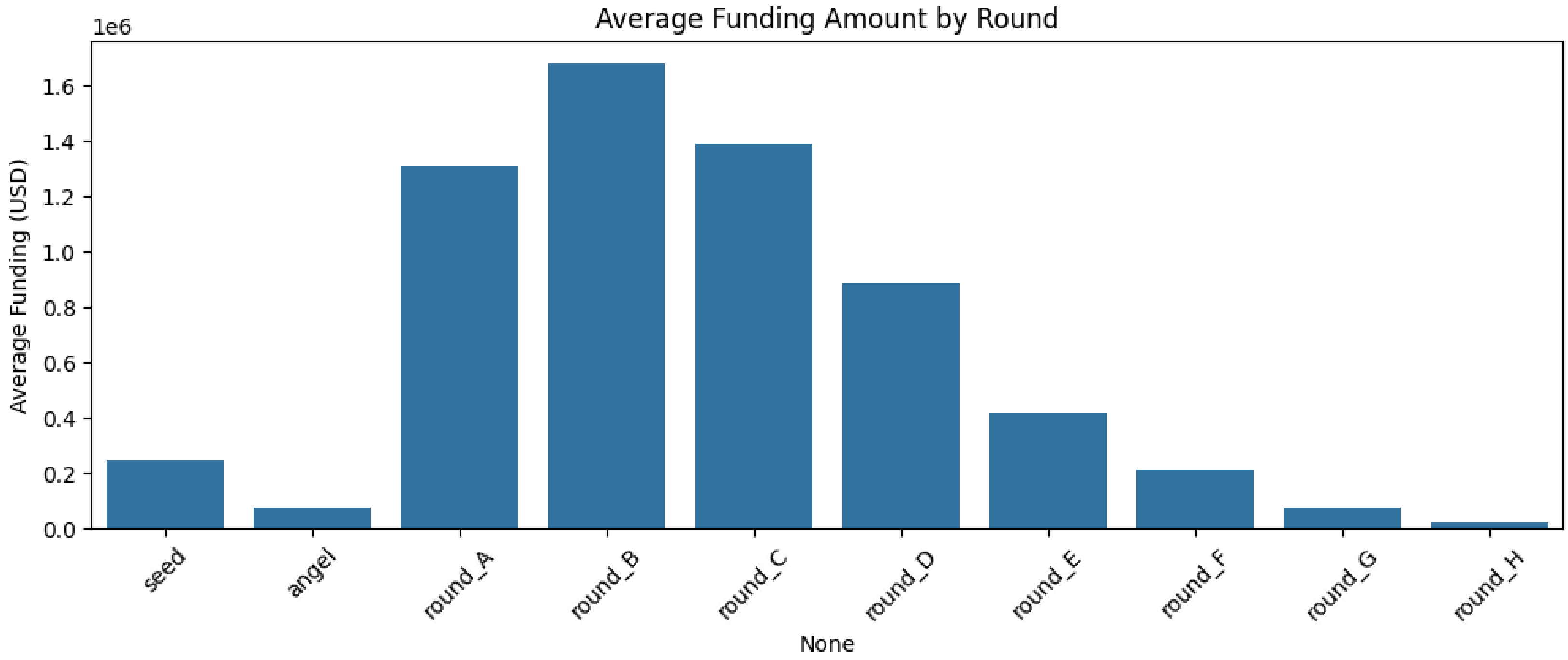
Observation

- Majority of the data points are clustered around lower funding amounts.
- A few extreme outliers with total funding over $20billion, withonenear30$ billion.
- The distribution is highly skewed, indicating large disparities in funding among startups.

- Most companies receive lower funding, while a small number secure significantly higher investments.

