# Project Brief :

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

## Business and Data Understanding :

X has two minor constraints for investments:

1. It wants to invest between 5 to 15 million USD per round of investment
2. It wants to invest only in English-speaking countries because of the ease of communication with the companies it would invest in.

These conditions will give you sufficient information for your initial analysis. Before getting to specific questions, let’s understand the problem and the data first.

### 1. What is the strategy?

X wants to invest where most other investors are investing. This pattern is often observed among early stage startup investors.

### 2. Where did we get the data from?

We have taken real investment data from crunchbase.com, so the insights you get may be incredibly useful. For this group project, we have divided the data into three files.

You have to use three main data tables for the entire analysis.

### 3. What is X business objective?

The business objectives and goals of data analysis are pretty straightforward.

#### 1. Business objective:

The objective is to identify the best sectors, countries, and a suitable investment type for making investments. The overall strategy is to invest where others are investing, implying that the 'best' sectors and countries are the ones 'where most investors are investing'.

#### 2. Goals of data analysis:

***Your goals are divided into three sub-goals:***

***Investment type analysis:***

Comparing the typical investment amounts in the venture, seed, angel, private equity etc. so that X can choose the type that is best suited for their strategy.

***Country analysis:***

Identifying the countries which have been the most heavily invested in the past. These will be X favourites as well.

***Sector analysis:***

Understanding the distribution of investments across the eight main sectors. (Note that we are interested in the eight 'main sectors' provided in the mapping file. The two files — companies and rounds2 — have numerous sub-sector names; hence, you will need to map each sub-sector to its main sector.)

In [9]:

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

# reading data files
# using encoding = "ISO-8859-1" to avoid pandas encoding error
rounds = pd.read_csv("rounds2.csv", encoding = "ISO-8859-1")
```

```
companies = pd.read_csv("companies.txt", sep="\t", encoding = "ISO-8859-1")
```

In [10]:

```
rounds.head(5)
```

Out[10]:

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

In [11]:

```
rounds.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 114949 entries, 0 to 114948
Data columns (total 6 columns):
company_permalink          114949 non-null object
funding_round_permalink    114949 non-null object
funding_round_type         114949 non-null object
funding_round_code         31140 non-null object
funded_at                  114949 non-null object
raised_amount_usd          94959 non-null float64
dtypes: float64(1), object(5)
memory usage: 5.3+ MB
```

In [12]:

```
rounds.shape
```

Out[12]:

```
(114949, 6)
```

The variables funding_round_code and raised_amount_usd contain some missing values, as shown above. We'll deal with them after we're done with understanding the data - column names, primary keys of tables etc.

In [13]:

```
companies.head()
```

Out[13]:

| | permalink | name | homepage_url | category_list | status | country_code | state_code | regio |
|---|---|---|---|---|---|---|---|---|
| **0** | /Organization/-Fame | #fame | http://livfame.com | Media | operating | IND | 16 | Mumb |
| **1** | /Organization/-Qounter | :Qounter | http://www.qounter.com | Application Platforms\|Real Time\|Social Network... | operating | USA | DE | DE - C |

|   | permalink | name | homepage_url | category_list | status | country_code | state_code | region |
|---|---|---|---|---|---|---|---|---|
| 2 | /Organization/-The-One-Of-Them-Inc- | (THE) ONE of THEM,Inc. | http://oneofthem.jp | Apps\|Games\|Mobile | operating | NaN | NaN | NaN |
| 3 | /Organization/0-6-Com | 0-6.com | http://www.0-6.com | Curated Web | operating | CHN | 22 | Beijing |
| 4 | /Organization/004-Technologies | 004 Technologies | http://004gmbh.de/en/004-interact | Software | operating | USA | IL | Spring<br>Illinois |

In [14]:

```
companies.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 66368 entries, 0 to 66367
Data columns (total 10 columns):
permalink        66368 non-null object
name             66367 non-null object
homepage_url     61310 non-null object
category_list    63220 non-null object
status           66368 non-null object
country_code     59410 non-null object
state_code       57821 non-null object
region           58338 non-null object
city             58340 non-null object
founded_at       51147 non-null object
dtypes: object(10)
memory usage: 5.1+ MB
```

In [15]:

```
companies.shape
```

Out[15]:

```
(66368, 10)
```

Ideally, the permalink column in the companies dataframe should be the unique_key of the table, having 66368 unique company names (links, or permalinks). Also, these 66368 companies should be present in the rounds file.

Let's first confirm that these 66368 permalinks (which are the URL paths of companies' websites) are not repeating in the column, i.e. they are unique.

## Data Cleaning

In [16]:

```
# identify the unique number of permalinks in companies
len(companies.permalink.unique())
```

Out[16]:

```
66368
```

Also, let's convert all the entries to lowercase (or uppercase) for uniformity.

In [17]:

```
# converting all permalinks to lowercase
companies['permalink'] = companies['permalink'].str.lower()
companies.head()
```

Out[17]:

|   | permalink | name | homepage_url | category_list | status | country_code | state_code | region |
|---|---|---|---|---|---|---|---|---|
|   | /organization/ |   |   |   |   |   |   |   |

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

In [18]:

```
# look at unique values again
len(companies.permalink.unique())
```

Out[18]:

```
66368
```

Thus, there are 66368 unique companies in the table and permalink is the unique primary key. Each row represents a unique company.

Let's now check whether all of these 66368 companies are present in the rounds file, and if some extra ones are present.

In [19]:

```
# look at unique company names in rounds df
# note that the column name in rounds file is different (company_permalink)
len(rounds.company_permalink.unique())
```

Out[19]:

```
90247
```

There seem to be 90247 unique values of company_permalink, whereas we expected only 66368. May be this is because of uppercase/lowercase issues.

Let's convert the column to lowercase and look at unique values again.

In [20]:

```
rounds['company_permalink'] = rounds['company_permalink'].str.lower()
```

In [21]:

```
len(rounds.company_permalink.unique())
```

Out[21]:

```
66370
```

There seem to be 2 extra permalinks in the rounds file which are not present in the companies file. Let's hope that this is a data quality issue, since if this were genuine, we have two companies whose investment round details are available but their metadata (company name, sector etc.) is not available in the companies table.

Let's have a look at the company permalinks which are in the 'rounds' file but not in 'companies'.

In [22]:

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

Out[22]:

| | company_permalink | funding_round_permalink | funding_round_type | funding_round_code | fur |
|---|---|---|---|---|---|
| **29597** | /organization/e-cãbica | /funding-round/8491f74869e4fe8ba9c378394f8fbdea | seed | NaN | 01-<br>20 |
| **31863** | /organization/energystone-games-çµç³æ¸¸æ | /funding-round/b89553f3d2279c5683ae93f45a21cfe0 | seed | NaN | 09-<br>20 |
| **45176** | /organization/huizuche-com-æ ç§ÿè½¦ | /funding-round/8f8a32dbeeb0f831a78702f83af78a36 | seed | NaN | 18-<br>20 |
| **58473** | /organization/magnet-tech-ç£ç³ç§æ | /funding-round/8fc91fbb32bc95e97f151dd0cb4166bf | seed | NaN | 16-<br>20 |
| **101036** | /organization/tipcat-interactive-æ²èÿä¿¡æ¯ç... | /funding-round/41005928a1439cb2d706a43cb661f60f | seed | NaN | 06-<br>20 |
| **109969** | /organization/weiche-tech-åè½¦ç§æ | /funding-round/f74e457f838b81fa0b29649740f186d8 | venture | A | 06-<br>20 |
| **113839** | /organization/zengame-ç¦æ¸¸ç§æ | /funding-round/6ba28fb4f3eadf5a9c6c81bc5dde6cdf | seed | NaN | 17-<br>20 |

All the permalinks have weird non-English characters. Let's see whether these characters are present in the original df as well.

In [23]:

