You work for Spark Funds, an asset management company. Spark Funds wants to make investments in a few companies. The CEO of Spark Funds wants to understand the global trends in investments so that she can take the investment decisions effectively.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import chardet as cd
```

CHECKPOINT 01 : DATA CLEANING

```
from google.colab import drive
```

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

Loading the DataSet

```
# Loading the Data sets
companies =  pd.read_csv("/content/drive/MyDrive/Other Drives/EvilFoxCorps Drive /Datasets/Investment Analys
#companies =  pd.read_csv("/content/companies.csv",encoding ='ISO-8859-1' )
companies.head()
```

	permalink	name	homepage_url	category_list	status	country_code	state_code	region	cit
0	/Organization/-Fame	#fame	http://livfame.com	Media	operating	IND	16	Mumbai	Mumba
1	/Organization/-Qounter	:Qounter	http://www.qounter.com	Application Platforms|Real Time|Social Network...	operating	USA	DE	DE - Other	Delawar Cit
2	/Organization/-The-One-Of-Them-Inc-	(THE) ONE of THEM,Inc.	http://oneofthem.jp	Apps|Games|Mobile	operating	NaN	NaN	NaN	Na
3	/Organization/0-6-Com	0-6.com	http://www.0-6.com	Curated Web	operating	CHN	22	Beijing	Beijin

```
rounds2 = pd.read_csv("/content/drive/MyDrive/Other Drives/EvilFoxCorps Drive /Datasets/Investment Analysis/
#rounds2 = pd.read_csv("/content/rounds2.csv",encoding = 'ISO-8859-1' )
rounds2.head
```

```
<bound method NDFrame.head of                                company_permalink  \
0                                  /organization/-fame
1                                /ORGANIZATION/-QOUNTER
2                                /organization/-qounter
3                          /ORGANIZATION/-THE-ONE-OF-THEM-INC-
4                                /organization/0-6-com
...                                               ...
114944                           /organization/zzzzapp-com
114945                           /ORGANIZATION/ZZZZAPP-COM
114946                             /organization/ãeron
114947                           /ORGANIZATION/Ã”ASYS-2
114948  /organization/ä°novatiff-reklam-ve-tanä±tä±m-h...

                        funding_round_permalink funding_round_type  \
0       /funding-round/9a01d05418af9f794eebff7ace91f638         venture
1       /funding-round/22dacff496eb7acb2b901dec1dfe5633         venture
2       /funding-round/b44fbb94153f6cdef13083530bb48030            seed
3       /funding-round/650b8f704416801069bb178a1418776b         venture
4       /funding-round/5727accaeaa57461bd22a9bdd945382d         venture
...                                                ...             ...
114944  /funding-round/8f6d25b8ee4199e586484d817bceda05  convertible_note
114945  /funding-round/ff1aa06ed5da186c84f101549035d4ae            seed
```

```
114946  /funding-round/59f4dce44723b794f21ded3daed6e4fe         venture
114947  /funding-round/35f09d0794651719b02bbfd859ba9ff5          seed
114948  /funding-round/af942869878d2cd788ef5189b435ebc4           grant

     funding_round_code   funded_at  raised_amount_usd
0                     B    5/1/2015        10000000.0
1                     A  14-10-2014               NaN
2                   NaN    1/3/2014         700000.0
3                     B  30-01-2014        3406878.0
4                     A  19-03-2008        2000000.0
...                 ...         ...              ...
114944              NaN    1/3/2014          41313.0
114945              NaN    1/5/2013          32842.0
114946                A    1/8/2014               NaN
114947              NaN    1/1/2015          18192.0
114948              NaN   1/10/2013          14851.0

[114949 rows x 6 columns]>
```

Checking Dataset and uniforming

```
1 companies['permalink'] = companies['permalink'].str.lower()
2 len(companies['permalink'].unique())
```

```
66368
```

```
1 companies.shape
```

```
(66368, 10)
```

```
1 rounds2.shape
```

```
(114949, 6)
```

```
1 rounds2['company_permalink'] = rounds2['company_permalink'].str.lower()
2 len(rounds2['company_permalink'].unique())
```

66370

```
1 pd.DataFrame(rounds2.describe())
```

	raised_amount_usd
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

1. Rounds2 Data have unique Data set of 90247 Unique Companies under the column [**company_permalink**], however that is because of the companies being names in higher or lower case which is being counted as a unique
2. Even after converting the same into lower class there seems to be additional 2 companies which seems to be unique to rounds 2

```
1 #checking the datasets again to verify if the companies in rounds2 are present in companies dataset
2 rounds2.loc[~ rounds2['company_permalink'].isin(companies['permalink']),:]
```

	company_permalink	funding_round_permalink	funding_round_type	funding_round_code	funded_at	raised_amount_u
29597	/organization/e-cãšbica	/funding-round/8491f74869e4fe8ba9c378394f8fbdea	seed	NaN	1/2/2015	N
31863	/organization/energystone-games-çµçŸ³æ„æ^	/funding-round/b89553f3d2279c5683ae93f45a21cfe0	seed	NaN	9/8/2014	N
45176	/organization/huizuche-com-æƒ çšÿè½¦	/funding-round/8f8a32dbeeb0f831a78702f83af78a36	seed	NaN	18-09-2014	N
58473	/organization/magnet-tech-ç£çŸ³ç§‘æŠ	/funding-round/8fc91fbb32bc95e97f151dd0cb4166bf	seed	NaN	16-08-2014	162558
	/organization/tipcat-	/funding-				