In [ ]:
```python
# Analyze progression through funding rounds
round_cols = ['seed', 'angel', 'round_A', 'round_B', 'round_C', 'round_D', 'round_E', 'round_F', 'round_G', 'round_H']
round_data = df[round_cols].mean()

plt.figure(figsize=(12, 4))
sns.barplot(x=round_data.index, y=round_data.values)
plt.title('Average Funding Amount by Round')
plt.ylabel('Average Funding (USD)')
plt.xticks(rotation=45)
plt.show()
```

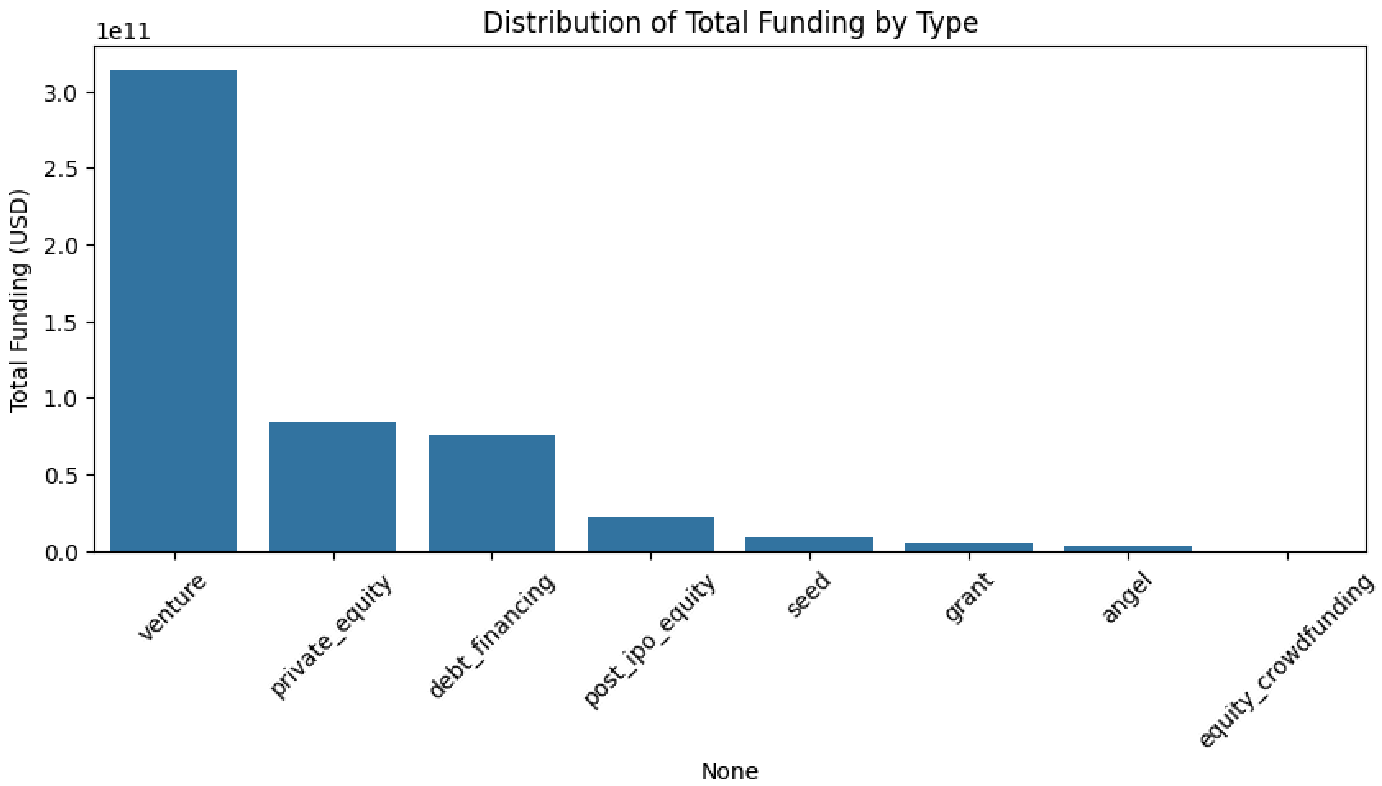


Observation

- Round B has the highest average funding, over $1.6 million.
- Rounds A and C also have significant funding, around $1.4millionand$1.2 million respectively.
- Funding decreases notably after Round C, with Round D having a lower average around $700,000.
- The funding significantly drops after Round D, with Rounds E to H showing relatively low average funding amounts.
- Seed and Angel rounds have the lowest average funding.

