```
!gdown 110tJaSocsCTvgqjJ3LyfzZOmhy8woTfP
```

```
Downloading...
From: https://drive.google.com/uc?id=110tJaSocsCTvgqjJ3LyfzZOmhy8woTfP
To: /content/investments_VC.csv
100% 12.5M/12.5M [00:00<00:00, 110MB/s]
```

```
import pandas as pd
```

```
dt = pd.read_csv("investments_VC.csv", encoding_errors='ignore')
```

```
dt
```

	permalink	name	homepage_url	category_list	market	funding_tota
0	/organization/waywire	#waywire	http://www.waywire.com	|Entertainment|Politics|Social Media|News|	News	17,
1	/organization/tv-communications	&TV Communications	http://enjoyandtv.com	|Games|	Games	40,
2	/organization/rock-your-paper	'Rock' Your Paper	http://www.rockyourpaper.org	|Publishing|Education|	Publishing	-
3	/organization/in-touch-network	(In)Touch Network	http://www.InTouchNetwork.com	|Electronics|Guides|Coffee|Restaurants|Music|i...	Electronics	15,
4	/organization/r-ranch-and-mine	-R- Ranch and Mine	NaN	|Tourism|Entertainment|Games|	Tourism	
...	...	...	...	...	...	
54289	NaN	NaN	NaN	NaN	NaN	
54290	NaN	NaN	NaN	NaN	NaN	
54291	NaN	NaN	NaN	NaN	NaN	
54292	NaN	NaN	NaN	NaN	NaN	
54293	NaN	NaN	NaN	NaN	NaN	

54294 rows × 39 columns

```
import numpy as np
```

```
dt[dt.isna().all(axis=1)]
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code	region	...	secondary_
49438	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	
49439	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	
49440	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	
49441	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	
49442	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	
...	...	...	...	...	...	...	...	...	...	...	...	
54289	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	
54290	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	
54291	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	
54292	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	
54293	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	

4856 rows × 39 columns

try to find suitable encode for the file

```
pip install chardet
```

```
Requirement already satisfied: chardet in /usr/local/lib/python3.10/dist-packages (5.2.0)
```

```
# --## method 1
# import chardet

# with open('investments_VC.csv', 'rb') as f:
#     result = chardet.detect(f.read())
#     encoding = result['encoding']
#     confidence = result['confidence']

# print(f"Encoding: {encoding}, Confidence: {confidence}")
```

```
# --## method 2
# import pandas as pd

# with open('investments_VC.csv', 'rb') as f:
#     result = chardet.detect(f.read())
#     encoding = result['encoding']

# print(f"File Encoding: {encoding}")
```

both method show suitable encodind - MacRoman

```
dt = pd.read_csv("investments_VC.csv", encoding='MacRoman')
```

```
dt
```

	permalink	name	homepage_url	category_list	market	funding_tota
0	/organization/waywire	#waywire	http://www.waywire.com	|Entertainment|Politics|Social Media|News|	News	17,
1	/organization/tv-communications	&TV Communications	http://enjoyandtv.com	|Games|	Games	40,
2	/organization/rock-your-paper	'Rock' Your Paper	http://www.rockyourpaper.org	|Publishing|Education|	Publishing	-
3	/organization/in-touch-network	(In)Touch Network	http://www.InTouchNetwork.com	|Electronics|Guides|Coffee|Restaurants|Music|i...	Electronics	15,
4	/organization/r-ranch-and-mine	-R- Ranch and Mine	NaN	|Tourism|Entertainment|Games|	Tourism	-
...	...	...	...	...	...	
54289	NaN	NaN	NaN	NaN	NaN	
54290	NaN	NaN	NaN	NaN	NaN	
54291	NaN	NaN	NaN	NaN	NaN	
54292	NaN	NaN	NaN	NaN	NaN	
54293	NaN	NaN	NaN	NaN	NaN	

54294 rows × 39 columns

```
len(dt[dt.isnull().all(axis=1)])
```

```
4856
```

The data is really missing instead of missing due to encoding

```
dt[~dt.isnull().all(axis=1)] ## non null data
```

	permalink	name	homepage_url	category_list	market	funding_to
0	/organization/waywire	#waywire	http://www.waywire.com	|Entertainment|Politics|Social Media|News|	News	1
1	/organization/tv-communications	&TV Communications	http://enjoyandtv.com	|Games|	Games	4
2	/organization/rock-your-paper	'Rock' Your Paper	http://www.rockyourpaper.org	|Publishing|Education|	Publishing	
3	/organization/in-touch-network	(In)Touch Network	http://www.InTouchNetwork.com	|Electronics|Guides|Coffee|Restaurants|Music|i...	Electronics	1
4	/organization/r-ranch-and-mine	-R- Ranch and Mine	NaN	|Tourism|Entertainment|Games|	Tourism	
...	...	...	...	...	...	
49433	/organization/zzish	Zzish	http://www.zzish.com	|Analytics|Gamification|Developer APIs|iOS|And...	Education	
49434	/organization/zznode-science-and-technology-co...	ZZNode Science and Technology	http://www.zznode.com	|Enterprise Software|	Enterprise Software	1
49435	/organization/zzzzapp-com	Zzzzapp Wireless ltd.	http://www.zzzzapp.com	|Web Development|Advertising|Wireless|Mobile|	Web Development	
49436	/organization/a-list-games	[a]list games	http://www.alistgames.com	|Games|	Games	9
49437	/organization/x	[x+1]	http://www.xplusone.com/	|Enterprise Software|	Enterprise Software	4,5

49438 rows × 39 columns

```
dt.isnull().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

Confirmed ~4856 rows are completly missing

round_B	4856

Data cleaning

round_E	4856

```
dt.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
```

```
## Clean Name Folder
dt['name'] = dt['name'].str.replace('[^a-zA-Z0-9\. ]', '', regex=True)
dt.head()
```

	permalink	name	homepage_url	category_list	market	funding_total_us
0	/organization/waywire	waywire	http://www.waywire.com	|Entertainment|Politics|Social Media|News|	News	17,50,00
1	/organization/tv-communications	TV Communications	http://enjoyandtv.com	|Games|	Games	40,00,00
2	/organization/rock-your-paper	Rock Your Paper	http://www.rockyourpaper.org	|Publishing|Education|	Publishing	40,00
3	/organization/in-touch-network	InTouch Network	http://www.InTouchNetwork.com	|Electronics|Guides|Coffee|Restaurants|Music|i...	Electronics	15,00,00
4	/organization/r-ranch-and-mine	R Ranch and Mine	NaN	|Tourism|Entertainment|Games|	Tourism	60,00

5 rows × 39 columns

```
## checking data type consistency for each row
def check_dtype_consistency(dt):
  return dt.apply(lambda x: x.apply(type).nunique() == 1)
check_dtype_consistency(dt)
```

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

```
# 🌟 🌟 🌟 🌟 🌟

def check_dtype_consistency(dt):
    """
    Checks if each row in a column has the same data type and counts null values.

    Args:
        df (pd.DataFrame): Input DataFrame.

    Returns:
        pd.DataFrame: DataFrame containing data type consistency and null value counts for each column.
    """
    result = dt.apply(lambda x: pd.Series({
        'dtype_consistent': x.apply(type).nunique() == 1,
        'null_count': x.isnull().sum(),
        'unique_dtypes': x.apply(type).unique()
    }))
    return result

rslt = check_dtype_consistency(dt)
```

```
rslt.T
```

	dtype_consistent	null_count	unique_dtypes
permalink	False	4856	[<class 'str'>, <class 'float'>]
name	False	4857	[<class 'str'>, <class 'float'>]
homepage_url	False	8305	[<class 'str'>, <class 'float'>]
category_list	False	8817	[<class 'str'>, <class 'float'>]
market	False	8824	[<class 'str'>, <class 'float'>]
funding_total_usd	False	4856	[<class 'str'>, <class 'float'>]
status	False	6170	[<class 'str'>, <class 'float'>]
country_code	False	10129	[<class 'str'>, <class 'float'>]
state_code	False	24133	[<class 'str'>, <class 'float'>]
region	False	10129	[<class 'str'>, <class 'float'>]
city	False	10972	[<class 'str'>, <class 'float'>]
funding_rounds	True	4856	[<class 'float'>]
founded_at	False	15740	[<class 'str'>, <class 'float'>]
founded_month	False	15812	[<class 'str'>, <class 'float'>]
founded_quarter	False	15812	[<class 'str'>, <class 'float'>]
founded_year	True	15812	[<class 'float'>]
first_funding_at	False	4856	[<class 'str'>, <class 'float'>]
last_funding_at	False	4856	[<class 'str'>, <class 'float'>]
seed	True	4856	[<class 'float'>]
venture	True	4856	[<class 'float'>]
equity_crowdfunding	True	4856	[<class 'float'>]
undisclosed	True	4856	[<class 'float'>]
convertible_note	True	4856	[<class 'float'>]
debt_financing	True	4856	[<class 'float'>]
angel	True	4856	[<class 'float'>]
grant	True	4856	[<class 'float'>]
private_equity	True	4856	[<class 'float'>]
post_ipo_equity	True	4856	[<class 'float'>]
post_ipo_debt	True	4856	[<class 'float'>]
secondary_market	True	4856	[<class 'float'>]
product_crowdfunding	True	4856	[<class 'float'>]

```
dt.head(1).T
```

	0
permalink	/organization/waywire
name	waywire
homepage_url	http://www.waywire.com
category_list	|Entertainment|Politics|Social Media|News|
market	News
funding_total_usd	17,50,000
status	acquired
country_code	USA
state_code	NY
region	New York City
city	New York
funding_rounds	1.0
founded_at	2012-06-01
founded_month	2012-06
founded_quarter	2012-Q2
founded_year	2012.0
first_funding_at	2012-06-30
last_funding_at	2012-06-30
seed	1750000.0
venture	0.0
equity_crowdfunding	0.0
undisclosed	0.0
convertible_note	0.0
debt_financing	0.0
angel	0.0
grant	0.0
private_equity	0.0
post_ipo_equity	0.0
post_ipo_debt	0.0
secondary_market	0.0
product_crowdfunding	0.0

```
dt.iloc[:,0]  ## checking permalink column
```

	permalink
0	/organization/waywire
1	/organization/tv-communications
2	/organization/rock-your-paper
3	/organization/in-touch-network
4	/organization/r-ranch-and-mine
...	...
54289	NaN
54290	NaN
54291	NaN
54292	NaN
54293	NaN

54294 rows × 1 columns

dtype: object

```
## delete columns with all null values
dt = dt[~dt.isnull().all(axis=1)]
```

```
len(dt)
```

```
49438
```

```
dt
```

	permalink	name	homepage_url	category_list	market	funding_to
0	/organization/waywire	waywire	http://www.waywire.com	|Entertainment|Politics|Social Media|News|	News	
1	/organization/tv-communications	TV Communications	http://enjoyandtv.com	|Games|	Games	
2	/organization/rock-your-paper	Rock Your Paper	http://www.rockyourpaper.org	|Publishing|Education|	Publishing	
3	/organization/in-touch-network	InTouch Network	http://www.InTouchNetwork.com	|Electronics|Guides|Coffee|Restaurants|Music|i...	Electronics	
4	/organization/r-ranch-and-mine	R Ranch and Mine	NaN	|Tourism|Entertainment|Games|	Tourism	
...	...	...	...	...	...	
49433	/organization/zzish	Zzish	http://www.zzish.com	|Analytics|Gamification|Developer APIs|iOS|And...	Education	
49434	/organization/zznode-science-and-technology-co...	ZZNode Science and Technology	http://www.zznode.com	|Enterprise Software|	Enterprise Software	
49435	/organization/zzzzapp-com	Zzzzapp Wireless ltd.	http://www.zzzzapp.com	|Web Development|Advertising|Wireless|Mobile|	Web Development	
49436	/organization/a-list-games	alist games	http://www.alistgames.com	|Games|	Games	
49437	/organization/x	x1	http://www.xplusone.com/	|Enterprise Software|	Enterprise Software	

49438 rows × 39 columns

```
check_dtype_consistency(dt).T
## which means null values crewate unnecessary unconsistency eith dtype float .
```

	dtype_consistent	null_count	unique_dtypes
permalink	True	0	[<class 'str'>]
name	False	1	[<class 'str'>, <class 'float'>]
homepage_url	False	3449	[<class 'str'>, <class 'float'>]
category_list	False	3961	[<class 'str'>, <class 'float'>]
market	False	3968	[<class 'str'>, <class 'float'>]
funding_total_usd	True	0	[<class 'str'>]
status	False	1314	[<class 'str'>, <class 'float'>]
country_code	False	5273	[<class 'str'>, <class 'float'>]
state_code	False	19277	[<class 'str'>, <class 'float'>]
region	False	5273	[<class 'str'>, <class 'float'>]
city	False	6116	[<class 'str'>, <class 'float'>]
funding_rounds	True	0	[<class 'float'>]
founded_at	False	10884	[<class 'str'>, <class 'float'>]
founded_month	False	10956	[<class 'str'>, <class 'float'>]
founded_quarter	False	10956	[<class 'str'>, <class 'float'>]
founded_year	True	10956	[<class 'float'>]
first_funding_at	True	0	[<class 'str'>]
last_funding_at	True	0	[<class 'str'>]
seed	True	0	[<class 'float'>]
venture	True	0	[<class 'float'>]
equity_crowdfunding	True	0	[<class 'float'>]
undisclosed	True	0	[<class 'float'>]
convertible_note	True	0	[<class 'float'>]
debt_financing	True	0	[<class 'float'>]
angel	True	0	[<class 'float'>]
grant	True	0	[<class 'float'>]
private_equity	True	0	[<class 'float'>]
post_ipo_equity	True	0	[<class 'float'>]
post_ipo_debt	True	0	[<class 'float'>]
secondary_market	True	0	[<class 'float'>]
product_crowdfunding	True	0	[<class 'float'>]

```
dt[dt['name'].isna()]
```

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

1 rows × 39 columns

```
## Treating Missing value in name column
dt['name'][dt['name'].isna()] = 'tellitin'
dt[dt['name'].isna()]
```

Show hidden output

```
#Missing values in  homepage url column
# dt[dt['homepage_url'].isna()]

dt[dt.iloc[:,2].isna()]
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	s
4	/organization/r-ranch-and-mine	R Ranch and Mine	NaN	|Tourism|Entertainment|Games|	Tourism	60,000	operating	USA	
11	/organization/1-4-all	14 All	NaN	|Entertainment|Games|Software|	Software	-	operating	USA	
14	/organization/1-618-technology	1.618 Technology	NaN	|Real Estate|	Real Estate	-	operating	USA	
29	/organization/10∞north	10North	NaN	|Fashion|	Fashion	-	operating	CAN	
42	/organization/121-rentals	121 Rentals	NaN	|Real Estate|	Real Estate	-	operating	USA	
...	...	...	...	...	...	...	...	...	
49341	/organization/zperfectgift	zPerfectGift	NaN	|Social Media|	Social Media	25,000	operating	USA	
49356	/organization/zuffle	Zuffle	NaN	NaN	NaN	1,32,351	operating	ITA	
49378	/organization/zumodrive	zumodrive	NaN	NaN	NaN	-	operating	NaN	
49390	/organization/zuse	Zuse	NaN	|Software|	Software	-	operating	USA	
49402	/organization/zwamy	Zwamy	NaN	NaN	NaN	3,87,794	operating	USA	

3449 rows × 39 columns

Treating Missing Values for homepage url column</>

Primary Retrieval with APIs: Use Rapidapi for most companies.

Fallback on Permalink Logic: For missing or ambiguous cases, use permalink parsing and HTTP validation.

Final Quality Check: Use a small, manual review for high-priority entries (e.g., large funding rounds or strategic markets).

🔴⚠

```
#  ⚠ ### api not working


# df = dt[dt.iloc[:,2].isna()]
# ## homepage url NaN values -> For maximum accuracy and scalability, a combination of API-based retrieval and permalink-based validation


# import http.client
# import json
# import requests
# from bs4 import BeautifulSoup
# import logging
# import time
# import random

# logging.basicConfig(filename='api_errors.log', level=logging.ERROR)

# # RapidAPI credentials
# api_key = "dc9dd430ecmsh5d0d1ca06aebb52p14d6a8jsn7c3f5daa1076"
# api_host = "crunchbase-scraper.p.rapidapi.com"


#     # List of User-Agents
# user_agents = [
#     'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/58.0.3029.110 Safari/537.3',
#     'Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/51.0.2704.103 Safari/537.36',
#     'Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:53.0) Gecko/20100101 Firefox/53.0'
# ]

# # Iterate through each company permalink
# for index, row in df.iterrows():
#     permalink = row['permalink']
```

```
#     permalink_url = f"https://www.crunchbase.com{permalink}"

#     # Choose random User-Agent
#     headers = {'User-Agent': random.choice(user_agents)}

#     # HTTP validation
#     try:
#         response = requests.head(permalink_url, headers=headers)
#         time.sleep(1)  # Add 1-second delay
#         if response.status_code == 200:
#             # Parse HTML for homepage URL
#             html_response = requests.get(permalink_url, headers=headers)
#             soup = BeautifulSoup(html_response.text, 'lxml')
#             homepage_url_tag = soup.find('link', rel='website')
#             if homepage_url_tag:
#                 homepage_url = homepage_url_tag['href']
#             else:
#                 homepage_url_tag = soup.find('a', href=True)
#                 if homepage_url_tag:
#                     homepage_url = homepage_url_tag['href']
#                 else:
#                     homepage_url = soup.select_one('meta[property="og:url"]')['content']
#             df.at[index, 'homepage_url'] = homepage_url
#         else:
#             logging.error(f"HTTP validation failed for {permalink}: {response.status_code}")
#             df.at[index, 'homepage_url'] = "HTTP validation failed"
#     except requests.exceptions.RequestException as e:
#         logging.error(f"Error processing {permalink}: {str(e)}")
#         df.at[index, 'homepage_url'] = "Request error"
#     except Exception as e:
#         logging.error(f"Unexpected error: {str(e)}")
#         df.at[index, 'homepage_url'] = "Unexpected error"
# # Save the updated DataFrame
# df.to_csv('updated_dataframe.csv', index=False)

```

```
dt
```

	permalink	name	homepage_url	category_list	market	funding_to
0	/organization/waywire	waywire	http://www.waywire.com	|Entertainment|Politics|Social Media|News|	News	
1	/organization/tv-communications	TV Communications	http://enjoyandtv.com	|Games|	Games	
2	/organization/rock-your-paper	Rock Your Paper	http://www.rockyourpaper.org	|Publishing|Education|	Publishing	
3	/organization/in-touch-network	InTouch Network	http://www.InTouchNetwork.com	|Electronics|Guides|Coffee|Restaurants|Music|i...	Electronics	
4	/organization/r-ranch-and-mine	R Ranch and Mine	NaN	|Tourism|Entertainment|Games|	Tourism	
...	...	...	...	...	...	
49433	/organization/zzish	Zzish	http://www.zzish.com	|Analytics|Gamification|Developer APIs|iOS|And...	Education	
49434	/organization/zznode-science-and-technology-co...	ZZNode Science and Technology	http://www.zznode.com	|Enterprise Software|	Enterprise Software	
49435	/organization/zzzzapp-com	Zzzzapp Wireless ltd.	http://www.zzzzapp.com	|Web Development|Advertising|Wireless|Mobile|	Web Development	
49436	/organization/a-list-games	alist games	http://www.alistgames.com	|Games|	Games	
49437	/organization/x	x1	http://www.xplusone.com/	|Enterprise Software|	Enterprise Software	4,9

49438 rows × 39 columns

Fallback system to combines permalink-based retrieval with a lightweight ML classifier

```
# from sklearn.ensemble import RandomForestClassifier
# from sklearn.feature_extraction.text import TfidfVectorizer
# from sklearn.model_selection import train_test_split

# # Load nonnull url dataset as labeled dataset
# df_lbl = dt[~dt.iloc[:,2].isna()]

# # Define features (X) and target variable (y)
# X = df_lbl[['permalink', 'name', 'category_list', ' market ', 'country_code', 'region', 'city']]
# y = df_lbl['homepage_url']

# # Split data into training and testing sets (80% for training, 20% for testing)
# from sklearn.model_selection import train_test_split
# X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
# # Create TF-IDF vectorizer
# vectorizer = TfidfVectorizer()

# # Fit vectorizer to training data and transform both sets
# X_train_tfidf = vectorizer.fit_transform(X_train['permalink'])

# X_test_tfidf = vectorizer.transform(X_test)

# # Train random forest classifier
# clf = RandomForestClassifier(n_estimators=100, random_state=42)

# clf.fit(X_train_tfidf, y_train)

# # Define fallback function
# def fallback_ml_classifier(company_name):
#     company_name_tfidf = vectorizer.transform([company_name])
#     predicted_url = clf.predict(company_name_tfidf)[0]
#     return predicted_url
```

```
# from sklearn.feature_extraction.text import TfidfVectorizer
# from sklearn.decomposition import TruncatedSVD
# from sklearn.preprocessing import OneHotEncoder
# from sklearn.pipeline import Pipeline
# from sklearn.compose import ColumnTransformer
# from sklearn.impute import SimpleImputer
# from sklearn.linear_model import LogisticRegression
# from sklearn.model_selection import train_test_split
# import pandas as pd

# # Load data
# df_lbl = dt[~dt['homepage_url'].isna()]  # Only rows with homepage URLs
# X = df_lbl[['permalink', 'name', 'category_list', ' market ', 'country_code', 'region', 'city']]
# y = df_lbl['homepage_url']

# # Handle missing values in text columns
# X['permalink'] = X['permalink'].fillna('unknown')
# X['name'] = X['name'].fillna('unknown')

# # Sample a subset of data to reduce memory usage (e.g., 10,000 rows)
# X_sample = X.sample(n=10000, random_state=42)
# y_sample = y[X_sample.index]

# # Split sampled data
# X_train, X_test, y_train, y_test = train_test_split(X_sample, y_sample, test_size=0.2, random_state=42)

# # Preprocessing pipelines
# # 1. TF-IDF for `permalink` column
# permalink_pipeline = Pipeline([
#     ('vectorizer', TfidfVectorizer(max_features=1000)),  # Reduce feature size
#     ('dim_reduction', TruncatedSVD(n_components=30))    # Further reduce dimensionality
# ])

# # 2. TF-IDF for `name` column
# name_pipeline = Pipeline([
#     ('vectorizer', TfidfVectorizer(max_features=1000)),
#     ('dim_reduction', TruncatedSVD(n_components=30))
# ])
```

```

# # 3. One-Hot Encoding for categorical columns
# categorical_features = ['category_list', 'country_code', 'region', 'city']
# categorical_pipeline = Pipeline([
#     ('imputer', SimpleImputer(strategy='constant', fill_value='unknown')),
#     ('encoder', OneHotEncoder(handle_unknown='ignore', sparse_output=True))  # Sparse output for efficiency
# ])

# # Combine pipelines in a ColumnTransformer
# preprocessor = ColumnTransformer([
#     ('permalink', permalink_pipeline, 'permalink'),
#     ('name', name_pipeline, 'name'),
#     ('cat', categorical_pipeline, categorical_features)
# ])

# # Final model pipeline
# model = Pipeline([
#     ('preprocessing', preprocessor),
#     ('classifier', LogisticRegression(max_iter=1000))  # Use a lightweight model
# ])

# # Train the model
# model.fit(X_train, y_train)

# # Test the model
# accuracy = model.score(X_test, y_test)
# print(f"Model Accuracy: {accuracy:.2f}")

```

Model Accuracy: 0.00

Failed to apply model

```
dt[dt['homepage_url'].isna()]
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code
4	/organization/r-ranch-and-mine	R Ranch and Mine	NaN	|Tourism|Entertainment|Games|	Tourism	60,000	operating	USA
11	/organization/1-4-all	14 All	NaN	|Entertainment|Games|Software|	Software	-	operating	USA
14	/organization/1-618-technology	1.618 Technology	NaN	|Real Estate|	Real Estate	-	operating	USA
29	/organization/10∞north	10North	NaN	|Fashion|	Fashion	-	operating	CAN
42	/organization/121-rentals	121 Rentals	NaN	|Real Estate|	Real Estate	-	operating	USA
...	...	...	...	...	...	...	...	...
49341	/organization/zperfectgift	zPerfectGift	NaN	|Social Media|	Social Media	25,000	operating	USA
49356	/organization/zuffle	Zuffle	NaN	NaN	NaN	1,32,351	operating	ITA
49378	/organization/zumodrive	zumodrive	NaN	NaN	NaN	-	operating	NaN
49390	/organization/zuse	Zuse	NaN	|Software|	Software	-	operating	USA
49402	/organization/zwamy	Zwamy	NaN	NaN	NaN	3,87,794	operating	USA

3449 rows × 39 columns

Generate Homepade url from company name

using webscrapper

```
df = dt[~dt['homepage_url'].notna()]
df.tail()
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code	re
49341	/organization/zperfectgift	zPerfectGift	NaN	|Social Media|	Social Media	25,000	operating	USA	UT	
49356	/organization/zuffle	Zuffle	NaN	NaN	NaN	1,32,351	operating	ITA	NaN	
49378	/organization/zumodrive	zumodrive	NaN	NaN	NaN	-	operating	NaN	NaN	
49390	/organization/zuse	Zuse	NaN	|Software|	Software	-	operating	USA	NY	
49402	/organization/zwamy	Zwamy	NaN	NaN	NaN	3,87,794	operating	USA	CA	S

5 rows × 39 columns

```
df.head()
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_
4	/organization/r-ranch-and-mine	R Ranch and Mine	NaN	|Tourism|Entertainment|Games|	Tourism	60,000	operating	USA	
11	/organization/1-4-all	14 All	NaN	|Entertainment|Games|Software|	Software	-	operating	USA	
14	/organization/1-618-technology	1.618 Technology	NaN	|Real Estate|	Real Estate	-	operating	USA	
29	/organization/10∞north	10North	NaN	|Fashion|	Fashion	-	operating	CAN	
42	/organization/121-rentals	121 Rentals	NaN	|Real Estate|	Real Estate	-	operating	USA	

5 rows × 39 columns

Using Webscrapping

```

# from googlesearch import search
# import time
# import random



# df['homepage_url'] = df['homepage_url'].fillna('')  # Ensure column exists

# # Function to fetch URL from Google search
# def fetch_url(company_name):
#     query = f"{company_name} official site"
#     try:
#         # Perform Google search and return the first result
#         for result in search(query, lang="en"):
#             return result  # Return the first result
#     except Exception as e:
#         print(f"Error fetching URL for {company_name}: {e}")
#         return None

# # Batch processing: Define batch size
# batch_size = 100
# total_rows = len(df)
# batches = [df.iloc[i:i + batch_size] for i in range(0, total_rows, batch_size)]

# for batch_index, batch in enumerate(batches):
#     print(f"Processing batch {batch_index + 1}/{len(batches)}")

#     for index, row in batch.iterrows():
#         if not row['homepage_url']:  # Only fetch for missing URLs
#             company_name = row['name']
#             if pd.isna(company_name) or not company_name.strip():
#                 df.at[index, 'homepage_url'] = None
#                 continue
```

```
#             url = fetch_url(company_name)
#             df.at[index, 'homepage_url'] = url
#             print(f"Fetched URL for {company_name}: {url}")

#         # Add a random short delay between individual requests within the batch
#         time.sleep(random.uniform(1, 2))

#     # Add a longer delay after each batch
#     print(f"Completed batch {batch_index + 1}. Pausing...")
#     time.sleep(random.uniform(20, 40))  # Pause for 20-40 seconds after each batch

# # Save updated data
# set
# df.to_csv("updated_dataset.csv", index=False)
# print("Homepage URL fetching complete!")