```
1 #checking the datasets again to verify if the companies in companies are present in rounds2 dataset
2 companies.loc[~ companies['permalink'].isin(rounds2['company_permalink']),:]
```

	permalink	name	homepage_url	category_list	status	country_code	state_code	region	city
16827	/organization/e-cãšbica	E CÃšBICA	NaN	NaN	operating	NaN	NaN	NaN	NaN
18197	/organization/energystone-games-çµçÿ³æ„æ^	EnergyStone Games çµçŸ³æ„æ^	NaN	Mobile Games|Online Gaming	closed	NaN	NaN	NaN	NaN
26139	/organization/huizuche-com-æƒ çšŸè½¦	Huizuche.com æƒ çšŸè½¦	http://huizuche.com	NaN	closed	NaN	NaN	NaN	NaN
58344	/organization/tipcat-interactive-æ²™è^Ÿä¿¡æ¯ç...	TipCat Interactive æ²™è^Ÿä¿¡æ¯ç§‘æŠ	http://www.tipcat.com	Mobile Games|Online Gaming	closed	NaN	NaN	NaN	NaN
65778	/organization/zengame-ç¦…æ„ç§‘æš	ZenGame ç¦… æ„ç§‘æŠ	http://www.zen-game.com	Internet|Mobile Games|Online Gaming	closed	NaN	NaN	NaN	NaN

Both Datasets seems to have symbol like english language, which is possible due to encoding issue

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

	company_permalink	funding_round_permalink	funding_round_type	funding_round_code	funded_at	raised_amount_u
77	/organization/10north	/funding-round/b41ff7de932f8b6e5bbeed3966c0ed6a	equity_crowdfunding	NaN	12/8/2014	Na
729	/organization/51wofang-	/funding-round/346b9180d276a74e0fbb2825e66c6f5b	venture	A	6/7/2015	5000000
2670	/organization/adslinked	/funding-round/449ae54bb63c768c232955ca6911dee4	seed	NaN	29-09-2014	100000
3166	/organization/aesthetic-everything-social-network	/funding-round/62593455f1a69857ed05d5734cc04132	equity_crowdfunding	NaN	12/10/2014	Na
3291	/organization/affluent-attach-club-2	/funding-round/626678bdf1654bc4df9b1b34647a4df1	seed	NaN	15-10-2014	100000
...	...	...	...	...	...	
110545	/organization/whodats-spaces	/funding-round/d5d6db3d1e6c54d71a63b3aa0c9278e6	seed	NaN	28-10-2014	30000
113839	/organization/zengame-	/funding-round/6ba28fb4f3eadf5a9c6c81bc5dde6cdf	seed	NaN	17-07-2010	Na
114946	/organization/eron	/funding-round/59f4dce44723b794f21ded3daed6e4fe	venture	A	1/8/2014	Na
114947	/organization/asys-2	/funding-round/35f09d0794651719b02bbfd859ba9ff5	seed	NaN	1/1/2015	1819
114948	/organization/novatiff-reklam-ve-tantm-hizmetl	/funding-round/af942869878d2cd788ef5189b435ebc4	grant	NaN	1/10/2013	1485

```
# companies present in companies df but not in rounds df
companies.loc[~companies['permalink'].isin(rounds2['company_permalink']), :]
```

	permalink	name	homepage_url	category_list	status	country_code	state_code	re
43	/organization/10â°north	10Â°North	NaN	Fashion	operating	CAN	ON	Tor
426	/organization/51wofang-æ— å¿§æ‘æ¿	51wofang æ— å¿§æ‘æ¿	http://www.51wofang.com	NaN	closed	NaN	NaN	
1506	/organization/adslinkedâ„¢ç	AdsLinkedâ„¢ç	http://www.adslinked.com	Advertising|Internet	operating	NaN	NaN	
1775	/organization/aesthetic-everythingâ®-social-ne...	Aesthetic EverythingÂ® Social Network	http://aestheticeverything.com/	Public Relations	operating	USA	CA	Ang
1834	/organization/affluent-attachã©-club-2	Affluent AttachÃ© Club	http://www.affluentattache.com/	Hospitality	operating	USA	CA	Ang
...	...	...	...	...	...	...	...	
63833	/organization/whodatâ™s-spaces	Whodatâ™s Spaces	NaN	Apps	operating	NaN	NaN	
65778	/organization/zengame-ç¦…æ„ç§‘æ	ZenGame ç¦… æ„ç§‘æ	http://www.zen-game.com	Internet|Mobile Games|Online Gaming	closed	NaN	NaN	
66365	/organization/ãeron	ÃERON	http://www.aeron.hu/	NaN	operating	NaN	NaN	
66366	/organization/ã¨asys-2	Ã¨asys	http://www.oasys.io/	Consumer Electronics|Internet of Things|Teleco...	operating	USA	CA	SF
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	

68 rows × 10 columns

Observation

The following can be observed from the Data Set

1. Companies have a data set of 66368 Unique Companies under the coumn [**Permalink**]
2. The Unique Companies present in the Rounds2 Data set after uniforming the data is 66368 [**Comany_Permalink**]
3. The issues which were observed in Rounds2 dataset were due to encoding error

Confirmation of Datasets

```
#unique values
print(len(companies.permalink.unique()))
print(len(rounds2.company_permalink.unique()))

#checking if columns present in rounds not present in companies
print(len(rounds2.loc[~rounds2['company_permalink'].isin(companies['permalink']),:]))
```

```
66368
66368
74
```

Missing Values Cleaning

```
print(companies.shape)
print(rounds2.shape)
```

```
(66368, 10)
(114949, 6)
```

Merging the two datasets

```
master_frame = pd.merge(left = companies,right = rounds2,how ='inner',left_on='permalink',right_on='company_per
```

2 master_frame.head()

	permalink	name	homepage_url	category_list	status	country_code	state_code	region	city	founded_at
0	/organization/-fame	#fame	http://livfame.com	Media	operating	IND	16	Mumbai	Mumbai	NaN
1	/organization/-qounter	:Qounter	http://www.qounter.com	Application Platforms|Real Time|Social Network...	operating	USA	DE	DE - Other	Delaware City	04-09-2014
2	/organization/-qounter	:Qounter	http://www.qounter.com	Application Platforms|Real Time|Social Network...	operating	USA	DE	DE - Other	Delaware City	04-09-2014
3	/organization/-the-one-of-them-inc-	(THE) ONE of THEM,Inc.	http://oneofthem.jp	Apps|Games|Mobile	operating	NaN	NaN	NaN	NaN	NaN
4	/organization/0-6-com	0-6.com	http://www.0-6.com	Curated Web	operating	CHN	22	Beijing	Beijing	01-01-2007

1 master_frame.info()

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 114875 entries, 0 to 114874
Data columns (total 16 columns):
 #   Column         Non-Null Count   Dtype 
---  ------         --------------   ----- 
 0   permalink      114875 non-null  object
 1   name           114874 non-null  object
 2   homepage_url   108749 non-null  object
 3   category_list  111488 non-null  object
 4   status         114875 non-null  object
 5   country_code   106238 non-null  object
 6   state_code     103972 non-null  object
 7   region         104749 non-null  object
```

```
 8   city                  104752 non-null  object
 9   founded_at             94387 non-null  object
 10  company_permalink      114875 non-null  object
 11  funding_round_permalink  114875 non-null  object
 12  funding_round_type     114875 non-null  object
 13  funding_round_code     31132 non-null  object
 14  funded_at              114875 non-null  object
 15  raised_amount_usd      94915 non-null  float64
dtypes: float64(1), object(15)
memory usage: 14.9+ MB
```