```
# looking at the indices with weird characters
rounds_original = pd.read_csv("rounds2.csv", encoding = "ISO-8859-1")
rounds_original.iloc[[29597, 31863, 45176, 58473], :]
```

Out[23]:

| | company_permalink | funding_round_permalink | funding_round_type | funding_round_co |
|---|---|---|---|---|
| **29597** | /ORGANIZATION/E-CÃBICA | /funding-round/8491f74869e4fe8ba9c378394f8fbdea | seed | NaN |
| **31863** | /ORGANIZATION/ENERGYSTONE-GAMES-ÇµÇ³Æ¸¸Æ | /funding-round/b89553f3d2279c5683ae93f45a21cfe0 | seed | NaN |
| **45176** | /organization/huizuche-com-æ ç§ÿè½¦ | /funding-round/8f8a32dbeeb0f831a78702f83af78a36 | seed | NaN |
| **58473** | /ORGANIZATION/MAGNET-TECH-Ç£Ç³Ç§Æ | /funding-round/8fc91fbb32bc95e97f151dd0cb4166bf | seed | NaN |

The company weird characters appear when you import the data file. To confirm whether these characters are actually present in the given data or whether python has introduced them while importing into pandas, let's have a look at the original CSV file in Excel.

Thus, this is most likely a data quality issue we have introduced while reading the data file into python. Specifically, this is most likely caused because of encoding.

First, let's try to figure out the encoding type of this file. Then we can try specifying the encoding type at the time of reading the file. The chardet library shows the encoding type of a file.

In [24]:

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

Out[24]:

| | company_permalink | funding_round_permalink | funding_round_type | funding_round_code | fu |
|---|---|---|---|---|---|
| | | /funding | | | 1 |

| | company_permalink | funding_round_permalink | funding_round_type | funding_round_code | f |
|---|---|---|---|---|---|
| **77** | /organization/10north | /funding-round/b41ff7de932f8b6e5bbeed3966c0ed6a | equity_crowdfunding | NaN | 2 |
| **729** | /organization/51wofang- | /funding-round/346b9180d276a74e0fbb2825e66c6f5b | venture | A | 0<br>2 |
| **2670** | /organization/adslinked | /funding-round/449ae54bb63c768c232955ca6911dee4 | seed | NaN | 2<br>2 |
| **3166** | /organization/aesthetic-everything-social-network | /funding-round/62593455f1a69857ed05d5734cc04132 | equity_crowdfunding | NaN | 1<br>2 |
| **3291** | /organization/affluent-attach-club-2 | /funding-round/626678bdf1654bc4df9b1b34647a4df1 | seed | NaN | 1<br>2 |
| **4568** | /organization/allgu-outlet | /funding-round/49e8a9b54ed19c8505ca92dc031a8e9c | venture | NaN | 1<br>2 |
| **8097** | /organization/asiansbook | /funding-round/3f243ab92b4fe397d41b4734a17ca5f0 | seed | NaN | 1<br>2 |
| **8652** | /organization/atlye-gri | /funding-round/75bdeacd95a647108aa4bc480e77894d | grant | NaN | 0<br>2 |
| **9784** | /organization/axgaz | /funding-round/511a41181aaf193bbd419babfb8d66e9 | venture | NaN | 0<br>2 |
| **14311** | /organization/boral-bikes-incorporated | /funding-round/be79575bf4b5b5d6fa64670800a3ca5e | seed | NaN | 2<br>2 |
| **14798** | /organization/brasil-oznio | /funding-round/4e0dd70413b121d23274187704e2d91b | seed | NaN | 0<br>2 |
| **14951** | /organization/bricopriv-com | /funding-round/c14e573c4cea05d355a20b5ba6b0d12d | undisclosed | NaN | 2<br>2 |
| **15384** | /organization/brv | /funding-round/978b27fe5c90372b11adbe33c75cdd03 | seed | NaN | 3<br>2 |
| **16018** | /organization/bhner-eh-gmbh | /funding-round/21fde9aaef25f3e1889b8662d7540eda | seed | NaN | 1<br>2 |
| **16624** | /organization/canal-da-pea | /funding-round/a16902fe6bd5e67a44dd222ac209fa7e | seed | NaN | 1<br>2 |
| **16839** | /organization/cappt | /funding-round/b2b88b247a67469bc8a2df8510078290 | convertible_note | NaN | 0<br>2 |
| **23210** | /organization/contrato-rpido | /funding-round/33fe7ab355ca20d8993d8dbe3dcd62d2 | seed | NaN | 0<br>2 |
| **23518** | /organization/cop-active-ltd | /funding-round/962d2256a1d52ffd07536d03cd90e5b3 | seed | NaN | 1<br>2 |
| **24932** | /organization/crme-ciseaux | /funding-round/0aeedf03903b6ff7a7c5d02871842588 | equity_crowdfunding | NaN | 0<br>2 |
| **24933** | /organization/crme-ciseaux | /funding-round/c05eea4d4b53eac40f29b4ea9ea67838 | equity_crowdfunding | NaN | 0<br>2 |
| **27110** | /organization/desafo-tctico | /funding-round/32109a690a936b7678359619734b8cf9 | convertible_note | NaN | 0<br>2 |
| **29597** | /organization/e-cbica | /funding-round/8491f74869e4fe8ba9c378394f8fbdea | seed | NaN | 0<br>2 |
| **31863** | /organization/energystone-games- | /funding-round/b89553f3d2279c5683ae93f45a21cfe0 | seed | NaN | 0<br>2 |
| **33069** | /organization/etool-io | /funding-round/3575ca572169fb7b320665767250355a | seed | NaN | 1<br>2 |