```

```
df = pd.read_csv('updated_dataset.csv')
```

```
df
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	sta
0	/organization/r-ranch-and-mine	R Ranch and Mine	https://r-ranchca.com/	|Tourism|Entertainment|Games|	Tourism	60,000	opera
1	/organization/1-4-all	14 All	https://www.android.com/android-14/	|Entertainment|Games|Software|	Software	-	opera
2	/organization/1-618-technology	1618 Technology	https://1618technologies.com/	|Real Estate|	Real Estate	-	opera
3	/organization/10∞north	10North	https://www.tenthavenuenorth.com/	|Fashion|	Fashion	-	opera
4	/organization/121-rentals	121 Rentals	https://www.station121dsm.com/	|Real Estate|	Real Estate	-	opera
...	...	...	...	...	...	...	
3444	/organization/zperfectgift	zPerfectGift	https://www.perfectgift.com/	|Social Media|	Social Media	25,000	opera
3445	/organization/zuffle	Zuffle	https://www.zufrie.com/?srsltid=AfmBOoooSLI1Oq...	NaN	NaN	1,32,351	opera
3446	/organization/zumodrive	zumodrive	https://en.wikipedia.org/wiki/ZumoDrive	NaN	NaN	-	opera
3447	/organization/zuse	Zuse	https://www.thezeusnetwork.com/	|Software|	Software	-	opera
3448	/organization/zwamy	Zwamy	https://www.zwaan.in/?srsltid=AfmBOoqOJ848wu5L...	NaN	NaN	3,87,794	opera

3449 rows × 39 columns

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

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code	reg
1598	/organization/kaldoora	Kaldoora	NaN	NaN	NaN	7,80,000	operating	USA	CA	SF A
3399	/organization/xochitl-so-shee-gold-mines	Xochitl SoShee Gold mines	NaN	|Manufacturing|	Manufacturing	-	operating	USA	CA	SF

2 rows × 39 columns

```
df['homepage_url'].map(type).unique()
```

```
array([<class 'str'>, <class 'float'>], dtype=object)
```

NaN value is <class 'float'> thats why can't use dropna()

```
df = df.drop(df.index[[1598, 3399]])
df[df['homepage_url'].isna()]
```

permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code	region	...	secondary_marke

0 rows × 39 columns

dt.loc[dt['homepage_url'].isna(), 'homepage_url'] = df['homepage_url'].values

```
dt_mrg = pd.merge(dt,df, on = 'name', how = 'left', suffixes = ('_dt','_df'))
```

```
dt_mrg[['name','homepage_url_dt','homepage_url_df']][dt_mrg["homepage_url_df"].notna()]
```

	name	homepage_url_dt	homepage_url_df
4	R Ranch and Mine	NaN	https://r-ranchca.com/
11	14 All	NaN	https://www.android.com/android-14/
29	10North	NaN	https://www.tenthavenuenorth.com/
42	121 Rentals	NaN	https://www.station121dsm.com/
59	16 Mile Solutions	NaN	https://www.crunchbase.com/organization/16-mil...
...	...	...	...
49345	zPerfectGift	NaN	https://www.perfectgift.com/
49360	Zuffle	NaN	https://www.zufrie.com/?srsltid=AfmBOoooSLI1Oq...
49382	zumodrive	NaN	https://en.wikipedia.org/wiki/ZumoDrive
49394	Zuse	NaN	https://www.thezeusnetwork.com/
49406	Zwamy	NaN	https://www.zwaan.in/?srsltid=AfmBOoqOJ848wu5L...

3463 rows × 3 columns

```
dt_mrg["homepage_url_dt"] = dt_mrg["homepage_url_dt"].fillna(dt_mrg["homepage_url_df"])
```

```
dt_mrg[dt_mrg["homepage_url_dt"].isna()]
```

	permalink_dt	name	homepage_url_dt	category_list_dt	market _dt	funding_total_usd _dt	status_dt	country_code_dt	st
14	/organization/1-618-technology	1.618 Technology	NaN	|Real Estate|	Real Estate	-	operating	USA	
22400	/organization/kaldoora	Kaldoora	NaN	NaN	NaN	7,80,000	operating	USA	
48287	/organization/xochitl-so-shee-gold-mines	Xochitl SoShee Gold mines	NaN	|Manufacturing|	Manufacturing	-	operating	USA	

3 rows × 77 columns

this merge will contain double column one for treated missing value and another (original)

_dt & _df

```
len(dt['homepage_url']), len(dt_mrg["homepage_url_dt"])
```

```
(49438, 49442)
```

```
dt_mrg = dt_mrg.loc[:, ~dt_mrg.columns.str.contains("_df")]
```

```
dt_mrg.columns =  dt_mrg.columns.str.replace("_dt", "")
```

```
dt_mrg.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')
```

```
dt_mrg.iloc[50:200, :]
## chking random data(url) manually
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_
50	/organization/12society	12Society	http://www.12Society.com	|E-Commerce|	E-Commerce	-	acquired	
51	/organization/1366-technologies	1366 Technologies	http://www.1366tech.com	|Manufacturing|	Manufacturing	6,64,50,000	operating	
52	/organization/139shop	139shop	http://www.139shop.com	|E-Commerce|	E-Commerce	-	operating	
53	/organization/13th-lab	13th Lab	http://13thlab.com	|iPad|Games|Augmented Reality|Computer Vision|...	Mobile	7,00,000	operating	
54	/organization/140-proof	140 Proof	http://140proof.com	|Interest Graph|Advertising|	Advertising	55,00,000	operating	
...	...	...	...	...	...	...	...	
195	/organization/3baysover	3BaysOver	http://www.3baysover.com	|Online Travel|Networking|Travel|	Online Travel	17,19,583	operating	
196	/organization/3c-plus	3C Plus	http://www.3cplusnow.com	|Communications Hardware|Internet|Web Developm...	Web Development	75,000	operating	
197	/organization/3ci	3CI	http://www.3ci.bz.it/en-US/Products.html	|Biotechnology|	Biotechnology	1,33,620	operating	
198	/organization/3clickemr-corporation	3ClickEMR Corporation	http://www.3clickemr.com	|Health Care|iPad|Consumer Electronics|Electro...	Electronic Health Records	5,00,000	closed	
199	/organization/3clogic	3CLogic	http://www.3clogic.com	|Customer Service|Software|	Customer Service	57,40,000	operating	

150 rows × 39 columns

```
# Checking for missing values and basic statistics of the dataset
missing_values = dt_mrg.isnull().sum()
statistics = dt_mrg.describe(include='all')
print(f"{missing_values}\n{'-'*20}\n{statistics}")
```

```
permalink                   0
name                        0
homepage_url                3
category_list            3964
 market                  3971
 funding_total_usd          0
status                   1314
country_code             5275
state_code              19279
region                   5275
city                     6118
funding_rounds              0
founded_at              10886
founded_month           10958
founded_quarter         10958
founded_year            10958
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  \
count                                           49442
unique                                          49436
top     /organization/treasure-valley-urology-services
freq                                                4
mean                                              NaN
std                                               NaN
min                                               NaN
25%                                               NaN
50%                                               NaN
75%                                               NaN
max                                               NaN

                                  name            homepage_url  \
count                            49442                   49439
unique                           49346                   49274
top     Treasure Valley Urology Services  https://tvurology.com/
```

Missing data treatment

Using :- Imputation : Mean/Median/Mode imputation

Interpolation : no need as used in time series

```
## homepage_url 3
dt_mrg[dt_mrg["homepage_url"].isna()]
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code
14	/organization/1-618-technology	1.618 Technology	NaN	|Real Estate|	Real Estate	-	operating	USA	FL
22400	/organization/kaldoora	Kaldoora	NaN	NaN	NaN	7,80,000	operating	USA	CA
48287	/organization/xochitl-so-shee-gold-mines	Xochitl SoShee Gold mines	NaN	|Manufacturing|	Manufacturing	-	operating	USA	CA

3 rows × 39 columns

🌟 🌟 🌟 🌟 🌟

dt_mrg["homepage_url"].dropna(inplace=True) inplace won't work here bcs

The dropna() function doesn't work in place when applied to a Series (like dt_mrg["homepage_url"]). Instead, it returns a new Series with the missing values removed

> **because dropna() is designed to work on DataFrames, not on Series.**

```
dt_mrg = dt_mrg.dropna(subset=["homepage_url"])
```

```
dt_mrg[dt_mrg["homepage_url"].isna()]
```

permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code	region	...	secondary_marke

0 rows × 39 columns

```
## category_list 3964 nulls
dt_mrg[['name',"category_list"," market "]][dt_mrg['category_list'].isna()]
```

	name	category_list	market
48	12Bis	NaN	NaN
61	16Lab Inc.	NaN	NaN
72	1d4 Pty	NaN	NaN
85	1RP Media	NaN	NaN
88	1st Choice Lawn Care	NaN	NaN
...	...	...	...
49360	Zuffle	NaN	NaN
49382	zumodrive	NaN	NaN
49393	Zursh	NaN	NaN
49401	Zuzher	NaN	NaN
49406	Zwamy	NaN	NaN

3963 rows × 3 columns

```
dt_mrg["category_list"][dt_mrg['category_list'].isna()] = 'Not Available'
```

Show hidden output

```
## category_list 3964 nulls
dt_mrg[" market "][dt_mrg[' market '].isna()] = dt_mrg['name'].values
```

Show hidden output

```
dt_mrg[dt_mrg[' market '].isin(["12Bis", "16Lab Inc.", "1d4 Pty", "1RP Media", "1st Choice Lawn Care"])]
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code
48	/organization/12bis	12Bis	http://12bis.com	Not Available	12Bis	1,30,636	operating	FRA	NaN
61	/organization/16lab-inc-	16Lab Inc.	http://www.16lab.net	Not Available	16Lab Inc.	-	operating	NaN	NaN
72	/organization/1d4-pty	1d4 Pty	http://www.immortaloutdoors.com	Not Available	1d4 Pty	40,000	operating	NOR	NaN
85	/organization/1rp-media	1RP Media	https://onerpm.com/	Not Available	1RP Media	-	operating	USA	CA
88	/organization/1st-choice-lawn-care	1st Choice Lawn Care	https://1stchoicelawncare.com/	Not Available	1st Choice Lawn Care	-	operating	USA	LA

5 rows × 39 columns

```
dt_mrg['status'].unique()
```

```
array(['acquired', 'operating', nan, 'closed'], dtype=object)
```

```
dt_mrg['status'].fillna('Unknown',inplace = True)
```

Show hidden output

```
dt_mrg['status'].unique()
```

```
array(['acquired', 'operating', 'Unknown', 'closed'], dtype=object)
```

```
# Display all columns
pd.set_option('display.max_columns', None)
dt_mrg[(dt_mrg['country_code'].isna()) & (dt_mrg['state_code'].isna()) &(dt_mrg['region'].isna()) &(dt_mrg['city'].isna())]
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	statu
7	/organization/0-6-com	06.com	http://www.0-6.com	|Curated Web|	Curated Web	20,00,000	operatir
18	/organization/1000-markets	1000 Markets	http://www.1000markets.com	|Marketplaces|Art|E-Commerce|	Marketplaces	5,00,000	acquire
24	/organization/100du-tv	100du.tv	http://www.100du.com	|Hospitality|	Hospitality	30,00,000	operatir
31	/organization/10sec	10sec	http://10s.ec/	|Social Commerce|E-Commerce|Mobile Commerce|	Mobile Commerce	16,00,000	operatir
36	/organization/115-network-disks	115 network disks	http://www.115.com/	|Education|	Education	-	operatir
...	...	...	...	...	...	...	
49393	/organization/zursh-2	Zursh	http://www.zursh.com/	Not Available	Zursh	25,000	operatir
49401	/organization/zuzher	Zuzher	http://www.zuzher.com/	Not Available	Zuzher	15,698	operatir
49411	/organization/zwoor-com	zwoor.com	http://www.zwoor.com	|Meeting Software|Polling|Surveys|Android|iPho...	Polling	1,00,000	operatir
49425	/organization/zympi	Zympi	http://www.zympi.com	|Networking|Web Hosting|Mobile|	Web Hosting	-	operatir
49440	/organization/a-list-games	alist games	http://www.alistgames.com	|Games|	Games	93,00,000	operatir

5275 rows × 39 columns

country_code 5275

state_code 19279

region 5275

city 6118

so i can impute some information state_code as well as city n but i need to fetch country code and region from outside (api/scraping) bcs there is no supportive data to do group by.

```
dt_mrg['country_code'][dt_mrg['country_code'].isna()] = 'Unavailable'
```

Show hidden output

```
dt_mrg['region'].fillna('unavailable',inplace=True)
```

Show hidden output

```
dt_mrg[dt_mrg['state_code'].isna()]
```

	permalink	name	homepage_url	category_list	market	funding_tota
2	/organization/rock-your-paper	Rock Your Paper	http://www.rockyourpaper.org	|Publishing|Education|	Publishing	
3	/organization/in-touch-network	InTouch Network	http://www.InTouchNetwork.com	|Electronics|Guides|Coffee|Restaurants|Music|i...	Electronics	15,
6	/organization/fox-networks	.Fox Networks	http://www.dotfox.com	|Advertising|	Advertising	49,
7	/organization/0-6-com	06.com	http://www.0-6.com	|Curated Web|	Curated Web	20,
9	/organization/01games-technology	01Games Technology	http://www.01games.hk/	|Games|	Games	
...	...	...	...	...	...	
49436	/organization/zytoprotec	Zytoprotec	http://www.zytoprotec.com	|Biotechnology|	Biotechnology	26,
49437	/organization/zzish	Zzish	http://www.zzish.com	|Analytics|Gamification|Developer APIs|iOS|And...	Education	3,
49438	/organization/zznode-science-and-technology-co...	ZZNode Science and Technology	http://www.zznode.com	|Enterprise Software|	Enterprise Software	15,
49439	/organization/zzzzapp-com	Zzzzapp Wireless ltd.	http://www.zzzzapp.com	|Web Development|Advertising|Wireless|Mobile|	Web Development	
49440	/organization/a-list-games	alist games	http://www.alistgames.com	|Games|	Games	93,

19279 rows × 39 columns

```
dt_mrg.groupby(['country_code', 'region', 'city'])['state_code']
```

```
<pandas.core.groupby.generic.SeriesGroupBy object at 0x78efdaca5060>
```

```
dt_mrg['state_code'].fillna(df.groupby(['country_code', 'region', 'city'])['state_code'].transform('first')).isna().sum()
```

```
18443
```

```
dt_mrg['state_code'].fillna(df.groupby(['country_code','state_code', 'region', 'city'])['state_code'].transform('first')).isna().sum()
```

```
18443
```

lets just treat city first

```
#dt_mrg[dt_mrg['city'].isna()].where(dt_mrg['country_code']=='Unavailable') instead
dt_mrg[(dt_mrg['region'] != 'unavailable') & (dt_mrg['city'].isna())]
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_c
52	/organization/139shop	139shop	http://www.139shop.com	|E-Commerce|	E-Commerce	-	operating	C
56	/organization/phoneuser-network	159.com	http://www.159.com	|Software|	Software	62,04,822	operating	C
72	/organization/1d4-pty	1d4 Pty	http://www.immortaloutdoors.com	Not Available	1d4 Pty	40,000	operating	N
125	/organization/24tidy	24tidy	http://24tidy.com	|Services|	Services	18,56,561	operating	C
180	/organization/360shop	360SHOP	http://www.360shop.com.cn	|Enterprise Software|	Enterprise Software	16,23,640	operating	C
...	...	...	...	...	...	...	...	
49107	/organization/zhijiang-jonway-automobile	Zhijiang Jonway Automobile	http://www.jonwayauto.com	|Automotive|	Automotive	2,90,30,000	acquired	C
49111	/organization/zhitu	Zhitu	http://www.imzhitu.com	|Photography|	Photography	2,40,727	operating	C
49121	/organization/zhui-xin	Zhui Xin	http://zhui.cn	|Software|	Software	16,23,640	operating	C
49358	/organization/zubka	Zubka	http://www.zubka.com	|Consulting|	Consulting	38,96,936	closed	(
49391	/organization/zurn-international-e-commerce-co...	Zurn	http://www.zurn.com	|Social Media|	Social Media	-	operating	C

843 rows × 39 columns

```
dt_mrg.loc[(dt_mrg['region'] != 'unavailable') & (dt_mrg['city'].isna()), 'city'] = dt_mrg['region']
```

```
dt_mrg['city'].isna().sum()
```

```
5275
```

before null values in city was 6118 now 5275.

```
# treat rest city col. nan values
```

```
dt_mrg['city'].fillna('Unknown',inplace = True)
```

Show hidden output

```
dt_mrg['city'].isna().sum()
```

```
0
```

```
dt_mrg['state_code'].isna().sum()
```

```
19279
```

```
#dt_mrg.groupby(['country_code','state_code', 'region', 'city'])['state_code'].apply(list)
## can't have state_code inside bcs want to find missing state_code so instead do
dt_mrg.groupby(['country_code','region','city']).apply(list)
```

```
<ipython-input-116-af83cbc69bac>:3: DeprecationWarning: DataFrameGroupBy.apply operated on the grouping columns. This behavior is deprec
  dt_mrg.groupby(['country_code','region','city']).apply(list)
```

			0
country_code	**region**	**city**	
ALB	**Tirana**	**Tirana**	[permalink, name, homepage_url, category_list,...
ARE	**Abu Dhabi**	**Abu Dhabi**	[permalink, name, homepage_url, category_list,...
	Dubai	**Dubai**	[permalink, name, homepage_url, category_list,...
ARG	**ARG - Other**	**ARG - Other**	[permalink, name, homepage_url, category_list,...
		Finca Elisa	[permalink, name, homepage_url, category_list,...
...	**...**	**...**	...
ZAF	**Johannesburg**	**Midrand**	[permalink, name, homepage_url, category_list,...
		Pretoria	[permalink, name, homepage_url, category_list,...
		Randburg	[permalink, name, homepage_url, category_list,...
	ZAF - Other	**Bloemfontein**	[permalink, name, homepage_url, category_list,...
ZWE	**Harare**	**Harare**	[permalink, name, homepage_url, category_list,...

4815 rows × 1 columns

dtype: object

---

group by logic also not working somehow

```
# def fill_missing_state_code(group):
#     # Get the unique non-null state codes in the group
#     state_codes = group['state_code'].dropna().unique()
#     if len(state_codes) == 1:  # If there's only one unique state_code
#         group['state_code'] = group['state_code'].fillna(state_codes[0])
#     return group
```

```
# dt_mrg = dt_mrg.groupby(['country_code', 'region', 'city'], group_keys=False).apply(fill_missing_state_code)
```

```
dt_mrg['state_code'].fillna(dt_mrg['region'], inplace=True)
```

Show hidden output

```
dt_mrg['founded_at'] = dt_mrg['founded_at'].fillna(dt_mrg['first_funding_at'])
```

```
dt_mrg['founded_month'] = dt_mrg['founded_month'].fillna(dt_mrg['founded_at'].str.slice(0,7))
```

```
dt_mrg['founded_year'] = dt_mrg['founded_year'].fillna(dt_mrg['founded_at'].str.slice(0,4))
```

```
from dateutil import parser

def fill_quarter(date_str, quarter_str):
    if pd.isnull(quarter_str) or quarter_str == '':
        date = parser.parse(date_str)
        quarter = (date.month - 1) // 3 + 1
        return f"{date.year}-Q{quarter}"
    else:
        return quarter_str

dt_mrg['founded_quarter'] = dt_mrg.apply(lambda row: fill_quarter(row['founded_at'], row['founded_quarter']), axis=1)
```

```
dt_mrg['founded_quarter']
```

	founded_quarter
0	2012-Q2
1	2010-Q2
2	2012-Q4
3	2011-Q2
4	2014-Q1
...	...
49437	2013-Q1
49438	2012-Q2
49439	2012-Q2
49440	2011-Q4
49441	1999-Q1

49439 rows × 1 columns

dtype: object

```
dt_mrg.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

Hence all missing values been treated.

round_B 0

Data cleaning continue...

round_D 0

```
# 🌟 🌟 🌟 🌟 🌟

def check_dtype_consistency(dt):
    result = dt.apply(lambda x: pd.Series({
        'dtype_consistent': x.apply(type).nunique() == 1,
        'null_count': x.isnull().sum(),
        'unique_dtypes': x.apply(type).unique()
    }))
    return result

rslt = check_dtype_consistency(dt_mrg)
```

```
rslt.T
```

	dtype_consistent	null_count	unique_dtypes
permalink	True	0	[<class 'str'>]
name	True	0	[<class 'str'>]
homepage_url	True	0	[<class 'str'>]
category_list	True	0	[<class 'str'>]
market	True	0	[<class 'str'>]
funding_total_usd	True	0	[<class 'str'>]
status	True	0	[<class 'str'>]
country_code	True	0	[<class 'str'>]
state_code	True	0	[<class 'str'>]
region	True	0	[<class 'str'>]
city	True	0	[<class 'str'>]
funding_rounds	True	0	[<class 'float'>]
founded_at	True	0	[<class 'str'>]
founded_month	True	0	[<class 'str'>]
founded_quarter	True	0	[<class 'str'>]
founded_year	False	0	[<class 'float'>, <class 'str'>]
first_funding_at	True	0	[<class 'str'>]
last_funding_at	True	0	[<class 'str'>]
seed	True	0	[<class 'float'>]
venture	True	0	[<class 'float'>]
equity_crowdfunding	True	0	[<class 'float'>]
undisclosed	True	0	[<class 'float'>]
convertible_note	True	0	[<class 'float'>]
debt_financing	True	0	[<class 'float'>]
angel	True	0	[<class 'float'>]
grant	True	0	[<class 'float'>]
private_equity	True	0	[<class 'float'>]
post_ipo_equity	True	0	[<class 'float'>]
post_ipo_debt	True	0	[<class 'float'>]
secondary_market	True	0	[<class 'float'>]
product_crowdfunding	True	0	[<class 'float'>]

```
dt_mrg.reset_index(drop=True,inplace=True)
```

round_B	True	0	[<class 'float'>]

```
dt_mrg[dt_mrg['round_A'].apply(lambda x: isinstance(x, str))]
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code	region	city	funding_rounds

round_F	True	0	[<class 'float'>]

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

```
rslt = check_dtype_consistency(dt_mrg)
rslt.T
```

	dtype_consistent	null_count	unique_dtypes
permalink	True	0	[<class 'str'>]
name	True	0	[<class 'str'>]
homepage_url	True	0	[<class 'str'>]
category_list	True	0	[<class 'str'>]
market	True	0	[<class 'str'>]
funding_total_usd	True	0	[<class 'str'>]
status	True	0	[<class 'str'>]
country_code	True	0	[<class 'str'>]
state_code	True	0	[<class 'str'>]
region	True	0	[<class 'str'>]
city	True	0	[<class 'str'>]
funding_rounds	True	0	[<class 'float'>]
founded_at	True	0	[<class 'str'>]
founded_month	True	0	[<class 'str'>]
founded_quarter	True	0	[<class 'str'>]
founded_year	True	0	[<class 'int'>]
first_funding_at	True	0	[<class 'str'>]
last_funding_at	True	0	[<class 'str'>]
seed	True	0	[<class 'float'>]
venture	True	0	[<class 'float'>]
equity_crowdfunding	True	0	[<class 'float'>]
undisclosed	True	0	[<class 'float'>]
convertible_note	True	0	[<class 'float'>]
debt_financing	True	0	[<class 'float'>]
angel	True	0	[<class 'float'>]
grant	True	0	[<class 'float'>]
private_equity	True	0	[<class 'float'>]
post_ipo_equity	True	0	[<class 'float'>]
post_ipo_debt	True	0	[<class 'float'>]
secondary_market	True	0	[<class 'float'>]
product_crowdfunding	True	0	[<class 'float'>]

Data is consistent now

round_B	True	0	[<class 'float'>]

```
## check data inconsistency, data with 'Unknown' str
dt_mrg[dt_mrg['round_A'].apply(lambda x: isinstance(x, str))]
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code	region	city	funding_rounds

Data Validation

```
dt_mrg.shape
```

```
(49439, 39)
```

```
dt_mrg.dtypes
```

	0
permalink	object
name	object
homepage_url	object
category_list	object
market	object
funding_total_usd	object
status	object
country_code	object
state_code	object
region	object
city	object
funding_rounds	float64
founded_at	object
founded_month	object
founded_quarter	object
founded_year	int64
first_funding_at	object
last_funding_at	object
seed	float64
venture	float64
equity_crowdfunding	float64
undisclosed	float64
convertible_note	float64
debt_financing	float64
angel	float64
grant	float64
private_equity	float64
post_ipo_equity	float64
post_ipo_debt	float64
secondary_market	float64
product_crowdfunding	float64

```
dt_mrg.describe(include='all').T
```

	count	unique	top	freq	mean	std	min	25%	50%	75%	
permalink	49439	49433	/organization/treasure-valley-urology-services	4	NaN	NaN	NaN	NaN	NaN	NaN	
name	49439	49343	Spire	4	NaN	NaN	NaN	NaN	NaN	NaN	
homepage_url	49439	49274	https://tvurology.com/	4	NaN	NaN	NaN	NaN	NaN	NaN	
category_list	49439	16676	Not Available	3963	NaN	NaN	NaN	NaN	NaN	NaN	
market	49439	4718	Software	4620	NaN	NaN	NaN	NaN	NaN	NaN	
funding_total_usd	49439	14617	-	8530	NaN	NaN	NaN	NaN	NaN	NaN	
status	49439	4	operating	41830	NaN	NaN	NaN	NaN	NaN	NaN	
country_code	49439	116	USA	28792	NaN	NaN	NaN	NaN	NaN	NaN	
state_code	49439	865	CA	9915	NaN	NaN	NaN	NaN	NaN	NaN	
region	49439	1090	SF Bay Area	6802	NaN	NaN	NaN	NaN	NaN	NaN	
city	49439	4315	Unknown	5275	NaN	NaN	NaN	NaN	NaN	NaN	
funding_rounds	49439.0	NaN	NaN	NaN	1.696252	1.294241	1.0	1.0	1.0	2.0	
founded_at	49439	3979	2012-01-01	2262	NaN	NaN	NaN	NaN	NaN	NaN	
founded_month	49439	486	2012-01	2504	NaN	NaN	NaN	NaN	NaN	NaN	
founded_quarter	49439	278	2012-Q1	3308	NaN	NaN	NaN	NaN	NaN	NaN	
founded_year	49439.0	NaN	NaN	NaN	2007.736584	20.050492	1.0	2006.0	2010.0	2012.0	
first_funding_at	49439	3914	2012-01-01	468	NaN	NaN	NaN	NaN	NaN	NaN	
last_funding_at	49439	3657	2013-01-01	387	NaN	NaN	NaN	NaN	NaN	NaN	
seed	49439.0	NaN	NaN	NaN	217329.047715	1056974.556024	0.0	0.0	0.0	25000.0	13
venture	49439.0	NaN	NaN	NaN	7500910.41801	28470853.194594	0.0	0.0	0.0	5000000.0	235
equity_crowdfunding	49439.0	NaN	NaN	NaN	6163.197455	199902.795474	0.0	0.0	0.0	0.0	2
undisclosed	49439.0	NaN	NaN	NaN	130218.647141	2981373.557515	0.0	0.0	0.0	0.0	29
convertible_note	49439.0	NaN	NaN	NaN	23363.625619	1432031.255253	0.0	0.0	0.0	0.0	30
debt_financing	49439.0	NaN	NaN	NaN	1888102.92332	138203168.837339	0.0	0.0	0.0	0.0	3007
angel	49439.0	NaN	NaN	NaN	65417.656324	658284.197147	0.0	0.0	0.0	0.0	6
grant	49439.0	NaN	NaN	NaN	162841.982787	5612031.289242	0.0	0.0	0.0	0.0	75
private_equity	49439.0	NaN	NaN	NaN	2074243.792633	31671994.34066	0.0	0.0	0.0	0.0	350
post_ipo_equity	49439.0	NaN	NaN	NaN	608861.333886	26783209.780422	0.0	0.0	0.0	0.0	470
post_ipo_debt	49439.0	NaN	NaN	NaN	443427.004612	34281342.872975	0.0	0.0	0.0	0.0	580
secondary_market	49439.0	NaN	NaN	NaN	38455.144522	3864421.539846	0.0	0.0	0.0	0.0	68
product_crowdfunding	49439.0	NaN	NaN	NaN	7074.984093	438213.257081	0.0	0.0	0.0	0.0	7

```
pd.reset_option('display.max_rows')
dt_mrg.reset_index(drop=True, inplace=True)
```

round_B	49439.0	NaN	NaN	NaN	1492860.957422	7472631.8928	0.0	0.0	0.0	0.0	54

```
dt_mrg
```

	permalink	name	homepage_url	category_list	market	funding_to
0	/organization/waywire	waywire	http://www.waywire.com	|Entertainment|Politics|Social Media|News|	News	
1	/organization/tv-communications	TV Communications	http://enjoyandtv.com	|Games|	Games	
2	/organization/rock-your-paper	Rock Your Paper	http://www.rockyourpaper.org	|Publishing|Education|	Publishing	
3	/organization/in-touch-network	InTouch Network	http://www.InTouchNetwork.com	|Electronics|Guides|Coffee|Restaurants|Music|i...	Electronics	
4	/organization/r-ranch-and-mine	R Ranch and Mine	https://r-ranchca.com/	|Tourism|Entertainment|Games|	Tourism	
...	...	...	...	...	...	
49434	/organization/zzish	Zzish	http://www.zzish.com	|Analytics|Gamification|Developer APIs|iOS|And...	Education	
49435	/organization/zznode-science-and-technology-co...	ZZNode Science and Technology	http://www.zznode.com	|Enterprise Software|	Enterprise Software	
49436	/organization/zzzzapp-com	Zzzzapp Wireless ltd.	http://www.zzzzapp.com	|Web Development|Advertising|Wireless|Mobile|	Web Development	
49437	/organization/a-list-games	alist games	http://www.alistgames.com	|Games|	Games	
49438	/organization/x	x1	http://www.xplusone.com/	|Enterprise Software|	Enterprise Software	4,

49439 rows × 39 columns

```
dt_mrg[dt_mrg['permalink'].duplicated()].sort_values('name')
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_c
24026	/organization/lightex-ltd-	Lightex Ltd.	https://www.lightexhire.com/about	Not Available	Lightex Ltd.	-	operating	Unavaila
24028	/organization/lightex-ltd--2	Lightex Ltd.	https://www.lightexhire.com/about	Not Available	Lightex Ltd.	5,00,000	operating	Unavaila
33940	/organization/prysm	Prysm	http://www.prysm.com	|Displays|Hardware + Software|	Displays	29,30,80,123	operating	U
44034	/organization/treasure-valley-urology-services	Treasure Valley Urology Services	https://tvurology.com/	|Biotechnology|	Biotechnology	3,32,194	operating	U
44035	/organization/treasure-valley-urology-services	Treasure Valley Urology Services	https://tvurology.com/	Not Available	Treasure Valley Urology Services	3,32,194	operating	U
44036	/organization/treasure-valley-urology-services	Treasure Valley Urology Services	https://tvurology.com/	Not Available	Treasure Valley Urology Services	3,32,194	operating	U