1 round(master_frame.isnull().sum()/len(master_frame)*100,2)

```
permalink                 0.00
name                      0.00
homepage_url              5.33
category_list             2.95
status                    0.00
country_code              7.52
state_code                9.49
region                    8.81
city                      8.81
founded_at               17.84
company_permalink         0.00
funding_round_permalink   0.00
funding_round_type        0.00
funding_round_code       72.90
funded_at                 0.00
raised_amount_usd        17.38
dtype: float64
```

1 master_frame.columns

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

Dropping the following field after taking into consideration their importance:

```
print("Before Dropping: ",master_frame.shape)
master_frame.drop(columns=['company_permalink','homepage_url','founded_at','state_code','region','city','funding_r
print("After Dropping: ",master_frame.shape)
```

```
Before Dropping:  (114875, 16)
After Dropping:  (114875, 9)
```

```
#Checking the Null Values again
round(master_frame.isnull().sum()/len(master_frame)*100,2)
```

```
permalink                 0.00
name                      0.00
category_list             2.95
status                    0.00
country_code              7.52
funding_round_permalink   0.00
funding_round_type        0.00
funded_at                 0.00
raised_amount_usd        17.38
dtype: float64
```

SInce Raised_Amount_USD is an important dataset, we need to treat the missing 17% of the missing data, we can look at dropping the dataset since its better to have a dataset with actual values than NAN, on top of which we are not losing much information

```
master_frame = master_frame[~np.isnan(master_frame['raised_amount_usd'])]
round(master_frame.isnull().sum()/len(master_frame)*100,2)
```

```
permalink                 0.00
name                      0.00
category_list             1.09
status                    0.00
country_code              6.14
funding_round_permalink   0.00
funding_round_type        0.00
funded_at                 0.00
```

```
raised_amount_usd         0.00
dtype: float64
```

Understanding the Country Code Bifurcation in the dataset. Country code can be considered as a type of category

```
master_frame['country_code'] = master_frame['country_code'].astype('category')
master_frame.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 94915 entries, 0 to 114874
Data columns (total 9 columns):
 #   Column                Non-Null Count  Dtype   
---  ------                --------------  -----   
 0   permalink             94915 non-null  object  
 1   name                  94914 non-null  object  
 2   category_list         93877 non-null  object  
 3   status                94915 non-null  object  
 4   country_code          89085 non-null  category
 5   funding_round_permalink  94915 non-null  object  
 6   funding_round_type    94915 non-null  object  
 7   funded_at             94915 non-null  object  
 8   raised_amount_usd     94915 non-null  float64 
dtypes: category(1), float64(1), object(7)
memory usage: 6.7+ MB
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """Entry point for launching an IPython kernel.
```

```
pd.set_option('display.max_rows',None)
pd.DataFrame(master_frame['country_code'].value_counts()/len(master_frame)*100)
```

	country_code
USA	65.369014
GBR	5.286836
CAN	2.756150
CHN	2.030238
IND	1.737344
FRA	1.525576
ISR	1.437075
ESP	1.131539
DEU	1.095717
AUS	0.683770
RUS	0.619502
IRL	0.593162
SWE	0.590002
SGP	0.575252
NLD	0.560502
JPN	0.510984
ITA	0.508876
BRA	0.506769
CHE	0.460412
KOR	0.455144
CHL	0.454091
FIN	0.402465
DNK	0.330822
ARG	0.312912

BEL	0.308697
HKG	0.263394
TUR	0.203340
NOR	0.201233

```
master_frame = master_frame[~pd.isnull(master_frame['country_code'])]
round(master_frame.isnull().sum()/len(master_frame)*100,2)
```

```
permalink                 0.00
name                      0.00
category_list             0.65
status                    0.00
country_code              0.00
funding_round_permalink   0.00
funding_round_type        0.00
funded_at                 0.00
raised_amount_usd         0.00
dtype: float64
```

```
#Imputing the category list as wel
master_frame = master_frame[~pd.isnull(master_frame['category_list'])]
round(master_frame.isnull().sum()/len(master_frame)*100,2)
```

```
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
```

LVA	0.061107

```
master_frame.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 88507 entries, 0 to 114874
Data columns (total 9 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   permalink               88507 non-null  object
 1   name                    88506 non-null  object
 2   category_list           88507 non-null  object
 3   status                  88507 non-null  object
 4   country_code            88507 non-null  category
 5   funding_round_permalink  88507 non-null  object
 6   funding_round_type      88507 non-null  object
 7   funded_at               88507 non-null  object
 8   raised_amount_usd       88507 non-null  float64
dtypes: category(1), float64(1), object(7)
memory usage: 6.3+ MB
```

SVN	0.027393

Observation:

1. We now have a Non-Null Data of 88507 Datasets which is **77%** of the orignal Data Set[114875]
2. We prefered imputing the **raised_amount_usd** column since it was important, and the data loss was at 17%.
   It did not make sense to fill the data set with median or mean or mode since it will hamper the sanctity of the data
3. we also imputed the **category_list** and **country_code** since they were 1% & 7% respectively and will not hamper the over all analysis due to data loss