| **37562** | /organization/freem | /funding-round/34d9e77d186da2e69b68f371a3f4926e | convertible_note | NaN | 1<br>2 |
| **37876** | /organization/frquentiel | /funding-round/b0778e93110786d599ce8a4788adda9d | undisclosed | NaN | 1<br>2 |
| **39460** | /organization/gesto-sade-e-tecnologia-3 | /funding-round/f56dcf159f31717b4b4ddd91f732e881 | venture | NaN | 0<br>2 |

|  | company_permalink | funding_round_permalink | funding_round_type | funding_round_code | f |
|---|---|---|---|---|---|
| **42221** | /organization/[illegible]-en-lnea | /funding-round/cb747fc6c90fa66ebc3ebe25d3377358 | equity_crowdfunding | NaN | 0 2 |
| **45176** | /organization/huizuche-com- | /funding-round/8f8a32dbeeb0f831a78702f83af78a36 | seed | NaN | 1 2 |
| **46281** | /organization/ignia-bienes-races | /funding-round/ad8b2bf09fda2a09dcdd2ed2a86d0d9c | venture | NaN | 0 2 |
| **...** | ... | ... | ... | ... | .. |
| **58474** | /organization/magnet-tech- | /funding-round/be2fb8789ec4e1902c2a7e1f7313ad3d | venture | A | 1 2 |
| **60960** | /organization/mercado-electrnico | /funding-round/e4a31c7eb3546cc2bc6eb0c8d05625d4 | private_equity | NaN | 1 2 |
| **62172** | /organization/ming-yazlm | /funding-round/b510e4822c0f1b4977cf988e4c5054d0 | grant | NaN | 0 2 |
| **63761** | /organization/monnier-frres | /funding-round/606655ce25b330ee620137c09af4ec21 | seed | NaN | 3 2 |
| **65471** | /organization/mdica-santa-carmen-2 | /funding-round/bd94fb319f6ec0e2021dcc5bfad03479 | seed | NaN | 0 2 |
| **73633** | /organization/patrofn | /funding-round/a49df2be4369c01a4a16b9356f5640dd | grant | NaN | 0 2 |
| **78497** | /organization/prodti-cz | /funding-round/e3e7909a3c46b470a35fcaf469bdbcae | seed | NaN | 0 2 |
| **79281** | /organization/przewietl-pl | /funding-round/9fe8ed2986d279646cbcdb72c6c5128a | seed | NaN | 0 2 |
| **85619** | /organization/salo-vip | /funding-round/6cc488b6bee6c0741491ef71b953dbc6 | grant | NaN | 0 2 |
| **85996** | /organization/satlite-distribuidora-de-petrleo | /funding-round/a99b1d7625b01c9af80d622e61aaa511 | venture | NaN | 1 2 |
| **90328** | /organization/skar-is | /funding-round/4569a387b074bdc097442f9753914289 | seed | NaN | 2 2 |
| **91932** | /organization/socit-internationale-de-plantati... | /funding-round/4f6d8e2551eb84c3b5c8234d19b63944 | venture | NaN | 2 2 |
| **97158** | /organization/slfar-studios | /funding-round/69c192af02bf6e34dd7210298cecdc3b | seed | NaN | 1 2 |
| **97571** | /organization/talentsigned | /funding-round/b5e611e4c9f4f4ac6b9df67628a90382 | seed | NaN | 2 2 |
| **99945** | /organization/the-vision-lab- | /funding-round/f0d1dc9c2fc5784065990cebcc4c3515 | convertible_note | NaN | 2 2 |
| **100588** | /organization/th-gii-di-ng | /funding-round/a49012d798c32ae6c113b8234bdcf804 | venture | NaN | 2 2 |
| **101036** | /organization/tipcat-interactive- | /funding-round/41005928a1439cb2d706a43cb661f60f | seed | NaN | 0 2 |
| **104092** | /organization/tximo | /funding-round/574466178e0b9e182f1e541c6313ea27 | venture | NaN | 0 2 |
| **104093** | /organization/to-conejo | /funding-round/f7559463034c712e16100e9466a93057 | convertible_note | NaN | 0 2 |
| **105508** | /organization/vacation-bnb | /funding-round/f9cc0781977926a4132d6a0f87b0f774 | seed | NaN | 0 2 |
| **108953** | /organization/v-de-txi | /funding-round/5fe845b41da2eaa8842feb65bb4d1f08 | seed | NaN | 1 2 |
| **108954** | /organization/vnder-sports-network | /funding-round/63e91de54ea7451b1e2e89ffa6d37443 | seed | NaN | 1 2 |
| **109968** | /organization/weiche-tech- | /funding-round/27b0cd2e0b75cbceb717343ea86c2c28 | angel | NaN | 1 2 |

| | company_permalink | funding_round_permalink | funding_round_type | funding_round_code | f |
|---|---|---|---|---|---|
| **109969** | /organization/weiche-tech- | /funding-round/f74e457f838b81fa0b29649740f186d8 | venture | A | 0<br>2 |
| **110516** | /organization/whites-holdings | /funding-round/7407f06542934e7dd12beaacc71e8fdc | undisclosed | NaN | 1<br>2 |
| **110545** | /organization/whodats-spaces | /funding-round/d5d6db3d1e6c54d71a63b3aa0c9278e6 | seed | NaN | 2<br>2 |
| **113839** | /organization/zengame- | /funding-round/6ba28fb4f3eadf5a9c6c81bc5dde6cdf | seed | NaN | 1<br>2 |
| **114946** | /organization/eron | /funding-round/59f4dce44723b794f21ded3daed6e4fe | venture | A | 0<br>2 |
| **114947** | /organization/asys-2 | /funding-round/35f09d0794651719b02bbfd859ba9ff5 | seed | NaN | 0<br>2 |
| **114948** | /organization/novatiff-reklam-ve-tantm-hizmetl... | /funding-round/af942869878d2cd788ef5189b435ebc4 | grant | NaN | 0<br>2 |

74 rows × 6 columns

This seems to work fine.

Let's now look at the number of unique values in rounds dataframe again.

In [26]:

```
# Look at unique values again
len(rounds.company_permalink.unique())
```

Out[26]:

```
66368
```

Now it makes sense - there are 66368 unique companies in both the rounds and companies dataframes.

It is possible that a similar encoding problems are present in the companies file as well. Let's look at the companies which are present in the companies file but not in the rounds file - if these have special characters, then it is most likely because the companies file is encoded (while rounds is not).

In [27]:

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

Out[27]:

| | permalink | name | homepage_url | category_list | sta |
|---|---|---|---|---|---|
| **43** | /organization/10â°north | 10Â°North | NaN | Fashion | ope |
| **426** | /organization/51wofang-æ å¿§æ æ¿ | 51wofang æ å¿§æ æ¿ | http://www.51wofang.com | NaN | clos |
| **1506** | /organization/adslinkedâ¢ | AdsLinkedâ¢ | http://www.adslinked.com | Advertising\|Internet | ope |
| **1775** | /organization/aesthetic-everythingâ®-social-ne... | Aesthetic EverythingÂ® Social Network | http://aestheticeverything.com/ | Public Relations | ope |
| **1834** | /organization/affluent-attachã©-club-2 | Affluent AttachÃ© Club | http://www.affluentattache.com/ | Hospitality | ope |
| **2556** | /organization/allgã¤u-outlet | AllgÃ¤u Outlet | http://allgaeuoutlet.de/ | Fashion | ope |
| **4567** | /organization/asiansbookâ¢ | Asiansbookâ¢ | http://www.asiansbook.com | Social Media\|Social Network Media | ope |
| **4903** | /organization/atã¶lye-gri | AtÃ¶lye Gri | http://www.atolyegri.com/ | Advertising | ope |

| | permalink | name | homepage_url | category_list | sta |
|---|---|---|---|---|---|
| **5490** | /organization/axã¨gaz | AxÃ¨gaz | http://www.axegaz.com/ | Energy | ope |
| **8131** | /organization/borã©al-bikes-incorporated | BorÃ©al Bikes Incorporated | http://www.borealbikes.com | Automotive\|Design\|Manufacturing | ope |
| **8385** | /organization/brasil-ozã´nio | Brasil OzÃ´nio | http://www.brasilozonio.com.br | Environmental Innovation\|Services\|Water | clos |
| **8477** | /organization/bricoprivã©-com | BricoprivÃ©.com | http://www.bricoprive.com/ | Product Design | ope |
| **8710** | /organization/bräv | BrÄv | http://brav.org/ | All Students | ope |
| **9095** | /organization/bã¶hner-eh-gmbh | BÃ¶hner-EH GmbH | http://www.eh-d.de/ | NaN | ope |
| **9441** | /organization/canal-da-peã§a | Canal da PeÃ§a S.A. | http://cdp.parts | NaN | ope |
| **9569** | /organization/capptãº | CapptÃº | http://www.capptu.com/ | Apps\|Communities\|Photography | ope |
| **13126** | /organization/contrato-rã¡pido | Contrato RÃ¡pido | http://www.contratorapido.com.br | Document Management\|Legal\|SaaS\|Software | ope |
| **13286** | /organization/copã©-active-ltd | CopÃ© Active Ltd. | http://www.copeactive.com/ | Active Lifestyle\|E-Commerce\|Health and Wellnes... | ope |
| **14078** | /organization/crã¨me-ciseaux | CrÃ¨me & Ciseaux | https://creme-ciseaux.com/ | NaN | clos |
| **15306** | /organization/desafão-tã¡ctico | DesafÃo TÃ¡ctico | http://desafiotactico.260mb.org/ | NaN | ope |
| **16827** | /organization/e-cãbica | E CÃBICA | NaN | NaN | ope |
| **18197** | /organization/energystone-games-çµçÿ³æ¸¸æ | EnergyStone Games çµç³æ¸¸æ | NaN | Mobile Games\|Online Gaming | clos |
| **18926** | /organization/etool-ioâ | eTool.ioÂ | http://www.eTool.io | Advertising | ope |
| **21578** | /organization/freemå | FreeMÅ | http://www.getfreemo.com | Mobile | ope |
| **21769** | /organization/frã©quentiel | FrÃ©quentiel | http://www.frequentiel.com/fr/accueil/ | NaN | ope |
| **22711** | /organization/gesto-saãºde-e-tecnologia-3 | Gesto SaÃºde e Tecnologia | http://www.gestosaude.com.br/ | NaN | ope |
| **24327** | /organization/grã¡fica-en-lã-nea | GrÃ¡fica en lÃ-nea | http://otw2.vsoft.cl | Manufacturing\|Software | ope |
| **26139** | /organization/huizuche-com-æ ç§è½¦ | Huizuche.com æ ç§è½¦ | http://huizuche.com | NaN | clos |
| **26818** | /organization/ignia-bienes-raãces | IGNIA Bienes RaÃces | http://www.casaspremin.com.mx | NaN | ope |
| **26892** | /organization/ikigã¼nde-com | IkigÃ¼nde.com | http://www.ikigunde.com/ | E-Commerce | ope |
| **...** | ... | ... | ... | ... | ... |
| **31693** | /organization/lawpã dã | LawPadi | http://lawpadi.com/ | Advice\|Internet\|Legal | clos |
| **33892** | /organization/magnet-tech-ç£çÿ³ç§æ | Magnet Tech ç£ç³ç§æ | http://www.buga.cn | Communications Hardware\|Families\|Hardware + So... | clos |
| **35353** | /organization/mercado-electrã´nico | Mercado EletrÃ´nico | http://www.me.com.br | Internet\|Outsourcing\|SaaS | clos |
| **36039** | /organization/ming-yazä±lä±m | Ming YazÄ±lÄ±m | http://www.ming.com.tr | Software | ope |
| **36917** | /organization/monnier-frã¨res | Monnier FrÃ¨res | http://www.monnierfreres.com/ | E-Commerce\|Fashion\|Lifestyle | acq |
| **37951** | /organization/mã©dica- | MÃ©dica Santa | http://www.medicasantacarmen.com | NaN | ope |

| | permalink | name | homepage_url | category_list | sta |
|---|---|---|---|---|---|
| | santa-carmen-2 | Carmen | | | |
| **42529** | /organization/patrofä°n | PatroFÄ°N | http://www.patrofin.com | Software | ope |
| **45338** | /organization/prodäti-cz | ProdÄti.cz | NaN | NaN | clos |
| **45760** | /organization/przeåwietl-pl | PrzeÅwietl.pl | https://przeswietl.pl/ | NaN | ope |
| **49431** | /organization/salã£o-vip | SalÃ£o VIP | http://www.salaovip.com.br/ | Beauty\|Internet\|Online Scheduling | ope |
| **49635** | /organization/satã©lite-distribuidora-de-petrã... | SatÃ©lite Distribuidora de PetrÃ³leo | NaN | NaN | ope |
| **52055** | /organization/skarã¸-is | SkarÃ¸ is | NaN | NaN | ope |
| **52994** | /organization/sociã©tã©-internationale-de-plan... | SociÃ©tÃ© Internationale de Plantations d'HÃ©v... | http://www.siph.com/ | Natural Resources\|Product Development Services... | clos |
| **56005** | /organization/sã³lfar-studios | SÃ³lfar Studios | http://www.solfar.com/ | Games | ope |
| **56251** | /organization/talentsignedâ¢ | TalentSignedâ¢ | http://www.talentsigned.com | Internet | ope |
| **57710** | /organization/the-vision-lab-â® | The Vision Lab Â® | http://www.thevisionlab.com | Crowdsourcing\|Enterprise Software\|Internet | ope |
| **58099** | /organization/thá°¿-giá»i-di-äá»ng | The Gioi Di Dong | https://www.thegioididong.com/ | NaN | ope |
| **58344** | /organization/tipcat-interactive-æ²èä¿¡æ¯ç... | TipCat Interactive æ²èä¿¡æ¯ç§æ | http://www.tipcat.com | Mobile Games\|Online Gaming | clos |
| **60102** | /organization/tã¡ximo | TÃ¡ximo | http://www.taximo.co/ | Automotive | ope |
| **60103** | /organization/tão-conejo | TÃo Conejo | http://tioconejo.net/ | Graphics\|Services | ope |
| **60981** | /organization/vacation-bnbâ¢ | Vacation BnBâ¢ | http://www.vacabnb.com | Tourism\|Travel\|Travel & Tourism | ope |
| **62884** | /organization/vã¡-de-tã¡xi | VÃ¡ de TÃ¡xi | http://www.vadetaxi.com.br | Internet\|Taxis\|Transportation | clos |
| **62885** | /organization/vã¼nder-sports-network | VÃ¼nder Sports Network | http://www.vundersports.com/ | Communities | ope |
| **63486** | /organization/weiche-tech-åè½¦ç§æ | Weiche Tech åè½¦ç§æ | http://www.weicheche.cn | NaN | ope |
| **63811** | /organization/whiteâs-holdings | Whiteâs Holdings | NaN | NaN | ope |
| **63833** | /organization/whodatâs-spaces | Whodatâs Spaces | NaN | Apps | ope |
| **65778** | /organization/zengame-ç¦æ¸¸ç§æ | ZenGame ç¦æ¸¸ç§æ | http://www.zen-game.com | Internet\|Mobile Games\|Online Gaming | clos |
| **66365** | /organization/äeron | ÄERON | http://www.aeron.hu/ | NaN | ope |
| **66366** | /organization/äasys-2 | Äasys | http://www.oasys.io/ | Consumer Electronics\|Internet of Things\|Teleco... | ope |
| **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 | ope |

68 rows × 10 columns

Thus, the companies df also contains special characters. Let's treat those as well.

In [28]:

In [28]:

```
# remove encoding from companies df
companies['permalink'] = companies.permalink.str.encode('utf-8').str.decode('ascii', 'ignore')
```

Let's now look at the companies present in the companies df but not in rounds df - ideally there should be none.

In [29]:

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

Out[29]:

| | permalink | name | homepage_url | category_list | status | country_code | state_code | region | city | founded_at |
|---|---|---|---|---|---|---|---|---|---|---|

Thus, the encoding issue seems resolved now. Let's write these (clean) dataframes into separate files so we don't have to worry about encoding problems again.

In [30]:

```
# write rounds file
rounds.to_csv("rounds_clean.csv", sep=',', index=False)

# write companies file
companies.to_csv("companies_clean.csv", sep='\t', index=False)
```

Now that we've treated the encoding problems (caused by special characters), let's complete the data cleaning process by treating missing values.

We'll read the clean csv files we created in the previous exercise.

In [31]:

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

# read the new, decoded csv files
rounds = pd.read_csv("rounds_clean.csv", encoding = "ISO-8859-1")
companies = pd.read_csv("companies_clean.csv", sep="\t", encoding = "ISO-8859-1")
```