Double-click (or enter) to edit

```
dt_mrg['permalink'].drop_duplicates(keep=False)
```

Show hidden output

```
dt_mrg[dt_mrg['permalink'].duplicated()]
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_c
24026	/organization/lightex-ltd-	Lightex Ltd.	https://www.lightexhire.com/about	Not Available	Lightex Ltd.	-	operating	Unavaila
24028	/organization/lightex-ltd--2	Lightex Ltd.	https://www.lightexhire.com/about	Not Available	Lightex Ltd.	5,00,000	operating	Unavaila
33940	/organization/prysm	Prysm	http://www.prysm.com	|Displays|Hardware + Software|	Displays	29,30,80,123	operating	U
44034	/organization/treasure-valley-urology-services	Treasure Valley Urology Services	https://tvurology.com/	|Biotechnology|	Biotechnology	3,32,194	operating	U
44035	/organization/treasure-valley-urology-services	Treasure Valley Urology Services	https://tvurology.com/	Not Available	Treasure Valley Urology Services	3,32,194	operating	U
44036	/organization/treasure-valley-urology-services	Treasure Valley Urology Services	https://tvurology.com/	Not Available	Treasure Valley Urology Services	3,32,194	operating	U

```
dt_mrg[dt_mrg['name']=='Treasure Valley Urology Services']
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_co
44033	/organization/treasure-valley-urology-services	Treasure Valley Urology Services	https://tvurology.com/	|Biotechnology|	Biotechnology	3,32,194	operating	USA	
44034	/organization/treasure-valley-urology-services	Treasure Valley Urology Services	https://tvurology.com/	|Biotechnology|	Biotechnology	3,32,194	operating	USA	
44035	/organization/treasure-valley-urology-services	Treasure Valley Urology Services	https://tvurology.com/	Not Available	Treasure Valley Urology Services	3,32,194	operating	USA	
44036	/organization/treasure-valley-urology-services	Treasure Valley Urology Services	https://tvurology.com/	Not Available	Treasure Valley Urology Services	3,32,194	operating	USA	

```
dt_mrg = dt_mrg.drop(index=[39385,39386])
```

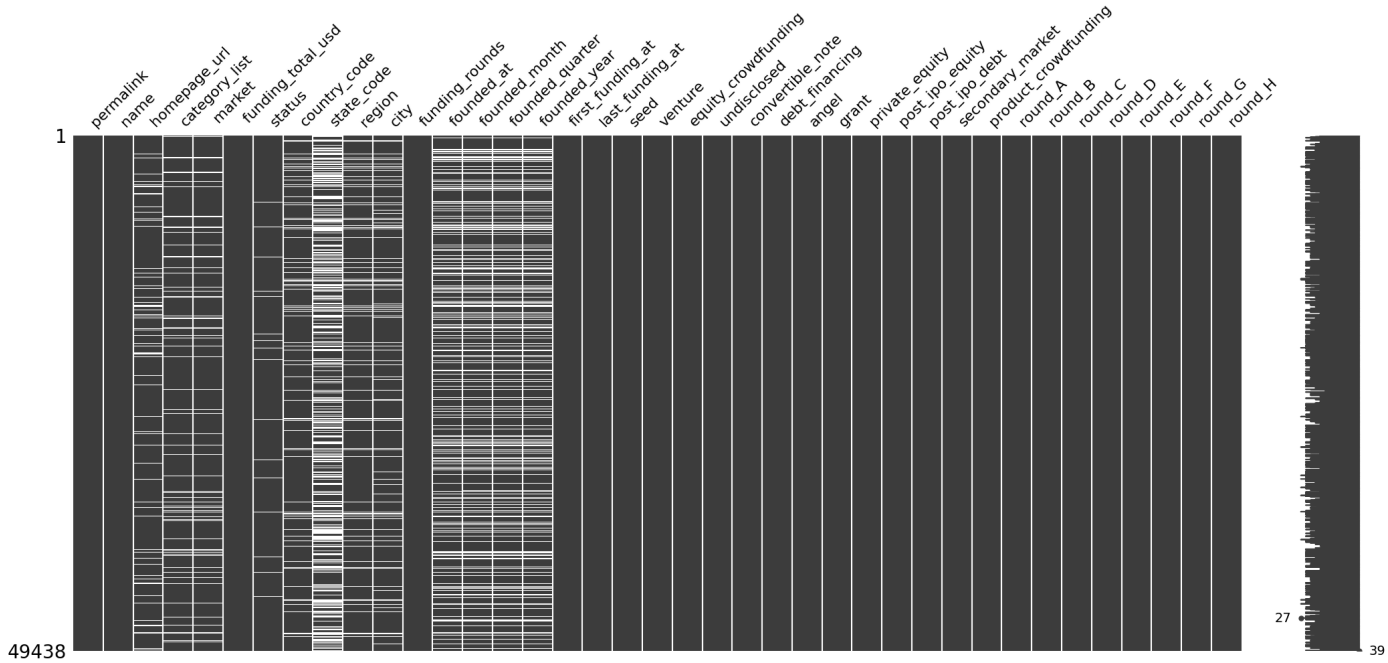
```
dt_mrg[dt_mrg['name']=='Treasure Valley Urology Services']
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_co
44033	/organization/treasure-valley-urology-services	Treasure Valley Urology Services	https://tvurology.com/	|Biotechnology|	Biotechnology	3,32,194	operating	USA	
44034	/organization/treasure-valley-urology-services	Treasure Valley Urology Services	https://tvurology.com/	|Biotechnology|	Biotechnology	3,32,194	operating	USA	
44035	/organization/treasure-valley-urology-services	Treasure Valley Urology Services	https://tvurology.com/	Not Available	Treasure Valley Urology Services	3,32,194	operating	USA	
44036	/organization/treasure-valley-urology-services	Treasure Valley Urology Services	https://tvurology.com/	Not Available	Treasure Valley Urology Services	3,32,194	operating	USA	

🌟🌟🌟 Visualize missing data using missingno btw dt and dt_mrg

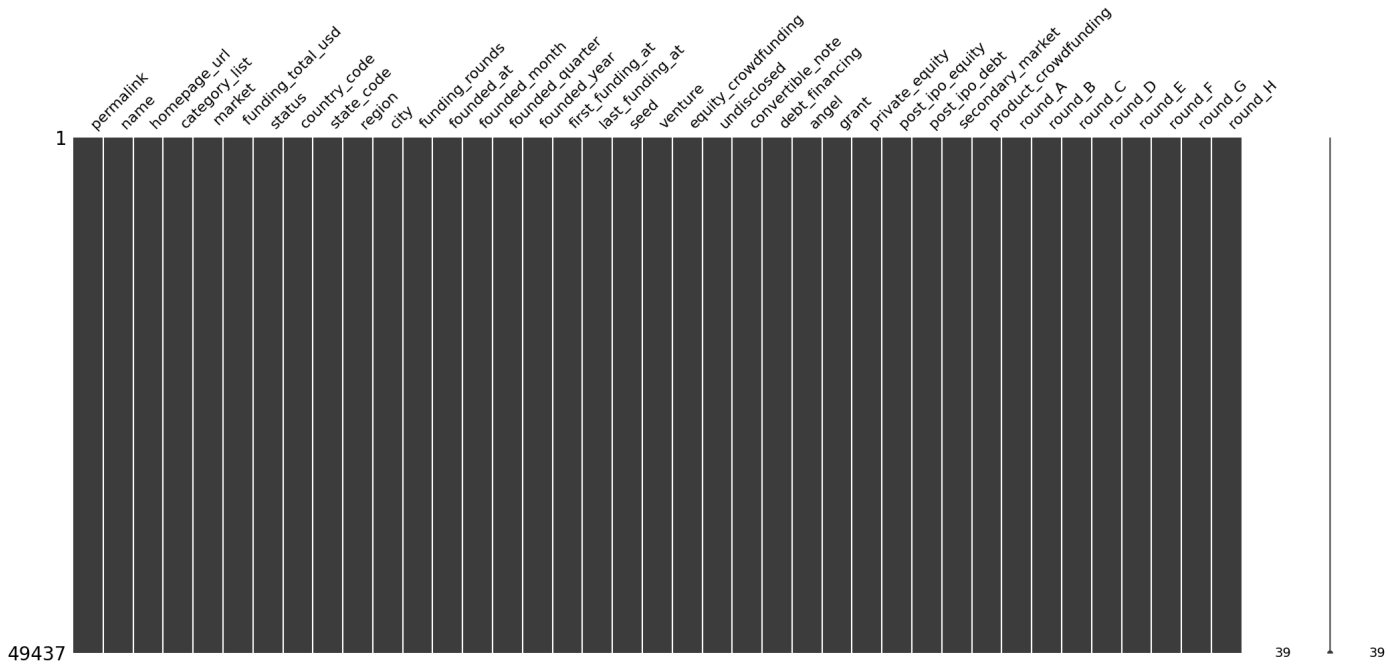
```
import missingno as msno
msno.matrix(dt)
```

```
<Axes: >
```




```
msno.matrix(dt_mrg)
```

```
<Axes: >
```



Quickly check for duplicate records once more at table level:

```
duplicates = dt_mrg[dt_mrg.duplicated()]
print(duplicates)
```

```
                                        permalink  \
24026                     /organization/lightex-ltd-
24028                    /organization/lightex-ltd--2
44034  /organization/treasure-valley-urology-services
44036  /organization/treasure-valley-urology-services

                                 name                      homepage_url  \
24026                     Lightex Ltd.  https://www.lightexhire.com/about
24028                     Lightex Ltd.  https://www.lightexhire.com/about
44034  Treasure Valley Urology Services             https://tvurology.com/
44036  Treasure Valley Urology Services             https://tvurology.com/

         category_list                          market   funding_total_usd   \
24026    Not Available                     Lightex Ltd.                -
24028    Not Available                     Lightex Ltd.           5,00,000
44034  |Biotechnology|                    Biotechnology           3,32,194
44036    Not Available  Treasure Valley Urology Services           3,32,194

           status country_code   state_code       region     city  \
24026  operating  Unavailable  unavailable  unavailable  Unknown
24028  operating  Unavailable  unavailable  unavailable  Unknown
44034  operating          USA           TX       Austin   Austin
44036  operating          USA           TX       Austin   Austin

       funding_rounds  founded_at founded_month founded_quarter  founded_year  \
24026             1.0  2011-04-01       2011-04         2011-Q2          2011
24028             1.0  2011-01-01       2011-01         2011-Q1          2011
44034             4.0  2004-01-01       2004-01         2004-Q1          2004
44036             1.0  2004-01-01       2004-01         2004-Q1          2004

      first_funding_at last_funding_at      seed   venture  \
24026       2011-04-01      2011-04-01       0.0       0.0
24028       2011-01-01      2011-01-01  500000.0       0.0
44034       2009-09-09      2014-05-05   45321.0  286873.0
44036       2014-07-18      2014-07-18   45321.0  286873.0

       equity_crowdfunding  undisclosed  convertible_note  debt_financing  \
```

```
24026                 0.0          0.0               0.0             0.0
24028                 0.0          0.0               0.0             0.0
44034                 0.0          0.0               0.0             0.0
44036                 0.0          0.0               0.0             0.0

       angel  grant  private_equity  post_ipo_equity  post_ipo_debt  \
24026    0.0    0.0            0.0              0.0            0.0
24028    0.0    0.0            0.0              0.0            0.0
44034    0.0    0.0            0.0              0.0            0.0
44036    0.0    0.0            0.0              0.0            0.0

       secondary_market  product_crowdfunding  round_A  round_B  round_C  \
24026               0.0                   0.0      0.0      0.0      0.0
24028               0.0                   0.0      0.0      0.0      0.0
44034               0.0                   0.0      0.0      0.0      0.0
44036               0.0                   0.0      0.0      0.0      0.0

       round_D  round_E  round_F  round_G  round_H
24026      0.0      0.0      0.0      0.0      0.0
24028      0.0      0.0      0.0      0.0      0.0
44034      0.0      0.0      0.0      0.0      0.0
```

Data Type Validation

Numeric columns contain numbers

```
dt_mrg.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')
```

```
numeric_columns = [' funding_total_usd ', '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']
for col in numeric_columns:
    print(f"{col} has non-numeric values:", dt_mrg[col].apply(lambda x: not isinstance(x, (int, float))).sum())
```

```
 funding_total_usd  has non-numeric values: 49437
seed has non-numeric values: 0
venture has non-numeric values: 0
equity_crowdfunding has non-numeric values: 0
undisclosed has non-numeric values: 0
convertible_note has non-numeric values: 0
debt_financing has non-numeric values: 0
angel has non-numeric values: 0
grant has non-numeric values: 0
private_equity has non-numeric values: 0
post_ipo_equity has non-numeric values: 0
post_ipo_debt has non-numeric values: 0
secondary_market has non-numeric values: 0
product_crowdfunding has non-numeric values: 0
round_A has non-numeric values: 0
round_B has non-numeric values: 0
round_C has non-numeric values: 0
round_D has non-numeric values: 0
round_E has non-numeric values: 0
round_F has non-numeric values: 0
round_G has non-numeric values: 0
round_H has non-numeric values: 0
```

```
dt_mrg[' funding_total_usd '] = pd.to_numeric(dt_mrg[' funding_total_usd '].str.replace(',', ''), errors='coerce')
```

chk Dates Format

```
print(dt_mrg[['founded_at', 'founded_month',
       'founded_quarter', 'founded_year', 'first_funding_at',
       'last_funding_at']].isna().sum())  # Check invalid dates
```

```
founded_at          0
founded_month       0
founded_quarter     0
founded_year        0
first_funding_at    0
last_funding_at     0
dtype: int64
```

Check Value Ranges and Uniqueness

```
# value Ranges
dt_mrg[' funding_total_usd '].min(),dt_mrg[' funding_total_usd '].max()
```

```
(1.0, 30079503000.0)
```

```
dt_mrg[dt_mrg[' funding_total_usd ']==14.0] ## interesting
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code	re
26507	/organization/miceit-co	MICEit.co	http://www.miceit.co	Not Available	MICEit.co	14.0	operating	JOR	Amman	An

```
dt_mrg[dt_mrg[' funding_total_usd ']==30079503000.0]
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_
45817	/organization/verizon	Verizon Communications	http://www.verizon.com/?lid=//global//residential	|Mobile|	Mobile	3.007950e+10	operating	USA	

```
## unique values
dt_mrg['country_code'].unique()  # Check unique categories
```

```
array(['USA', 'EST', 'GBR', 'ARG', 'Unavailable', 'HKG', 'CHL', 'DEU',
       'FRA', 'CHN', 'CAN', 'AUS', 'ROM', 'NLD', 'SWE', 'RUS', 'DNK',
       'IND', 'SGP', 'NOR', 'BEL', 'IRL', 'ITA', 'ISR', 'ESP', 'THA',
       'NZL', 'CZE', 'CHE', 'BRA', 'HUN', 'JPN', 'BWA', 'KOR', 'NGA',
       'FIN', 'TUR', 'CRI', 'PRT', 'TWN', 'KHM', 'COL', 'UKR', 'LTU',
       'ZAF', 'AUT', 'PHL', 'ISL', 'BGR', 'URY', 'HRV', 'KEN', 'MEX',
       'JOR', 'VNM', 'GHA', 'PER', 'POL', 'IDN', 'PAN', 'LVA', 'ALB',
       'UGA', 'LBN', 'GRC', 'ARE', 'PAK', 'EGY', 'SVK', 'LUX', 'MYS',
       'BHS', 'ARM', 'DZA', 'MDA', 'TUN', 'NIC', 'TZA', 'CYP', 'NPL',
       'BHR', 'CMR', 'SRB', 'SAU', 'CYM', 'BRN', 'SLV', 'ECU', 'MLT',
       'SVN', 'LAO', 'TTO', 'MAR', 'MMR', 'BGD', 'DOM', 'BMU', 'LIE',
       'MOZ', 'GTM', 'AZE', 'MCO', 'ZWE', 'UZB', 'OMN', 'BLR', 'JEY',
       'JAM', 'KWT', 'MUS', 'CIV', 'SOM', 'MKD', 'GIB', 'SYC', 'MAF'],
      dtype=object)
```

Outlier Detection

```
dt_mrg[['funding_rounds', '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']] =  dt_mrg[['funding_rounds', 'seed', 'venture', 'equity_crowd
        'private_equity', 'post_ipo_equity', 'post_ipo_debt','secondary_market', 'product_crowdfunding', 'round_A', 'round_B','round_C',
```

```
dt_mrg['founded_at'] = pd.to_datetime(dt_mrg['founded_at'], errors='coerce')
dt_mrg['founded_at'].info()
```

```
<class 'pandas.core.series.Series'>
Index: 49437 entries, 0 to 49438
Series name: founded_at
Non-Null Count  Dtype
```

```
--------------  -----
49432 non-null  datetime64[ns]
dtypes: datetime64[ns](1)
memory usage: 772.5 KB
```

```
dt_mrg[' funding_total_usd '].fillna(0, inplace=True)
```

Show hidden output

```
dt_mrg[dt_mrg['founded_at'].isna()]
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_co
38	/organization/fitfrnd-2	12 Labs	http://www.12labs.com/	|Personal Health|Health and Wellness|	Personal Health	0.0	operating	Unavailal
18500	/organization/harvard-university	Harvard University	http://harvard.edu	|Education|	Education	90000000.0	operating	U
36309	/organization/rotor	Rotor	http://rotor.ie/	|Developer Tools|Music|Video|	Video	0.0	operating	I
38070	/organization/shopboostr	Shopboostr	http://shopboostr.de/	|SaaS|Reviews and Recommendations|E-Commerce|	SaaS	0.0	operating	Unavailal
42916	/organization/the-urban-roosters	The Urban Roosters	http://www.urbanroosters.com	Not Available	The Urban Roosters	0.0	operating	Unavailal

```
dt[dt['name']=='Harvard University']
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code	reg
18501	/organization/harvard-university	Harvard University	http://harvard.edu	|Education|	Education	9,00,00,000	operating	USA	MA	Bos

```
dt[dt['name']=='Rotor']
```

	permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code	region	cit
36309	/organization/rotor	Rotor	http://rotor.ie/	|Developer Tools|Music|Video|	Video	-	operating	IRL	NaN	Dublin	Dubl

```
dt_mrg.loc[dt_mrg['name'] == 'Harvard University', :] = dt_mrg.loc[dt_mrg['name'] == 'Harvard University'].fillna(pd.Period('1636-09-08',
```

Show hidden output

```
dt_mrg.reset_index(inplace=True, drop=True)
```

```
dt_mrg.drop(dt_mrg.index[32543], inplace=True)
```

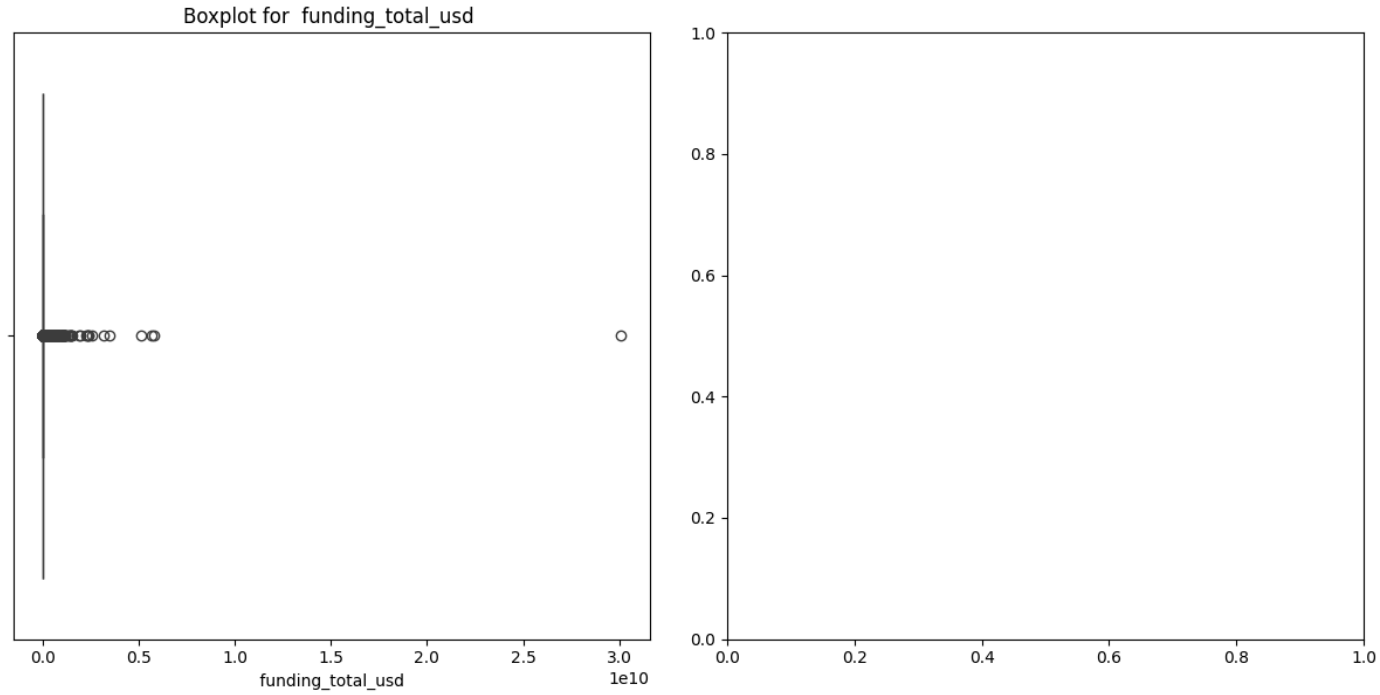
```
## filter numeric cols
dt_mrg.select_dtypes(include=['int64', 'float64']).columns
```

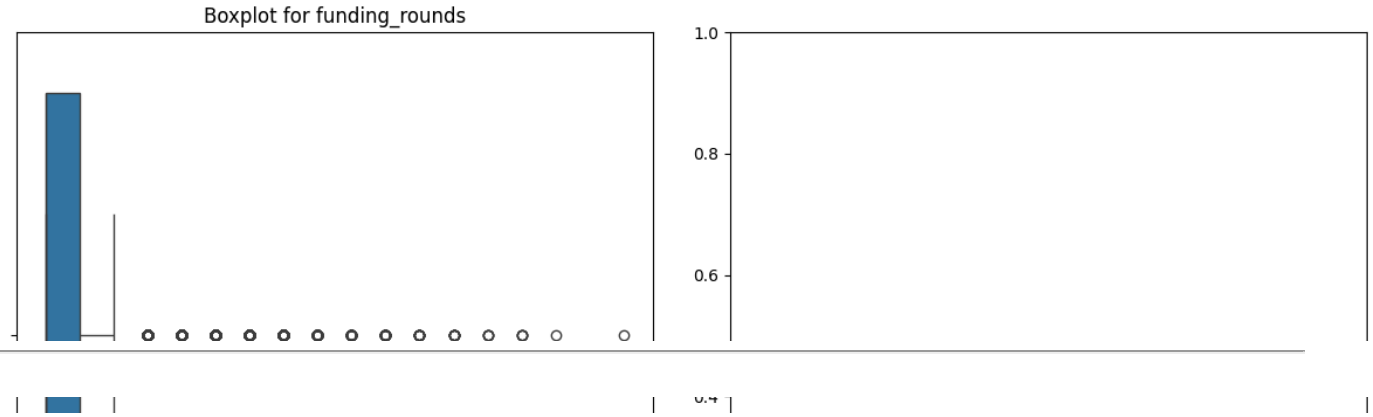
```
Index([' funding_total_usd ', 'funding_rounds', 'founded_year', 'seed',
       'venture', 'equity_crowdfunding', 'undisclosed', 'convertible_note',
       'debt_financing', 'angel', 'grant', 'private_equity', 'post_ipo_equity',
       'post_ipo_debt', 'secondary_market', 'product_crowdfunding', 'round_A',
       'round_B', 'round_C', 'round_D', 'round_E', 'round_F', 'round_G',
       'round_H'],
      dtype='object')
```

```
import matplotlib.pyplot as plt
import seaborn as sns

# Select numeric columns
numeric_columns = [' funding_total_usd ', 'funding_rounds', 'seed', 'venture',
       'equity_crowdfunding', 'undisclosed', 'convertible_note',
```

```
        'debt_financing', 'angel', 'grant', 'private_equity', 'post_ipo_equity',
        'post_ipo_debt', 'secondary_market', 'product_crowdfunding']

# Create subplot fig
fig, axes = plt.subplots(nrows=len(numeric_columns), ncols=2, figsize=(12, 6*len(numeric_columns)))


for i, column in enumerate(numeric_columns):
    # Boxplot for outlier detection
    sns.boxplot(dt_mrg[column], ax=axes[i, 0],orient='h')
    axes[i, 0].set_title(f'Boxplot for {column}')


 # lower and upper limits for outlier detection
Q1 = dt_mrg[' funding_total_usd '].quantile(0.25)
Q3 = dt_mrg[' funding_total_usd '].quantile(0.75)
IQR = Q3 - Q1
lower_limit = Q1 - 1.5 * IQR
upper_limit = Q3 + 1.5 * IQR



# Identify outliers
outliers = dt_mrg[' funding_total_usd '][(dt_mrg[' funding_total_usd '] < lower_limit) | (dt_mrg[' funding_total_usd '] > upper_limit)]
print(f'Outliers for  funding_total_usd{outliers}')


sns.scatterplot(x=dt_mrg[' funding_total_usd ']*len(outliers), y=outliers, ax=axes[i, 1])
axes[i, 1].set_title(f'Outliers for funding_total_usd')
# Layout so plots do not overlap
fig.tight_layout()

plt.show()
```

```
Outliers for  funding_total_usd27        35000000.0
36        20000000.0
50        66450000.0
62        84440319.0
66       200000000.0
             ...
49419     56200000.0
49425    866550786.0
49426     25000000.0
49428     34275015.0
49436     45000000.0
Name:  funding_total_usd , Length: 6959, dtype: float64
```





Consistency Checks

Cross-column Logic: Validate logical consistency between columns

```
dt_mrg[dt_mrg['last_funding_at'] < dt_mrg['first_funding_at']]
```

permalink	name	homepage_url	category_list	market	funding_total_usd	status	country_code	state_code	region	city	funding_r

Aggregate Validation: Ensure aggregates match expectations:

```
dt_mrg.head(5)
```

	permalink	name	homepage_url	category_list	market	funding_total_us
0	/organization/waywire	waywire	http://www.waywire.com	|Entertainment|Politics|Social Media|News|	News	1750000.
1	/organization/tv-communications	TV Communications	http://enjoyandtv.com	|Games|	Games	4000000.
2	/organization/rock-your-paper	Rock Your Paper	http://www.rockyourpaper.org	|Publishing|Education|	Publishing	40000.
3	/organization/in-touch-network	InTouch Network	http://www.InTouchNetwork.com	|Electronics|Guides|Coffee|Restaurants|Music|i...	Electronics	1500000.
4	/organization/r-ranch-and-mine	R Ranch and Mine	https://r-ranchca.com/	|Tourism|Entertainment|Games|	Tourism	60000.

Boxplot for venture

```
dt_mrg[['undisclosed','convertible_note','debt_financing',"angel","grant","private_equity","post_ipo_equity","post_ipo_debt",   "secondar
```

	0
undisclosed	[0.0, 4912393.0, 3248511.0, 60000000.0, 161054...
convertible_note	[0.0, 1750000.0, 1540000.0, 205000.0, 270000.0...
debt_financing	[0.0, 2050000.0, 1147879.0, 3000000.0, 77500.0...
angel	[0.0, 100000.0, 500000.0, 619494.0, 1200000.0,...
grant	[0.0, 20000.0, 225000.0, 3000000.0, 2500000.0,...
private_equity	[0.0, 200000000.0, 325000000.0, 85000000.0, 50...
post_ipo_equity	[0.0, 100000000.0, 736000000.0, 6100000.0, 983...
post_ipo_debt	[0.0, 150000000.0, 15000000.0, 35000000.0, 500...
secondary_market	[0.0, 7343789.0, 210000000.0, 126700000.0, 619...
product_crowdfunding	[0.0, 4545754.0, 929212.0, 1260000.0, 15003.0,...

dtype: object

```
dt_mrg = dt_mrg.rename(columns={' funding_total_usd ':'funding_total_usd'})
```

venture

