Checkpoint 1

Data Cleaning

In [1]:

```
# importing all required libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

In [2]:

```
#importing the csv files
companies=pd.read_csv("companies.txt",sep="\t", encoding = "ISO-8859-1")
rounds2=pd.read_csv("rounds2.csv", encoding = "ISO-8859-1")
```

In [3]:

```
rounds2.head()
```

Out[3]:

	company_permalink	funding_round_permalink	funding_round_type	funding_round_code	funded_at	raised_amount_
0	/organization/-fame	/funding-round/9a01d05418af9f794eebff7ace91f638	venture	B	05-01-2015	1000000
1	/ORGANIZATION/-QOUNTER	/funding-round/22dacff496eb7acb2b901dec1dfe5633	venture	A	14-10-2014	N
2	/organization/-qounter	/funding-round/b44fbb94153f6cdef13083530bb48030	seed	NaN	01-03-2014	70000
3	/ORGANIZATION/-THE-ONE-OF-THEM-INC-	/funding-round/650b8f704416801069bb178a1418776b	venture	B	30-01-2014	340687
4	/organization/0-6-com	/funding-round/5727accaeaa57461bd22a9bdd945382d	venture	A	19-03-2008	200000

In [4]:

```
companies.head()
```

Out[4]:

	permalink	name	homepage_url	category_list	status	country_code	state_code	region	
0	/Organization/-Fame	#fame	http://livfame.com	Media	operating	IND	16	Mumbai	Mum
1	/Organization/-Qounter	:Qounter	http://www.qounter.com	Application Platforms|Real Time|Social Network...	operating	USA	DE	DE - Other	Delaw
2	/Organization/-The-One-Of-Them-Inc-	(THE) ONE of THEM,Inc.	http://oneofthem.jp	Apps|Games|Mobile	operating	NaN	NaN	NaN	N
3	/Organization/0-6-Com	0-6.com	http://www.0-6.com	Curated Web	operating	CHN	22	Beijing	Bei
4	/Organization/004-Technologies	004 Technologies	http://004gmbh.de/en/004-interact	Software	operating	USA	IL	Springfield, Illinois	Champa

In [5]:

```
len(rounds2["company_permalink"].unique())
```

Out[5]:

```
90247
```

In [6]:

```
len(companies["permalink"].unique())
```

Out[6]:

```
66368
```

Number of Unique companies in rounds2 dataframe are 90247 and number of Unique companies in companies dataframe are 6638. Also,Permalink can be used as the Unique key for each company.

In [7]:

```
companies['permalink']=companies['permalink'].apply(lambda x: x.lower())
companies.head()
```

Out[7]:

	permalink	name	homepage_url	category_list	status	country_code	state_code	region	c
0	/organization/-fame	#fame	http://livfame.com	Media	operating	IND	16	Mumbai	Mum
1	/organization/-qounter	:Qounter	http://www.qounter.com	Application Platforms|Real Time|Social Network...	operating	USA	DE	DE - Other	Delaw
2	/organization/-the-one-of-them-inc-	(THE) ONE of THEM,Inc.	http://oneofthem.jp	Apps|Games|Mobile	operating	NaN	NaN	NaN	N
3	/organization/0-6-com	0-6.com	http://www.0-6.com	Curated Web	operating	CHN	22	Beijing	Beij
4	/organization/004-technologies	004 Technologies	http://004gmbh.de/en/004-interact	Software	operating	USA	IL	Springfield, Illinois	Champa

In [8]:

```
rounds2["company_permalink"]=rounds2['company_permalink'].apply(lambda x: x.lower())
rounds2.head()
```

Out[8]:

	company_permalink	funding_round_permalink	funding_round_type	funding_round_code	funded_at	raised_amount_
0	/organization/-fame	/funding-round/9a01d05418af9f794eebff7ace91f638	venture	B	05-01-2015	1000000
1	/organization/-qounter	/funding-round/22dacff496eb7acb2b901dec1dfe5633	venture	A	14-10-2014	N
2	/organization/-qounter	/funding-round/b44fbb94153f6cdef13083530bb48030	seed	NaN	01-03-2014	70000
3	/organization/-the-one-of-them-inc-	/funding-round/650b8f704416801069bb178a1418776b	venture	B	30-01-2014	340687
4	/organization/0-6-com	/funding-round/5727accaeaa57461bd22a9bdd945382d	venture	A	19-03-2008	200000

Checking if number of unique companies in rounds2 DF and companies DF are same

In [9]:

```
len(rounds2['company_permalink'].unique())
```

Out[9]:

66370

In [10]:

```
len(companies['permalink'].unique())
```

Out[10]:

66368

Yes there are 2 extra companies in rounds2 DF as compared to companies DF.

Lets Check for companies present in rounds2 DF but not in companies DF and remove them.

In [11]:

```
rounds2.loc[~rounds2['company_permalink'].isin(companies['permalink']),:]
```

Out[11]:

	company_permalink	funding_round_permalink	funding_round_type	funding_round_code	funded_at	raised_
29597	/organization/e-cãbica	/funding-round/8491f74869e4fe8ba9c378394f8fbdea	seed	NaN	01-02-2015	
31863	/organization/energystone-games-çµç³æ¸æ	/funding-round/b89553f3d2279c5683ae93f45a21cfe0	seed	NaN	09-08-2014	
45176	/organization/huizuche-com-æ ç§ÿè½¦	/funding-round/8f8a32dbeeb0f831a78702f83af78a36	seed	NaN	18-09-2014	
58473	/organization/magnet-tech-ç£ç³ç§æ	/funding-round/8fc91fbb32bc95e97f151dd0cb4166bf	seed	NaN	16-08-2014	
101036	/organization/tipcat-interactive-æ²èÿä¿¡æ¯ç...	/funding-round/41005928a1439cb2d706a43cb661f60f	seed	NaN	06-09-2010	
109969	/organization/weiche-tech-âè½¦ç§æ	/funding-round/f74e457f838b81fa0b29649740f186d8	venture	A	06-09-2015	
113839	/organization/zengame-ç¦æ¸ç§æ	/funding-round/6ba28fb4f3eadf5a9c6c81bc5dde6cdf	seed	NaN	17-07-2010	

As we see in rounds2 Df that there are weird characters at different indices. These weird characters are not present in actual CSV file. So this problem is because of improper decoding of csv file while importing.

After searching a lot and trying all compaitible decodings(none of them worked) it can be concluded that it has multiple encoding. Searching Stackoverflow we got into the conclusion that this problem can be solved with the following code:

In [12]:

```
rounds2['company_permalink'] = rounds2.company_permalink.str.encode('utf-8').str.decode('ascii', 'ignore')
rounds2.loc[~rounds2['company_permalink'].isin(companies['permalink']), :]
```

Out[12]:

	company_permalink	funding_round_permalink	funding_round_type	funding_round_code	funded_at	raised_a
77	/organization/10north	/funding-round/b41ff7de932f8b6e5bbeed3966c0ed6a	equity_crowdfunding	NaN	12-08-2014	
729	/organization/51wofang-	/funding-round/346b9180d276a74e0fbb2825e66c6f5b	venture	A	06-07-2015	
2670	/organization/adslinked	/funding-round/449ae54bb63c768c232955ca6911dee4	seed	NaN	29-09-2014	
	/organization/aesthetic-	/funding-			12-10-	

	company_permalink	funding_round_permalink	funding_round_type	funding_round_code	funded_at	raised_a
3166	everything-social-network	round/62593455f1[illegible]	equity_crowdfunding	NaN	[illegible]	
3291	/organization/affluent-attach-club-2	/funding-round/626678bdf1654bc4df9b1b34647a4df1	seed	NaN	15-10-2014	
...	...	...	...	...	...	
110545	/organization/whodats-spaces	/funding-round/d5d6db3d1e6c54d71a63b3aa0c9278e6	seed	NaN	28-10-2014	
113839	/organization/zengame-	/funding-round/6ba28fb4f3eadf5a9c6c81bc5dde6cdf	seed	NaN	17-07-2010	
114946	/organization/eron	/funding-round/59f4dce44723b794f21ded3daed6e4fe	venture	A	01-08-2014	
114947	/organization/asys-2	/funding-round/35f09d0794651719b02bbfd859ba9ff5	seed	NaN	01-01-2015	
114948	/organization/novatiff-reklam-ve-tantm-hizmetl...	/funding-round/af942869878d2cd788ef5189b435ebc4	grant	NaN	01-10-2013	

74 rows × 6 columns

Now everything seems fine. There are no special characters left in rounds2 DF

In [13]:

```
len(rounds2['company_permalink'].unique())
```

Out[13]:

```
66368
```

In [14]:

```
len(companies['permalink'].unique())
```

Out[14]:

```
66368
```

As we can see, after cleaning, now the number of unique companies in both rounds2 and companies df is same.

Let's check if companies DF also have special character.

In [15]:

```
companies.loc[~companies['permalink'].isin(rounds2['company_permalink']),:]
```

Out[15]:

	permalink	name	homepage_url	category_list	status	country_code	state_code	regi
43	/organization/10â°north	10Â°North	NaN	Fashion	operating	CAN	ON	Toro
426	/organization/51wofang-æ å¿§ææ¿	51wofang æ å¿§ææ¿	http://www.51wofang.com	NaN	closed	NaN	NaN	N
1506	/organization/adslinkedâ¢	AdsLinkedâ¢	http://www.adslinked.com	Advertising|Internet	operating	NaN	NaN	N
1775	/organization/aesthetic-everythingâ®-social-ne...	Aesthetic EverythingÂ® Social Network	http://aestheticeverything.com/	Public Relations	operating	USA	CA	L Ange
1834	/organization/affluent-attachã©-club-2	Affluent AttachÃ© Club	http://www.affluentattache.com/	Hospitality	operating	USA	CA	L Ange
...	...	...	...	...	...	...	...	
63833	/organization/whodatâs-spaces	Whodatâs Spaces	NaN	Apps	operating	NaN	NaN	N
				Internet|Mobile				

	permalink	name	homepage_url	category_list	status	country_code	state_code	regi
65778	/organization/zengame-çÍæ¸¸çŠæ	ZenGame çÍæ¸¸çŠæ	http://www.zen-game.com	Internet|Mobile Games|Online Gaming	closed	NaN	NaN	N
66365	/organization/ãeron	ÃERON	http://www.aeron.hu/	NaN	operating	NaN	NaN	N
66366	/organization/ãasys-2	Ãasys	http://www.oasys.io/	Consumer Electronics|Internet of Things|Teleco...	operating	USA	CA	SF E A
66367	/organization/ä°novatiff-reklam-ve-tanä±tä±m-h...	Ä°novatiff Reklam ve TanÄ±tÄ±m Hizmetleri Tic	http://inovatiff.com	Consumer Goods|E-Commerce|Internet	operating	NaN	NaN	N

68 rows × 10 columns

As we can see companies df also has special characters. We will need to filter this df also.

In [16]:

```
companies['permalink'] = companies.permalink.str.encode('utf-8').str.decode('ascii', 'ignore')
companies.loc[~companies['permalink'].isin(rounds2['company_permalink']),:]
```

Out[16]:

permalink	name	homepage_url	category_list	status	country_code	state_code	region	city	founded_at

Now companies df is also clean . As we see, now there are no companies in companies df that are not present in rounds2 df. Also ,there are no companies in rounds2 df that are not present in companies df.

Now lets create a seperate csv file for both csv so that we dont have deal with encoding everytime we work.

In [17]:

```
rounds2.to_csv("rounds2_clean.csv",sep=",",index=False)
companies.to_csv("companies_clean.csv",sep="\t",index=False)
```

In [18]:

```
rounds2_clean=pd.read_csv("rounds2_clean.csv")
rounds2_clean.head()
```

Out[18]:

	company_permalink	funding_round_permalink	funding_round_type	funding_round_code	funded_at	raised_amount_
0	/organization/-fame	/funding-round/9a01d05418af9f794eebff7ace91f638	venture	B	05-01-2015	1000000
1	/organization/-qounter	/funding-round/22dacff496eb7acb2b901dec1dfe5633	venture	A	14-10-2014	N
2	/organization/-qounter	/funding-round/b44fbb94153f6cdef13083530bb48030	seed	NaN	01-03-2014	70000
3	/organization/-the-one-of-them-inc-	/funding-round/650b8f704416801069bb178a1418776b	venture	B	30-01-2014	340687
4	/organization/0-6-com	/funding-round/5727accaeaa57461bd22a9bdd945382d	venture	A	19-03-2008	200000

In [19]:

```
rounds2_clean.rename(columns={'company_permalink':'permalink'}, inplace=True)
```

Renamed company_permalink column to permalink so that I don't face any ambiguity during merge of both DF(rounds2_clean and companies_clean),

In [20]:

In [20]:

```
rounds2_clean.head()
```

Out[20]:

	permalink	funding_round_permalink	funding_round_type	funding_round_code	funded_at	raised_amount_usd
0	/organization/-fame	/funding-round/9a01d05418af9f794eebff7ace91f638	venture	B	05-01-2015	10000000.0
1	/organization/-qounter	/funding-round/22dacff496eb7acb2b901dec1dfe5633	venture	A	14-10-2014	NaN
2	/organization/-qounter	/funding-round/b44fbb94153f6cdef13083530bb48030	seed	NaN	01-03-2014	700000.0
3	/organization/-the-one-of-them-inc-	/funding-round/650b8f704416801069bb178a1418776b	venture	B	30-01-2014	3406878.0
4	/organization/0-6-com	/funding-round/5727accaeaa57461bd22a9bdd945382d	venture	A	19-03-2008	2000000.0

Checking the number of missing values in permalink column of both the DFs

In [21]:

```
rounds2_clean.isnull().sum()
```

Out[21]:

```
permalink                    0
funding_round_permalink      0
funding_round_type           0
funding_round_code       83809
funded_at                    0
raised_amount_usd        19990
dtype: int64
```

In [22]:

```
companies_clean=pd.read_csv("companies_clean.csv", sep="\t", encoding = "ISO-8859-1")
companies_clean.isnull().sum()
```

Out[22]:

```
permalink            0
name                 1
homepage_url      5058
category_list     3148
status               0
country_code      6958
state_code        8547
region            8030
city              8028
founded_at       15221
dtype: int64
```

There are no missing values in permalink of both df. So we are good to merge these 2 in a new df called master

In [23]:

```
master=pd.merge(companies_clean,rounds2_clean)
master.head(25)
```

Out[23]:

	permalink	name	homepage_url	category_list	status	country_code	st
0	/organization/-fame	#fame	http://livfame.com	Media	operating	IND	

	permalink	name	homepage_url	category_list	status	country_code	st
1	/organization/-qounter	:Qounter	http://www.qounter.com	Application Platforms|Real Time|Social Network...	operating	USA	
2	/organization/-qounter	:Qounter	http://www.qounter.com	Application Platforms|Real Time|Social Network...	operating	USA	
3	/organization/-the-one-of-them-inc-	(THE) ONE of THEM,Inc.	http://oneofthem.jp	Apps|Games|Mobile	operating	NaN	
4	/organization/0-6-com	0-6.com	http://www.0-6.com	Curated Web	operating	CHN	
5	/organization/004-technologies	004 Technologies	http://004gmbh.de/en/004-interact	Software	operating	USA	
6	/organization/01games-technology	01Games Technology	http://www.01games.hk/	Games	operating	HKG	
7	/organization/0ndine-biomedical-inc	Ondine Biomedical Inc.	http://ondinebio.com	Biotechnology	operating	CAN	
8	/organization/0ndine-biomedical-inc	Ondine Biomedical Inc.	http://ondinebio.com	Biotechnology	operating	CAN	
9	/organization/0xdata	H2O.ai	http://h2o.ai/	Analytics	operating	USA	
10	/organization/0xdata	H2O.ai	http://h2o.ai/	Analytics	operating	USA	
11	/organization/0xdata	H2O.ai	http://h2o.ai/	Analytics	operating	USA	
12	/organization/0xdata	H2O.ai	http://h2o.ai/	Analytics	operating	USA	
13	/organization/1	One Inc.	http://whatis1.com	Mobile	operating	USA	
14	/organization/1	One Inc.	http://whatis1.com	Mobile	operating	USA	
15	/organization/1	One Inc.	http://whatis1.com	Mobile	operating	USA	
16	/organization/1-2-3-listo	1,2,3 Listo	http://www.123listo.com	E-Commerce	operating	CHL	
17	/organization/1-4-all	1-4 All	NaN	Entertainment|Games|Software	operating	USA	
18	/organization/1-618-technology	1.618 Technology	http://www.Homeandcondogallery.com	Networking|Real Estate|Web Hosting	operating	USA	
19	/organization/1-800-dentist	1-800-DENTIST	http://www.1800dentist.com	Health and Wellness	operating	USA	
20	/organization/1-800-doctors	1-800-DOCTORS	http://1800doctors.com	Health and Wellness	operating	USA	
21	/organization/1-800-publicrelations-inc-	1-800-PublicRelations, Inc.	http://www.1800publicrelations.com	Internet Marketing|Media|Public Relations	operating	USA	
22	/organization/1-mainstream	1 Mainstream	http://www.1mainstream.com	Apps|Cable|Distribution|Software	acquired	USA	
23	/organization/1-of-99	1 of 99	NaN	Entertainment|Games	operating	USA	
24	/organization/10-20-media	10-20 Media	http://www.10-20media.com	E-Commerce	operating	USA	

Checkpoint 2

Funding Type Analysis

In [24]:

```
master.columns
```

Out[24]:

```
Index(['permalink', 'name', 'homepage_url', 'category_list', 'status',
       'country_code', 'state_code', 'region', 'city', 'founded_at',
       'funding_round_permalink', 'funding_round_type', 'funding_round_code',
       'funded_at', 'raised_amount_usd'],
      dtype='object')
```

In [25]:

```
master.isnull().sum()
```

Out[25]:

```
permalink                    0
name                         1
homepage_url              6134
category_list             3410
status                       0
country_code              8678
state_code               10946
region                   10167
city                     10164
founded_at               20521
funding_round_permalink      0
funding_round_type           0
funding_round_code       83809
funded_at                    0
raised_amount_usd        19990
dtype: int64
```

Finding the fraction of missing values(Column wise)

In [26]:

```
100*(master.isnull().sum())/len(master.index)
```

Out[26]:

```
permalink                 0.000000
name                      0.000870
homepage_url              5.336280
category_list             2.966533
status                    0.000000
country_code              7.549435
state_code                9.522484
region                    8.844792
city                      8.842182
founded_at               17.852265
funding_round_permalink   0.000000
funding_round_type        0.000000
funding_round_code       72.909725
funded_at                 0.000000
raised_amount_usd        17.390321
dtype: float64
```

Let's drop the columns which have too much missing values or are not necessary for our analysis

In [27]:

```
master=master.drop(['funding_round_code','homepage_url','founded_at','state_code','region','city']
,axis=1)
```

In [28]:

```
master.head()
```

Out[28]:

	permalink	name	category_list	status	country_code	funding_round_permalink	funding_round_type
0	/organization/-fame	#fame	Media	operating	IND	/funding-round/9a01d05418af9f794eebff7ace91f638	venture
1	/organization/-		Application Platforms|Real		USA	/funding-	

	permalink	name	category_list	status	country_code	funding_round_permalink	funding_round_type
1	qounter	:Qounter	Network...	operating	USA	round/22dacff496eb7acb2b901dec1dfe5633	venture
2	/organization/-qounter	:Qounter	Application Platforms|Real Time|Social Network...	operating	USA	/funding-round/b44fbb94153f6cdef13083530bb48030	seed
3	/organization/-the-one-of-them-inc-	(THE) ONE of THEM,Inc.	Apps|Games|Mobile	operating	NaN	/funding-round/650b8f704416801069bb178a1418776b	venture
4	/organization/0-6-com	0-6.com	Curated Web	operating	CHN	/funding-round/5727accaeaa57461bd22a9bdd945382d	venture

Let's see again how much NaN values are left (column wise)

In [29]:

```
100*(master.isnull().sum())/len(master.index)
```

Out[29]:

```
permalink                   0.000000
name                        0.000870
category_list               2.966533
status                      0.000000
country_code                7.549435
funding_round_permalink     0.000000
funding_round_type          0.000000
funded_at                   0.000000
raised_amount_usd          17.390321
dtype: float64
```

In [30]:

```
master['raised_amount_usd'].describe()
```

Out[30]:

```
count    9.495900e+04
mean     1.042687e+07
std      1.148212e+08
min      0.000000e+00
25%      3.225000e+05
50%      1.680511e+06
75%      7.000000e+06
max      2.127194e+10
Name: raised_amount_usd, dtype: float64
```

Let's drop the rows in raised_amount_usd having NaN values

In [31]:

```
master.dropna(subset=['raised_amount_usd'], how='all', inplace=True)
```

In [32]:

```
100*(master.isnull().sum())/len(master.index)
```

Out[32]:

```
permalink                   0.000000
name                        0.001053
category_list               1.099422
status                      0.000000
country_code                6.161607
funding_round_permalink     0.000000
funding_round_type          0.000000
funded_at                   0.000000
raised_amount_usd           0.000000
dtype: float64
```

Now raised_amount_usd is filtered and it contains no more NaN values

In [33]:

```
master['country_code'].value_counts()
```

Out[33]:

```
USA    62049
GBR     5019
CAN     2616
CHN     1927
IND     1649
       ...
SOM        1
GRD        1
MKD        1
QAT        1
PRY        1
Name: country_code, Length: 134, dtype: int64
```

We saw in code lin 31 that there are almost 6.16% percent of rows missing the country_code. As this is very small % so we can delete these rows also.

In [34]:

```
master.dropna(subset=['country_code'], how='all', inplace=True)
100*(master.isnull().sum())/len(master.index)
```

Out[34]:

```
permalink                 0.000000
name                      0.001122
category_list             0.649773
status                    0.000000
country_code              0.000000
funding_round_permalink   0.000000
funding_round_type        0.000000
funded_at                 0.000000
raised_amount_usd         0.000000
dtype: float64
```

So all NaN rows of country_code are deleted. We notice here that category_list and name are also having a very small % of NaN values. So lets also remove these too.

In [35]:

```
master.dropna(subset=['category_list'], how='all', inplace=True)
100*(master.isnull().sum())/len(master.index)
```

Out[35]:

```
permalink                 0.00000
name                      0.00113
category_list             0.00000
status                    0.00000
country_code              0.00000
funding_round_permalink   0.00000
funding_round_type        0.00000
funded_at                 0.00000
raised_amount_usd         0.00000
dtype: float64
```

In [36]:

```
master.dropna(subset=['name'], how='all', inplace=True)
100*(master.isnull().sum())/len(master.index)
```

Out[36]:

```
permalink                  0.0
name                       0.0
category_list              0.0
status                     0.0
country_code               0.0
funding_round_permalink    0.0
funding_round_type         0.0
funded_at                  0.0
raised_amount_usd          0.0
dtype: float64
```

In [37]:

```
master.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 88528 entries, 0 to 114947
Data columns (total 9 columns):
permalink                  88528 non-null object
name                       88528 non-null object
category_list              88528 non-null object
status                     88528 non-null object
country_code               88528 non-null object
funding_round_permalink    88528 non-null object
funding_round_type         88528 non-null object
funded_at                  88528 non-null object
raised_amount_usd          88528 non-null float64
dtypes: float64(1), object(8)
memory usage: 6.8+ MB
```

Now there's no NaN rows. Everything is filtered. Lets create a new clean df for future reference.

In [38]:

```
master.to_csv("master_df.csv",sep=",",index=False)
```

In [39]:

```
df=pd.read_csv("master_df.csv")
```

In [40]:

```
df.head()
```

Out[40]:

	permalink	name	category_list	status	country_code	funding_round_permalink	funding_round_t
0	/organization/-fame	#fame	Media	operating	IND	/funding-round/9a01d05418af9f794eebff7ace91f638	ven
1	/organization/-qounter	:Qounter	Application Platforms|Real Time|Social Network...	operating	USA	/funding-round/b44fbb94153f6cdef13083530bb48030	s
2	/organization/0-6-com	0-6.com	Curated Web	operating	CHN	/funding-round/5727accaeaa57461bd22a9bdd945382d	ven
3	/organization/01games-technology	01Games Technology	Games	operating	HKG	/funding-round/7d53696f2b4f607a2f2a8cbb83d01839	undisclo
4	/organization/0ndine-biomedical-inc	Ondine Biomedical Inc.	Biotechnology	operating	CAN	/funding-round/2b9d3ac293d5cdccbecff5c8cb0f327d	s

We need to keep only the 4 funding types mentioned in Checkpoint so let's remove the

unwanted ones

In [41]:

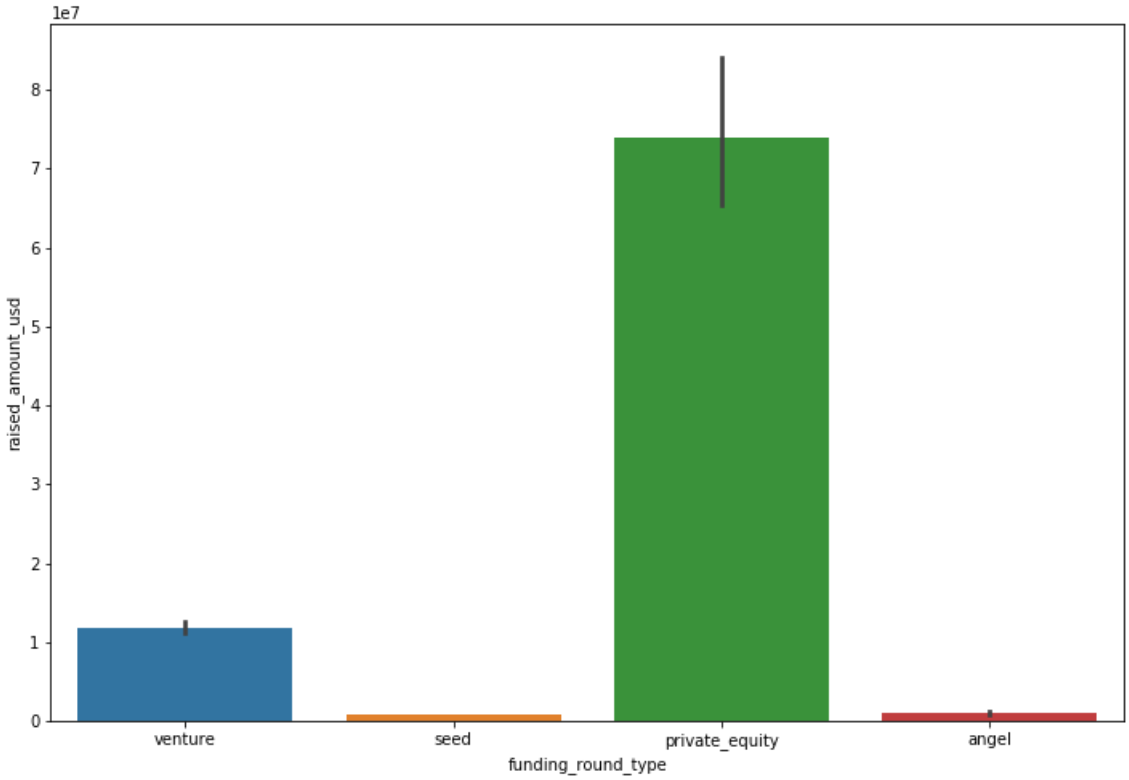
```
df=df[(df['funding_round_type']=="venture" )|
       (df['funding_round_type']=="seed") |
       (df['funding_round_type']=="angel") |
       (df['funding_round_type']=="private_equity")]
```

In [42]:

```
plt.figure(figsize=(12,8))
sns.barplot(x="funding_round_type",y="raised_amount_usd",data=df)
```

Out[42]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1dc844c8cc8>
```



In [43]:

```
df.describe()
```

Out[43]:

	raised_amount_usd
count	7.512300e+04
mean	9.519601e+06
std	7.792829e+07
min	0.000000e+00
25%	4.708105e+05
50%	2.000000e+06
75%	8.000000e+06
max	1.760000e+10

Checkpoint 2.2 asks for the most representative value of the investment amount for each of the four funding types

It can be represented either through mean or median . So let's check the mean and median of different funding types

In [44]:

```
df.groupby("funding_round_type").mean()
```

Out[44]:

funding_round_type	raised_amount_usd
angel	9.715739e+05
private_equity	7.393849e+07
seed	7.478279e+05
venture	1.172422e+07

In [45]:

```
df.groupby("funding_round_type").median()
```

Out[45]:

funding_round_type	raised_amount_usd
angel	414906.0
private_equity	20000000.0
seed	300000.0
venture	5000000.0

We see that the difference between the mean and median of all 4 funding types is very high.

Also, It is given in the question that Spark Funds want to invest between 5M and 15 M $

Let's assume mean to be the most representative value of the investment amount for each of the four types

Upon giving a close look to the both mean and median of the 4 funding types, we see that mean of venture 11.72 M and median of venture 5M falls under the Spark's criteria of 5M-15M USD . So, by the analysis so far, venture seems to be most suitable investment for Spark Funds.

Checkpoint 3

Country Analysis

Lets first quickly have a look at our rows and columns of df DataFrame

In [46]:

```
df.head()
```

Out[46]:

	permalink	name	category_list	status	country_code	funding_round_permalink	funding_round_type
						/funding-	

	permalink	name	category_list	status	country_code	funding_round_permalink	funding_round_type
0	/organization/-fame	#fame	Media	operating	IND	round/9a01d05418af9794ebdad7cec1d209	venture
1	/organization/-qounter	:Qounter	Application Platforms|Real Time|Social Network...	operating	USA	/funding-round/b44fbb94153f6cdef13083530bb48030	seed
2	/organization/0-6-com	0-6.com	Curated Web	operating	CHN	/funding-round/5727accaeaa57461bd22a9bdd945382d	venture
4	/organization/0ndine-biomedical-inc	Ondine Biomedical Inc.	Biotechnology	operating	CAN	/funding-round/2b9d3ac293d5cdccbecff5c8cb0f327d	seed
5	/organization/0ndine-biomedical-inc	Ondine Biomedical Inc.	Biotechnology	operating	CAN	/funding-round/954b9499724b946ad8c396a57a5f3b72	venture

Now filter the df for only venture type funding

In [47]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
df = df[df['funding_round_type']=="venture"]
country_wise_total = df.groupby('country_code')['raised_amount_usd'].sum().sort_values(ascending=False)
country_wise_total
```

Out[47]:

```
country_code
USA    4.200680e+11
CHN    3.933892e+10
GBR    2.007281e+10
IND    1.426151e+10
CAN    9.482218e+09
           ...
MCO    6.570000e+05
SAU    5.000000e+05
CMR    3.595610e+05
GTM    3.000000e+05
MMR    2.000000e+05
Name: raised_amount_usd, Length: 97, dtype: float64
```

Lets see the top 9 countries with highest amount of investment as asked in checkpoint 3.1.

In [48]:

```
top_9_countries = country_wise_total[:9]
top_9_countries
```

Out[48]:

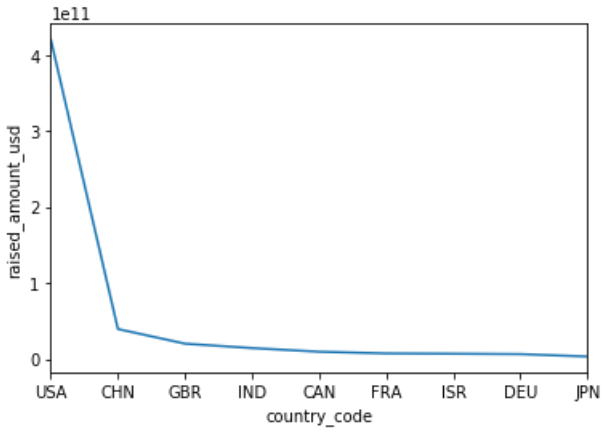
```
country_code
USA    4.200680e+11
CHN    3.933892e+10
GBR    2.007281e+10
IND    1.426151e+10
CAN    9.482218e+09
FRA    7.226851e+09
ISR    6.854350e+09
DEU    6.306922e+09
JPN    3.167647e+09
Name: raised_amount_usd, dtype: float64
```

In [49]:

```
top_9_countries.plot()
plt.ylabel("raised_amount_usd")
```

Out[49]:

```
Text(0, 0.5, 'raised_amount_usd')
```



We can clearly see that excluding China, USA-Great Britain and India are top 3 maximum invested english speaking countries.

In [50]:

```
# filtering for the top three countries
df = df[(df.country_code=='USA') | (df.country_code=='GBR') | (df.country_code=='IND')]
df.head()
```

Out[50]:

	permalink	name	category_list	status	country_code	funding_round_permalink	fu
0	/organization/-fame	#fame	Media	operating	IND	/funding-round/9a01d05418af9f794eebff7ace91f638	
7	/organization/0xdata	H2O.ai	Analytics	operating	USA	/funding-round/3bb2ee4a2d89251a10aaa735b1180e44	
8	/organization/0xdata	H2O.ai	Analytics	operating	USA	/funding-round/ae2a174c06517c2394aed45006322a7e	
9	/organization/0xdata	H2O.ai	Analytics	operating	USA	/funding-round/e1cfcbe1bdf4c70277c5f29a3482f24e	
15	/organization/1-mainstream	1 Mainstream	Apps|Cable|Distribution|Software	acquired	USA	/funding-round/b952cbaf401f310927430c97b68162ea	

In [51]:

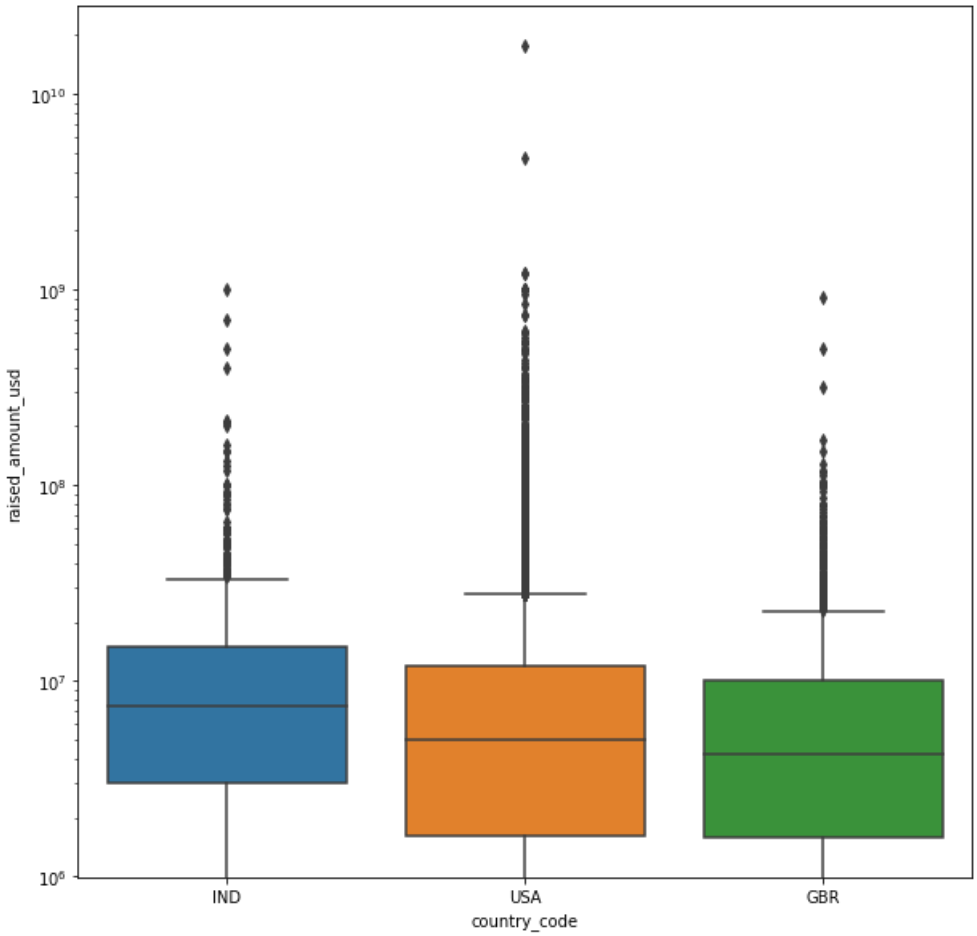
```
# filtered df has about 38803 observations
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 38803 entries, 0 to 88517
Data columns (total 9 columns):
permalink                38803 non-null object
name                     38803 non-null object
category_list            38803 non-null object
status                   38803 non-null object
country_code             38803 non-null object
funding_round_permalink  38803 non-null object
funding_round_type       38803 non-null object
funded_at                38803 non-null object
raised_amount_usd        38803 non-null float64
dtypes: float64(1), object(8)
memory usage: 3.0+ MB
```

In [52]:

```
plt.figure(figsize=(10, 10))
```

```
sns.boxplot(x='country_code', y='raised_amount_usd', data=df)
plt.yscale('log')
plt.show()
```



Checkpoint 4

Sector Analysis 1

In [53]:

```
df.loc[:,'main_category']=df['category_list'].apply(lambda x: x.split('|')[0])
df.head()
```

Out[53]:

	permalink	name	category_list	status	country_code	funding_round_permalink	fu
0	/organization/-fame	#fame	Media	operating	IND	/funding-round/9a01d05418af9f794eebff7ace91f638	
7	/organization/0xdata	H2O.ai	Analytics	operating	USA	/funding-round/3bb2ee4a2d89251a10aaa735b1180e44	
8	/organization/0xdata	H2O.ai	Analytics	operating	USA	/funding-round/ae2a174c06517c2394aed45006322a7e	
9	/organization/0xdata	H2O.ai	Analytics	operating	USA	/funding-round/e1cfcbe1bdf4c70277c5f29a3482f24e	
15	/organization/1-mainstream	1 Mainstream	Apps|Cable|Distribution|Software	acquired	USA	/funding-round/b952cbaf401f310927430c97b68162ea	

In [54]:

```
# drop the category_list column
df = df.drop('category_list', axis=1)
df.head()
```

Out[54]:

	permalink	name	status	country_code	funding_round_permalink	funding_round_type	funded_at
0	/organization/-fame	#fame	operating	IND	/funding-round/9a01d05418af9f794eebff7ace91f638	venture	05-01-2015
7	/organization/0xdata	H2O.ai	operating	USA	/funding-round/3bb2ee4a2d89251a10aaa735b1180e44	venture	09-11-2015
8	/organization/0xdata	H2O.ai	operating	USA	/funding-round/ae2a174c06517c2394aed45006322a7e	venture	03-01-2013
9	/organization/0xdata	H2O.ai	operating	USA	/funding-round/e1cfcbe1bdf4c70277c5f29a3482f24e	venture	19-07-2014
15	/organization/1-mainstream	1 Mainstream	acquired	USA	/funding-round/b952cbaf401f310927430c97b68162ea	venture	17-03-2015

In [55]:

```
# read mapping file
mapping = pd.read_csv("mapping.csv", sep=",")
mapping.head()
```

Out[55]:

	category_list	Automotive & Sports	Blanks	Cleantech / Semiconductors	Entertainment	Health	Manufacturing	News, Search and Messaging	Others	Social, Finance, Analytics, Advertising
0	NaN	0	1	0	0	0	0	0	0	0
1	3D	0	0	0	0	0	1	0	0	0
2	3D Printing	0	0	0	0	0	1	0	0	0
3	3D Technology	0	0	0	0	0	1	0	0	0
4	Accounting	0	0	0	0	0	0	0	0	1

In [56]:

```
mapping.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 688 entries, 0 to 687
Data columns (total 10 columns):
category_list                            687 non-null object
Automotive & Sports                      688 non-null int64
Blanks                                   688 non-null int64
Cleantech / Semiconductors               688 non-null int64
Entertainment                            688 non-null int64
Health                                   688 non-null int64
Manufacturing                            688 non-null int64
News, Search and Messaging               688 non-null int64
Others                                   688 non-null int64
Social, Finance, Analytics, Advertising  688 non-null int64
dtypes: int64(9), object(1)
memory usage: 53.9+ KB
```

In [57]:

```
# missing values in mapping file
mapping.isnull().sum()
```

Out[57]:

```
category_list                            1
Automotive & Sports                      0
Blanks                                   0
Cleantech / Semiconductors               0
Entertainment                            0
Health                                   0
Manufacturing                            0
```