CHECKPOINT 02: FUNDING TYPE ANALYSIS

Data Preparation

```
1 master_frame.head(2)
```

	permalink	name	category_list	status	country_code	funding_round_permalink	funding_round_type	funded_at	r
0	/organization/-fame	#fame	Media	operating	IND	/funding-round/9a01d05418af9f794eebff7ace91f638	venture	5/1/2015	
2	/organization/-qounter	:Qounter	Application Platforms|Real Time|Social	operating	USA	/funding-round/b44fbb94153f6cdef13083530bb48030	seed	1/3/2014	

```
rounds2 = master_frame.groupby('funding_round_type')['raised_amount_usd'].median()
pd.DataFrame(rounds2.sort_values(ascending= False))

```

funding_round_type	raised_amount_usd
secondary_market	45850000.0
private_equity	20000000.0
post_ipo_debt	19900000.0
post_ipo_equity	12262852.5
venture	5000000.0
undisclosed	1100000.0
debt_financing	1096653.0
angel	414906.0
convertible_note	300000.0
seed	300000.0
grant	225000.0
product_crowdfunding	211500.0
equity_crowdfunding	85000.0
non_equity_assistance	60000.0
ZMB	0.002107

```
rounds2 = round(rounds2/10e5,2)
rounds2
```

```
funding_round_type
angel                0.41
convertible_note     0.30
debt_financing       1.10
equity_crowdfunding  0.08
grant                0.22
non_equity_assistance  0.06
post_ipo_debt       19.90
post_ipo_equity     12.26
private_equity      20.00
product_crowdfunding  0.21
secondary_market    45.85
seed                 0.30
undisclosed          1.10
venture              5.00
Name: raised_amount_usd, dtype: float64
```

```
master_ = master_frame[(master_frame.funding_round_type=='venture') | (master_frame.funding_round_type == 'an
```

```
master_.head(2)
```

	permalink	name	category_list	status	country_code	funding_round_permalink	fun
0	/organization/-fame	#fame	Media	operating	IND	/funding-round/9a01d05418af9f794eebff7ace91f638	
2	/organization/-qounter	:Qounter	Application Platforms|Real Time|Social Network...	operating	USA	/funding-round/b44fbb94153f6cdef13083530bb48030	

```
master_.shape
```

(75111, 9)

The above represents the final data with only the requisite funding types [**master_**]

MKD 0.001054

Funding Analysis

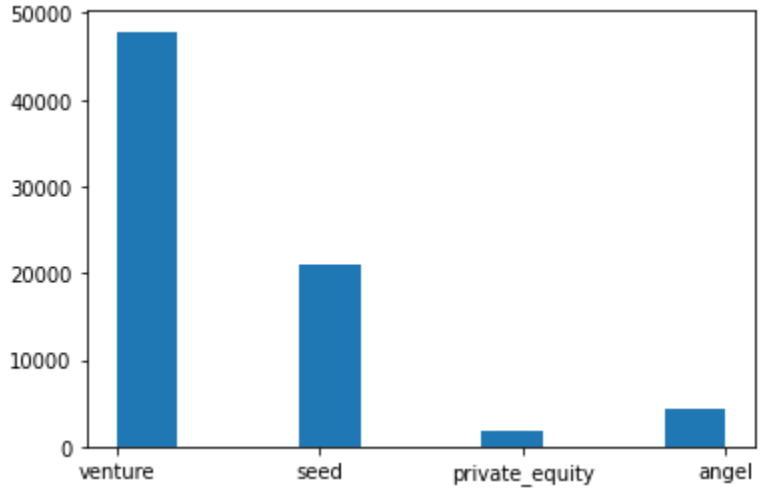
PSE 0.001054

Analysis to understand which investment type is best suited for our needs

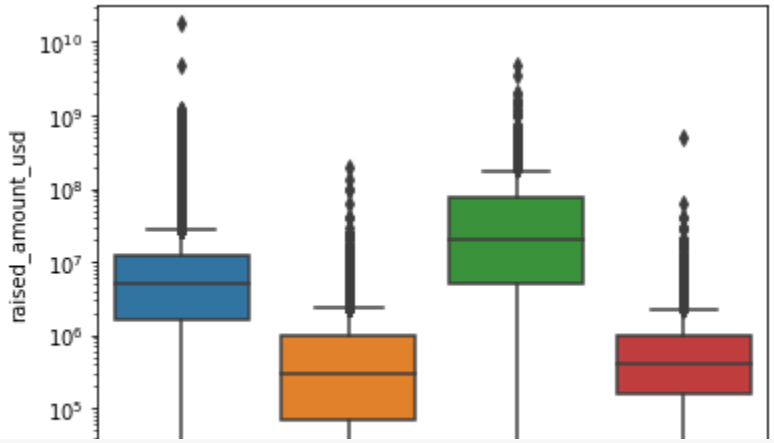
KNA 0.001054

```
plt.hist(master_['funding_round_type'])
```

```
(array([47804.,     0.,     0., 21087.,     0.,     0.,  1820.,     0.,
            0.,  4400.]),
 array([0. , 0.3, 0.6, 0.9, 1.2, 1.5, 1.8, 2.1, 2.4, 2.7, 3. ]),
 <a list of 10 Patch objects>)
```




```
#Comparing Summary Statistics across the 4 categories
sns.boxplot(x = 'funding_round_type', y = 'raised_amount_usd' , data = master_)
plt.yscale('log')
plt.show()
```




```
#comparing the median investment amount across the types
pd.DataFrame(master_.groupby('funding_round_type')['raised_amount_usd'].median().sort_values(ascending= Fals
```

funding_round_type	raised_amount_usd
private_equity	20000000.0
venture	5000000.0
angel	414906.0
seed	300000.0

Since Sparks funds are seeking investment between 05 to 15 mil dollars. venture capital investment will be most suitable

Observation:

1. We found the 4 types of investment types which are the most representative value of the investment with thier respective amounts 2. Out of which the one suiting our needs is the Venture Capital Investments as it fits the budget which we have i.e. 5-15 million dollars

Country Analysis

Spark Funds wants to invest in countries with the highest amount of funding for the chosen investment type. This is a part of its broader strategy to invest where most investments are occurring.