```
dt_mrg['funding_total_usd'].name
```

```
'funding_total_usd'
```

```
dt_mrg[dt_mrg['funding_total_usd'] >= (dt_mrg['seed'] + dt_mrg['venture'] + dt_mrg['equity_crowdfunding'] + dt_mrg['product_crowdfunding'
```

	permalink	name	homepage_url	category_list	market	funding_to
0	/organization/waywire	waywire	http://www.waywire.com	|Entertainment|Politics|Social Media|News|	News	1
1	/organization/tv-communications	TV Communications	http://enjoyandtv.com	|Games|	Games	4
2	/organization/rock-your-paper	Rock Your Paper	http://www.rockyourpaper.org	|Publishing|Education|	Publishing	
3	/organization/in-touch-network	InTouch Network	http://www.InTouchNetwork.com	|Electronics|Guides|Coffee|Restaurants|Music|i...	Electronics	1
4	/organization/r-ranch-and-mine	R Ranch and Mine	https://r-ranchca.com/	|Tourism|Entertainment|Games|	Tourism	
...	...	...	...	...	...	
49432	/organization/zzish	Zzish	http://www.zzish.com	|Analytics|Gamification|Developer APIs|iOS|And...	Education	
49433	/organization/zznode-science-and-technology-co...	ZZNode Science and Technology	http://www.zznode.com	|Enterprise Software|	Enterprise Software	1
49434	/organization/zzzzapp-com	Zzzzapp Wireless ltd.	http://www.zzzzapp.com	|Web Development|Advertising|Wireless|Mobile|	Web Development	
49435	/organization/a-list-games	alist games	http://www.alistgames.com	|Games|	Games	9
49436	/organization/x	x1	http://www.xplusone.com/	|Enterprise Software|	Enterprise Software	45

49436 rows × 39 columns

Data Profiling

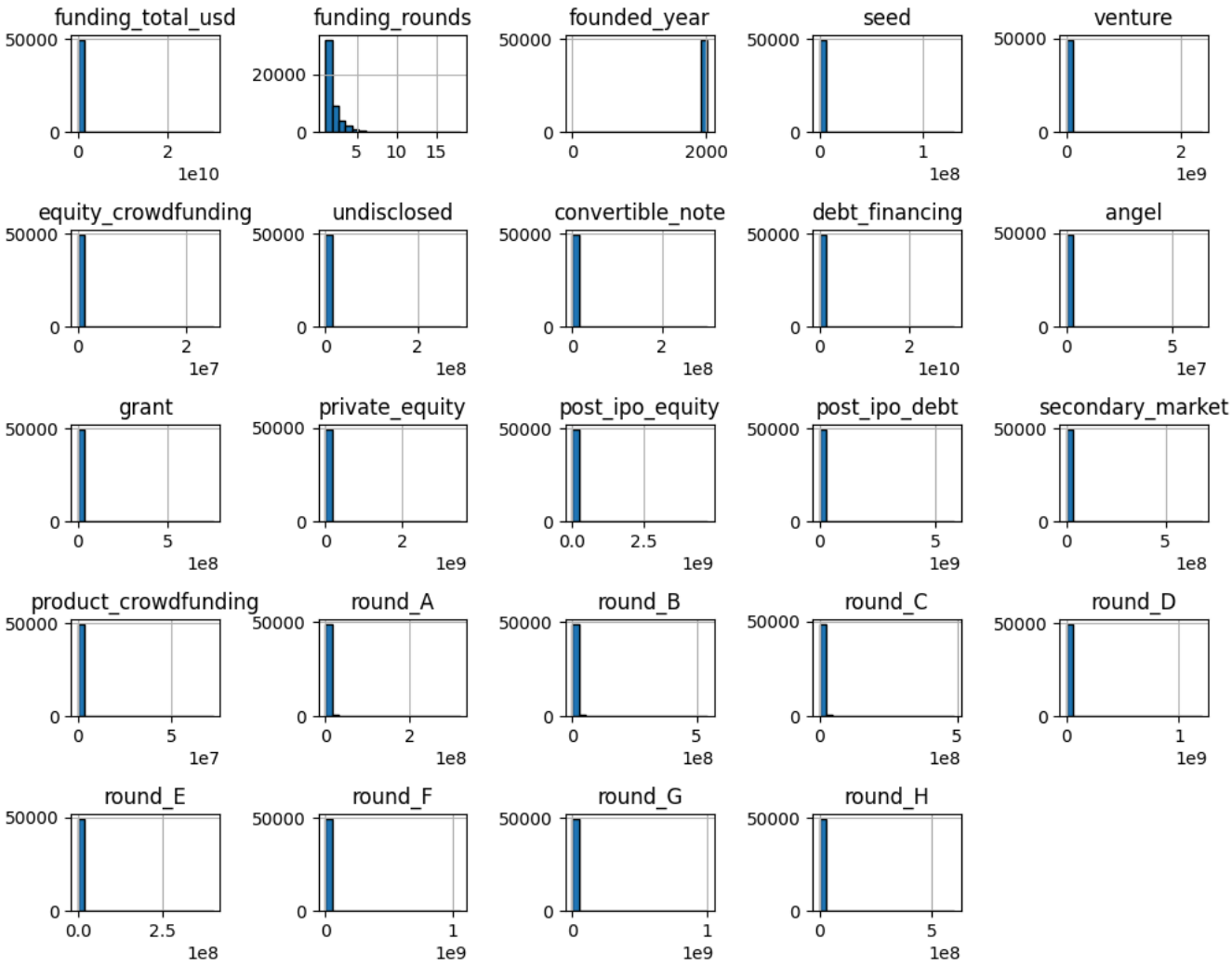
```
pip install pandas-profiling sweetviz
```

Show hidden output

```
import sweetviz
```

Data Distributions

```
dt_mrg.hist(figsize=(10, 8), bins=20, edgecolor='black')
plt.tight_layout()
plt.show()
```



Category Counts

```
dt_mrg.select_dtypes(include=['object']).columns
```

```
Index(['permalink', 'name', 'homepage_url', 'category_list', ' market ',
       'status', 'country_code', 'state_code', 'region', 'city', 'founded_at',
       'founded_month', 'founded_quarter', 'first_funding_at',
       'last_funding_at'],
      dtype='object')
```

```
dt_mrg['status'].value_counts()
```

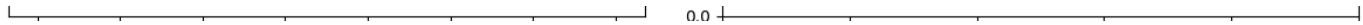
status	count
operating	41827
acquired	3692
closed	2603
Unknown	1314

dtype: int64

```
categorical_cols = dt_mrg.select_dtypes(include=['object']).columns
for col in categorical_cols:
    print(f" {'='*200}\n {col}:\n {'|'*200}", df[col].value_counts())
```

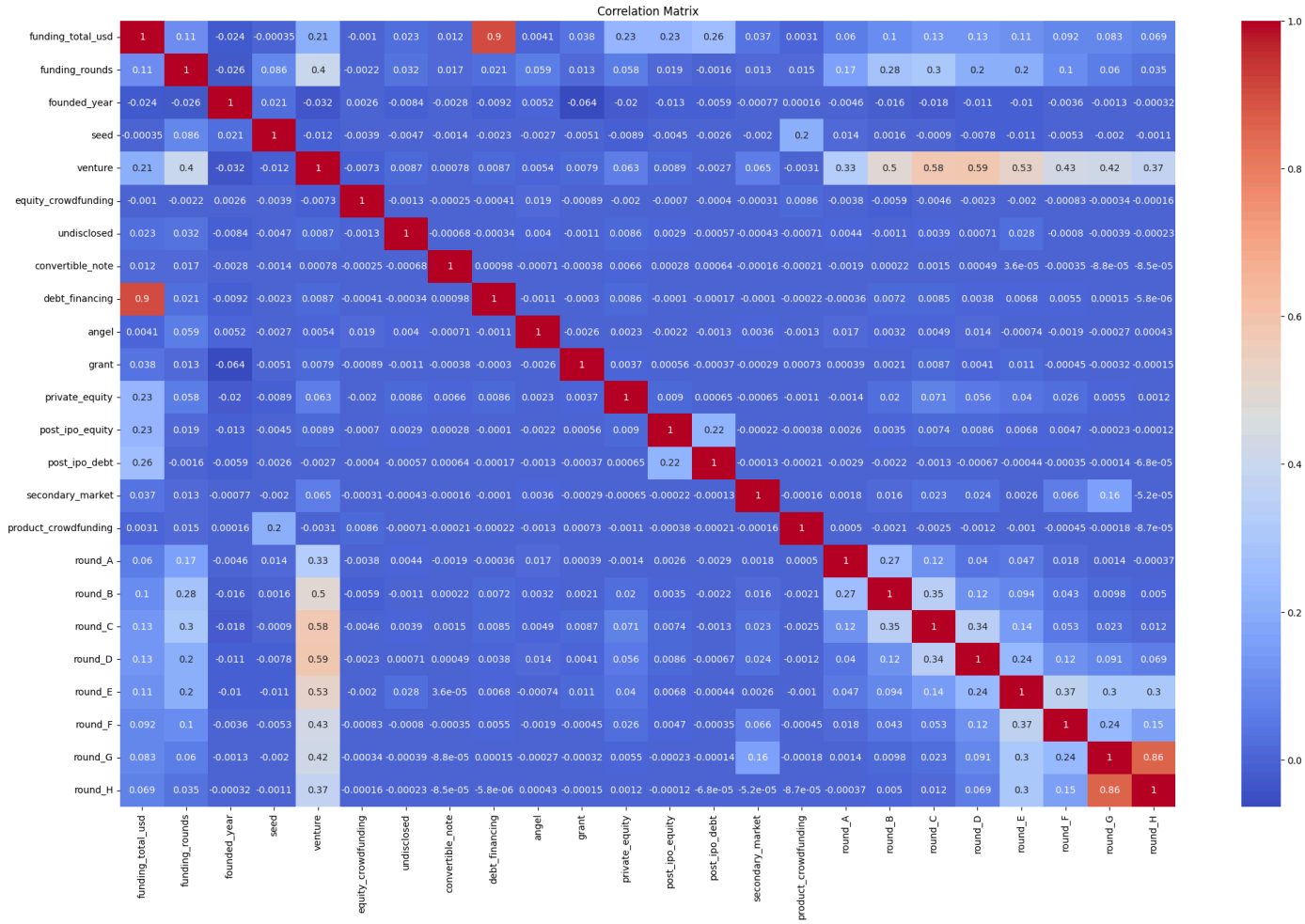
Show hidden output

Correlation Analysis




```
# Ensure no NaN values in 'funding_total_usd'
dt_mrg['funding_total_usd'] = dt_mrg['funding_total_usd'].fillna(0)

# Create pivot table for the heatmap
pivot = dt_mrg.pivot_table(
    values='funding_total_usd',
    index='region',
    columns='market',
    aggfunc='mean',
    fill_value=0
)

# Check if the pivot table is empty
if pivot.empty:
    print("The pivot table is empty. Check your data for inconsistencies.")
else:
    # Plot the heatmap
    plt.figure(figsize=(16, 8))
    sns.heatmap(
        pivot,
        cmap='coolwarm',
        linewidths=0.5,
        cbar_kws={'label': 'Average Funding (USD)'},
        xticklabels=True,
        yticklabels=True
    )
    plt.title('Funding Distribution Across Regions and Markets', fontsize=14)
    plt.xlabel('Market', fontsize=12)
    plt.ylabel('Region', fontsize=12)
    plt.xticks(rotation=45, fontsize=10)
    plt.yticks(fontsize=10)
    plt.tight_layout()
    plt.show()
```

```
plt.figure(figsize=(25, 15))
corr = dt_mrg.select_dtypes(include=['integer', 'float']).corr()
sns.heatmap(corr, annot=True, cmap='coolwarm')
plt.title("Correlation Matrix")
plt.show()
```



1. **Moderate to High Correlations:**
   - There are several pairs of variables exhibiting moderate to high correlations, indicating strong relationships between them. Some of these include:
     - `funding_total_usd` and `funding_rounds` (0.71)
     - `venture` and `round_D` (0.59)
     - `private_equity` and `post_ipo_equity` (0.55)
     - `round_B` and `round_C` (0.54)
   - These relationships suggest that companies with higher funding rounds tend to raise more money, and those receiving venture funding often proceed to later rounds like Series D. Private equity investment often precedes post-IPO equity.
2. **Weak Correlations:**
   - Many variables have weak or negligible correlations, indicating limited relationships. This is evident in the large number of cells with values close to zero.
   - This suggests that many variables are not strongly related to each other, or their relationships are complex and not captured by simple correlations.

Specific Trends

1. **Funding and Rounds:**
   - As mentioned, there's a strong positive correlation between `funding_total_usd` and `funding_rounds`. This implies that companies that raise more money tend to do so in multiple rounds.
2. **Investment Stages:**
   - The correlations between different funding rounds (e.g., `round_A`, `round_B`, etc.) suggest a progression in investment stages. Companies that successfully raise Series A funding are more likely to move on to Series B, and so on.
3. **Funding Types:**
   - Correlations between `venture`, `private_equity`, and `post_ipo_equity` indicate that these investment types often occur in sequence. Companies that receive venture funding may later attract private equity or go public.

Limitations

- **Correlation does not imply causation:** While the correlation matrix reveals relationships between variables, it cannot establish causality. We can't definitively say that one variable causes another to change.

Automated Profiling

```
dt_mrg['founded_at'] = pd.to_datetime(dt_mrg['founded_at'], errors='coerce')
dt_mrg['founded_at'].dtype
```

```
dtype('<M8[ns]')
```

```
dt_mrg['founded_at'].info()
```

```
<class 'pandas.core.series.Series'>
Index: 49436 entries, 0 to 49436
Series name: founded_at
Non-Null Count  Dtype
--------------  -----
49431 non-null  datetime64[ns]
dtypes: datetime64[ns](1)
memory usage: 1.8 MB
```

```
cols_to_convert = ['founded_month','founded_year', 'first_funding_at','last_funding_at']
dt_mrg[cols_to_convert] = dt_mrg[cols_to_convert].apply(pd.to_datetime, errors='coerce')
```

```
report = sweetviz.analyze(dt_mrg)
report.show_html("sweetviz_report.html")  # Generates a detailed report
```

```
Done! Use 'show' commands to display/save.   [100%]   00:04 -> (00:00 left)
Report sweetviz_report.html was generated! NOTEBOOK/COLAB USERS: the web browser MAY not pop up, regardless, the report IS saved in your
```

```
report.show_notebook()
```

3 homepage_url

VALUES:	49,436 (100%)
MISSING:	---
DISTINCT:	49,271 (>99%)

4	<1%	https://tvurology.com/
4	<1%	https://www.lightexhire.com/about
2	<1%	http://ayibang.com
2	<1%	http://www.comsenz.com
2	<1%	http://www.lekiosk.com
2	<1%	http://rang.com
2	<1%	http://www.cirqle.nl/
49,418	>99%	(Other)

4 category_list

VALUES:	49,436 (100%)
MISSING:	---
DISTINCT:	16,673 (34%)

3,963	8%	Not Available
3,650	7%	|Software|
3,598	7%	|Biotechnology|
1,263	3%	|E-Commerce|
1,211	2%	|Mobile|
1,120	2%	|Curated Web|
1,109	2%	|Clean Technology|
33,522	68%	(Other)

5 market

VALUES:	49,436 (100%)
MISSING:	---
DISTINCT:	4,718 (10%)

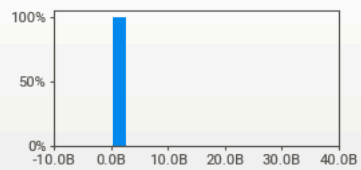
4,620	9%	Software
3,689	7%	Biotechnology
1,983	4%	Mobile
1,805	4%	E-Commerce
1,655	3%	Curated Web
1,280	3%	Enterprise Software
1,207	2%	Health Care
33,197	67%	(Other)

6 funding_total_usd

VALUES:	49,436 (100%)
MISSING:	---
DISTINCT:	14,615 (30%)
ZEROES:	8,530 (17%)

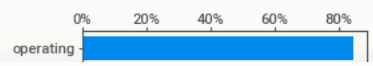
MAX	30.1B
95%	0.1B
Q3	0.0B
AVG	0.0B
MEDIAN	0.0B
Q1	0.0B
5%	0.0B
MIN	0.0B

RANGE	30.1B
IQR	6.7M
STD	153.6M
VAR	23579.8T
KURT.	29,855
SKEW	156
SUM	650.9B



7 status

VALUES:	49,436 (100%)
MISSING:	---



EDA

Total Funding by Category

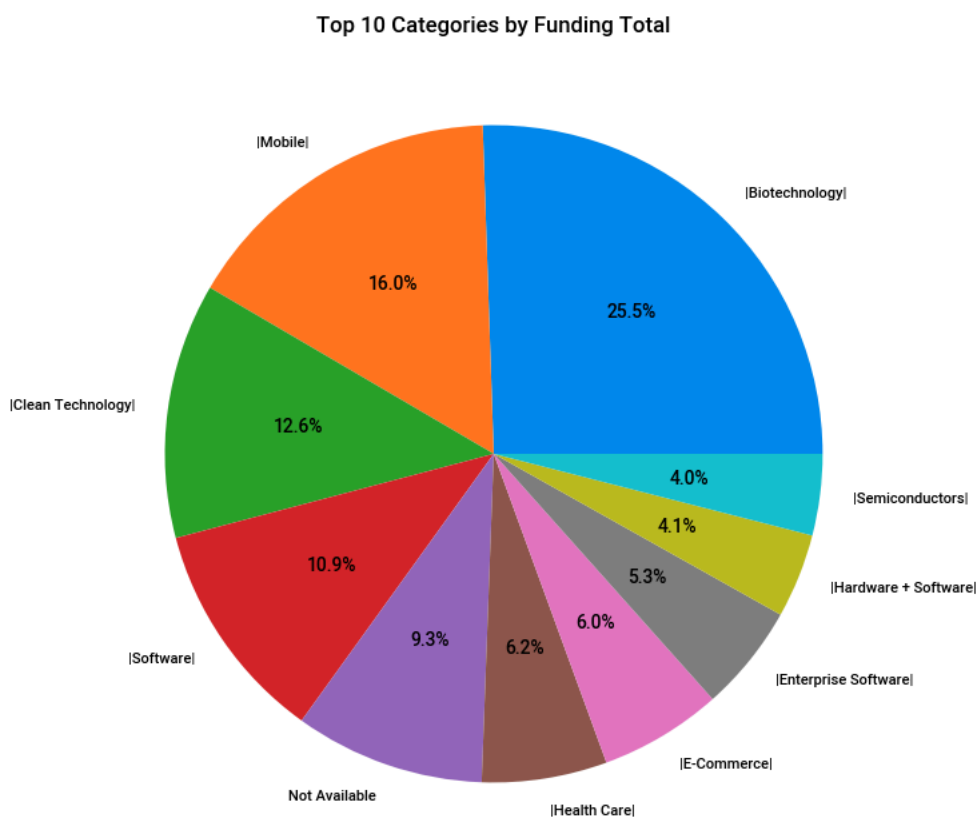
```
dt_mrg.groupby('category_list')['funding_total_usd'].sum().sort_values(ascending=False)
```

category_list	funding_total_usd
|Biotechnology|	6.982689e+10
|Mobile|	4.383988e+10
|Clean Technology|	3.433949e+10
|Software|	2.993001e+10
Not Available	2.556072e+10
...	...
|Data Centers|Data Center Infrastructure|Semiconductors|Optical Communications|	0.000000e+00
|DOD/Military|World Domination|Networking|Content|Social Media|	0.000000e+00
|Information Technology|Cloud Computing|Design|	0.000000e+00
|Information Technology|Computers|	0.000000e+00
|mHealth|Mobile|Health and Wellness|Health Care|	0.000000e+00

16673 rows × 1 columns

dtype: float64

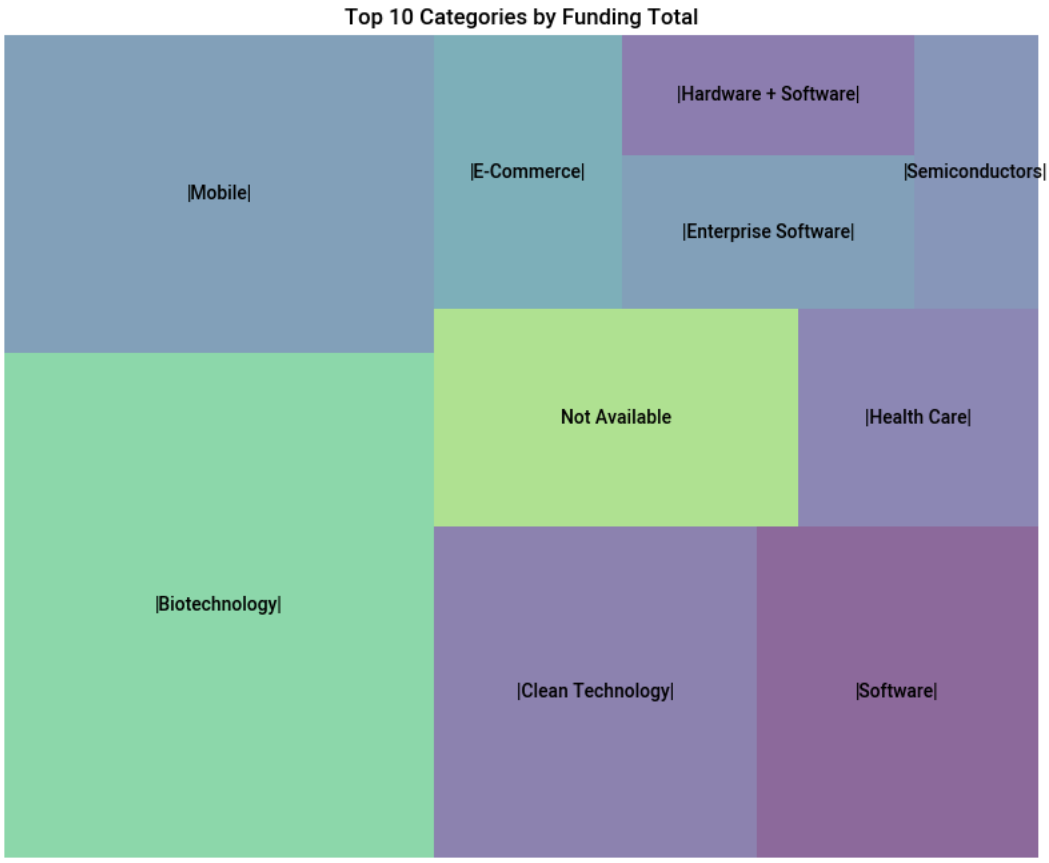
```
## Top 10 categories with the highest funding totals
top_10_categories = dt_mrg.groupby('category_list')['funding_total_usd'].sum().nlargest(10)

# pie chart
plt.figure(figsize=(10, 8))
plt.pie(top_10_categories.values, labels=top_10_categories.index, autopct='%1.1f%%')
plt.title('Top 10 Categories by Funding Total')
plt.show()
```



```
pip install squarify
```

```
Collecting squarify
  Downloading squarify-0.4.4-py3-none-any.whl.metadata (600 bytes)
Downloading squarify-0.4.4-py3-none-any.whl (4.1 kB)
Installing collected packages: squarify
Successfully installed squarify-0.4.4
```

```
import squarify

# treemap
plt.figure(figsize=(10, 8))
squarify.plot(sizes=top_10_categories.values, label=top_10_categories.index, alpha=0.6)
plt.axis('off')
plt.title('Top 10 Categories by Funding Total')
plt.show()
```



Average Funding by cat

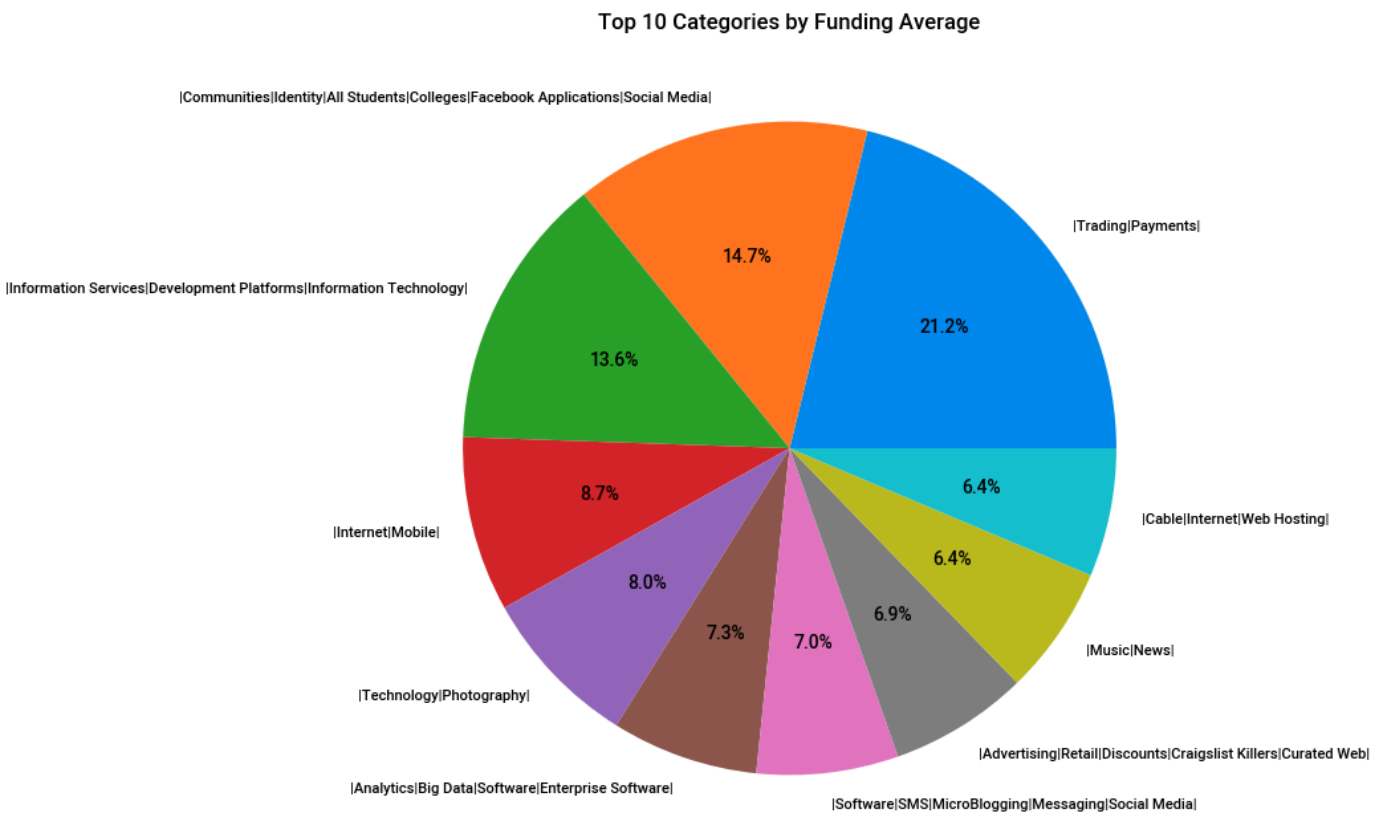
```
dt_mrg.groupby('category_list')['funding_total_usd'].mean()
```

category_list	funding_total_usd
Not Available	6.449840e+06
|3D Printing|	5.600000e+05
|3D Printing|3D|	0.000000e+00
|3D Printing|Consumer Electronics|	1.860000e+05
|3D Printing|Hardware + Software|	2.230000e+07
...	...
|mHealth|Health Care Information Technology|	2.766720e+05
|mHealth|Health Care|	2.500000e+04
|mHealth|Location Based Services|SaaS|Enterprise Software|	4.000000e+05
|mHealth|Mobile|	6.193280e+05
|mHealth|Mobile|Health and Wellness|Health Care|	0.000000e+00

16673 rows × 1 columns

dtype: float64

```
## Top 10 categories with the highest funding totals
top_10_categories = dt_mrg.groupby('category_list')['funding_total_usd'].mean().nlargest(10)

# pie chart
plt.figure(figsize=(10, 8))
plt.pie(top_10_categories.values, labels=top_10_categories.index, autopct='%1.1f%%')
plt.title('Top 10 Categories by Funding Average')
plt.show()
```



Total funding by market

```
## Top performing
dt_mrg.groupby(' market ')['funding_total_usd'].sum().sort_values(ascending=False).head(10)
```

market	funding_total_usd
Biotechnology	7.337328e+10
Mobile	4.947011e+10
Software	4.222348e+10
Clean Technology	3.761994e+10
Health Care	2.946608e+10
E-Commerce	2.240839e+10
Enterprise Software	1.929719e+10
Finance	1.737230e+10
Hardware + Software	1.281293e+10
Advertising	1.266066e+10

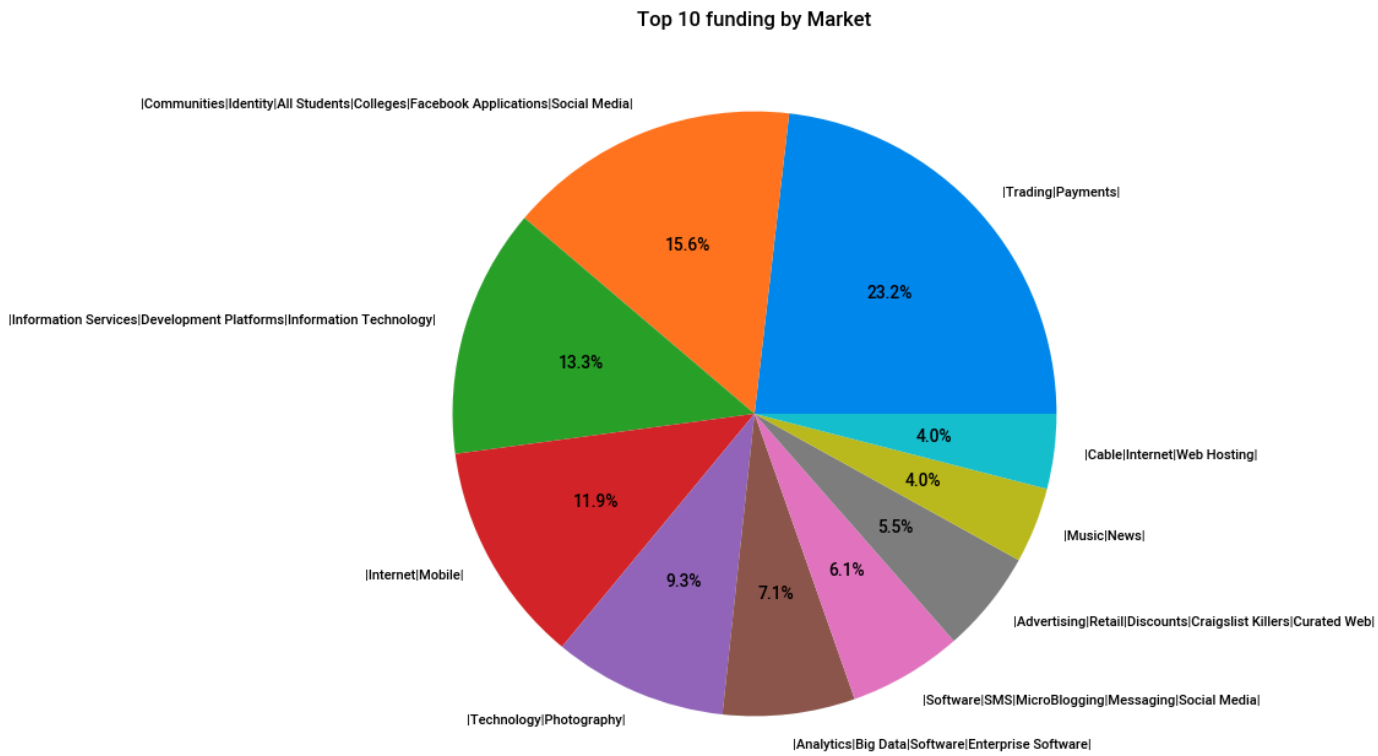
dtype: float64