```
News, Search and Messaging                  0
Others                                      0
Social, Finance, Analytics, Advertising     0
dtype: int64
```

In [58]:

```
# remove the row with missing values
mapping = mapping[~pd.isnull(mapping['category_list'])]
mapping.isnull().sum()
```

Out[58]:

```
category_list                               0
Automotive & Sports                         0
Blanks                                      0
Cleantech / Semiconductors                  0
Entertainment                               0
Health                                      0
Manufacturing                               0
News, Search and Messaging                  0
Others                                      0
Social, Finance, Analytics, Advertising     0
dtype: int64
```

In [59]:

```
# converting common columns to lowercase
mapping['category_list'] = mapping['category_list'].str.lower()
df['main_category'] = df['main_category'].str.lower()
```

In [60]:

```
mapping.head()
```

Out[60]:

	category_list	Automotive & Sports	Blanks	Cleantech / Semiconductors	Entertainment	Health	Manufacturing	News, Search and Messaging	Others	Social, Finance, Analytics, Advertising
1	3d	0	0	0	0	0	1	0	0	0
2	3d printing	0	0	0	0	0	1	0	0	0
3	3d technology	0	0	0	0	0	1	0	0	0
4	accounting	0	0	0	0	0	0	0	0	1
5	active lifestyle	0	0	0	0	1	0	0	0	0

In [61]:

```
df.head()
```

Out[61]:

	permalink	name	status	country_code	funding_round_permalink	funding_round_type	funded_at
0	/organization/-fame	#fame	operating	IND	/funding-round/9a01d05418af9f794eebff7ace91f638	venture	05-01-2015
7	/organization/0xdata	H2O.ai	operating	USA	/funding-round/3bb2ee4a2d89251a10aaa735b1180e44	venture	09-11-2015
8	/organization/0xdata	H2O.ai	operating	USA	/funding-round/ae2a174c06517c2394aed45006322a7e	venture	03-01-2013
9	/organization/0xdata	H2O.ai	operating	USA	/funding-round/e1cfcbe1bdf4c70277c5f29a3482f24e	venture	19-07-2014
15	/organization/1-mainstream	1 Mainstream	acquired	USA	/funding-round/b952cbaf401f310927430c97b68162ea	venture	17-03-2015

In [62]:

```
mapping['category_list'].head(35)
```

Out[62]:

```
1                                    3d
2                             3d printing
3                           3d technology
4                              accounting
5                          active lifestyle
6                              ad targeting
7                         advanced materials
8                           adventure travel
9                              advertising
10                  advertising exchanges
11                   advertising networks
12                  advertising platforms
13                                 advice
14                               aerospace
15                             agriculture
16                  air pollution control
17                               algorithms
18                             all markets
19                            all students
20                    alter0tive medicine
21                                 alumni
22                               a0lytics
23                                android
24                                angels
25                             animal feed
26             anything capital intensive
27                           app discovery
28                           app marketing
29                              app stores
30    application performance monitoring
31                  application platforms
32                                  apps
33                             aquaculture
34                            architecture
35                               archiving
Name: category_list, dtype: object
```

In [63]:

```
df[~df['main_category'].isin(mapping['category_list'])]
```

Out[63]:

	permalink	name	status	country_code	funding_round_permalink	funding_round_type	funded_at
7	/organization/0xdata	H2O.ai	operating	USA	/funding-round/3bb2ee4a2d89251a10aaa735b1180e44	venture	09-11-2015
8	/organization/0xdata	H2O.ai	operating	USA	/funding-round/ae2a174c06517c2394aed45006322a7e	venture	03-01-2013
9	/organization/0xdata	H2O.ai	operating	USA	/funding-round/e1cfcbe1bdf4c70277c5f29a3482f24e	venture	19-07-2014
47	/organization/100plus	100Plus	acquired	USA	/funding-round/b5facb0d9dea2f0352b5834892c88c53	venture	02-11-2011
136	/organization/1world-online	1World Online	operating	USA	/funding-round/32936e588a134502712877150198a0b3	venture	13-08-2015
...	...	...	...	...	...	...	...
88269	/organization/zoopla	Zoopla	ipo	GBR	/funding-round/98da1f441a55c9a9629a256828923e38	venture	19-01-2009
88290	/organization/zopa	Zopa	operating	GBR	/funding-round/2a55d435c3433d8f903526c050c19361	venture	20-03-2007
88291	/organization/zopa	Zopa	operating	GBR	/funding-round/4b0740cb83da8d2af9d221e5455f8923	venture	01-03-2006
88292	/organization/zopa	Zopa	operating	GBR	/funding-round/54dbfbd899caf7d1d4b2b7676065f303	venture	01-07-2006
88293	/organization/zopa	Zopa	operating	GBR	/funding-round/720b9f244c1f4d4fed63361d3bb0aa22	venture	01-01-2005

	permalink	name	status	country_code	funding_round_permalink	funding_round_type	funded_at

2616 rows × 9 columns

In [64]:

```
mapping[~mapping['category_list'].isin(df['main_category'])]
```

Out[64]:

	category_list	Automotive & Sports	Blanks	Cleantech / Semiconductors	Entertainment	Health	Manufacturing	News, Search and Messaging	Others	Social, Finance, Analytics, Advertising
16	air pollution control	0	0	1	0	0	0	0	0	0
20	alter0tive medicine	0	0	0	0	1	0	0	0	0
22	a0lytics	0	0	0	0	0	0	0	0	1
33	aquaculture	0	0	1	0	0	0	0	0	0
49	b2b express delivery	0	0	0	0	0	0	0	0	1
...	...	...	...	...	...	...	...	...	...	...
670	virtual workforces	0	0	0	1	0	0	0	0	0
672	waste ma0gement	0	0	1	0	0	0	0	0	0
682	weddings	0	0	0	1	0	0	0	0	0
683	wholesale	0	0	0	0	0	0	0	1	0
686	women	0	0	0	0	0	0	0	1	0

175 rows × 10 columns

Now there's no row with missing values but we see that the in values of category_list "0" appears in place of "na". Lets first replace "0" with "na" wherever required.

In [65]:

```
# replacing '0' with 'na'
mapping['category_list'] = mapping['category_list'].apply(lambda x: x.replace('0', 'na'))
print(mapping['category_list'])
```

```
1                    3d
2              3d printing
3            3d technology
4              accounting
5       active lifestyle
              ...
683             wholesale
684     wine and spirits
685              wireless
686                 women
687         young adults
Name: category_list, Length: 687, dtype: object
```

Now we need to merge mapping with df . So, lets make make common columns lower case

In [66]:

```
# merge the dfs
df = pd.merge(df, mapping, how='inner', left_on='main_category', right_on='category_list')
df.head()
```

Out[66]:

	permalink	name	status	country_code	funding_round_permalink	funding_round_type	funded_at	raised_
0	/organization/-fame	#fame	operating	IND	/funding-round/9a01d05418af9f794eebff7ace91f638	venture	05-01-2015	
1	/organization/90min	90min	operating	GBR	/funding-round/21a2cbf6f2fb2a1c2a61e04bf930dfe6	venture	06-10-2015	
2	/organization/90min	90min	operating	GBR	/funding-round/bd626ed022f5c66574b1afe234f3c90d	venture	07-05-2013	
3	/organization/90min	90min	operating	GBR	/funding-round/fd4b15e8c97ee2ffc0acccdbe1a98810	venture	26-03-2014	
4	/organization/all-def-digital	All Def Digital	operating	USA	/funding-round/452a2342fe720285c3b92e9bd927d9ba	venture	06-08-2014	

In [67]:

```
# let's drop the category_list column since it is the same as main_category
df = df.drop('category_list', axis=1)
df.head()
```

Out[67]:

	permalink	name	status	country_code	funding_round_permalink	funding_round_type	funded_at	raised_
0	/organization/-fame	#fame	operating	IND	/funding-round/9a01d05418af9f794eebff7ace91f638	venture	05-01-2015	
1	/organization/90min	90min	operating	GBR	/funding-round/21a2cbf6f2fb2a1c2a61e04bf930dfe6	venture	06-10-2015	
2	/organization/90min	90min	operating	GBR	/funding-round/bd626ed022f5c66574b1afe234f3c90d	venture	07-05-2013	
3	/organization/90min	90min	operating	GBR	/funding-round/fd4b15e8c97ee2ffc0acccdbe1a98810	venture	26-03-2014	
4	/organization/all-def-digital	All Def Digital	operating	USA	/funding-round/452a2342fe720285c3b92e9bd927d9ba	venture	06-08-2014	

In [68]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 38788 entries, 0 to 38787
Data columns (total 18 columns):
permalink                                38788 non-null object
name                                     38788 non-null object
status                                   38788 non-null object
country_code                             38788 non-null object
funding_round_permalink                  38788 non-null object
funding_round_type                       38788 non-null object
funded_at                                38788 non-null object
raised_amount_usd                        38788 non-null float64
main_category                            38788 non-null object
Automotive & Sports                      38788 non-null int64
Blanks                                   38788 non-null int64
Cleantech / Semiconductors               38788 non-null int64
Entertainment                            38788 non-null int64
Health                                   38788 non-null int64
Manufacturing                            38788 non-null int64
News, Search and Messaging               38788 non-null int64
Others                                   38788 non-null int64
Social, Finance, Analytics, Advertising  38788 non-null int64
dtypes: float64(1), int64(9), object(8)
memory usage: 5.6+ MB
```

You'll notice that the columns representing the main category in the mapping file are originally in the 'wide' format - Automotive & Sports, Cleantech / Semiconductors etc. They contain the value '1' if the company belongs to that category, else 0. This is quite redundant. We can as well have a column named 'sub-category' having these values. Let's convert the df into the long format from the current wide format. First, we'll store the 'value variables' (those which are to be melted) in an array. The rest will then be the 'index variables'.

In [69]:

```
value_vars = df.columns[9:18]
id_vars=df.columns[0:9]
print(id_vars)
print(value_vars)
```

```
Index(['permalink', 'name', 'status', 'country_code',
       'funding_round_permalink', 'funding_round_type', 'funded_at',
       'raised_amount_usd', 'main_category'],
      dtype='object')
Index(['Automotive & Sports', 'Blanks', 'Cleantech / Semiconductors',
       'Entertainment', 'Health', 'Manufacturing',
       'News, Search and Messaging', 'Others',
       'Social, Finance, Analytics, Advertising'],
      dtype='object')
```

In [70]:

```
long_df = pd.melt(df,id_vars=list(id_vars),value_vars=list(value_vars))
long_df.head()
```

Out[70]:

	permalink	name	status	country_code	funding_round_permalink	funding_round_type	funded_at	raised_
0	/organization/-fame	#fame	operating	IND	/funding-round/9a01d05418af9f794eebff7ace91f638	venture	05-01-2015	
1	/organization/90min	90min	operating	GBR	/funding-round/21a2cbf6f2fb2a1c2a61e04bf930dfe6	venture	06-10-2015	
2	/organization/90min	90min	operating	GBR	/funding-round/bd626ed022f5c66574b1afe234f3c90d	venture	07-05-2013	
3	/organization/90min	90min	operating	GBR	/funding-round/fd4b15e8c97ee2ffc0acccdbe1a98810	venture	26-03-2014	
4	/organization/all-def-digital	All Def Digital	operating	USA	/funding-round/452a2342fe720285c3b92e9bd927d9ba	venture	06-08-2014	

Keeping rows with value=1

In [71]:

```
long_df=long_df[long_df['value']==1]
```

In [72]:

```
long_df.head(35)
```

Out[72]:

	permalink	name	status	country_code	funding_round_permalink	funding_round_type	fu
25828	/organization/3d-robotics	3D Robotics	operating	USA	/funding-round/2785595770e91ab8fd4854ef125ec563	venture	
25829	/organization/3d-robotics	3D Robotics	operating	USA	/funding-round/7ca0d4dc119b6d65eebfb352c3544542	venture	
25830	/organization/3d-robotics	3D Robotics	operating	USA	/funding-round/d6221c11246b0a536ee2cadd9fcf54d3	venture	
25831	/organization/3d-robotics	3D Robotics	operating	USA	/funding-round/ff3c1d1ae1c3486d775095b093d99b58	venture	

	permalink	name	status	country_code	funding_round_permalink	funding_round_type	fu
25832	/organization/cabk-productions	Cabk Productions	operating	USA	/funding-round/156e4fbce54aca39a8be9a1a2fa1fb77	venture	
25833	/organization/dronedeploy	DroneDeploy	operating	USA	/funding-round/bdf644f3fa66533c048719bf0d000893	venture	
25834	/organization/droneshield	DroneShield	operating	USA	/funding-round/0935e9fee6d86b49420da74cf4a3a94e	venture	
25835	/organization/ehang	Ehang	operating	USA	/funding-round/3ffe5bfadb0a64d2d3c931d6a98c5683	venture	
25836	/organization/ehang	Ehang	operating	USA	/funding-round/cf1321bcd5745aade7e99eedaaa26ded	venture	
25837	/organization/yuneec-apv	Yuneec APV	operating	USA	/funding-round/ebb2406162ab04029c9d0c940ecd982e	venture	
25856	/organization/3floz-com	3FLOZ	operating	USA	/funding-round/db1213a3ff5f9e74f756e4b5c6772f5a	venture	
25857	/organization/bang-networks	Bang Networks	operating	USA	/funding-round/ba025fbc8bc3ca77ea945b61c4d21724	venture	
25858	/organization/dvdplay	DVDPlay	closed	USA	/funding-round/19cba6123538b83a006903f2ef76338e	venture	
25859	/organization/softwear-automation	SoftWear Automation	operating	USA	/funding-round/9309894f722f8ce4d65d8b18f0831e57	venture	
27336	/organization/4home	4Home	acquired	USA	/funding-round/3801a81d9c5b5a3d6be0e2c18b1ef09c	venture	
27337	/organization/4home	4Home	acquired	USA	/funding-round/9919b9adaadcfd3f3e9ee52ee14a7fdb	venture	
27338	/organization/4home	4Home	acquired	USA	/funding-round/d976a3a9eaed96cbae0bd6c2158e2b35	venture	
27339	/organization/4home	4Home	acquired	USA	/funding-round/dd581bca505c94ccda21dab6a117a3df	venture	
27340	/organization/4home	4Home	acquired	USA	/funding-round/e3a19950b347b628c80938b8958a7c39	venture	
27341	/organization/4home	4Home	acquired	USA	/funding-round/ff37a5cbb584d51ac44288341836d520	venture	
27342	/organization/additech	Additech	operating	USA	/funding-round/6f06ad0022ccad7a54241c334dc55d25	venture	
27343	/organization/additech	Additech	operating	USA	/funding-round/c1c50ebc27ce45adbdd21e0b121fd23a	venture	
27344	/organization/agm-automotive	AGM Automotive	operating	USA	/funding-round/0ac85ce267380a3fd4a7e0cea153dfe1	venture	
27345	/organization/airbiquity	Airbiquity	operating	USA	/funding-round/a32d7bb9953596c010b81e1b44f2018c	venture	
27346	/organization/airbiquity	Airbiquity	operating	USA	/funding-round/c6909a12d18862ebd5173dd1ee6abd6a	venture	
27347	/organization/ani-technologies	Ola	operating	IND	/funding-round/1e2b54335e2a41d8d7db25b7c11db399	venture	
27348	/organization/ani-technologies	Ola	operating	IND	/funding-round/3722a5bf71ee371f98e83fe2dd04596d	venture	
27349	/organization/ani-technologies	Ola	operating	IND	/funding-round/b6d53e0d0ecf4b720d5a8306e20d97fd	venture	
27350	/organization/ani-technologies	Ola	operating	IND	/funding-round/bbce7c1d8470d24a5b05375a1e58a34e	venture	
27351	/organization/ani-technologies	Ola	operating	IND	/funding-round/d585974a6ae7ca30ff102a0691ab2c1b	venture	
27352	/organization/ani-technologies	Ola	operating	IND	/funding-round/e0e7c05049288bed3a9abf6741d7b6f4	venture	
27353	/organization/ansible	Ansible	acquired	USA	/funding-round/2692caf147ec410d38a509c2499902c6	venture	
27354	/organization/aptera	Aptera	closed	USA	/funding-round/74dc54cf94102e9620e19a561104ba2b	venture	
27355	/organization/aptera	Aptera	closed	USA	/funding-round/7738f883d6188485957de3e3f0cf9228	venture	
27356	/organization/aptera	Aptera	closed	USA	/funding-round/cad865d67775a4373b36dc40d937ce58	venture	

In [73]:

```
len(long_df)
```

Out[73]:

```
38788
```

In [74]:

```
long_df = long_df.rename(columns={'variable': 'sector'})
```

In [75]:

```
long_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 38788 entries, 25828 to 349075
Data columns (total 11 columns):
permalink                  38788 non-null object
name                       38788 non-null object
status                     38788 non-null object
country_code               38788 non-null object
funding_round_permalink    38788 non-null object
funding_round_type         38788 non-null object
funded_at                  38788 non-null object
raised_amount_usd          38788 non-null float64
main_category              38788 non-null object
sector                     38788 non-null object
value                      38788 non-null int64
dtypes: float64(1), int64(1), object(9)
memory usage: 3.6+ MB
```

The dataframe now contains only venture type investments in countries USA, IND and GBR, and we have mapped each company to one of the eight main sectors (named 'sector' in the dataframe). We can now compute the sector-wise number(count) and the amount of investment in the three countries.