In [32]:

```
# quickly verify that there are 66368 unique companies in both
# and that only the same 66368 are present in both files

# unqiue values
print(len(companies.permalink.unique()))
print(len(rounds.company_permalink.unique()))

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

```
66368
66368
0
```

## Missing Value Treatment

Let's now move to missing value treatment.

Let's have a look at the number of missing values in both the dataframes.

In [33]:

```
# missing values in companies df
companies.isnull().sum()
```

Out[33]:

```
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
```

In [34]:

```
# missing values in rounds df
rounds.isnull().sum()
```

Out[34]:

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

Since there are no misisng values in the permalink or company_permalink columns, let's merge the two and then work on the master dataframe.

In [35]:

```
# merging the two dfs
master = pd.merge(companies, rounds, how="inner", left_on="permalink", right_on="company_permalink"
)
master.head()
```

Out[35]:

| | permalink | name | homepage_url | category_list | status | country_code | state_code | region | city |
|---|---|---|---|---|---|---|---|---|---|
| **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 | Delav City |
| **2** | /organization/-qounter | :Qounter | http://www.qounter.com | Application Platforms\|Real Time\|Social Network... | operating | USA | DE | DE - Other | Delav City |
| **3** | /organization/-the-one-of-them-inc- | (THE) ONE of THEM,Inc. | http://oneofthem.jp | Apps\|Games\|Mobile | operating | NaN | NaN | NaN | NaN |
| **4** | /organization/0-6-com | 0-6.com | http://www.0-6.com | Curated Web | operating | CHN | 22 | Beijing | Beijin |

Since the columns company_permalink and permalink are the same, let's remove one of them.

In [36]:

```
# print column names
master.columns
```

Out[36]:

```
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')
```

In [37]:

```
# removing redundant columns
master =  master.drop(['company_permalink'], axis=1)
```

In [38]:

```
# look at columns after dropping
master.columns
```

Out[38]:

```
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')
```

Let's now look at the number of missing values in the master df.

In [39]:

```
# column-wise missing values
master.isnull().sum()
```

Out[39]:

```
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
```

Let's look at the fraction of missing values in the columns.

In [41]:

```
# summing up the missing values (column-wise) and displaying fraction of NaNs
round(100*(master.isnull().sum()/len(master.index)), 2)
```

Out[41]:

```
permalink                   0.00
name                        0.00
homepage_url                5.34
category_list               2.97
status                      0.00
country_code                7.55
state_code                  9.52
```

```
region                      8.84
city                        8.84
founded_at                 17.85
funding_round_permalink     0.00
funding_round_type          0.00
funding_round_code         72.91
funded_at                   0.00
raised_amount_usd          17.39
dtype: float64
```

Clearly, the column funding_round_code is useless (with about 73% missing values). Also, for the business objectives given, the columns homepage_url, founded_at, state_code, region and city need not be used.

Thus, let's drop these columns.

In [42]:

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

Out[42]:

| | permalink | name | category_list | status | country_code | funding_round_permalink | fundi |
|---|---|---|---|---|---|---|---|
| **0** | /organization/-fame | #fame | Media | operating | IND | /funding-round/9a01d05418af9f794eebff7ace91f638 | ventu |
| **1** | /organization/-qounter | :Qounter | Application Platforms\|Real Time\|Social Network... | operating | USA | /funding-round/22dacff496eb7acb2b901dec1dfe5633 | ventu |
| **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 | ventu |
| **4** | /organization/0-6-com | 0-6.com | Curated Web | operating | CHN | /funding-round/5727accaeaa57461bd22a9bdd945382d | ventu |

In [43]:

```
master.shape
```

Out[43]:

```
(114949, 9)
```

In [44]:

```
# summing up the missing values (column-wise) and displaying fraction of NaNs
round(100*(master.isnull().sum()/len(master.index)), 2)
```

Out[44]:

```
permalink                   0.00
name                        0.00
category_list               2.97
status                      0.00
country_code                7.55
funding_round_permalink     0.00
funding_round_type          0.00
funded_at                   0.00
raised_amount_usd          17.39
dtype: float64
```

Note that the column raised_amount_usd is an important column, since that is the number we want to analyse (compare, means, sum etc.). That needs to be carefully treated.

Also, the column country_code will be used for country-wise analysis, and category_list will be used to merge the dataframe with the main categories.

Let's first see how we can deal with missing values in raised_amount_usd.

The mean is somewhere around USD 10 million, while the median is only about USD 1m. The min and max values are also miles apart.

In general, since there is a huge spread in the funding amounts, it will be inappropriate to impute it with a metric such as median or mean. Also, since we have quite a large number of observations, it is wiser to just drop the rows.

Let's thus remove the rows having NaNs in raised_amount_usd.

In [45]:

```
# removing NaNs in raised_amount_usd
master = master[~np.isnan(master['raised_amount_usd'])]
round(100*(master.isnull().sum()/len(master.index)), 2)
```

Out[45]:

```
permalink                  0.00
name                       0.00
category_list              1.10
status                     0.00
country_code               6.16
funding_round_permalink    0.00
funding_round_type         0.00
funded_at                  0.00
raised_amount_usd          0.00
dtype: float64
```

Let's now look at the column country_code. To see the distribution of the values for categorical variables, it is best to convert them into type 'category'.

In [48]:

```
country_codes = master['country_code'].astype('category')
# displaying frequencies of each category
country_codes.value_counts()
```

Out[48]:

```
USA    62049
GBR     5019
CAN     2616
CHN     1927
IND     1649
FRA     1451
ISR     1364
ESP     1074
DEU     1042
AUS      649
RUS      588
IRL      563
SWE      560
SGP      546
NLD      532
JPN      485
ITA      483
BRA      483
CHE      437
KOR      432
CHL      432
FIN      382
DNK      314
ARG      297
BEL      293
```

```
HKG       250
TUR       196
NOR       191
BGR       190
MEX       189
         ...
KHM         2
DOM         2
MAR         2
MAF         2
KWT         2
NIC         2
ZMB         2
KAZ         2
TUN         2
SOM         1
SYC         1
SEN         1
TGO         1
QAT         1
UZB         1
PSE         1
PRY         1
OMN         1
DMA         1
BLM         1
MNE         1
MKD         1
BRB         1
LAO         1
IRN         1
HND         1
GRD         1
GGY         1
DZA         1
KNA         1
Name: country_code, Length: 134, dtype: int64
```

By far, the most number of investments have happened in American countries. We can also see the fractions.

In [49]:

```
# viewing fractions of counts of country_codes
100*(master['country_code'].value_counts()/len(master.index))
```

Out[49]:

```
USA    65.342937
GBR     5.285439
CAN     2.754873
CHN     2.029297
IND     1.736539
FRA     1.528028
ISR     1.436409
ESP     1.131014
DEU     1.097316
AUS     0.683453
RUS     0.619215
IRL     0.592887
SWE     0.589728
SGP     0.574985
NLD     0.560242
JPN     0.510747
BRA     0.508641
ITA     0.508641
CHE     0.460199
CHL     0.454933
KOR     0.454933
FIN     0.402279
DNK     0.330669
ARG     0.312767
BEL     0.308554
HKG     0.263272
TUR     0.206405
NOR     0.201139
```