---

1. Sparks fund wants to invest in countries which are **English Speaking**

```
#Spark Funds wants to see the top nine countries
#which have received the highest total funding (across ALL sectors for the chosen investment type)
#Sorting out the country which have been funded by Venture Investments
master_venture = master_[master_.funding_round_type =='venture']
country_wise_total = pd.DataFrame(round((master_venture.groupby('country_code')['raised_amount_usd'].sum().s
country_wise_total
```

country_code	raised_amount_usd
USA	4200.68
CHN	393.39
GBR	200.73
IND	142.62
CAN	94.82
FRA	72.08
ISR	68.54
DEU	63.06
JPN	31.68
SWE	31.46
NLD	29.04
CHE	28.02
SGP	27.94
ESP	18.28
BRA	17.86
IRL	16.68

```
#Showing of Top09 Companies
top_9 = country_wise_total [:9]
top_9
```

country_code	raised_amount_usd
USA	4200.68
CHN	393.39
GBR	200.73
IND	142.62
CAN	94.82
FRA	72.08

Among the top 09 Countries to whom venture investment has been funded, the top 03 English speaking countries are USA, GBR and IND

```
master_venture.head(2)
```

	permalink	name	category_list	status	country_code	funding_round_permalink	fundi
0	/organization/-fame	#fame	Media	operating	IND	/funding-round/9a01d05418af9f794eebff7ace91f638	
4	/organization/0-6-com	0-6.com	Curated Web	operating	CHN	/funding-round/5727accaeaa57461bd22a9bdd945382d	

```
master_venture['country_code'] = master_venture['country_code'].astype('object')
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """Entry point for launching an IPython kernel.
```

```
#Filtering the top 03 Countries where maximum Investments have taken place
```

```
master_venture = master_venture[
                        (master_venture['country_code']== 'USA') |
                        (master_venture['country_code']== 'GBR') |
                        (master_venture['country_code']== 'IND')
                        ]

```

```
print(master_venture.country_code.unique())
```

```
['IND' 'USA' 'GBR']
```

```
master_venture['index_cc'] = master_venture['country_code']
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """Entry point for launching an IPython kernel.
```

```
master_venture
```

	permalink	name	
0	/organization/-fame	#fame	
10	/organization/0xdata	H2O.ai	
11	/organization/0xdata	H2O.ai	
12	/organization/0xdata	H2O.ai	
22	/organization/1-mainstream	1 Mainstream	Apps|Cable|Distr
28	/organization/10-minutes-with	10 Minutes With	
34	/organization/1000memories	1000memories	
38	/organization/1000museums-com	1000museums.com	
39	/organization/1000museums-com	1000museums.com	
41	/organization/1000museums-com	1000museums.com	
44	/organization/1000museums-com	1000museums.com	
59	/organization/100health	Redox	Health Care|Health Care Informa
61	/organization/100plus	100Plus	
62	/organization/1010data	1010data	

```
1 ccr = round((master_venture.groupby('country_code')['raised_amount_usd'].sum())/10e7,2)
```

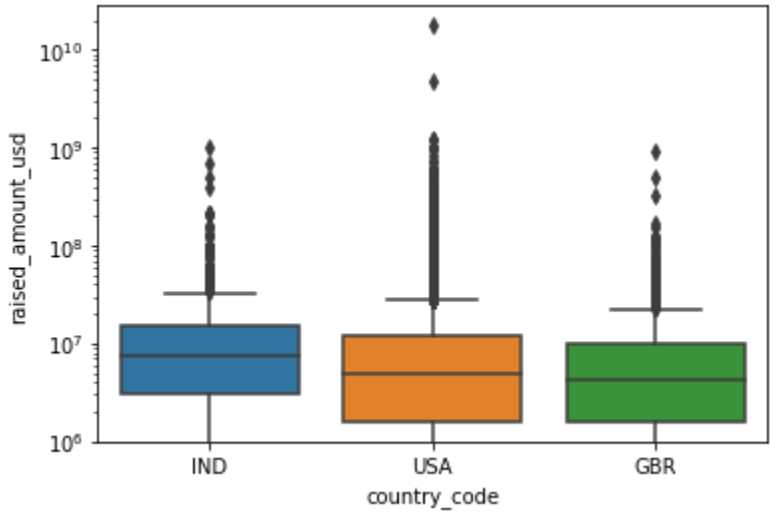
```
ccr = pd.DataFrame(ccr.sort_values(ascending = False))
ccr = ccr[0:3]
ccr
```

country_code	raised_amount_usd
USA	4200.68
GBR	200.73
IND	142.62

109	/organization/1366-technologies	1366 Technologies

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

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f0211706710>
```



Observation

The top 3 english speaking countries are [**USA**] , [**GBR**] ,[**IND**]

Sector Analysis

Sector Analysis

```
mapping  = pd.read_csv('/content/drive/MyDrive/Other Drives/EvilFoxCorps Drive /Datasets/Investment Analysis/
#mapping  = pd.read_csv('/content/mapping.csv')
```

```
mapping.head()
```

	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

```
mapping.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 688 entries, 0 to 687
Data columns (total 10 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   category_list                    687 non-null    object
 1   Automotive & Sports              688 non-null    int64
 2   Blanks                           688 non-null    int64
 3   Cleantech / Semiconductors       688 non-null    int64
```

```
 4   Entertainment                       688 non-null    int64
 5   Health                              688 non-null    int64
 6   Manufacturing                       688 non-null    int64
 7   News, Search and Messaging          688 non-null    int64
 8   Others                              688 non-null    int64
 9   Social, Finance, Analytics, Advertising  688 non-null    int64
dtypes: int64(9), object(1)
memory usage: 53.9+ KB
```