In [ ]:
```python
#Funding Type Comparison:
funding_types = ['seed', 'venture', 'equity_crowdfunding', 'angel', 'grant', 'private_equity', 'post_ipo_equity', 'debt_financing']
funding_distribution = df[funding_types].sum().sort_values(ascending=False)

plt.figure(figsize=(10, 4))
sns.barplot(x=funding_distribution.index, y=funding_distribution.values)
plt.title('Distribution of Total Funding by Type')
plt.ylabel('Total Funding (USD)')
plt.xticks(rotation=45)
plt.show()
```



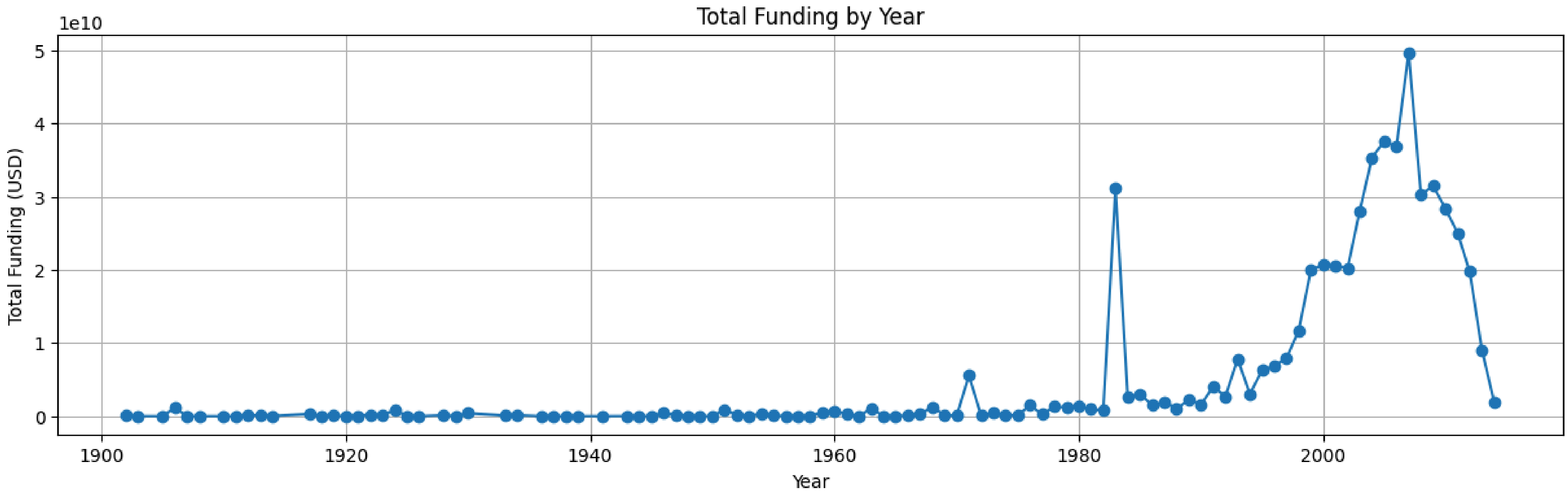
Observation

- Venture funding dominates, surpassing $300 billion.
- Private equity is second, around $100 billion.
- Debt financing and post-IPO equity have moderate funding.
- Seed, grants, and angel investments contribute small amounts.
- Most funding is concentrated in venture capital and private equity.