```
BGR     0.200086
MEX     0.199033
          ...
ZMB     0.002106
DOM     0.002106
CIV     0.002106
TUN     0.002106
ALB     0.002106
MCO     0.002106
BAH     0.002106
KWT     0.002106
KHM     0.002106
UZB     0.001053
BLM     0.001053
QAT     0.001053
GGY     0.001053
OMN     0.001053
MKD     0.001053
SYC     0.001053
HND     0.001053
PRY     0.001053
SOM     0.001053
KNA     0.001053
TGO     0.001053
GRD     0.001053
MNE     0.001053
LAO     0.001053
SEN     0.001053
BRB     0.001053
DMA     0.001053
PSE     0.001053
DZA     0.001053
IRN     0.001053
Name: country_code, Length: 134, dtype: float64
```

Now, we can either delete the rows having country_code missing (about 6% rows), or we can impute them by USA. Since the number 6 is quite small, and we have a decent amount of data, it may be better to just remove the rows.

Note that np.isnan does not work with arrays of type 'object', it only works with native numpy type (float). Thus, you can use pd.isnull() instead.

In [50]:

```
# removing rows with missing country_codes
master = master[~pd.isnull(master['country_code'])]

# look at missing values
round(100*(master.isnull().sum()/len(master.index)), 2)
```

Out[50]:

```
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
```

Note that the fraction of missing values in the remaining dataframe has also reduced now - only 0.65% in category_list. Let's thus remove those as well.

Note Optionally, you could have simply let the missing values in the dataset and continued the analysis. There is nothing wrong with that. But in this case, since we will use that column later for merging with the 'main_categories', removing the missing values will be quite convenient (and again - we have enough data).

In [51]:

```
# removing rows with missing category_list values
master = master[~pd.isnull(master['category_list'])]
master = master[~pd.isnull(master['category_list'])]

# look at missing values
round(100*(master.isnull().sum()/len(master.index)), 2)
```

Out[51]:

```
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 [52]:

```
# writing the clean dataframe to an another file
master.to_csv("master_df.csv", sep=',', index=False)
```

In [53]:

```
# look at the master df info for number of rows etc.
master.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 88529 entries, 0 to 114947
Data columns (total 9 columns):
permalink                  88529 non-null object
name                       88528 non-null object
category_list              88529 non-null object
status                     88529 non-null object
country_code               88529 non-null object
funding_round_permalink    88529 non-null object
funding_round_type         88529 non-null object
funded_at                  88529 non-null object
raised_amount_usd          88529 non-null float64
dtypes: float64(1), object(8)
memory usage: 6.8+ MB
```

In [54]:

```
# after missing value treatment, approx 77% observations are retained
100*(len(master.index) / len(rounds.index))
```

Out[54]:

```
77.01589400516751
```

## Data Analysis

In this section, we'll conduct the three types of analysis - funding type, country analysis, and sector analysis.

## Funding Type Analysis

Let's compare the funding amounts across the funding types. Also, we need to impose the constraint that the investment amount should be between 5 and 15 million USD. We will choose the funding type such that the average investment amount falls in this range.

In [55]:

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

df = pd.read_csv("master_df.csv", sep=",", encoding="ISO-8859-1")
df = pd.read_csv("master_df.csv", sep=",", encoding="ISO-8859-1")
df.head()
```

Out[55]:

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

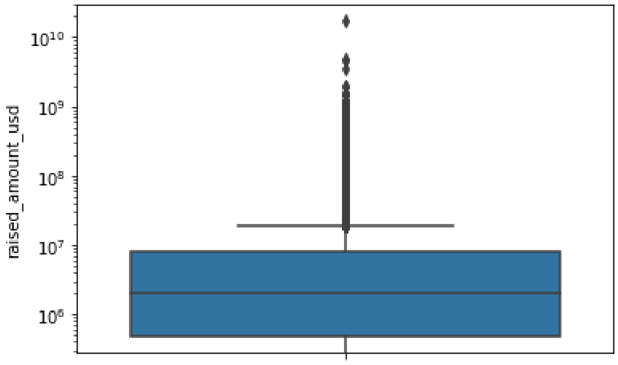
In [56]:

```
# first, let's filter the df so it only contains the four specified funding types
df = df[(df.funding_round_type == "venture") |
        (df.funding_round_type == "angel") |
        (df.funding_round_type == "seed") |
        (df.funding_round_type == "private_equity") ]
```

Now, we have to compute a representative value of the funding amount for each type of investment. We can either choose the mean or the median - let's have a look at the distribution of raised_amount_usd to get a sense of the distribution of data.

In [57]:

```
# distribution of raised_amount_usd
sns.boxplot(y=df['raised_amount_usd'])
plt.yscale('log')
plt.show()
```



In [58]:

```
# summary metrics
df['raised_amount_usd'].describe()
```

Out[58]:

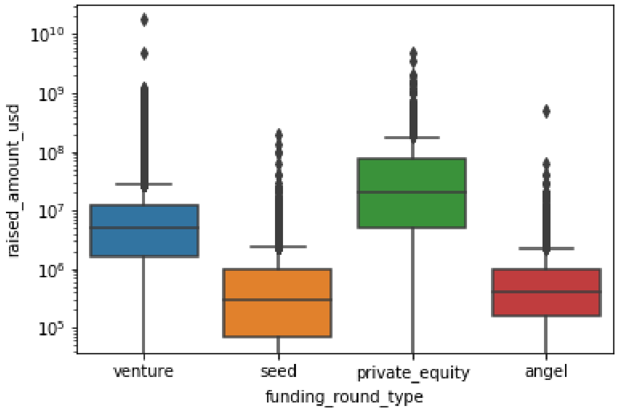
```
count    7.512400e+04
mean     9.519475e+06
std      7.792778e+07
min      0.000000e+00
25%      4.705852e+05
50%      2.000000e+06
```

```
75%      8.000000e+06
max      1.760000e+10
Name: raised_amount_usd, dtype: float64
```

Note that there's a significant difference between the mean and the median - USD 9.5m and USD 2m. Let's also compare the summary stats across the four categories.

In [59]:

```
# comparing summary stats across four categories
sns.boxplot(x='funding_round_type', y='raised_amount_usd', data=df)
plt.yscale('log')
plt.show()
```



In [60]:

```
# compare the mean and median values across categories
df.pivot_table(values='raised_amount_usd', columns='funding_round_type', aggfunc=[np.median, np.mea
n])
```

Out[60]:

| | median | | | | mean | | | |
|---|---|---|---|---|---|---|---|---|
| funding_round_type | angel | private_equity | seed | venture | angel | private_equity | seed | venture |
| raised_amount_usd | 414906.0 | 20000000.0 | 300000.0 | 5000000.0 | 971573.891136 | 7.393849e+07 | 747793.682484 | 1.172422 |

Note that there's a large difference between the mean and the median values for all four types. For type private equity, for e.g. the median is about 20m while the mean is about 70m.

Thus, the choice of the summary statistic will drastically affect the decision (of the investment type). Let's choose median, since there are quite a few extreme values pulling the mean up towards them - but they are not the most 'representative' values.

In [61]:

```
# compare the median investment amount across the types
df.groupby('funding_round_type')['raised_amount_usd'].median().sort_values(ascending=False)
```

Out[61]:

```
funding_round_type
private_equity    20000000.0
venture            5000000.0
angel               414906.0
seed                300000.0
Name: raised_amount_usd, dtype: float64
```

The median investment amount for type 'private_equity' is approx. USD 20m, which is beyond Teclov' range of 5-15m. The median of 'venture' type is about USD 5m, which is suitable for them. The average amounts of angel and seed types are lower than their range.

Thus, 'venture' type investment will be most suited to them.

## Country Analysis

Let's now compare the total investment amounts across countries. Note that we'll filter the data for only the 'venture' type investments and then compare the 'total investment' across countries

In [62]:

```
# filter the df for private equity type investments
df = df[df.funding_round_type=="venture"]