```
# Worst performing ⚠
dt_mrg.groupby(' market ')['funding_total_usd'].sum().nsmallest(10)
```

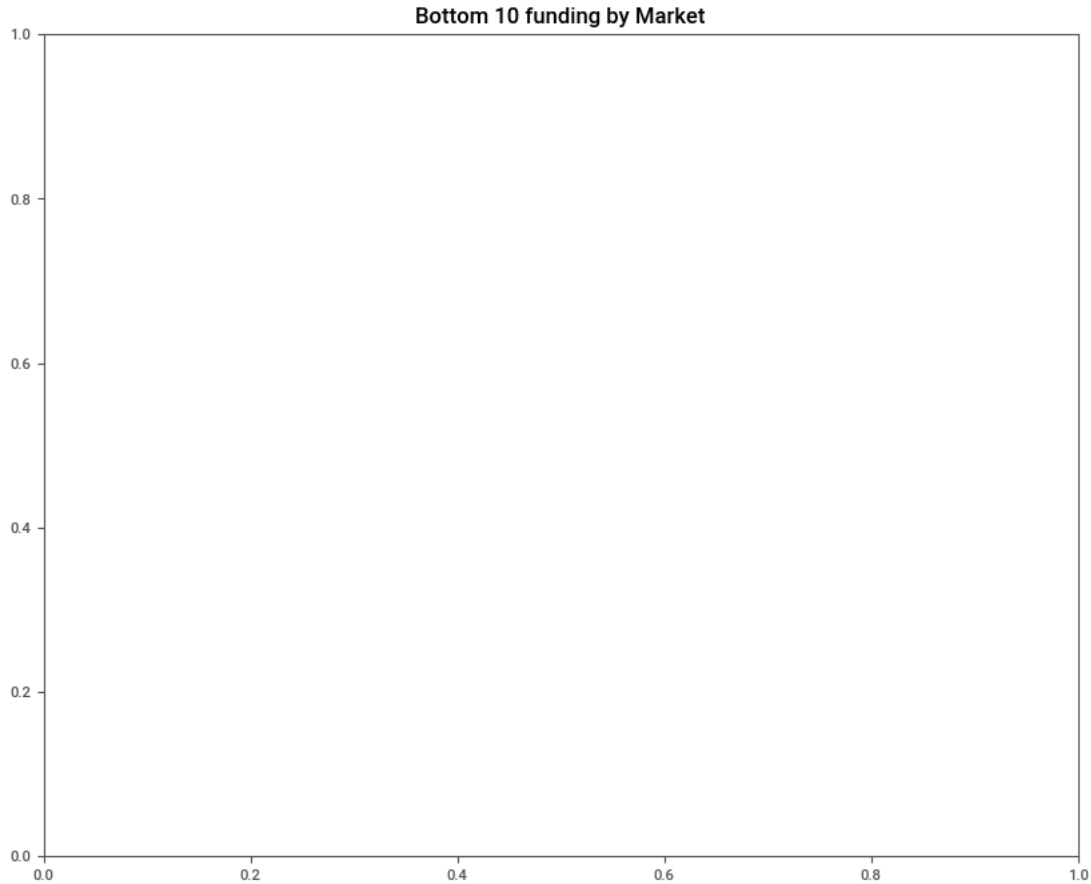
market	funding_total_usd
Advanced Materials	0.0
BPO Services	0.0
Debt Collecting	0.0
Dietary Supplements	0.0
EDA Tools	0.0
Green Building	0.0
Independent Music Labels	0.0
Lasers	0.0
Lotteries	0.0
Low Bid Auctions	0.0

dtype: float64

```
## Top 10 categories with the highest funding totals
top_10_market = dt_mrg.groupby(' market ')['funding_total_usd'].sum().nlargest(10)

# pie chart
plt.figure(figsize=(10, 8))
plt.pie(top_10_market.values, labels=top_10_categories.index, autopct='%1.1f%%')
plt.title('Top 10 funding by Market')
plt.show()

bottom_10_market = dt_mrg.groupby(' market ')['funding_total_usd'].sum().nsmallest(10)
plt.figure(figsize=(10, 8))
if bottom_10_market.sum() > 0:
    plt.pie(bottom_10_market.values, labels=bottom_10_market.index, autopct='%1.1f%%')
else:
    print("\033[1m\033[91mThe sum of the values is zero. Cannot plot.\033[0m")
plt.title('Bottom 10 funding by Market')
plt.show()
```

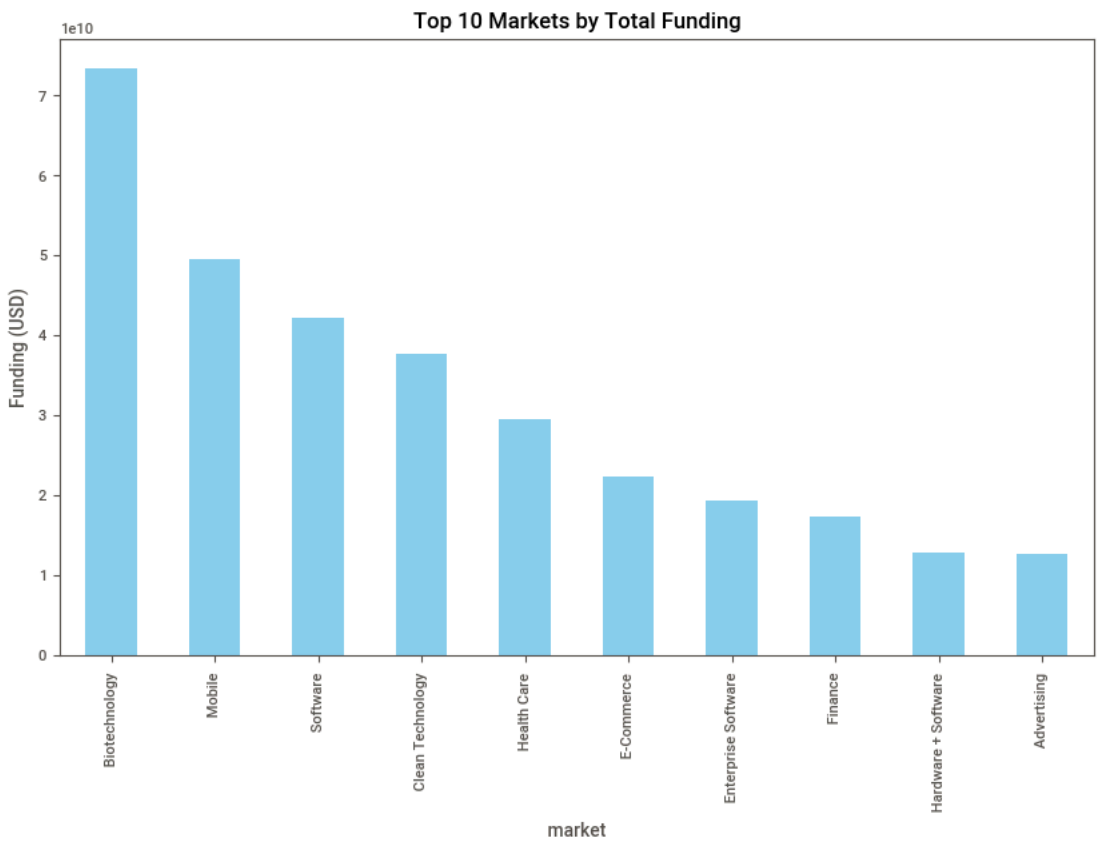



```
The sum of the values is zero. Cannot plot.
```

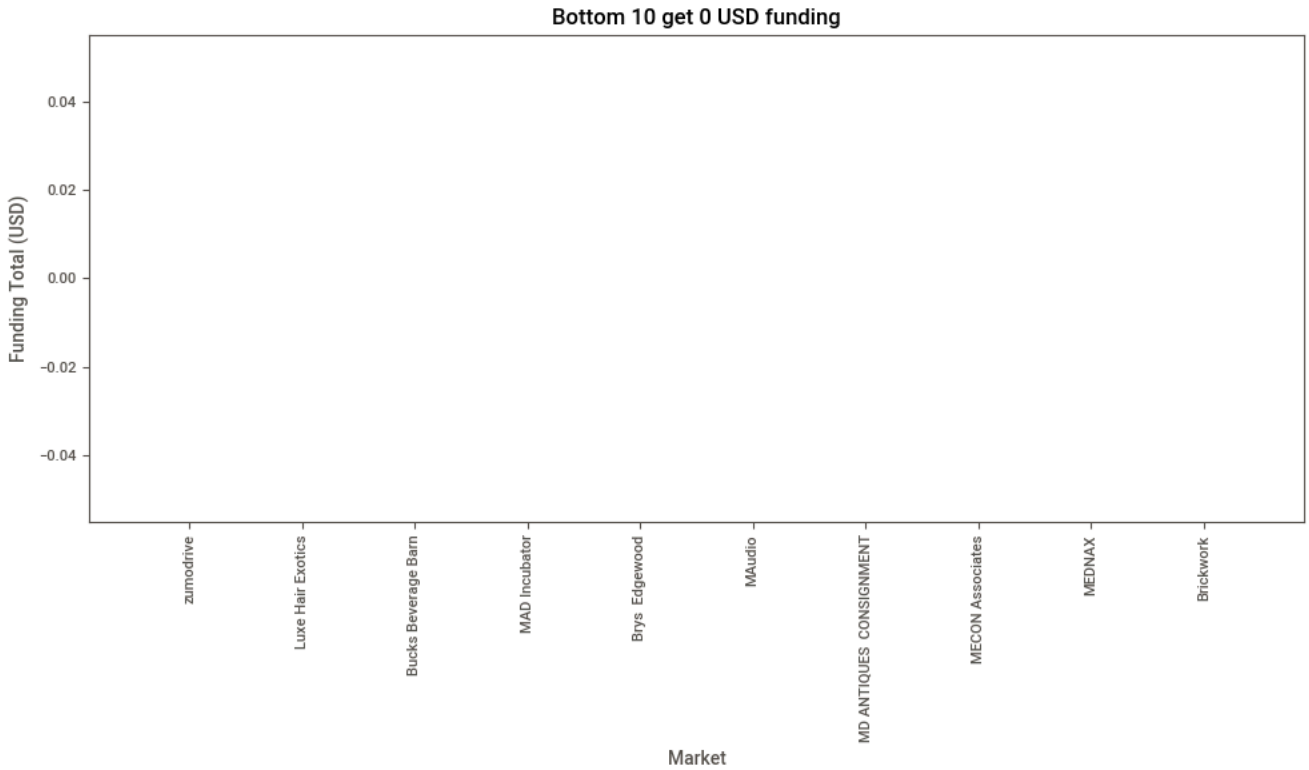


Distribution of funding across different market through bar plot

```
## Top 10
import matplotlib.pyplot as plt
market_funding = dt_mrg.groupby(' market ')['funding_total_usd'].sum().sort_values(ascending=False).head(10)
market_funding.plot(kind='bar', figsize=(10, 6), color='skyblue', title='Top 10 Markets by Total Funding')
plt.ylabel('Funding (USD)')
plt.show()
```




```
market_funding = dt_mrg.groupby(' market ')['funding_total_usd'].sum().sort_values(ascending=True).reset_index()
# Plot the top 10 markets by funding
plt.figure(figsize=(10, 6))
plt.bar(market_funding[' market '].head(10), market_funding['funding_total_usd'].head(10))
plt.xlabel('Market')
plt.ylabel('Funding Total (USD)')
plt.title('Bottom 10 get 0 USD funding')
plt.xticks(rotation=90)
plt.tight_layout()
plt.show()
```



No of funding rounds, relation with market/segments.

```
dt_mrg.groupby(' market ')['funding_rounds'].sum().sort_values(ascending=False)
```

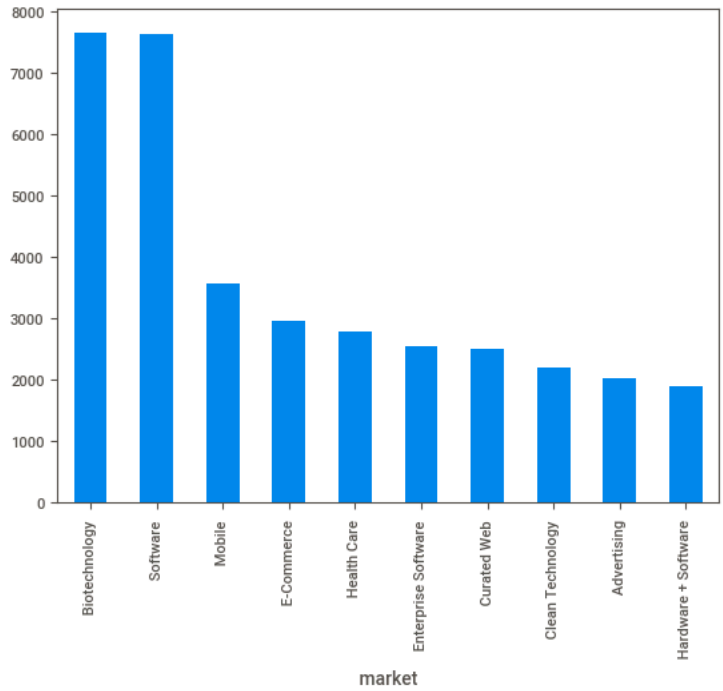
market	funding_rounds
Biotechnology	7656.0
Software	7633.0
Mobile	3570.0
E-Commerce	2958.0
Health Care	2777.0
...	...
Global CIO	1.0
Global Education Learning	1.0
Global Employment Solutions	1.0
Global Green Capitals Corporation	1.0
zumodrive	1.0

4718 rows × 1 columns

dtype: float64

```
dt_mrg.groupby(' market ')['funding_rounds'].sum().nlargest(10).plot(kind='bar')
```

```
<Axes: xlabel=' market '>
```




```
## by segments
dt_mrg.groupby('category_list')['funding_rounds'].sum().sort_values(ascending=False).head(10)
```

category_list	funding_rounds
|Biotechnology|	7433.0
|Software|	5824.0
Not Available	4433.0
|Mobile|	2105.0
|E-Commerce|	1977.0
|Clean Technology|	1976.0
|Health Care|	1941.0
|Enterprise Software|	1695.0
|Hardware + Software|	1669.0
|Curated Web|	1642.0

dtype: float64

```
dt_mrg.groupby('category_list')['funding_rounds'].sum().sort_values(ascending=True).head(10)
```

category_list	funding_rounds
|mHealth|Mobile|Health and Wellness|Health Care|	1.0
|Health Care|Medical|	1.0
|Sales and Marketing|Lead Generation|B2B|Architecture|Construction|Analytics|	1.0
|Sales and Marketing|Internet|Training|Teachers|Local Search|Search|Entertainment|Health and Wellness|Tutoring|Services|Local|Education|Online Rental|Games|	1.0
|Sales and Marketing|Internet|Startups|Social Media|	1.0
|Health Care|Medical|Pharmaceuticals|	1.0
|Health Care|Medical|Pharmaceuticals|Bio-Pharm|	1.0
|Sales and Marketing|Lead Generation|Software|Design|Solar|Clean Technology|	1.0
|Sales and Marketing|Internet|Mobile|Social Network Media|Advertising|	1.0
|Health Care|Medical|Therapeutics|	1.0

dtype: float64

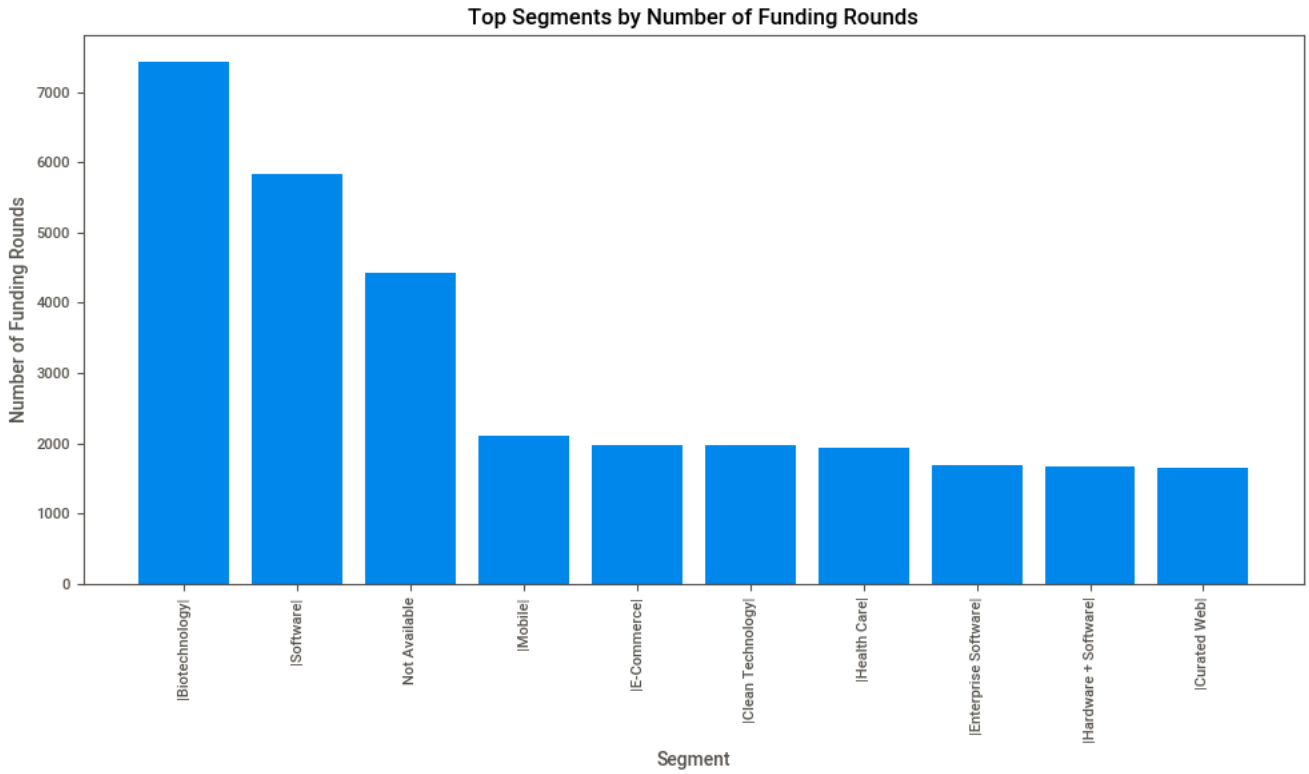
No of funding rounds in each segment.

```
seg =  dt_mrg.groupby('category_list')['funding_rounds'].sum().sort_values(ascending=False).reset_index()
seg
```

	category_list	funding_rounds
0	|Biotechnology|	7433.0
1	|Software|	5824.0
2	Not Available	4433.0
3	|Mobile|	2105.0
4	|E-Commerce|	1977.0
...	...	...
16668	|Networking|Content|Web Hosting|Virtualization...	1.0
16669	|Networking|Crowdsourcing|Crowdfunding|Finance|	1.0
16670	|Digital Media|Advertising|E-Commerce|Social M...	1.0
16671	|Digital Media|Advertising|E-Commerce|	1.0
16672	|mHealth|Mobile|Health and Wellness|Health Care|	1.0

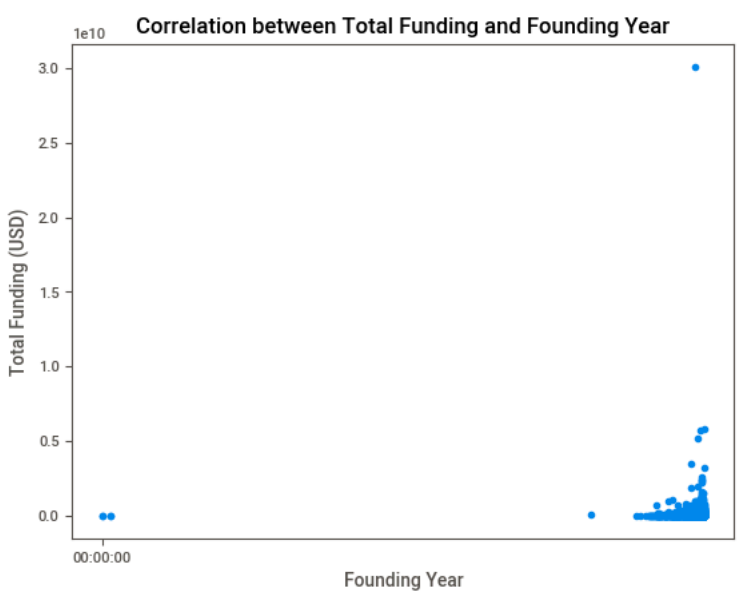
16673 rows × 2 columns

```
# Visualize the top segments by the number of funding rounds
plt.figure(figsize=(10, 6))
plt.bar( seg['category_list'].head(10), seg['funding_rounds'].nlargest(10))
plt.xlabel('Segment')
plt.ylabel('Number of Funding Rounds')
plt.title('Top Segments by Number of Funding Rounds')
plt.xticks(rotation=90)
plt.tight_layout()
plt.show()
```



Correlation b/w total funding vs founded year.

```
plt.scatter(dt_mrg['founded_year'], dt_mrg['funding_total_usd'])
plt.xlabel('Founding Year')
plt.ylabel('Total Funding (USD)')
plt.title('Correlation between Total Funding and Founding Year')
plt.show()
```




```
dt_mrg['funding_total_usd'].corr(dt_mrg['founded_year'])
```

```
-0.02436306579700567
```

Common Funding Sectorwise(market).

```
fnd = dt_mrg.groupby([' market ','seed', 'venture', 'equity_crowdfunding',
       'undisclosed', 'convertible_note', 'debt_financing', 'angel', 'grant',
       'private_equity', 'post_ipo_equity', 'post_ipo_debt',
       'secondary_market', 'product_crowdfunding']).size().reset_index()
fnd = fnd.drop(0,axis=1)
```

```
fnd.set_index(' market ',inplace=True)
```

```
fnd.loc[' 3D '].apply(lambda x: x[x != 0].count()).idxmax()
```

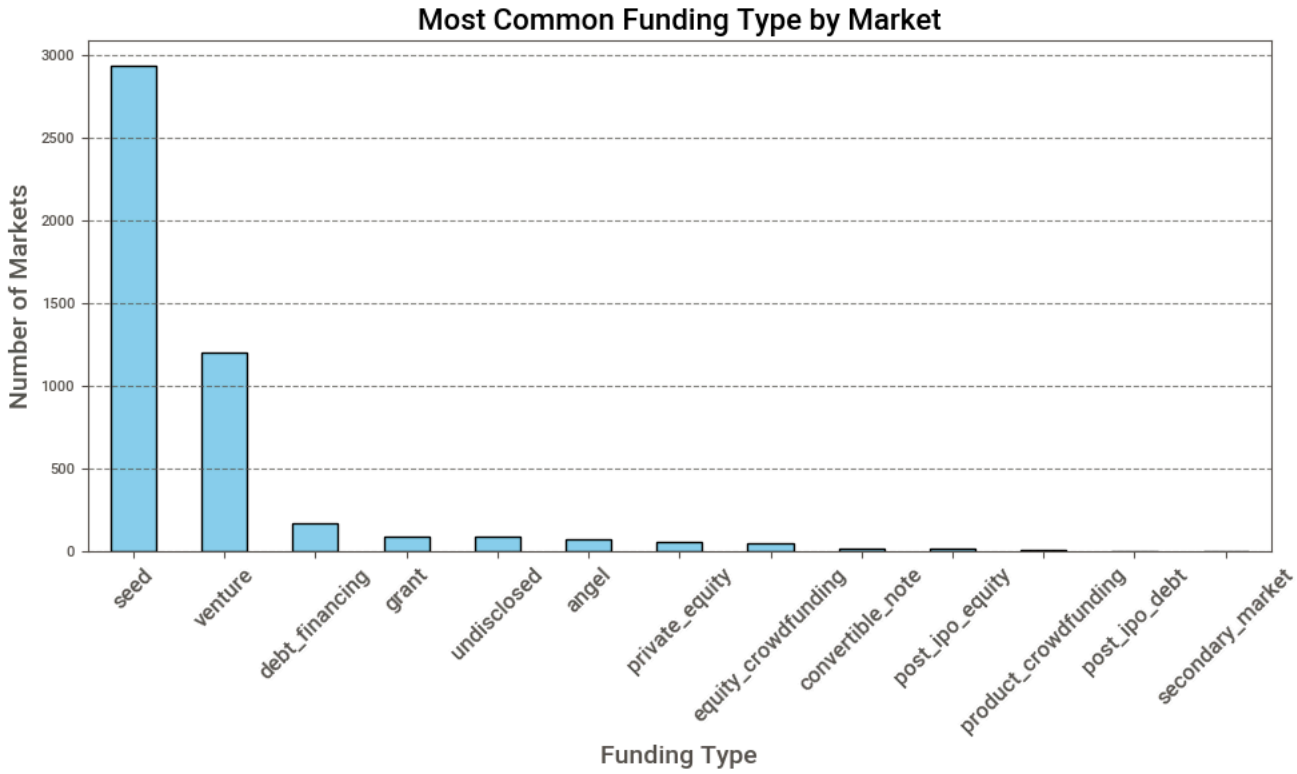
```
'seed'
```

```
# for market in fnd.index.unique():
#   fnd.loc[market].apply(lambda x: x[x != 0].count()).idxmax()
```

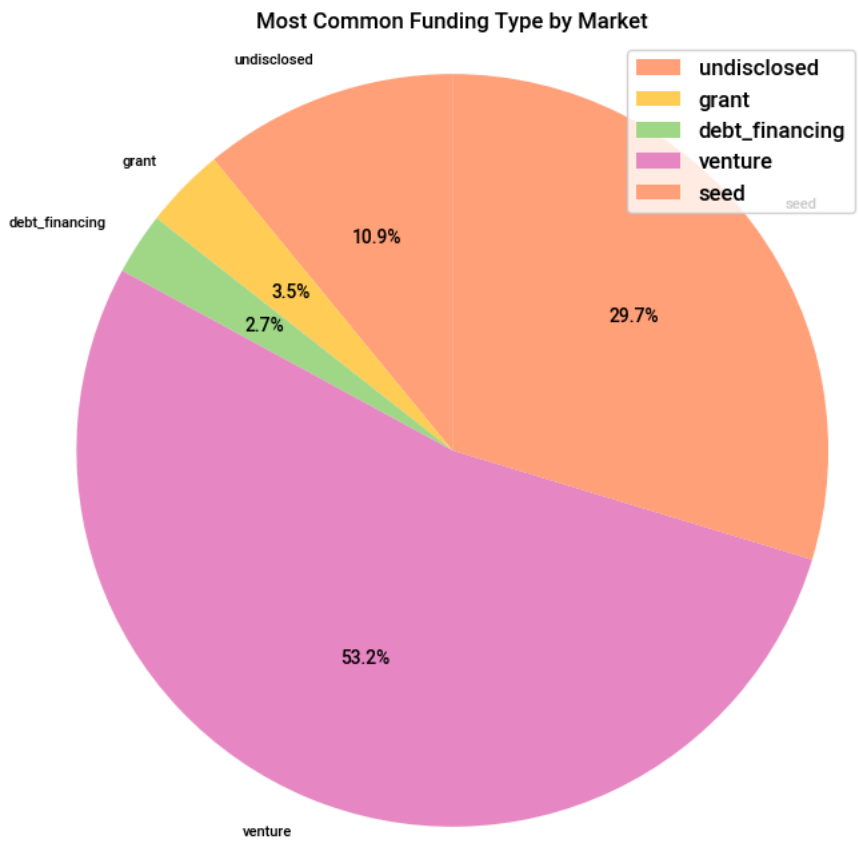
```
for market in fnd.index.unique():
    # Extract the DataFrame for the specific market
    market_df = fnd.loc[[market]]