In [ ]:
```python
#Time-Series Analysis:
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, 4))
plt.plot(yearly_funding['founded_year'], yearly_funding['funding_total_usd'], marker='o')
```

```python
plt.title('Total Funding by Year')
plt.xlabel('Year')
plt.ylabel('Total Funding (USD)')
plt.grid(True)
plt.show()
```



Observation

- Pre-1980: Funding remained relatively low and steady, with few noticeable spikes.
- 1980s: There is a minor spike in funding during this period, possibly related to changes in economic policies or growth in technology sectors.
- Late 1990s to Early 2000s: A sharp increase in funding is observed, likely due to the dot-com boom, where tech companies received massive investments.
- List item
- 2000-2010: A steep decline in funding follows after the early 2000s, possibly corresponding to the dot-com crash and the economic downturn.
- Post-2010: Another rise in funding, followed by a drop-off towards the end of the dataset. This could relate to the growth of new tech sectors or the global financial crisis in 2008 and recovery thereafter.

Insights ⚡

- **Market Concentration:** Biotechnology, Mobile, and Software are the top-funded markets, indicating strong investor confidence in these sectors. This suggests a focus on innovation-driven and technology-intensive industries.
- **Geographical Disparity:** There's a significant concentration of funding in major tech hubs, particularly the San Francisco Bay Area and New York City. This highlights the importance of location in accessing venture capital.
- **Funding Evolution:** The startup funding landscape has evolved dramatically since the 1980s, with significant spikes during the dot-com boom and post-2010 period, reflecting changing economic conditions and technological advancements.
- **Funding Round Dynamics:** While later rounds (B and C) tend to have higher average funding, there's a decrease in funding amounts for rounds D and beyond. This suggests a "funnel" effect where fewer companies reach later stages but those that do can secure significant investments.
- **Funding Type Preference:** Venture capital dominates the funding landscape, followed by private equity. This indicates a preference for high-risk, high-reward investments in the startup ecosystem.
- **Debt Financing Impact:** There's a strong correlation between debt financing and total funding, suggesting that companies leveraging debt alongside equity can secure higher overall funding.
- **Outlier Effect:** The presence of significant outliers in funding amounts highlights the potential for extraordinary success in the startup world, but also underscores the extreme variability in outcomes.

Recommendations 📊

For Entrepreneurs:

- Focus on high-potential sectors like Biotechnology, Mobile, and Software to align with investor interests.
- Consider relocating to major tech hubs to increase access to funding opportunities.
- Plan for a strategic mix of funding types, including debt financing, to maximize total funding potential.
- Prepare for a potential decrease in funding availability in later rounds and plan accordingly.

For Investors:

- Diversify portfolios across top-funded sectors to balance risk and potential returns.
- Look beyond traditional tech hubs for undervalued opportunities in emerging startup ecosystems.
- Consider the potential of debt financing as a complementary strategy to equity investments.
- Pay attention to economic cycles and adjust investment strategies accordingly, given the historical volatility in funding trends.

For Policymakers:

- Develop initiatives to support the growth of startup ecosystems outside of major tech hubs to distribute economic benefits more evenly.
- Create policies that encourage diverse funding types, including debt financing options for startups.
- Support sectors showing high growth potential, like Biotechnology and Clean Technology, through targeted programs and incentives.

For Startup Accelerators and Incubators:

- Tailor programs to prepare startups for the realities of funding round dynamics, especially the challenges of securing later-stage funding.
- Foster connections with a diverse range of funding sources, including venture capital, private equity, and debt financing options.
- Provide education on strategic location choices and their impact on funding accessibility.

General Strategy:

- Recognize the high variability in startup outcomes and plan for multiple scenarios.
- Stay informed about market trends and economic conditions that can impact funding availability.
- For extraordinary success, study outlier cases to understand factors contributing to their exceptional funding achievements.

These insights and recommendations provide a comprehensive view of the startup funding landscape, offering valuable guidance for all stakeholders in the ecosystem to navigate the complexities of startup financing and increase the chances of success.

By

Malarvizhi K