# group by country codes and compare the total funding amounts
country_wise_total = df.groupby('country_code')['raised_amount_usd'].sum().sort_values(ascending=Fa
lse)
print(country_wise_total)
```

```
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
SWE    3.145857e+09
NLD    2.903876e+09
CHE    2.801560e+09
SGP    2.793918e+09
ESP    1.827622e+09
BRA    1.785818e+09
IRL    1.669286e+09
RUS    1.570426e+09
AUS    1.319029e+09
DNK    1.228311e+09
FIN    1.043200e+09
BEL    1.030840e+09
NOR    9.536361e+08
KOR    8.919883e+08
MYS    8.830588e+08
HKG    7.812670e+08
TWN    6.239795e+08
AUT    5.833607e+08
TUR    5.590975e+08
ITA    4.882894e+08
NZL    4.483164e+08
           ...
KWT    1.400000e+07
LIE    1.309172e+07
MNE    1.220000e+07
SVN    1.201751e+07
BGR    1.130000e+07
KAZ    1.100000e+07
GRC    1.074378e+07
BAH    8.900000e+06
TTO    8.500000e+06
SVK    8.241062e+06
BGD    7.002000e+06
LBN    6.455000e+06
GGY    3.960000e+06
TUN    3.920000e+06
SEN    2.860000e+06
HRV    2.633669e+06
UGA    2.500000e+06
PER    2.469270e+06
BWA    2.250000e+06
LAO    2.100000e+06
PAN    2.100000e+06
MAR    1.600000e+06
MUS    1.500000e+06
PRI    1.441901e+06
ECU    9.658500e+05
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
```

Let's now extract the top 9 countries from country_wise_total.

In [63]:

```
# top 9 countries
top_9_countries = country_wise_total[:9]
top_9_countries
```

Out[63]:

```
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
```

Among the top 9 countries, USA, GBR and IND are the top three English speaking countries. Let's filter the dataframe so it contains only the top 3 countries.

In [64]:

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

Out[64]:

| | permalink | name | category_list | status | country_code | funding_round_permalink |
|---|---|---|---|---|---|---|
| **0** | /organization/-fame | #fame | Media | operating | IND | /funding-round/9a01d05418af9f794eebff7 |
| **7** | /organization/0xdata | H2O.ai | Analytics | operating | USA | /funding-round/3bb2ee4a2d89251a10aaa |
| **8** | /organization/0xdata | H2O.ai | Analytics | operating | USA | /funding-round/ae2a174c06517c2394aed |
| **9** | /organization/0xdata | H2O.ai | Analytics | operating | USA | /funding-round/e1cfcbe1bdf4c70277c5f29 |
| **15** | /organization/1-mainstream | 1 Mainstream | Apps\|Cable\|Distribution\|Software | acquired | USA | /funding-round/b952cbaf401f310927430c |

In [65]:

```
# filtered df has about 38800 observations
df.info()
```

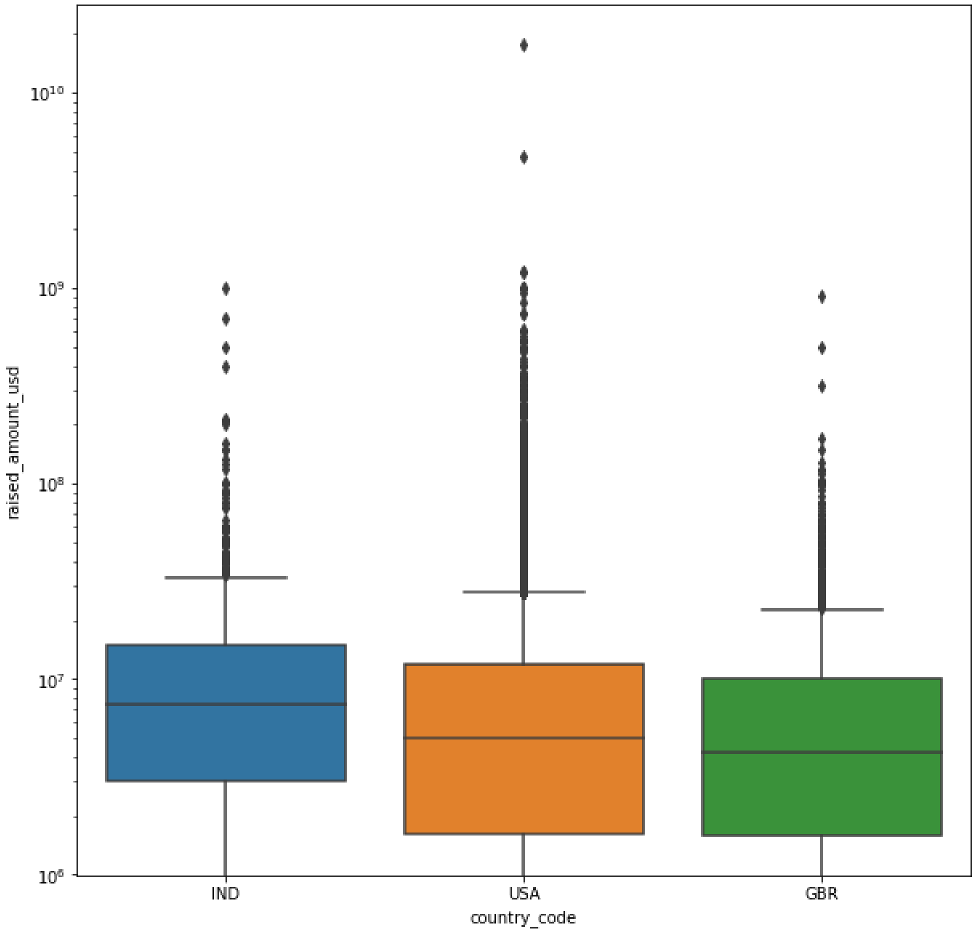
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 38803 entries, 0 to 88518
Data columns (total 9 columns):
permalink                  38803 non-null object
name                       38803 non-null object
category_list              38803 non-null object
status                     38803 non-null 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
```

One can visually analyse the distribution and the total values of funding amount.

In [66]:

```
# boxplot to see distributions of funding amount across countries
plt.figure(figsize=(10, 10))
sns.boxplot(x='country_code', y='raised_amount_usd', data=df)
plt.yscale('log')
plt.show()
```



Now, we have shortlisted the investment type (venture) and the three countries. Let's now choose the sectors.

## Sector Analysis

First, we need to extract the main sector using the column category_list. The category_list column contains values such as 'Biotechnology|Health Care' - in this, 'Biotechnology' is the 'main category' of the company, which we need to use.

Let's extract the main categories in a new column.

In [67]:

```
# extracting the main category
df.loc[:, 'main_category'] = df['category_list'].apply(lambda x: x.split("|")[0])
df.head()
```

Out[67]:

| | permalink | name | category_list | status | country_code | funding_round_permalink |
|---|---|---|---|---|---|---|
| **0** | /organization/-fame | #fame | Media | operating | IND | /funding- |

round/9a01d05418af9f794eebff7

| | permalink | name | category_list | status | country_code | funding_round_permalink |
|---|---|---|---|---|---|---|
| **7** | /organization/0xdata | H2O.ai | Analytics | operating | USA | /funding-round/3bb2ee4a2d89251a10aaa7 |
| **8** | /organization/0xdata | H2O.ai | Analytics | operating | USA | /funding-round/ae2a174c06517c2394aed4 |
| **9** | /organization/0xdata | H2O.ai | Analytics | operating | USA | /funding-round/e1cfcbe1bdf4c70277c5f29 |
| **15** | /organization/1-mainstream | 1 Mainstream | Apps\|Cable\|Distribution\|Software | acquired | USA | /funding-round/b952cbaf401f310927430c9 |

We can now drop the category_list column.