    # Count non-zero values in each column
    non_zero_counts = (market_df != 0).sum()

    # Identify the column with the maximum non-zero count
    max_column = non_zero_counts.idxmax()

    print(f"Market: {market}, Funding Type with most frequent funding : {max_column}")

```

Show hidden output

```

non_zero_counts_per_market = (fnd != 0).groupby(fnd.index).sum()
max_funding_types = non_zero_counts_per_market.idxmax(axis=1)

# Count occurrences of each funding type being the most common
funding_type_counts = max_funding_types.value_counts()

#  Plot the results
import matplotlib.pyplot as plt

plt.figure(figsize=(10, 6))
funding_type_counts.plot(kind='bar', color='skyblue', edgecolor='black')

#  Add titles and labels
plt.title("Most Common Funding Type by Market", fontsize=16)
plt.xlabel("Funding Type", fontsize=14)
plt.ylabel("Number of Markets", fontsize=14)
plt.xticks(rotation=45, fontsize=12)
plt.grid(axis='y', linestyle='--', alpha=0.7)

# the chart
plt.tight_layout()
plt.show()

```

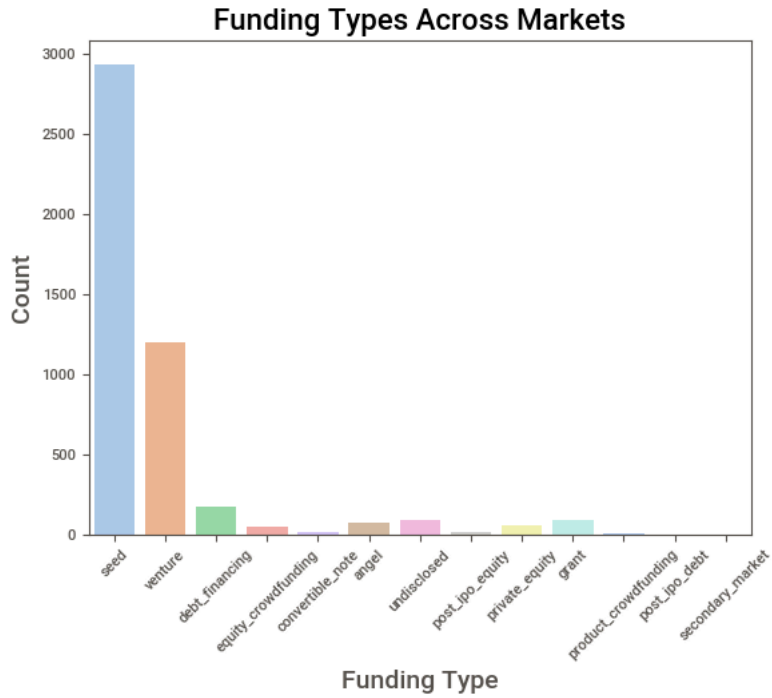



```
funding_type_counts = {"undisclosed": 321, "grant": 102, "debt_financing": 78, "venture": 1562, "seed": 871}
colors = ['#FFA07A', '#FFCE56', '#A0DB8B', '#E78AC3']  # Adjust colors as desired
plt.figure(figsize=(8, 8))
plt.pie(funding_type_counts.values(), labels=funding_type_counts.keys(), autopct='%1.1f%%', startangle=90, colors=colors)
plt.title("Most Common Funding Type by Market")
plt.axis('equal')
plt.legend(fontsize=12)
plt.show()
```




```
sns.countplot(data=max_funding_types.reset_index(), x=max_funding_types.values, palette="pastel")
plt.title("Funding Types Across Markets", fontsize=16)
plt.xlabel("Funding Type", fontsize=14)
plt.ylabel("Count", fontsize=14)
plt.xticks(rotation=45)
plt.show()
```

```
<ipython-input-213-cae4c19d5293>:1: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend

  sns.countplot(data=max_funding_types.reset_index(), x=max_funding_types.values, palette="pastel")
```




```
fnd.loc[(fnd != 0).sum(axis=1).nlargest(10).index].shape
```

```
(4220, 13)
```

```
funding_type_counts.index
```

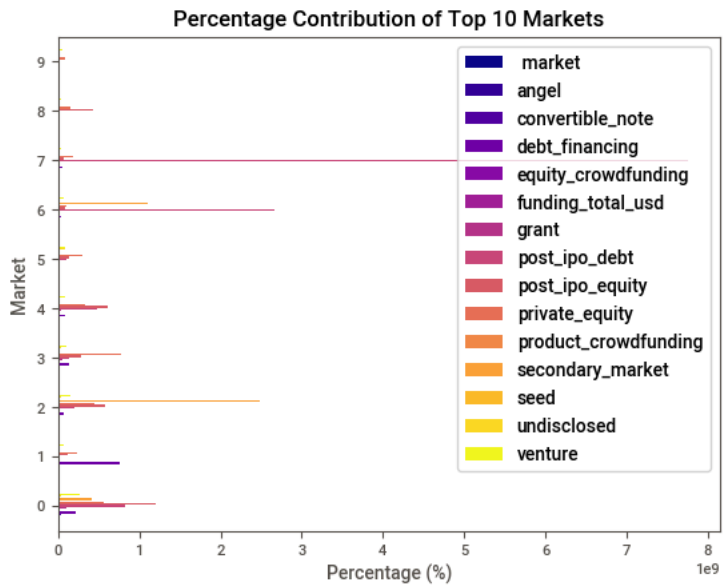
```
Index(['seed', 'venture', 'debt_financing', 'grant', 'undisclosed', 'angel',
       'private_equity', 'equity_crowdfunding', 'convertible_note',
       'post_ipo_equity', 'product_crowdfunding', 'post_ipo_debt',
       'secondary_market'],
      dtype='object')
```

```
dt_mrg.groupby(' market ')[list(funding_type_counts.index) + ['funding_total_usd']].sum().sort_values(by='funding_total_usd', ascending=F
```

	market	seed	venture	debt_financing	grant	undisclosed	angel	private_equity	equity_crowdfunding	c
0	Biotechnology	369675623.0	5.019133e+10	9.011835e+09	1.119789e+09	343539593.0	37731235.0	7.660596e+09	4629000.0	
1	Mobile	514764189.0	1.363873e+10	3.120950e+10	7.654236e+07	107386249.0	254899057.0	3.244764e+09	404419.0	
2	Software	853923812.0	2.913889e+10	3.105574e+09	4.644912e+07	343248211.0	216444395.0	6.031324e+09	20108116.0	
3	Clean Technology	103524817.0	1.944058e+10	5.616456e+09	5.474119e+08	410471772.0	15855672.0	1.048354e+10	11119016.0	
4	Health Care	237266260.0	1.821725e+10	3.689940e+09	3.661290e+08	116603429.0	23839913.0	4.422248e+09	1025000.0	
5	E-Commerce	431517930.0	1.517819e+10	1.138501e+09	6.637122e+06	778457363.0	197600996.0	4.138223e+09	10240642.0	
6	Enterprise Software	215007213.0	1.375703e+10	1.286473e+09	1.735125e+07	74266137.0	84499122.0	1.381407e+09	46553.0	
7	Finance	196794202.0	6.365423e+09	1.912794e+09	3.836244e+07	130562716.0	128242436.0	2.506866e+09	15256829	
8	Hardware + Software	165001642.0	8.407020e+09	6.063892e+08	6.985015e+07	112698215.0	28138908.0	1.999143e+09	0.0	
9	Advertising	278569117.0	1.052964e+10	3.944597e+08	8.099820e+05	104143006.0	67587874.0	1.258010e+09	2178501.0	

```
top_10_markets = dt_mrg.groupby(' market ')[list(funding_type_counts.index) + ['funding_total_usd']].sum().sort_values(by='funding_total_
percentage_contribution = (top_10_markets / non_zero_counts_per_market.sum()) * 100
percentage_contribution.plot(kind='barh', colormap='plasma')
```

```
plt.title("Percentage Contribution of Top 10 Markets")
plt.xlabel("Percentage (%)")
plt.ylabel("Market")
plt.show()
```

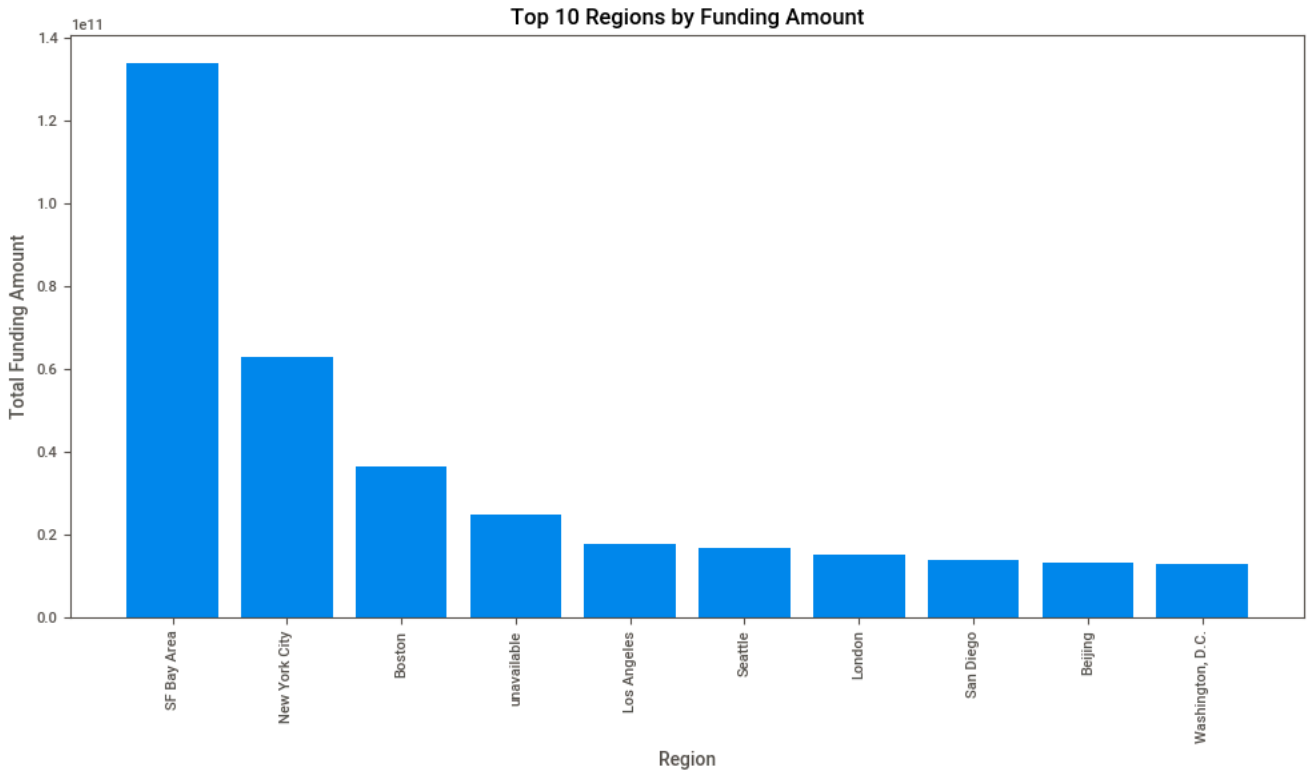


Funding based on Region

```
regional_funding = dt_mrg.groupby('region')['funding_total_usd'].sum().reset_index()
regional_funding = regional_funding.sort_values('funding_total_usd', ascending=False)
print(regional_funding.head(10))
```

```
                region  funding_total_usd
836        SF Bay Area       1.339018e+11
699      New York City       6.300848e+10
121             Boston       3.653691e+10
1087       unavailable       2.486025e+10
570        Los Angeles       1.783554e+10
876            Seattle       1.672701e+10
568             London       1.511296e+10
858          San Diego       1.381026e+10
89             Beijing       1.325378e+10
1038  Washington, D.C.       1.280293e+10
```

```
# Sort the data in descending order of total funding amount
top_10_regions = regional_funding.nlargest(10, 'funding_total_usd')

# Plot the top 10 regions
plt.figure(figsize=(10, 6))
plt.bar(top_10_regions['region'], top_10_regions['funding_total_usd'])
plt.xlabel('Region')
plt.ylabel('Total Funding Amount')
plt.title('Top 10 Regions by Funding Amount')
plt.xticks(rotation=90)  # Rotate x-axis labels for better readability
plt.tight_layout()  # Ensure labels fit within the plot area
plt.show()
```



Niche segment

```
niche_segments = dt_mrg['category_list'].value_counts()
print(niche_segments)
```

```
category_list
Not Available                                                                    3963
|Software|                                                                       3650
|Biotechnology|                                                                  3598
|E-Commerce|                                                                     1263
|Mobile|                                                                         1211
                                                                                 ...
|Social Network Media|Crowdsourcing|Crowdfunding|Gift Card|Curated Web|            1
|Fashion|Digital Media|Marketplaces|E-Commerce|                                   1
|Advertising|Web Development|App Marketing|Enterprises|Cloud Computing|Enterprise Software|   1
|Web Design|Software|Web Tools|Web Development|Enterprise Software|              1
|Web Development|Advertising|Wireless|Mobile|                                    1
Name: count, Length: 16673, dtype: int64
```

```
niche_segment_returns = dt_mrg.groupby('category_list')['funding_total_usd'].mean()
print(niche_segment_returns)
```

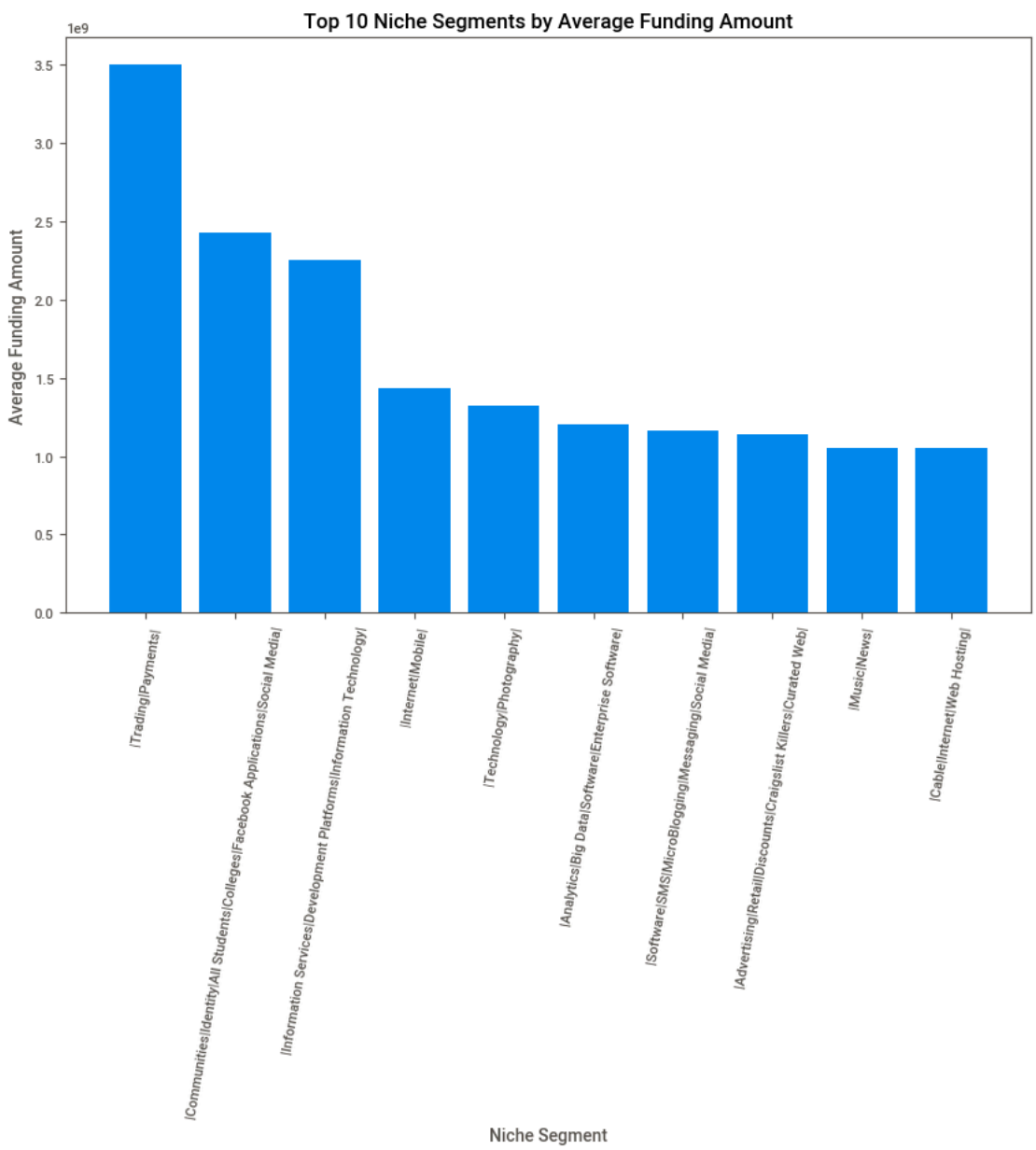
```
category_list
Not Available                                          6.449840e+06
|3D Printing|                                          5.600000e+05
|3D Printing|3D|                                       0.000000e+00
|3D Printing|Consumer Electronics|                     1.860000e+05
|3D Printing|Hardware + Software|                      2.230000e+07
                                                           ...
|mHealth|Health Care Information Technology|           2.766720e+05
|mHealth|Health Care|                                  2.500000e+04
|mHealth|Location Based Services|SaaS|Enterprise Software|   4.000000e+05
|mHealth|Mobile|                                       6.193280e+05
|mHealth|Mobile|Health and Wellness|Health Care|       0.000000e+00
Name: funding_total_usd, Length: 16673, dtype: float64
```

```
niche_segment_returns_sorted = niche_segment_returns.sort_values(ascending=False)
print(niche_segment_returns_sorted.head(10))
```

```
category_list
|Trading|Payments|                                                          3.500000e+09
|Communities|Identity|All Students|Colleges|Facebook Applications|Social Media|   2.425700e+09
|Information Services|Development Platforms|Information Technology|          2.250000e+09
```

```
|Internet|Mobile|                                                          1.431605e+09
|Technology|Photography|                                                    1.320000e+09
|Analytics|Big Data|Software|Enterprise Software|                          1.201000e+09
|Software|SMS|MicroBlogging|Messaging|Social Media|                        1.160167e+09
|Advertising|Retail|Discounts|Craigslist Killers|Curated Web|              1.137000e+09
|Music|News|                                                               1.055750e+09
|Cable|Internet|Web Hosting|                                               1.050000e+09
Name: funding_total_usd, dtype: float64
```

```
niche_segment_returns_sorted = niche_segment_returns.sort_values(ascending=False)
top_10_niche_segments = niche_segment_returns_sorted.head(10)
plt.figure(figsize=(10, 6))
plt.bar(top_10_niche_segments.index, top_10_niche_segments.values)
plt.xlabel('Niche Segment')
plt.ylabel('Average Funding Amount')
plt.title('Top 10 Niche Segments by Average Funding Amount')
plt.xticks(rotation=80)  # Rotate x-axis labels for better readability
plt.show()
```



Economic data, like GDP growth rates, unemployment rates, or venture capital availability.

```
!pip install fredapi
```

```
Collecting fredapi
  Downloading fredapi-0.5.2-py3-none-any.whl.metadata (5.0 kB)
```

```
Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-packages (from fredapi) (2.2.2)
Requirement already satisfied: numpy>=1.22.4 in /usr/local/lib/python3.10/dist-packages (from pandas->fredapi) (1.26.4)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.10/dist-packages (from pandas->fredapi) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas->fredapi) (2024.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.10/dist-packages (from pandas->fredapi) (2024.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.8.2->pandas->fredapi) (1.16.
Downloading fredapi-0.5.2-py3-none-any.whl (11 kB)
Installing collected packages: fredapi
Successfully installed fredapi-0.5.2
```

```
import yfinance as yf
from fredapi import Fred
fred = Fred(api_key='2e28afbde15bca02b906f3c2158f39f1')
gdp_growth_rate = fred.get_series('GDPC1')
unemployment_rate = fred.get_series('UNRATE')
```

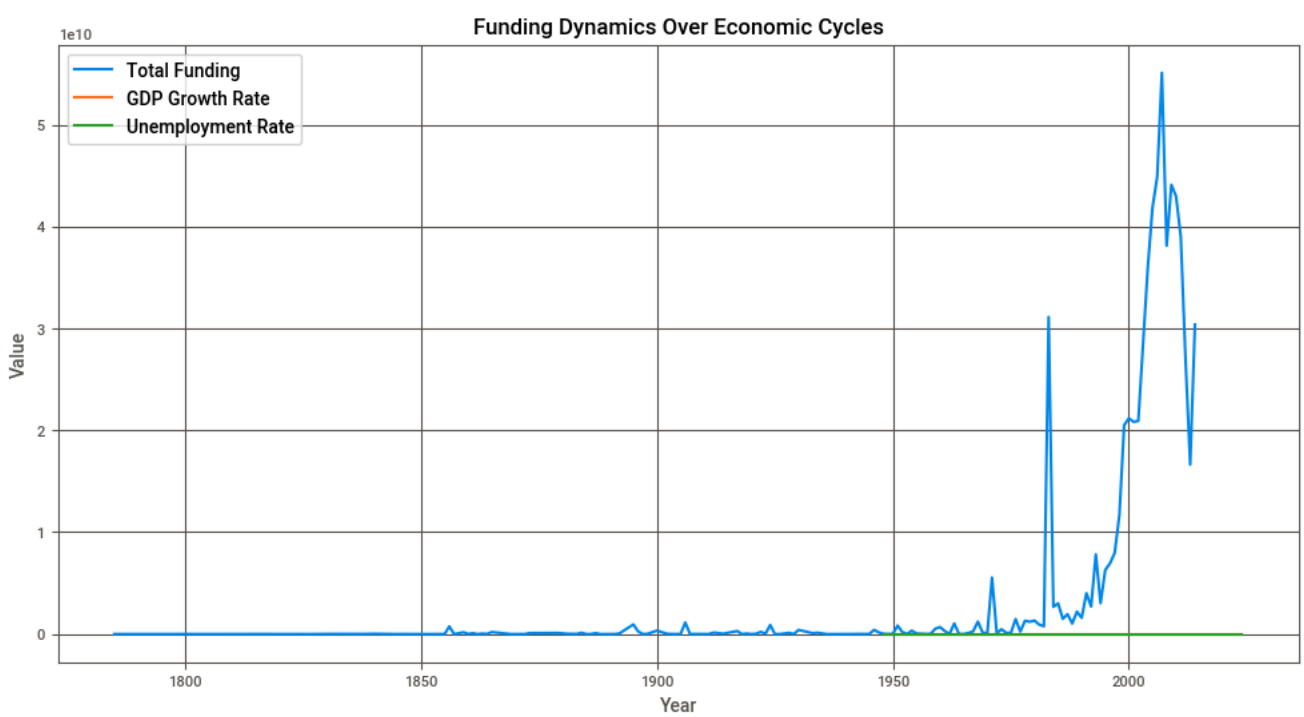
```
gdp_growth_rate
```

Show hidden output

```
unemployment_rate
```

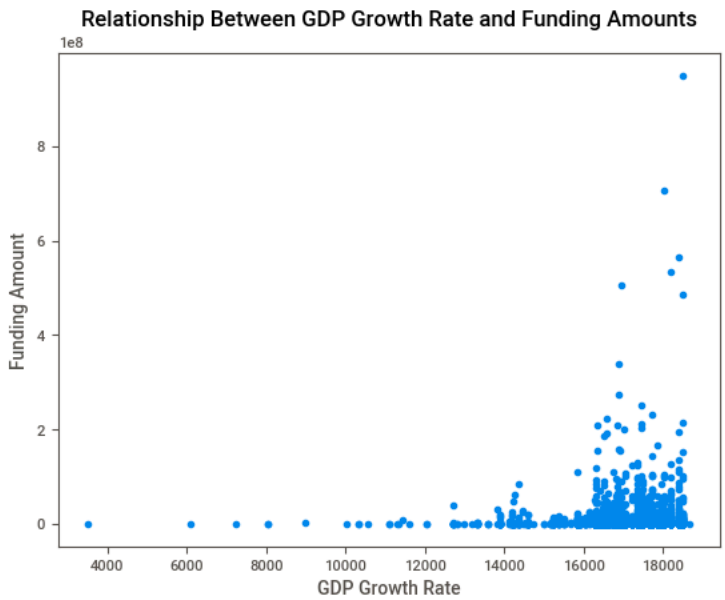
Show hidden output

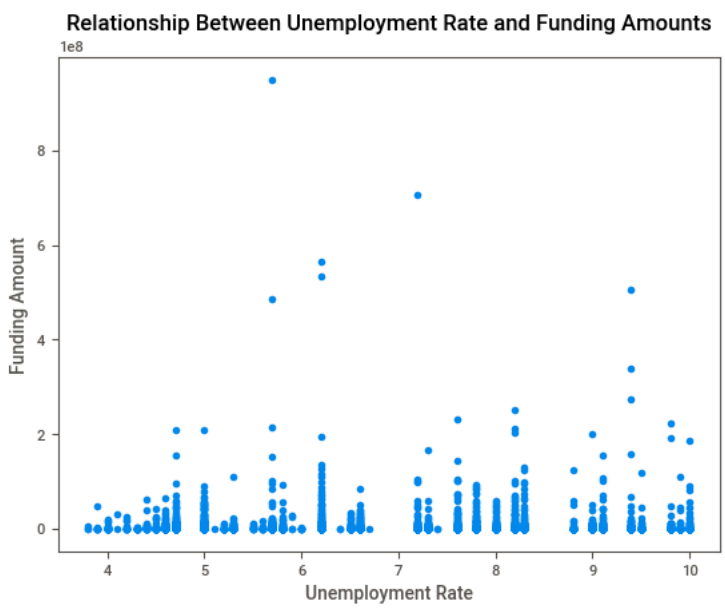
```
funding_by_year = dt_mrg.groupby('founded_year')['funding_total_usd'].sum()
plt.figure(figsize=(12, 6))
plt.plot(funding_by_year.index, funding_by_year.values, label='Total Funding')
plt.plot(economic_data['year'], economic_data['gdp_growth_rate'], label='GDP Growth Rate')
plt.plot(economic_data['year'], economic_data['unemployment_rate'], label='Unemployment Rate')
plt.title('Funding Dynamics Over Economic Cycles')
plt.xlabel('Year')
plt.ylabel('Value')
plt.legend()
plt.grid(True)
plt.show()
```




```
from fredapi import Fred
fred = Fred(api_key='2e28afbde15bca02b906f3c2158f39f1')
data
gdp_growth_rate = fred.get_series('GDPC1')
gdp_growth_rate = gdp_growth_rate.reset_index()
gdp_growth_rate = gdp_growth_rate.rename(columns={'index': 'date', 0: 'gdp_growth_rate'})
unemployment_rate = fred.get_series('UNRATE')
unemployment_rate = unemployment_rate.reset_index()
unemployment_rate = unemployment_rate.rename(columns={'index': 'date', 0: 'unemployment_rate'})
```