```
#Extracting the mainsector using the column category list
master_venture.loc[:,'main_category'] = master_venture['category_list'].apply(lambda x : x.split('|')[0])
master_venture.head(2)
```

	permalink	name	category_list	status	country_code	funding_round_permalink	funding_round_type	funded_at
0	/organization/-fame	#fame	Media	operating	IND	/funding-round/9a01d05418af9f794eebff7ace91f638	venture	5/1/2015
10	/organization/0xdata	H2O.ai	Analytics	operating	USA	/funding-round/3bb2ee4a2d89251a10aaa735b1180e44	venture	9/11/2015

```
#Dropping the Category list column
master_venture = master_venture.drop('category_list',axis = 1)
master_venture.head(2)
```

	permalink	name	status	country_code	funding_round_permalink	funding_round_type	funded_at	raised_amou
0	/organization/-fame	#fame	operating	IND	/funding-round/9a01d05418af9f794eebff7ace91f638	venture	5/1/2015	1000
10	/organization/0xdata	H2O.ai	operating	USA	/funding-round/3bb2ee4a2d89251a10aaa735b1180e44	venture	9/11/2015	2000

```
1 mapping.isnull().sum()
```

```
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
```

```
1 mapping = mapping[~pd.isnull(mapping['category_list'])]
2 mapping.isnull().sum()
```

```
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
```

Merging the **mapping** file with the main dataframe**(master_venture)**

converting the common column into lower case

```
1 mapping['category_list'] =  mapping['category_list'].str.lower()
2 master_venture['main_category'] = master_venture['main_category'].str.lower()
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
```

Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """Entry point for launching an IPython kernel.

```
1 mapping.head()
```

	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

```
1 #values in main_category columns in df which are not in the category in mapping file
2 master_venture[~master_venture['main_category'].isin(mapping['category_list'])]
```

	permalink	name	status	country_code	funding_round_permalink	fu
10	/organization/0xdata	H2O.ai	operating	USA	/funding-round/3bb2ee4a2d89251a10aaa735b1180e44	
11	/organization/0xdata	H2O.ai	operating	USA	/funding-round/ae2a174c06517c2394aed45006322a7e	
12	/organization/0xdata	H2O.ai	operating	USA	/funding-round/e1cfcbe1bdf4c70277c5f29a3482f24e	
61	/organization/100plus	100Plus	acquired	USA	/funding-round/b5facb0d9dea2f0352b5834892c88c53	
197	/organization/1world-online	1World Online	operating	USA	/funding-round/32936e588a134502712877150198a0b3	
198	/organization/1world-online	1World Online	operating	USA	/funding-round/4e30bd5c85d8163239a3479ec979647a	
199	/organization/1world-online	1World Online	operating	USA	/funding-round/a349bfd7a8d48cfc8b9fdb79480dea7f	
255	/organization/24-7-card	24/7 Card	closed	USA	/funding-round/0c38194ff2035185c96155dfad18f3bd	
820	/organization/6th-wave-innovations-corporation	6th Wave Innovations Corporation	operating	USA	/funding-round/75d128ac40f9e541a1a11786a47c2952	
830	/organization/7-billion-people	7 Billion People	closed	USA	/funding-round/58959ed2be7b14abd6beeb20c9eb17ca	
871	/organization/7park-data	7Park Data	operating	USA	/funding-round/64ddc56c450048911859956eade79cfa	
1004	/organization/9lenses	9Lenses	operating	USA	/funding-round/b27a23a29eb8207f78b60e1f64332832	
1005	/organization/9lenses	9Lenses	operating	USA	/funding-round/b58dcac20e96077aa9f6adf595f3b0fd	
1006	/organization/9lenses	9Lenses	operating	USA	/funding-round/ec22e2c9cac79e78da4c1325db5759d0	
1047	/organization/a-little-world	A LITTLE WORLD	operating	IND	/funding-round/18d88f82ed392b1409975b81f2e8b2fb	

```
# values in the category_list column which are not in main_category column
mapping[~mapping['category_list'].isin(master_venture['main_category'])]
```

	category_list	Automotive & Sports	Blanks	Cleantech / Semiconductors	Entertainment	Health	Manufacturing	News, Search and Messaging	Ot
16	air pollution control	0	0	1	0	0	0	0	
20	alter0tive medicine	0	0	0	0	1	0	0	
22	a0lytics	0	0	0	0	0	0	0	
33	aquaculture	0	0	1	0	0	0	0	
49	b2b express delivery	0	0	0	0	0	0	0	

```
1 mapping['category_list'] = mapping['category_list'].apply(lambda x:x.replace('0','na'))
2 print(mapping['category_list'])
```

```
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    alternative medicine
21              alumni
22           analytics
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
36	art
37	artificial intelligence
38	artists globally
39	assisitive technology
40	assisted living
41	auctions
42	audio
43	audiobooks
44	augmented reality
45	auto
46	automated kiosk
47	automotive
48	b2b
49	b2b express delivery
50	babies
51	baby accessories
52	baby boomers
53	baby safety
54	banking
55	batteries
56	beauty
57	bicycles
58	big data

/funding-

```
#Merging the two datasets together
df = pd.merge(master_venture,mapping,how = 'inner',left_on='main_category',right_on = 'category_list')
df.head()
```

	permalink	name	status	country_code	funding_round_permalink	funding_round_t
0	/organization/-fame	#fame	operating	IND	/funding-round/9a01d05418af9f794eebff7ace91f638	vent
1	/organization/90min	90min	operating	GBR	/funding-round/21a2cbf6f2fb2a1c2a61e04bf930dfe6	vent
2	/organization/90min	90min	operating	GBR	/funding-round/bd626ed022f5c66574b1afe234f3c90d	vent
3	/organization/90min	90min	operating	GBR	/funding-round/fd4b15e8c97ee2ffc0acccdbe1a98810	vent
4	/organization/all-def-digital	All Def Digital	operating	USA	/funding-round/452a2342fe720285c3b92e9bd927d9ba	vent

/funding-

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

permalink	name	status	country_code	funding_round_permalink	funding_round_t
				/funding	:9aefe79dbad639f4d9eba

1 df.info()