Checkpoint 5

Sector Analysis 2

We need to keep investment within the range of 5M $-15M $

In [76]:

```
df = long_df[(long_df['raised_amount_usd'] >= 5000000) & (long_df['raised_amount_usd'] <= 15000000)
]
df.head(25)
```

Out[76]:

	permalink	name	status	country_code	funding_round_permalink	funding_round_type	fu
25828	/organization/3d-robotics	3D Robotics	operating	USA	/funding-round/2785595770e91ab8fd4854ef125ec563	venture	
25829	/organization/3d-robotics	3D Robotics	operating	USA	/funding-round/7ca0d4dc119b6d65eebfb352c3544542	venture	
25832	/organization/cape-productions	Cape Productions	operating	USA	/funding-round/156e4fbce54aca39a8be9a1a2fa1fb77	venture	
25833	/organization/dronedeploy	DroneDeploy	operating	USA	/funding-round/bdf644f3fa66533c048719bf0d000893	venture	
25836	/organization/ehang	Ehang	operating	USA	/funding-round/cf1321bcd5745aade7e99eedaaa26ded	venture	
25857	/organization/bang-networks	Bang Networks	operating	USA	/funding-round/ba025fbc8bc3ca77ea945b61c4d21724	venture	
27343	/organization/additech	Additech	operating	USA	/funding-round/c1c50ebc27ce45adbdd21e0b121fd23a	venture	

	permalink	name	status	country_code	funding_round_permalink	funding_round_type	fu
27344	/organization/agm-automotive	AGM Automotive	operating	USA	/funding-round/0ac85ce267380a3fd4a7e0cea153dfe1	venture	
27346	/organization/airbiquity	Airbiquity	operating	USA	/funding-round/c6909a12d18862ebd5173dd1ee6abd6a	venture	
27347	/organization/ani-technologies	Ola	operating	IND	/funding-round/1e2b54335e2a41d8d7db25b7c11db399	venture	
27353	/organization/ansible	Ansible	acquired	USA	/funding-round/2692caf147ec410d38a509c2499902c6	venture	
27354	/organization/aptera	Aptera	closed	USA	/funding-round/74dc54cf94102e9620e19a561104ba2b	venture	
27359	/organization/ather-energy	Ather Energy	operating	IND	/funding-round/a3782f52b69e60629bcf7866ca8b1eca	venture	
27361	/organization/atieva	Atieva	operating	USA	/funding-round/6a5a9a2ff0c547710ac0387f87f1e343	venture	
27362	/organization/autoamerica	AutoAmerica	operating	USA	/funding-round/c456e0cd9471cc166f783ae1d131aeb4	venture	
27365	/organization/automile-ab	Automile	operating	USA	/funding-round/a380b558208f7edf23c3a49b290c7f96	venture	
27367	/organization/autopilot	Autopilot	closed	USA	/funding-round/9839633997e7c33fcfb4db546b99319c	venture	
27368	/organization/autoquake	Autoquake	acquired	GBR	/funding-round/067d143de46ec298cfa1893682f9911a	venture	
27369	/organization/autoquake	Autoquake	acquired	GBR	/funding-round/4c8372dfdea687c5f5fbab39b3e44dab	venture	
27370	/organization/autoquake	Autoquake	acquired	GBR	/funding-round/721aefa6f7e5bc71eb9d744359941958	venture	
27371	/organization/autoquake	Autoquake	acquired	GBR	/funding-round/a4d5080cbda34c2ef4295d8fbe4e9ad5	venture	
27373	/organization/beepi	Beepi	operating	USA	/funding-round/87d70c9019e13a2ba690b5b0f7c1f65a	venture	
27375	/organization/beepi	Beepi	operating	USA	/funding-round/8c04f7031be7fc7d215f7605de96934b	venture	
27381	/organization/brammo	Brammo	operating	USA	/funding-round/4d3f9611c76831d92e4a738570f8edb1	venture	
27382	/organization/brammo	Brammo	operating	USA	/funding-round/a3333a30934491d522d1735f7090af79	venture	

In [77]:

```
# groupby country, sector and compute the count and sum
d1=df[df["country_code"]=='USA'].groupby(['country_code',
'sector']).raised_amount_usd.agg(['count', 'sum'])
```

In [78]:

```
d1.head(10)
```

Out[78]:

		count	sum
country_code	sector		
USA	Automotive & Sports	167	1.454104e+09
	Cleantech / Semiconductors	2350	2.163343e+10
	Entertainment	591	5.099198e+09
	Health	909	8.211859e+09
	Manufacturing	799	7.258553e+09
	News, Search and Messaging	1583	1.397157e+10
	Others	2950	2.632101e+10
	Social, Finance, Analytics, Advertising	2714	2.380738e+10

In [79]:

```
d1["count"].sum()
```

Out[79]:

```
12063
```

In [80]:

```
d1["sum"].sum()
```

Out[80]:

```
107757097294.0
```

So, Total number(count) of investment in USA is 12063 and sum is $ 107757097294.0 .

In [81]:

```
d2=df[df["country_code"]=='IND'].groupby(['country_code',
'sector']).raised_amount_usd.agg(['count', 'sum'])
d2.head(10)
```

Out[81]:

		count	sum
country_code	sector		
IND	Automotive & Sports	13	1.369000e+08
	Cleantech / Semiconductors	20	1.653800e+08
	Entertainment	33	2.808300e+08
	Health	19	1.677400e+08
	Manufacturing	21	2.009000e+08
	News, Search and Messaging	52	4.338345e+08
	Others	110	1.013410e+09
	Social, Finance, Analytics, Advertising	60	5.505496e+08

In [82]:

```
d2["count"].sum()
```

Out[82]:

```
328
```

In [83]:

```
d2["sum"].sum()
```

Out[83]:

```
2949543602.0
```

So, Total number(count) of investment in India is 328 and sum is $ 2949543602.0.

In [84]:

```
d3=df[df["country_code"]=='GBR'].groupby(['country_code',
'sector']).raised_amount_usd.agg(['count', 'sum'])
d3.head(10)
```

Out[84]:

country_code	sector	count	sum
GBR	Automotive & Sports	16	1.670516e+08
	Cleantech / Semiconductors	130	1.163990e+09
	Entertainment	56	4.827847e+08
	Health	24	2.145375e+08
	Manufacturing	42	3.619403e+08
	News, Search and Messaging	73	6.157462e+08
	Others	147	1.283624e+09
	Social, Finance, Analytics, Advertising	133	1.089404e+09

In [85]:

```
d3["count"].sum()
```

Out[85]:

```
621
```

In [86]:

```
d3["sum"].sum()
```

Out[86]:

```
5379078691.0
```

So, Total number(count) of investment in GBR is 621 and sum is $ 5379078691.00 .

So by far our analysis, top country in terms of the number of investments and total amount invested is USA. The sectors 'Others', 'Social,FInance, Analytics and Advertising' and 'Cleantech/Semiconductors' are top 3 most heavily invested sectors.

So if an investment has to be made, it must be made in 'Others'. If you don't want to consider 'Others' then 'Social,FInance, Analytics and Advertising' is the next best option.

----------XXX----------