```
unemployment_rate = unemployment_rate.rename(columns={'index': 'date', 0: 'unemployment_rate'})
economic_data = pd.merge(gdp_growth_rate, unemployment_rate, on='date')
dt_mrg = dt_mrg.rename(columns={'last_funding_at': 'date'})
dt_mrg['date'] = pd.to_datetime(dt_mrg['date'], errors='coerce')
economic_data['date'] = pd.to_datetime(economic_data['date'], errors='coerce')
dt_mrg = dt_mrg.dropna(subset=['date'])
economic_data = economic_data.dropna(subset=['date'])
merged_data = pd.merge(dt_mrg, economic_data, on='date', how='inner')
plt.scatter(merged_data['gdp_growth_rate'], merged_data['funding_total_usd'])
plt.xlabel('GDP Growth Rate')
plt.ylabel('Funding Amount')
plt.title('Relationship Between GDP Growth Rate and Funding Amounts')
plt.show()
plt.scatter(merged_data['unemployment_rate'], merged_data['funding_total_usd'])
plt.xlabel('Unemployment Rate')
plt.ylabel('Funding Amount')
plt.title('Relationship Between Unemployment Rate and Funding Amounts')
plt.show()
```

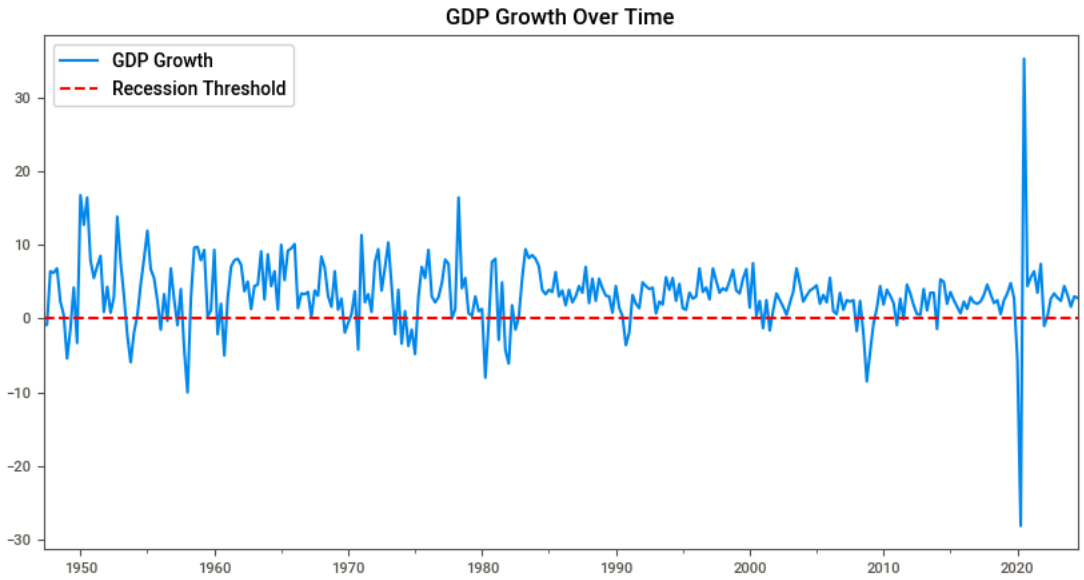


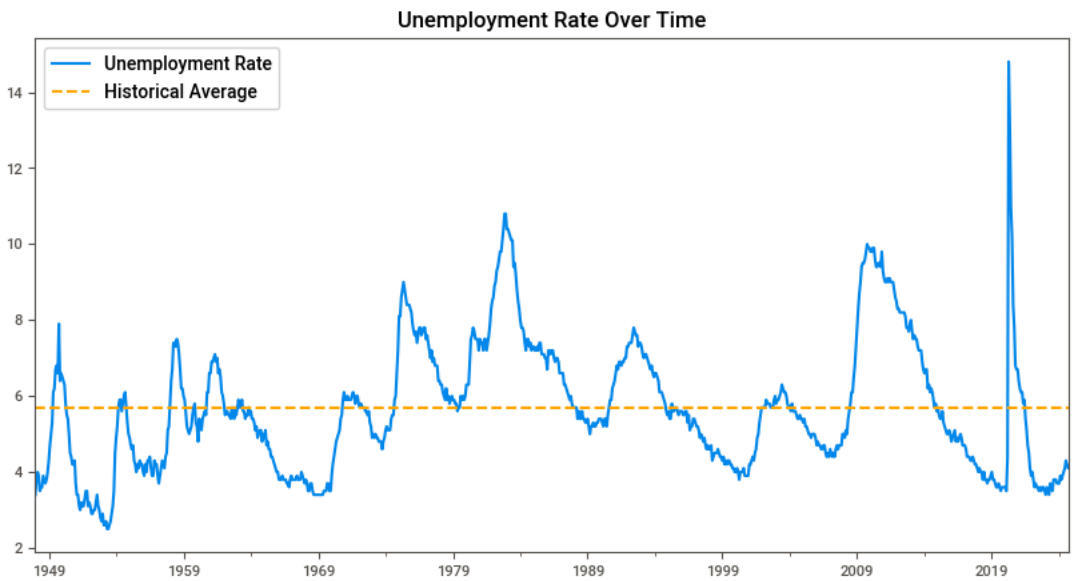


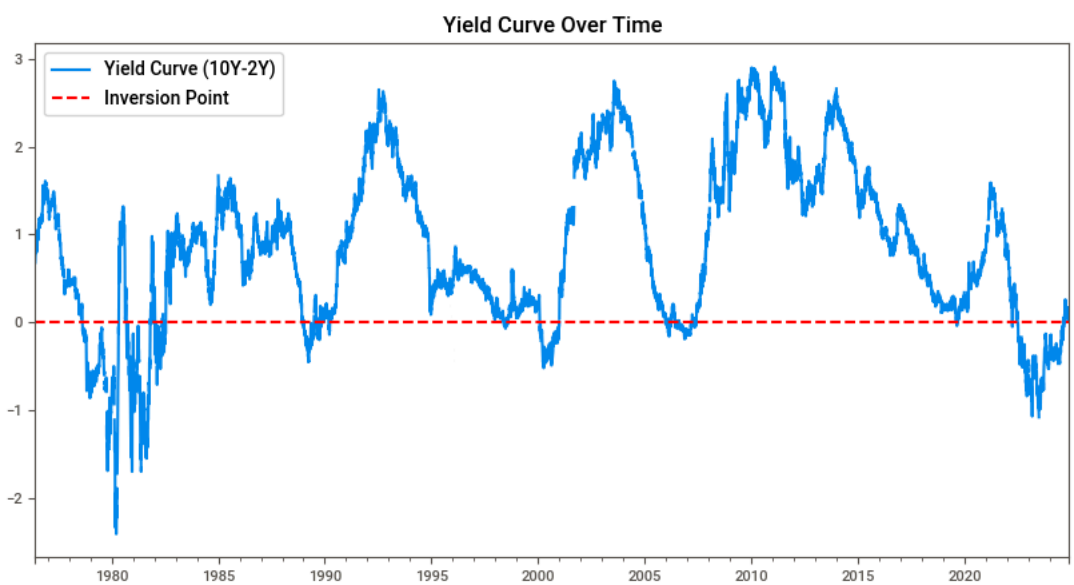
Receission or not, by indexes

```
from fredapi import Fred
fred = Fred(api_key='2e28afbde15bca02b906f3c2158f39f1')
gdp_growth = fred.get_series('A191RL1Q225SBEA')  # Real GDP Growth Rate
unemployment_rate = fred.get_series('UNRATE')    # Unemployment Rate
```

```
yield_curve = fred.get_series('T10Y2Y')          # 10-Year Treasury Minus 2-Year Treasury
consumer_confidence = fred.get_series('UMCSENT') # Consumer Confidence Index
plt.figure(figsize=(10, 5))
gdp_growth.plot(label='GDP Growth')
plt.axhline(0, color='red', linestyle='--', label='Recession Threshold')
plt.title('GDP Growth Over Time')
plt.legend()
plt.show()
plt.figure(figsize=(10, 5))
unemployment_rate.plot(label='Unemployment Rate')
plt.title('Unemployment Rate Over Time')
plt.axhline(unemployment_rate.mean(), color='orange', linestyle='--', label='Historical Average')
plt.legend()
plt.show()
plt.figure(figsize=(10, 5))
yield_curve.plot(label='Yield Curve (10Y-2Y)')
plt.axhline(0, color='red', linestyle='--', label='Inversion Point')
plt.title('Yield Curve Over Time')
plt.legend()
plt.show()
```







GDP growth over time insight:-

Gross domestic product development diagram shows an unpredictable example with times of extension and compression. The red line probably shows a downturn limit, with a few cases of the economy plunging beneath this level. This proposes a recurrent nature of financial

development, with highs and lows over the long haul.

Unempoyment Rate:-

The graph shows the joblessness rate over the long haul, with a flat line addressing the verifiable normal. The joblessness rate varies fundamentally, with times of high joblessness sprinkled with times of lower joblessness. The new pattern shows a lessening in the joblessness rate, which is beneath the verifiable normal. This proposes that the work market has been working on lately.

Yield curve over Time:-

The graph shows the yield bend spread (10-year Depository yield short 2-year Depository yield) after some time. The red line addresses the reversal point, where the yield bend rearranges (momentary rates surpass long haul rates). The graph shows numerous occasions of reversal, which are in many cases thought about a proactive factor of financial downturn.

Funding velocity, funding_total_amount/total_no_years

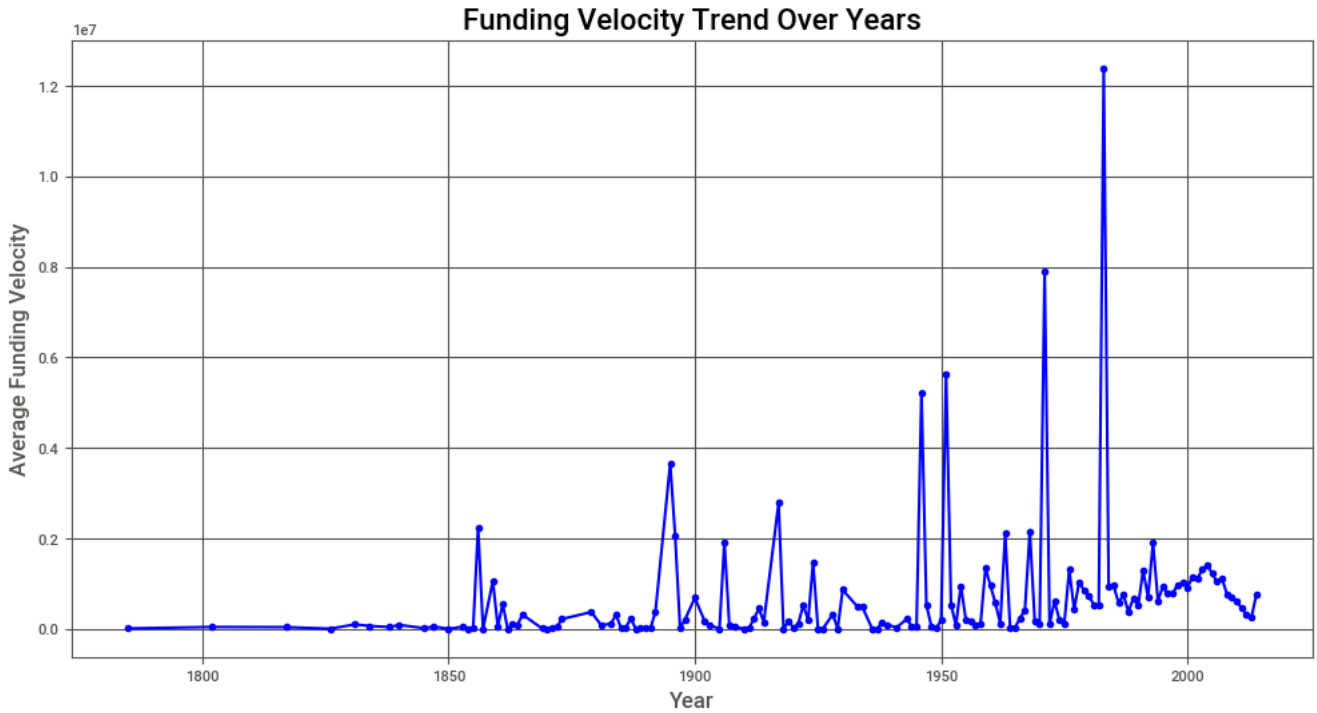
```
# Calculate total_no_years
dt_mrg['total_no_years'] = (dt_mrg['date'] - dt_mrg['founded_at']).dt.days / 365
```

```
# Calculate funding velocity
dt_mrg['funding_velocity'] = dt_mrg['funding_total_usd'] / dt_mrg['total_no_years']
```

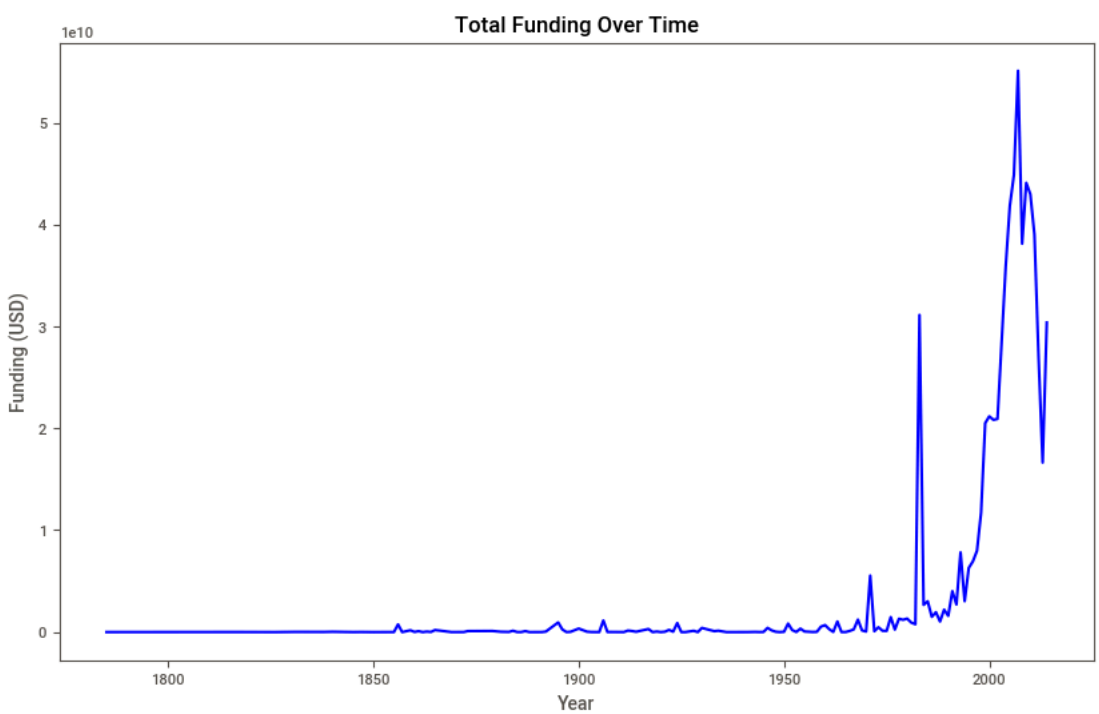
```
from datetime import datetime

current_date = datetime.now()


dt_mrg['years_since_founded'] = (current_date - dt_mrg['founded_at']).dt.days / 365.25


dt_mrg['funding_velocity'] = dt_mrg['funding_total_usd'] / dt_mrg['years_since_founded']

# Drop any rows with invalid data (e.g., negative years or zero years)
dt_mrg = dt_mrg[dt_mrg['years_since_founded'] > 0]

# Display the first few rows of the DataFrame to check
print(dt_mrg[['name', 'funding_total_usd', 'years_since_founded', 'funding_velocity']].head())

```

```
                 name  funding_total_usd  years_since_founded  funding_velocity
0            waywire          1750000.0            12.501027     139988.501971
1  TV Communications          4000000.0            14.494182     275972.799396
2    Rock Your Paper            40000.0            12.098563       3306.177868
3    InTouch Network          1500000.0            13.670089     109728.620068
4   R Ranch and Mine            60000.0            10.915811       5496.613995
```

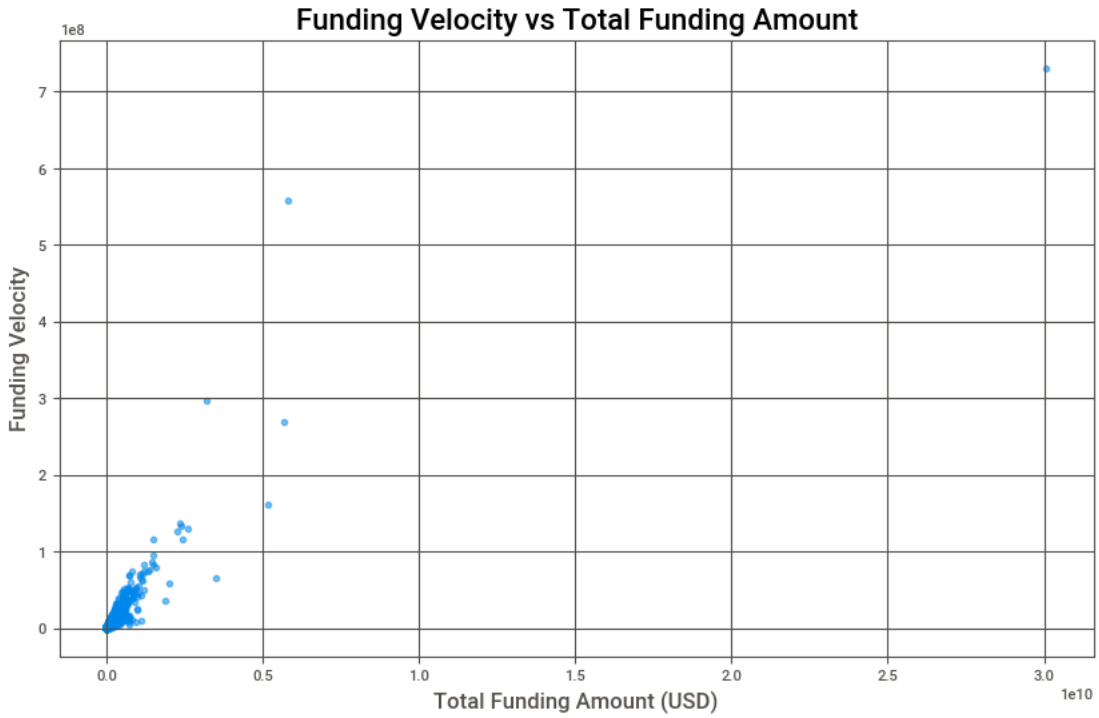
```
dt_mrg['founded_year'] = dt_mrg['founded_at'].dt.year
avg_funding_velocity_by_year = dt_mrg.groupby('founde_year')['funding_velocity'].mean()
plt.figure(figsize=(12, 6))
plt.plot(avg_funding_velocity_by_year.index, avg_funding_velocity_by_year.values, color='b', marker='o')
plt.title('Funding Velocity Trend Over Years', fontsize=16)
plt.xlabel('Year', fontsize=12)
plt.ylabel('Average Funding Velocity', fontsize=12)
plt.grid(True)
plt.show()
```




```
yearly_funding = dt_mrg.groupby('founded_year')['funding_total_usd'].sum()
yearly_funding.plot(kind='line', figsize=(10, 6), title='Total Funding Over Time', color='blue')
plt.ylabel('Funding (USD)')
plt.xlabel('Year')
plt.show()
```

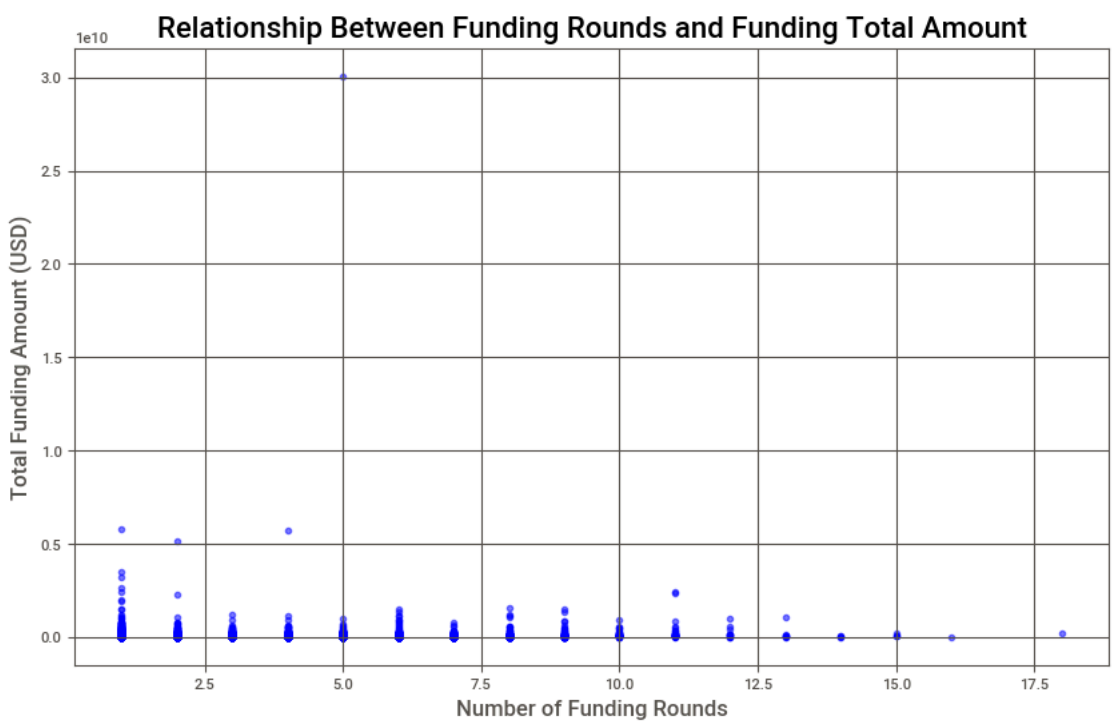



```
# Scatter plot to show relationship between Funding Velocity and Funding Amount
plt.figure(figsize=(10, 6))
plt.scatter(dt_mrg['funding_total_usd'], dt_mrg['funding_velocity'], alpha=0.5)
plt.title('Funding Velocity vs Total Funding Amount', fontsize=16)
plt.xlabel('Total Funding Amount (USD)', fontsize=12)
plt.ylabel('Funding Velocity', fontsize=12)
plt.grid(True)
plt.show()
```




```
print(dt_mrg[['funding_total_usd', 'funding_rounds']].head())
# Plotting Funding vs Funding Rounds (as a proxy for valuation)
plt.figure(figsize=(10, 6))
plt.scatter(dt_mrg['funding_rounds'], dt_mrg['funding_total_usd'], alpha=0.5, color='b')
plt.title('Relationship Between Funding Rounds and Funding Total Amount', fontsize=16)
plt.xlabel('Number of Funding Rounds', fontsize=12)
plt.ylabel('Total Funding Amount (USD)', fontsize=12)
plt.grid(True)
plt.show()
```

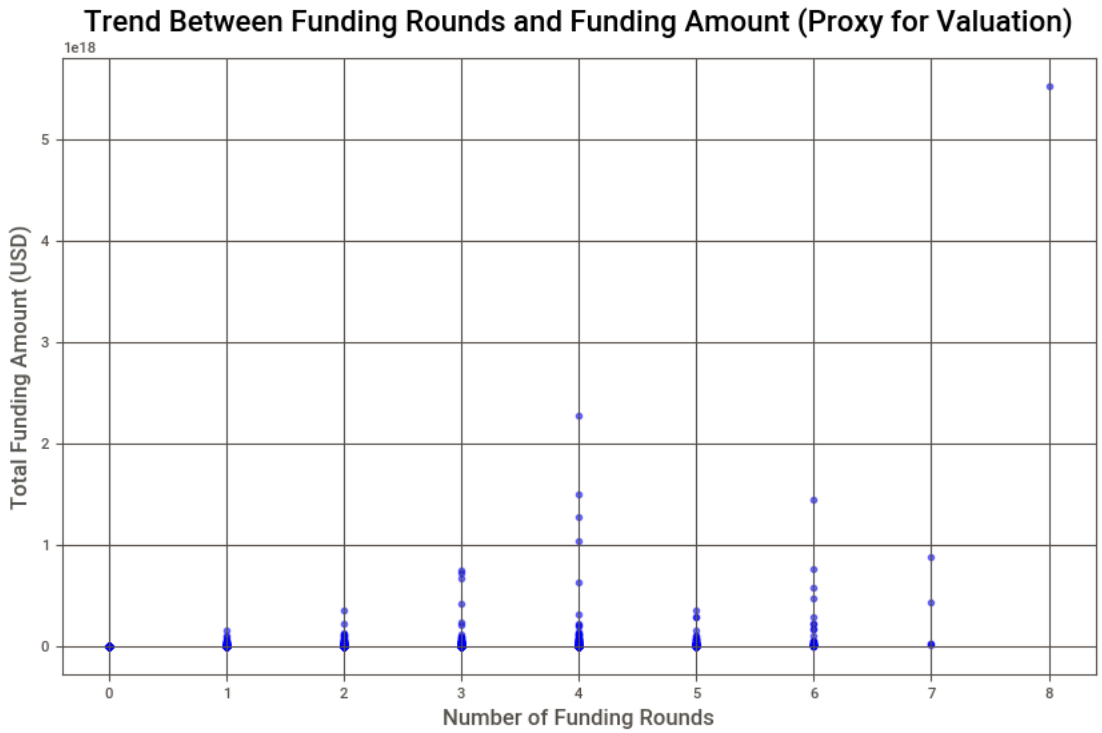
```
   funding_total_usd  funding_rounds
0          1750000.0             1.0
1          4000000.0             2.0
2            40000.0             1.0
3          1500000.0             1.0
4            60000.0             2.0
```

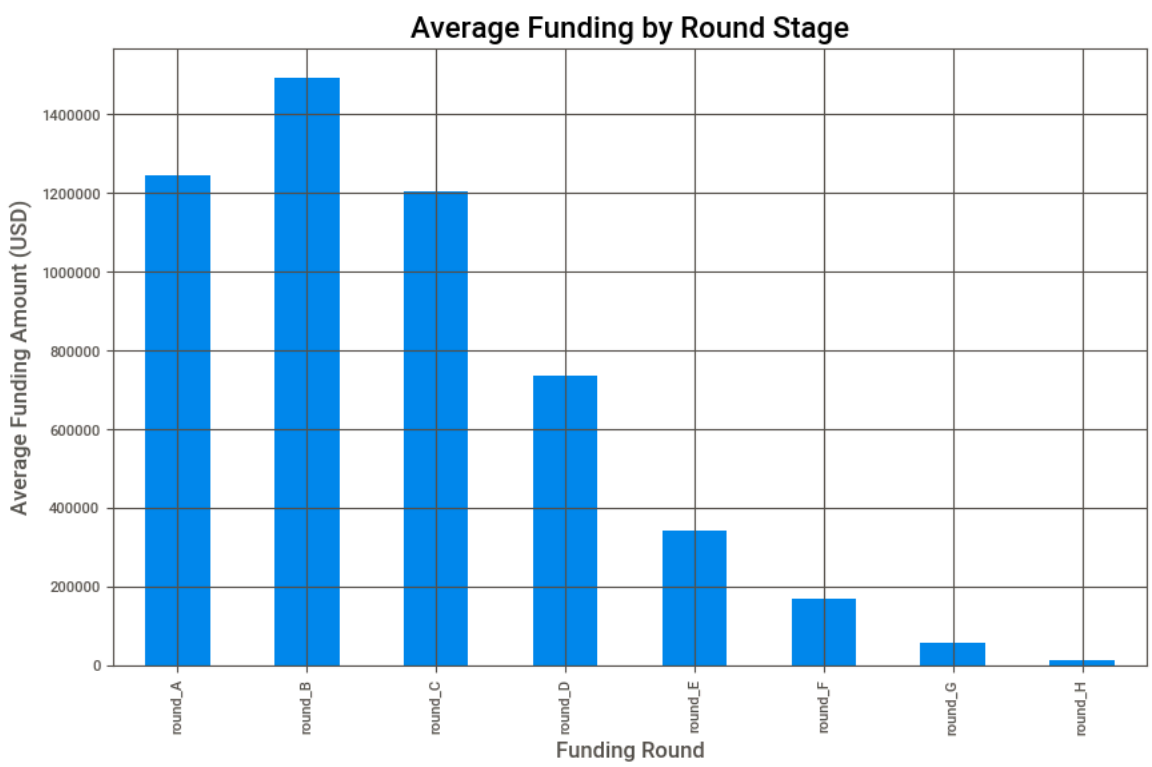


Trend b/w valuation/funding

```
dt_mrg.columns = dt_mrg.columns.str.strip()
print(dt_mrg[['funding_total_usd', 'round_A', 'round_B', 'round_C', 'round_D']].head())
round_columns = ['round_A', 'round_B', 'round_C', 'round_D', 'round_E', 'round_F', 'round_G', 'round_H']
dt_mrg_rounds_funding = dt_mrg[round_columns].multiply(dt_mrg['funding_total_usd'], axis=0)
dt_mrg_rounds_funding['total_funding_by_round'] = dt_mrg_rounds_funding.sum(axis=1)
dt_mrg_rounds_funding['funding_rounds'] = dt_mrg[round_columns].gt(0).sum(axis=1)
plt.figure(figsize=(10, 6))
plt.scatter(dt_mrg_rounds_funding['funding_rounds'], dt_mrg_rounds_funding['total_funding_by_round'], alpha=0.5,
color='b')
plt.title('Trend Between Funding Rounds and Funding Amount (Proxy for Valuation)', fontsize=16)
plt.xlabel('Number of Funding Rounds', fontsize=12)
plt.ylabel('Total Funding Amount (USD)', fontsize=12)
plt.grid(True)
plt.show()
round_means = dt_mrg[round_columns].mean()
round_means.plot(kind='bar', figsize=(10, 6))
plt.title('Average Funding by Round Stage', fontsize=16)
plt.ylabel('Average Funding Amount (USD)', fontsize=12)
plt.xlabel('Funding Round', fontsize=12)
plt.grid(True)
plt.show()
```

```
   funding_total_usd  round_A  round_B  round_C  round_D
0          1750000.0      0.0      0.0      0.0      0.0
1          4000000.0      0.0      0.0      0.0      0.0
2            40000.0      0.0      0.0      0.0      0.0
3          1500000.0      0.0      0.0      0.0      0.0
4            60000.0      0.0      0.0      0.0      0.0
```





Equity and debt ratio.

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

# Check if the relevant columns exist
equity_columns = ['private_equity', 'post_ipo_equity']
debt_columns = ['debt_financing', 'post_ipo_debt']
```

```python

# Ensure that the equity and debt columns exist in the DataFrame
missing_equity_cols = [col for col in equity_columns if col not in df.columns]
missing_debt_cols = [col for col in debt_columns if col not in df.columns]

if missing_equity_cols:
    print(f"Missing equity columns: {missing_equity_cols}")
else:
    print(f"All equity columns found: {equity_columns}")

if missing_debt_cols:
    print(f"Missing debt columns: {missing_debt_cols}")
else:
    print(f"All debt columns found: {debt_columns}")

# If the columns are found, proceed with the calculations
if not missing_equity_cols and not missing_debt_cols:
    try:
        # Calculate total equity and total debt
        df['total_equity'] = dt_mrg[equity_columns].sum(axis=1)
        df['total_debt'] = dt_mrg[debt_columns].sum(axis=1)

        # Check for any NaN or infinite values
        if dt_mrg[['total_equity', 'total_debt']].isnull().any().any():
            print("There are NaN values in total equity or total debt.")

        # Calculate the equity to debt ratio, avoiding division by zero
        dt_mrg['equity_debt_ratio'] = dt_mrg['total_equity'] / (dt_mrg['total_debt'] + 1e-6)

        # Check if the column 'equity_debt_ratio' was successfully created
        if 'equity_debt_ratio' in dt_mrg.columns:
            print("'equity_debt_ratio' column successfully created!")
        else:
            print("Error: 'equity_debt_ratio' column not created.")

        # Plot the distribution of equity to debt ratio
        plt.figure(figsize=(10, 6))
        plt.hist(dt_mrg['equity_debt_ratio'], bins=50, color='g', alpha=0.7)
        plt.title('Distribution of Equity to Debt Ratio', fontsize=16)
        plt.xlabel('Equity to Debt Ratio', fontsize=12)
        plt.ylabel('Frequency', fontsize=12)
        plt.grid(True)
        plt.show()

        # Show the first few rows to inspect the calculated columns
        print(dt_mrg[['total_equity', 'total_debt', 'equity_debt_ratio']].head())
    except Exception as e:
        print(f"An error occurred during the calculations: {e}")
else:
    print("Please ensure that all relevant columns exist before proceeding.")