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 38788 entries, 0 to 38787
Data columns (total 19 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   permalink                       38788 non-null  object
 1   name                            38788 non-null  object
 2   status                          38788 non-null  object
 3   country_code                    38788 non-null  object
 4   funding_round_permalink         38788 non-null  object
 5   funding_round_type              38788 non-null  object
 6   funded_at                       38788 non-null  object
 7   raised_amount_usd               38788 non-null  float64
 8   index_cc                        38788 non-null  object
 9   main_category                   38788 non-null  object
 10  Automotive & Sports             38788 non-null  int64
 11  Blanks                          38788 non-null  int64
 12  Cleantech / Semiconductors      38788 non-null  int64
 13  Entertainment                   38788 non-null  int64
 14  Health                          38788 non-null  int64
 15  Manufacturing                   38788 non-null  int64
 16  News, Search and Messaging      38788 non-null  int64
 17  Others                          38788 non-null  int64
 18  Social, Finance, Analytics, Advertising  38788 non-null  int64
dtypes: float64(1), int64(9), object(9)
memory usage: 5.9+ MB
```

						/funding-
	/organization/aibit	aibit	acquired	CRR		

1 df.isnull().sum()

```
permalink                  0
name                       0
status                     0
country_code               0
funding_round_permalink    0
```

```
funding_round_type                  0
funded_at                           0
raised_amount_usd                   0
index_cc                            0
main_category                       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
```

```
df.shape
```

```
(38788, 19)
```

The dataset seems to be in WIDE Format, which we need to convert into a long format

```
#Storing the values and ID_Variables in two seperate arrays

#storing the value variables in one series
value_vars = df.select_dtypes(include='int64')
id_vars = df.select_dtypes(exclude = 'int64')
#value_vars = df.columns[9:18]
#id_vars = np.setdiff1d(df.columns,value_vars)

print(value_vars.columns,"\n")
print(id_vars.columns)
```

```
Index(['Automotive & Sports', 'Blanks', 'Cleantech / Semiconductors',
       'Entertainment', 'Health', 'Manufacturing',
       'News, Search and Messaging', 'Others',
       'Social, Finance, Analytics, Advertising'],
```

```
      dtype='object')

Index(['permalink', 'name', 'status', 'country_code',
       'funding_round_permalink', 'funding_round_type', 'funded_at',
       'raised_amount_usd', 'index_cc', 'main_category'],
      dtype='object')
```

```
1 #Converting into LONG
2 long_df = pd.melt(df,id_vars=list(id_vars),value_vars=list(value_vars))
3 long_df.head()
```

	permalink	name	status	country_code	funding_round_permalink	funding_round_t
0	/organization/-fame	#fame	operating	IND	/funding-round/9a01d05418af9f794eebff7ace91f638	vent
1	/organization/90min	90min	operating	GBR	/funding-round/21a2cbf6f2fb2a1c2a61e04bf930dfe6	vent
2	/organization/90min	90min	operating	GBR	/funding-round/bd626ed022f5c66574b1afe234f3c90d	vent
3	/organization/90min	90min	operating	GBR	/funding-round/fd4b15e8c97ee2ffc0acccdbe1a98810	vent
4	/organization/all-def-digital	All Def Digital	operating	USA	/funding-round/452a2342fe720285c3b92e9bd927d9ba	vent

```
1 # renaming the 'variable' column
2 long_df = long_df.rename(columns={'variable': 'sector'})
```

```
1 long_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 349092 entries, 0 to 349091
```

```
Data columns (total 12 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   permalink             349092 non-null  object
 1   name                  349092 non-null  object
 2   status                349092 non-null  object
 3   country_code          349092 non-null  object
 4   funding_round_permalink  349092 non-null  object
 5   funding_round_type    349092 non-null  object
 6   funded_at             349092 non-null  object
 7   raised_amount_usd     349092 non-null  float64
 8   index_cc              349092 non-null  object
 9   main_category         349092 non-null  object
 10  sector                349092 non-null  object
 11  value                 349092 non-null  int64
dtypes: float64(1), int64(1), object(10)
memory usage: 32.0+ MB
```

The dataframe now contains only venture type investments in countries USA, IND, GBR and the same has been mapped to the 8 sectors in the dataframe

```
1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 38788 entries, 0 to 38787
Data columns (total 19 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   permalink                     38788 non-null  object
 1   name                          38788 non-null  object
 2   status                        38788 non-null  object
 3   country_code                  38788 non-null  object
 4   funding_round_permalink       38788 non-null  object
 5   funding_round_type            38788 non-null  object
 6   funded_at                     38788 non-null  object
 7   raised_amount_usd             38788 non-null  float64
 8   index_cc                      38788 non-null  object
 9   main_category                 38788 non-null  object
 10  Automotive & Sports           38788 non-null  int64
 11  Blanks                        38788 non-null  int64
 12  Cleantech / Semiconductors    38788 non-null  int64
```