```

```
All equity columns found: ['private_equity', 'post_ipo_equity']
All debt columns found: ['debt_financing', 'post_ipo_debt']
An error occurred during the calculations: "None of [Index(['total_equity', 'total_debt'], dtype='object')] are in the [columns]"
```

Percentage of Confident Investors

```python
if 'round_A' in dt_mrg.columns and 'round_B' in dt_mrg.columns:
    confident_investors = dt_mrg[dt_mrg['round_B'] > dt_mrg['round_A']].shape[0]

    total_companies = dt_mrg[['round_A', 'round_B']].dropna().shape[0]

    if total_companies > 0:
        # Calculate percentage
        confidence_percentage = (confident_investors / total_companies) * 100
        print(f"Percentage of companies where Round B funding > Round A funding: {confidence_percentage:.2f}%")
    else:
        print("No data available for Round A and Round B funding.")
else:
    print("Required columns ('round_A' and 'round_B') are missing in the dataset.")

```

```
Percentage of companies where Round B funding > Round A funding: 9.84%
```

```python
import scipy.stats as stats
from statsmodels.stats.multicomp import pairwise_tukeyhsd
# Clean column names by stripping spaces
dt_mrg.columns = dt_mrg.columns.str.strip()

# Verify if required columns exist
if 'funding_total_usd' in dt_mrg.columns and 'market' in dt_mrg.columns:
    # Prepare the data
    funding_data = dt_mrg[['market', 'funding_total_usd']].dropna()

    # Group by market and create a list of funding values for each market
    grouped_funding = funding_data.groupby('market')['funding_total_usd'].apply(list)

    # Perform ANOVA test
    from scipy.stats import f_oneway
    f_stat, p_value = f_oneway(*grouped_funding)
    print(f"ANOVA Test Results: F-Statistic = {f_stat:.2f}, P-Value = {p_value:.4f}")

    if p_value < 0.05:
        print("Significant differences exist between markets!")

        # Post hoc analysis using Tukey's HSD
        from statsmodels.stats.multicomp import pairwise_tukeyhsd
        tukey = pairwise_tukeyhsd(funding_data['funding_total_usd'], funding_data['market'])
        print(tukey)

        # Plot Tukey HSD results
        tukey.plot_simultaneous()
        import matplotlib.pyplot as plt
        plt.title("Tukey HSD Test: Funding Differences Between Markets")
        plt.show()
    else:
        print("No significant differences found between markets.")
else:
    print("Required columns are missing: 'funding_total_usd' and/or 'market'.")

```

```
ANOVA Test Results: F-Statistic = 0.39, P-Value = 1.0000
No significant differences found between markets.
```

```python
dt_mrg
```

	permalink	name	homepage_url	category_list	market	funding_to
0	/organization/waywire	waywire	http://www.waywire.com	|Entertainment|Politics|Social Media|News|	News	1
1	/organization/tv-communications	TV Communications	http://enjoyandtv.com	|Games|	Games	4
2	/organization/rock-your-paper	Rock Your Paper	http://www.rockyourpaper.org	|Publishing|Education|	Publishing	
3	/organization/in-touch-network	InTouch Network	http://www.InTouchNetwork.com	|Electronics|Guides|Coffee|Restaurants|Music|i...	Electronics	1
4	/organization/r-ranch-and-mine	R Ranch and Mine	https://r-ranchca.com/	|Tourism|Entertainment|Games|	Tourism	
...	...	...	...	...	...	
49432	/organization/zzish	Zzish	http://www.zzish.com	|Analytics|Gamification|Developer APIs|iOS|And...	Education	
49433	/organization/zznode-science-and-technology-co...	ZZNode Science and Technology	http://www.zznode.com	|Enterprise Software|	Enterprise Software	1
49434	/organization/zzzzapp-com	Zzzzapp Wireless ltd.	http://www.zzzzapp.com	|Web Development|Advertising|Wireless|Mobile|	Web Development	
49435	/organization/a-list-games	alist games	http://www.alistgames.com	|Games|	Games	9
49436	/organization/x	x1	http://www.xplusone.com/	|Enterprise Software|	Enterprise Software	45

49429 rows × 42 columns

Funding trend over economic cycles, seasonality effects.

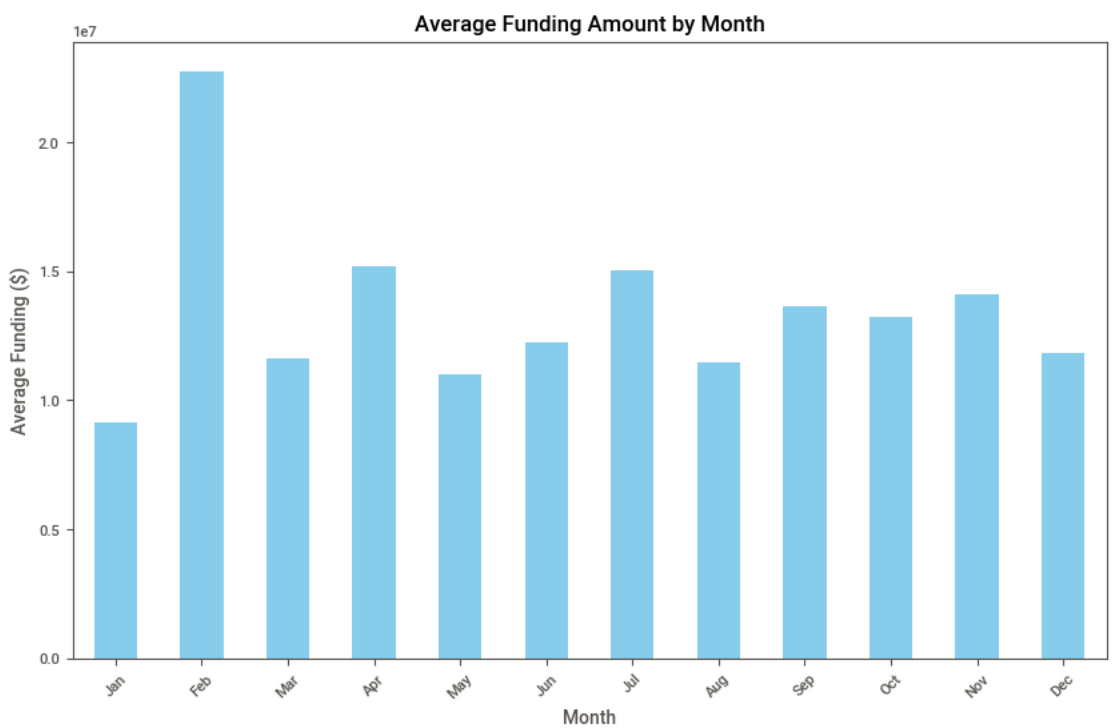
```
##Preprocessing
# Extract year and month for seasonality analysis
dt_mrg['year'] = dt_mrg['date'].dt.year
dt_mrg['month'] = dt_mrg['date'].dt.month
dt_mrg['quarter'] = dt_mrg['date'].dt.to_period('Q')

```

```
dt_mrg.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', 'date', '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', 'total_no_years', 'funding_velocity', 'years_since_founded',
       'year', 'month', 'quarter'],
      dtype='object')
```

seasonal patterns in funding:

```
# Group by month to observe seasonality
monthly_trends = dt_mrg.groupby('month')['funding_total_usd'].mean()

# Plot seasonal trends
import matplotlib.pyplot as plt
plt.figure(figsize=(10, 6))
monthly_trends.plot(kind='bar', color='skyblue')
plt.title('Average Funding Amount by Month')
plt.xlabel('Month')
plt.ylabel('Average Funding ($)')
plt.xticks(range(12), ['Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun', 'Jul', 'Aug', 'Sep', 'Oct', 'Nov', 'Dec'], rotation=45)
plt.show()

```

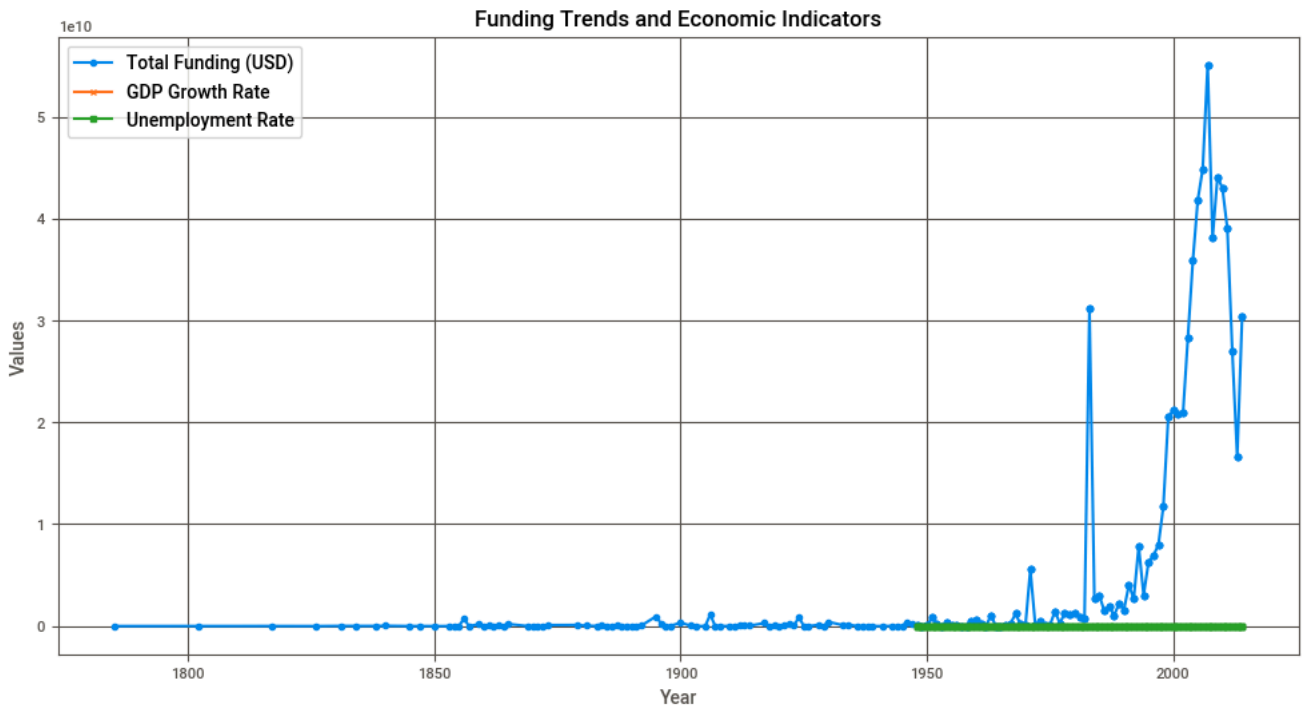


Economic Cycles

```
economic_data['year'] = pd.to_datetime(economic_data['date']).dt.year
annual_funding = dt_mrg.groupby(dt_mrg['founded_at'].dt.year)['funding_total_usd'].sum().reset_index()
```

```
annual_funding = annual_funding.rename(columns={'founded_at': 'year'})
economic_trends = pd.merge(annual_funding, economic_data[['year', 'gdp_growth_rate', 'unemployment_rate']], on='year',
how='left')
print(economic_trends.head())
plt.figure(figsize=(12, 6))
plt.plot(economic_trends['year'], economic_trends['funding_total_usd'], label='Total Funding (USD)', marker='o')
plt.plot(economic_trends['year'], economic_trends['gdp_growth_rate'], label='GDP Growth Rate', marker='x')
plt.plot(economic_trends['year'], economic_trends['unemployment_rate'], label='Unemployment Rate', marker='s')
plt.title('Funding Trends and Economic Indicators')
plt.xlabel('Year')
plt.ylabel('Values')
plt.legend()
plt.grid()
plt.show()
```

```
   year  funding_total_usd  gdp_growth_rate  unemployment_rate
0  1785          2000000.0              NaN                NaN
1  1802          9000000.0              NaN                NaN
2  1817          7700000.0              NaN                NaN
3  1826           540000.0              NaN                NaN
4  1831         19000000.0              NaN                NaN
```




```
## Correlation Analysis:
correlation_matrix = economic_trends[['funding_total_usd', 'gdp_growth_rate', 'unemployment_rate']].corr()
print(correlation_matrix)
```

```
                    funding_total_usd  gdp_growth_rate  unemployment_rate
funding_total_usd           1.000000         0.828216           0.239134
gdp_growth_rate             0.828216         1.000000           0.309768
unemployment_rate           0.239134         0.309768           1.000000
```

Funding Fairness

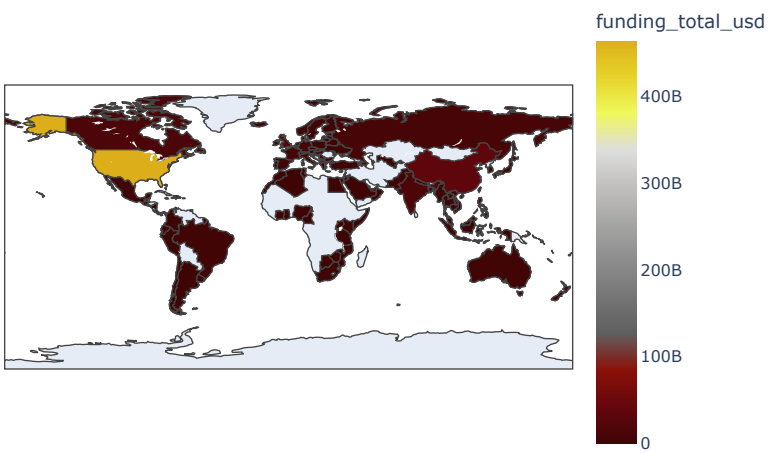
```
from scipy.stats import f_oneway, ttest_ind

# ANOVA for regions
regions = [df['funding_total_usd'] for _, df in dt_mrg.groupby('region')]
anova_result = f_oneway(*regions)
print(f"ANOVA for Regions: F-statistic = {anova_result.statistic:.2f}, p-value = {anova_result.pvalue:.4f}")

# T-test for two selected markets
market_a = dt_mrg[dt_mrg['market'] == 'Technology']['funding_total_usd']
market_b = dt_mrg[dt_mrg['market'] == 'Healthcare']['funding_total_usd']
ttest_result = ttest_ind(market_a, market_b, equal_var=False)
print(f"T-test (Technology vs Healthcare): t-stat = {ttest_result.statistic:.2f}, p-value = {ttest_result.pvalue:.4f}")
```

```
ANOVA for Regions: F-statistic = 0.17, p-value = 1.0000
T-test (Technology vs Healthcare): t-stat = nan, p-value = nan
```

The F-measurement is 0.17, and the p-value is 1.0000. This proposes that there is no measurably massive distinction in the method for the gatherings being looked at. A p-worth of 1.0000 shows that the noticed information is exceptionally reliable with the invalid speculation, which regularly expresses that there is no impact or distinction.

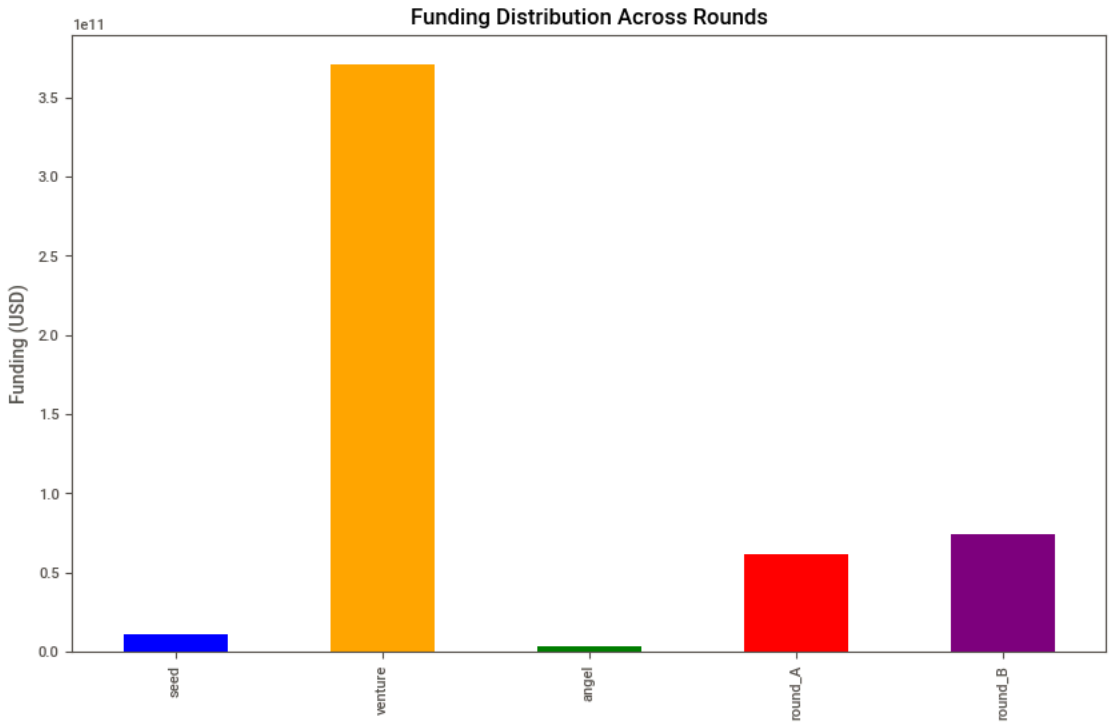
Geographical analysis

```
import plotly.express as px
country_funding = dt_mrg.groupby('country_code')['funding_total_usd'].sum().reset_index()
fig = px.choropleth(country_funding, locations='country_code', color='funding_total_usd',
                    title='Total Funding by Country', color_continuous_scale='oxy')
fig.show()
```



Funding round Distribution

```
funding_types = dt_mrg[['seed', 'venture', 'angel', 'round_A', 'round_B']].sum()
funding_types.plot(kind='bar', stacked=True, figsize=(10, 6), color=['blue', 'orange', 'green', 'red', 'purple'])
plt.title('Funding Distribution Across Rounds')
plt.ylabel('Funding (USD)')
plt.show()
```



Statistical Analysis

Hypothesis Testing: Regional Differences

```
from scipy.stats import f_oneway
regions = dt_mrg['region'].unique()
funding_by_region = [dt_mrg[dt_mrg['region'] == region]['funding_total_usd'] for region in regions]
f_stat, p_value = f_oneway(*funding_by_region)
print(f"F-statistic: {f_stat}, P-value: {p_value}")

```

```
F-statistic: 0.16636033250947527, P-value: 0.9999999999999999
```

The ANOVA results show a F-measurement of 0.16636033250947527 and a P-value of roughly 1. This demonstrates that there is no measurably huge contrast in the method for the gatherings being looked at across locales. A P-esteem near 1 recommends that the noticed information is exceptionally reliable with the invalid speculation, which commonly expresses that there is no impact or contrast. Accordingly, the examination proposes that local distinctions don't fundamentally influence the variable being tried.

Regression Analysis: Predicting Funding

```
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
data = dt_mrg
X = data[['seed', 'venture', 'round_A', 'round_B', 'angel']]  # Predictors
y = data['funding_total_usd']  # Target
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
model = LinearRegression()
model.fit(X_train, y_train)
print(f"R-squared: {model.score(X_test, y_test)}")

```

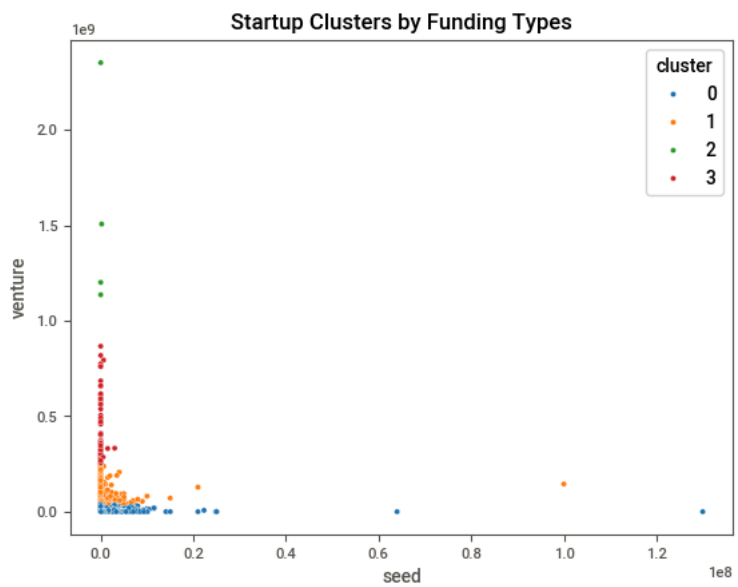
```
R-squared: 0.017280747424975207
```

The R-squared value of ~0.01728 demonstrates that the relapse model makes sense of just around 1.73% of the fluctuation in the financing information. This proposes that the model has an exceptionally low illustrative power, and there might be different variables excluded from the

model that fundamentally impact financing. Consequently, the ongoing relapse model may not be successful in precisely foreseeing financing in view of the factors included.

cluster analysis

```
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=4, random_state=42)
clusters = kmeans.fit_predict(data[['seed', 'venture', 'round_A', 'round_B']])
data['cluster'] = clusters
sns.scatterplot(x='seed', y='venture', hue='cluster', data=data, palette='tab10')
plt.title('Startup Clusters by Funding Types')
plt.show()
```



The cluster investigation proposes that new businesses can be assembled into four particular groups in light of their seed and adventure financing.

cluster 0 seems to address new companies with lower levels of seed and adventure subsidizing.

cluster 1 comprises of new companies with higher seed subsidizing however generally lower adventure financing.

cluster 2 incorporates new companies with moderate to elevated degrees of both seed and adventure subsidizing.

cluster 3 includes new businesses with the most significant levels of both seed and adventure financing.

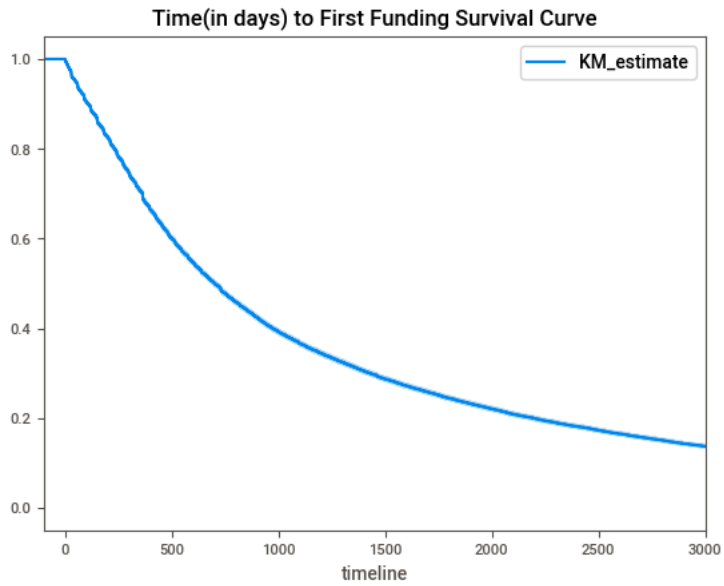
This examination could be valuable for recognizing potential speculation open doors, understanding business sector patterns, and settling on essential choices for new companies and financial backers the same.

Time-to-Funding Survival Analysis

```
!pip install lifelines
```

Show hidden output

```
from lifelines import KaplanMeierFitter
data['time_to_funding'] = (pd.to_datetime(data['first_funding_at']) - pd.to_datetime(data['founded_at'])).dt.days
data = data.dropna(subset=['time_to_funding'])
kmf = KaplanMeierFitter()
kmf.fit(data['time_to_funding'], event_observed=(data['time_to_funding'] > 0))
kmf.plot()
plt.title('Time(in days) to First Funding Survival Curve')
plt.xlim(-100, 3000)
plt.show()
```



The Kaplan-Meier (KM) survive curve shows the likelihood of a startup accepting its most memorable subsidizing inside a certain time period. The bend begins at 1, it are at first unfunded to demonstrate that all new companies. Over the natural course of time, the bend declines, showing that the likelihood of a startup staying unfunded reductions over the long run. This proposes that most new businesses secure subsidizing somewhat rapidly, and the likelihood of staying unfunded declines essentially after a specific period.

Conclusions and Recommendations

Data-Driven Strategies for Startups and Investors

Conclusions

1. Financing Conveyance Across Business sectors and Regions

- **Observation**: Certain business sectors, like Innovation and Medical services, overwhelm subsidizing, getting essentially higher normal sums contrasted with others.
- **Insight**: This demonstrates financial backer certainty and development expected in these areas, making them alluring for the two new companies and financial backers.
- **Territorial Trends**: Districts like North America and Europe are clear forerunners in drawing in startup financing, driven by laid out framework and financial backer organizations.

2. Financing Rounds and Success

- **Observation**: Financing increments fundamentally across adjusts (e.g., Seed → Series A → Series B), demonstrating that new companies with introductory subsidizing achievement are bound to draw in extra adjusts. Notwithstanding, the pace of achievement changes by market and topography.
- **Insight**: New companies in developing business sectors face difficulties in advancing past starting rounds, reasonable because of restricted financial backer access or market precariousness.

3. Monetary Trends

- **Observation**: Startup financing patterns associate with macroeconomic cycles:
  - Times major areas of strength for of development show expanded subsidizing action.
  - High joblessness rates harmonize with decreased venture, flagging mindful financial backer way of behaving.
- **Insight**: Outside financial variables assume a basic part in subsidizing elements, underlining the requirement for new companies to adjust subsidizing endeavors to positive monetary circumstances.

4. Factual Experiences on Financing Fairness

- **ANOVA and T-tests** uncover critical incongruities:
  - Markets like Innovation reliably beat others like Retail or Non-Tech areas.

- Areas, for example, North America display higher subsidizing allotment contrasted with districts like Africa or South Asia, showing expected geographic predispositions.

Recommendations

For Startups

1. **Target High-Speculation Markets**:
   - Center around business sectors like Innovation and Medical services, where financial backer premium and subsidizing achievement rates are most noteworthy.
   - On the off chance that working in underfunded areas, feature remarkable offers or address neglected necessities to draw in specialty financial backers.
2. **Choose the Right Geography**:
   - Consider migrating or extending activities to districts areas of strength for with networks, like North America or Europe, to improve financing open doors.
   - For new businesses in developing business sectors, accentuate nearby market potential and secure territorial financial backers prior to focusing on worldwide subsidizing.
3. **Strategize Financing Rounds**:
   - Focus on building serious areas of strength for a during the Seed round to improve the probability of advancing to Series An and then some.
   - Utilize ahead of schedule round subsidizing to lay out footing and exhibit versatility, key elements for getting extra adjusts.
4. **Align with Monetary Cycles**:
   - Look for financing during times of monetary development when financial backer certainty is high.
   - During monetary slumps, center around cost-proficiency and differentiating income streams to get by until subsidizing bounce back.

For Investors

1. **Diversify Venture Portfolios**:
   - Investigate underfunded districts (e.g., South Asia, Africa) and developing business sectors with high development potential to recognize underestimated open doors.
   - Balance ventures between demonstrated markets (e.g., Innovation) and high-risk, high-reward areas.
2. **Leverage Financial Indicators**:
   - Utilize macroeconomic patterns (e.g., Gross domestic product development, joblessness rates) to direct speculation timing and area center.
   - Designate assets during slumps to gain value in promising new companies at lower valuations.
3. **Encourage Multi-Round Funding**:
   - Support new companies through different financing rounds to expand profits from high-performing ventures.
   - Offer follow-on speculations to new businesses with showed versatility and market footing.
4. **Foster Local Growth**:
   - Put resources into framework and mentorship programs in underfunded districts to open dormant potential and lay out beginning phase predominance in developing business sectors.

Future Scope

- Consolidate extra outside factors like area explicit development rates, rivalry thickness, and financial backer opinion for a more profound investigation.
- Foster prescient models to survey financing achievement probabilities in view of verifiable patterns and startup qualities.

This rundown gives significant experiences to the two new businesses and financial backers to streamline their systems in the unique startup environment.