```
 13  Entertainment                       38788 non-null  int64
 14  Health                              38788 non-null  int64
 15  Manufacturing                       38788 non-null  int64
 16  News, Search and Messaging          38788 non-null  int64
 17  Others                              38788 non-null  int64
 18  Social, Finance, Analytics, Advertising  38788 non-null  int64
dtypes: float64(1), int64(9), object(9)
memory usage: 5.9+ MB
```

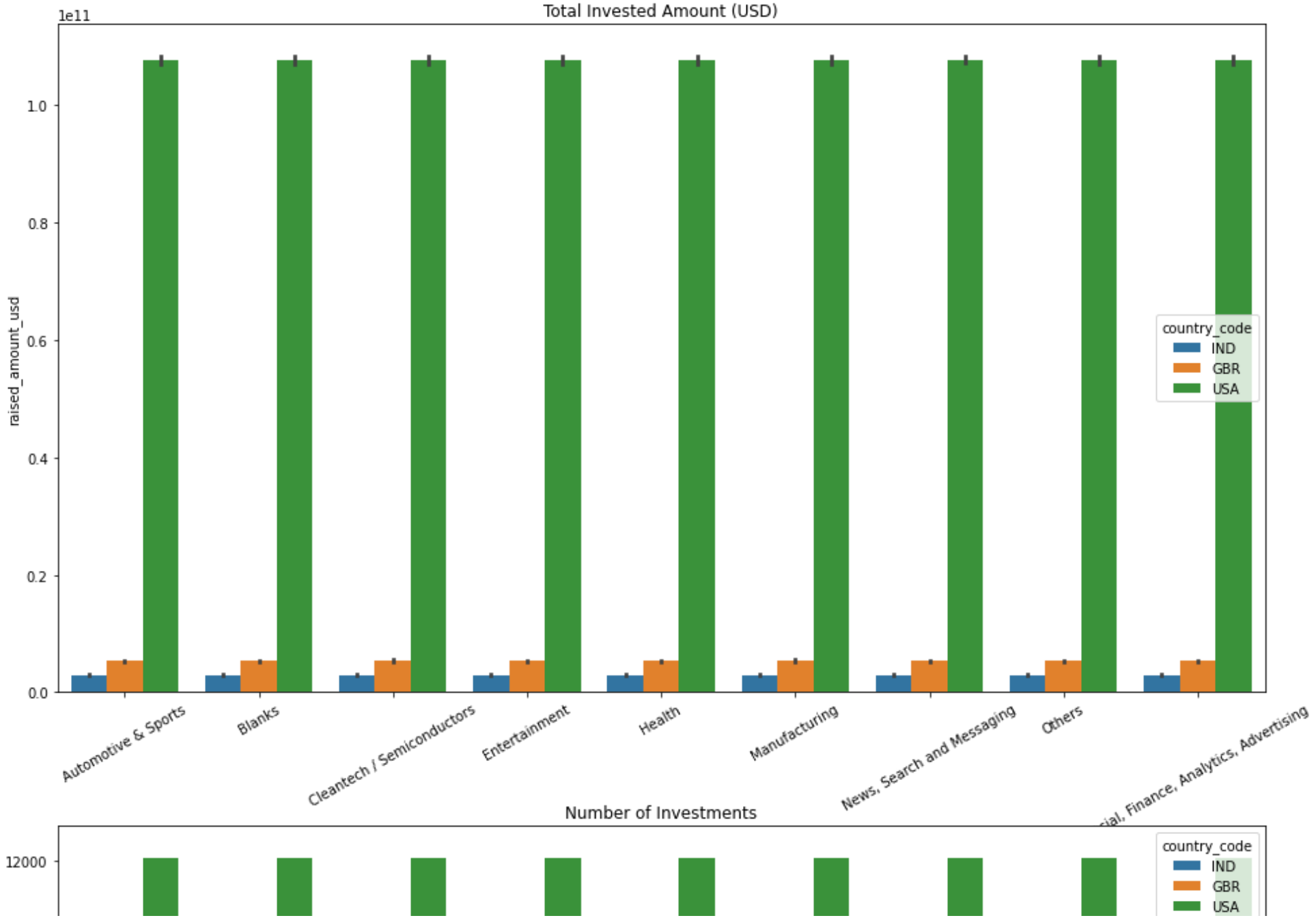
```
#Summarising the sector wise number and sum of venture investments across three countries

#Creating a investment filter in between range 5 and 15 mil

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

```
#groupby country sector and compute thr sum and count

df.groupby(['country_code','sector']).raised_amount_usd.agg(['count','sum'])
```

		count	sum
country_code	sector		
GBR	Automotive & Sports	621	5.379079e+09
	Blanks	621	5.379079e+09
	Cleantech / Semiconductors	621	5.379079e+09
	Entertainment	621	5.379079e+09
	Health	621	5.379079e+09
	Manufacturing	621	5.379079e+09
	News, Search and Messaging	621	5.379079e+09
	Others	621	5.379079e+09
	Social, Finance, Analytics, Advertising	621	5.379079e+09
IND	Automotive & Sports	328	2.949544e+09
	Blanks	328	2.949544e+09
	Cleantech / Semiconductors	328	2.949544e+09
	Entertainment	328	2.949544e+09
	Health	328	2.949544e+09
	Manufacturing	328	2.949544e+09
	News, Search and Messaging	328	2.949544e+09

```
#Plotting Sector wise count and sum of investments in the three countries
plt.figure(figsize=(16, 20))

plt.subplot(2, 1, 1)
p = sns.barplot(x='sector', y='raised_amount_usd', hue='country_code', data=df, estimator=np.sum)
p.set_xticklabels(p.get_xticklabels(),rotation=30)
plt.title('Total Invested Amount (USD)')
```

```
8
9  plt.subplot(2, 1, 2)
10 q = sns.countplot(x='sector', hue='country_code', data=df)
11 q.set_xticklabels(q.get_xticklabels(),rotation=30)
12 plt.title('Number of Investments')
13
14
15 plt.show()
```



Observation:

We can observe that the top country is USA for Investment with others , social finance analytics and advertising cleantech/semiconductors being the heavily invested ones

FINAL OBSERVATION ⚓:

1. we deduced the Spark Funds Investment Requirement to [**Venture Capital Investments**] which will be funding between 05 - 15 Million Dollars.
2. We deduced that amongst the top 10 countries being heavily invested in, the Top 3 countries which spoke English to fit the requirement criteria were **[USA] , [IND] , [GBR]**
3. We finally deduced that out of the 8 Sectors, the Sector under **[Others]** was the top investment selection amongst the countries

```
1
2
3
4
```

9095	/organization/autopilot-2	Autopilot	operating	USA	/funding-round/6f6c2e3b8856ea85d93349f9f6bac16c
9096	/organization/autopilot-2	Autopilot	operating	USA	/funding-round/8e31864c733235798a591a8d47f8720b
9097	/organization/autopilot-2	Autopilot	operating	USA	/funding-round/956a2493b26b37813aabdeab1cf1b88c
9098	/organization/autopilot-2	Autopilot	operating	USA	/funding-round/d347cad0cc64aa00bb3f112dda5f86f2
9162	/organization/avadhi-finance-and-technology	Avadhi Finance and Technology	operating	USA	/funding-round/3e117e3b970d090856a5a0af44dc0dcd
9224	/organization/avant-credit	Avant	operating	USA	/funding-round/22d3eba9a031effe2bede348d8a8be9e