In [68]:

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

Out[68]:

| | permalink | name | status | country_code | funding_round_permalink | funding_round_ty |
|---|---|---|---|---|---|---|
| **0** | /organization/-fame | #fame | operating | IND | /funding-round/9a01d05418af9f794eebff7ace91f638 | venture |
| **7** | /organization/0xdata | H2O.ai | operating | USA | /funding-round/3bb2ee4a2d89251a10aaa735b1180e44 | venture |
| **8** | /organization/0xdata | H2O.ai | operating | USA | /funding-round/ae2a174c06517c2394aed45006322a7e | venture |
| **9** | /organization/0xdata | H2O.ai | operating | USA | /funding-round/e1cfcbe1bdf4c70277c5f29a3482f24e | venture |
| **15** | /organization/1-mainstream | 1 Mainstream | acquired | USA | /funding-round/b952cbaf401f310927430c97b68162ea | venture |

Now, we'll read the mapping.csv file and merge the main categories with its corresponding column

In [71]:

```
# read mapping file
mapping = pd.read_csv("mapping.csv", sep=",")
mapping.head(10)
```

Out[71]:

| | category_list | Automotive & Sports | Blanks | Cleantech / Semiconductors | Entertainment | Health | Manufacturing | News, Search and Messaging | Others | Social, Finance, Analytics Advertisi |
|---|---|---|---|---|---|---|---|---|---|---|
| **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 |
| **5** | Active Lifestyle | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| **6** | Ad Targeting | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **7** | Advanced | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |

|   | category_list | Automotive & Sports | Blanks | Cleantech / Semiconductors | Entertainment | Health | Manufacturing | News, Search and Messaging | Others | Social, Finance, Analytics, Advertisi |
|---|---|---|---|---|---|---|---|---|---|---|
| 7 | Materials | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 8 | Adventure Travel | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | Advertising | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

Firstly, let's get rid of the missing values since we'll not be able to merge those rows anyway.

In [72]:

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

Out[72]:

```
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 [73]:

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

Out[73]:

```
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
```

Now, since we need to merge the mapping file with the main dataframe (df), let's convert the common column to lowercase in both.

In [75]:

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

Out[75]:

|   | category_list | Automotive & Sports | Blanks | Cleantech / Semiconductors | Entertainment | Health | Manufacturing | News, Search and Messaging | Others | Social, Finance, Analytics, Advertisi |
|---|---|---|---|---|---|---|---|---|---|---|
| 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 |

| | category_list | Automotive & Sports | Blanks | Cleantech / Semiconductors | Entertainment | Health | Manufacturing | News, Search and | Others | Social, Finance, Analytics |
|---|---|---|---|---|---|---|---|---|---|---|
| **4** | accounting | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **5** | active lifestyle | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |

In [76]:

```
df.head()
```

Out[76]:

| | permalink | name | status | country_code | funding_round_permalink | funding_round_ty |
|---|---|---|---|---|---|---|
| **0** | /organization/-fame | #fame | operating | IND | /funding-round/9a01d05418af9f794eebff7ace91f638 | venture |
| **7** | /organization/0xdata | H2O.ai | operating | USA | /funding-round/3bb2ee4a2d89251a10aaa735b1180e44 | venture |
| **8** | /organization/0xdata | H2O.ai | operating | USA | /funding-round/ae2a174c06517c2394aed45006322a7e | venture |
| **9** | /organization/0xdata | H2O.ai | operating | USA | /funding-round/e1cfcbe1bdf4c70277c5f29a3482f24e | venture |
| **15** | /organization/1-mainstream | 1 Mainstream | acquired | USA | /funding-round/b952cbaf401f310927430c97b68162ea | venture |

Let's have a look at the category_list column of the mapping file. These values will be used to merge with the main df.

In [77]:

```
mapping['category_list']
```

Out[77]:

```
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
                           ...
658                              usability
659                 user experience design
660                         user interface
661                              utilities
662                vending and concessions
663                        venture capital
```

```
664                          veteri0ry
665                              video
666                    video conferencing
667                        video games
668                     video on demand
669                     video streaming
670                   virtual workforces
671                               voip
672                     waste ma0gement
673                              watch
674                              water
675                   water purification
676                          wearables
677                       web browsers
678                         web design
679                    web development
680                        web hosting
681                          web tools
682                           weddings
683                          wholesale
684                    wine and spirits
685                           wireless
686                              women
687                       young adults
Name: category_list, Length: 687, dtype: object
```

To be able to merge all the main_category values with the mapping file's category_list column, all the values in the main_category column should be present in the category_list column of the mapping file.

Let's see if this is true.

In [78]:

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

Out[78]:

| | permalink | name | status | country_code | funding_round_permalink | funding |
|---|---|---|---|---|---|---|
| **7** | /organization/0xdata | H2O.ai | operating | USA | /funding-round/3bb2ee4a2d89251a10aaa735b1180e44 | venture |
| **8** | /organization/0xdata | H2O.ai | operating | USA | /funding-round/ae2a174c06517c2394aed45006322a7e | venture |
| **9** | /organization/0xdata | H2O.ai | operating | USA | /funding-round/e1cfcbe1bdf4c70277c5f29a3482f24e | venture |
| **47** | /organization/100plus | 100Plus | acquired | USA | /funding-round/b5facb0d9dea2f0352b5834892c88c53 | venture |
| **136** | /organization/1world-online | 1World Online | operating | USA | /funding-round/32936e588a134502712877150198a0b3 | venture |
| **137** | /organization/1world-online | 1World Online | operating | USA | /funding-round/4e30bd5c85d8163239a3479ec979647a | venture |
| **138** | /organization/1world-online | 1World Online | operating | USA | /funding-round/a349bfd7a8d48cfc8b9fdb79480dea7f | venture |
| **187** | /organization/24-7-card | 24/7 Card | closed | USA | /funding-round/0c38194ff2035185c96155dfad18f3bd | venture |
| **590** | /organization/6th-wave-innovations-corporation | 6th Wave Innovations Corporation | operating | USA | /funding-round/75d128ac40f9e541a1a11786a47c2952 | venture |
| **597** | /organization/7-billion-people | 7 Billion People | closed | USA | /funding-round/58959ed2be7b14abd6beeb20c9eb17ca | venture |
| **629** | /organization/7park-data | 7Park Data | operating | USA | /funding-round/64ddc56c450048911859956eade79cfa | venture |
| **723** | /organization/9lenses | 9Lenses | operating | USA | /funding- | venture |

| | permalink | name | status | country_code | funding_round_permalink | funding |
|---|---|---|---|---|---|---|
| **723** | /organization/9lenses | 9Lenses | operating | USA | round/b27a23a29eb8207f78b60e1f64332832 | venture |
| **724** | /organization/9lenses | 9Lenses | operating | USA | /funding-round/b58dcac20e96077aa9f6adf595f3b0fd | venture |
| **725** | /organization/9lenses | 9Lenses | operating | USA | /funding-round/ec22e2c9cac79e78da4c1325db5759d0 | venture |
| **753** | /organization/a-little-world | A LITTLE WORLD | operating | IND | /funding-round/18d98f82ed392b1609975b81f3e8b3fb | venture |
| **803** | /organization/abacast-inc | Abacast | acquired | USA | /funding-round/4abfb5502126b436ad34f8454f880cdc | venture |
| **932** | /organization/aboutone | AboutOne | operating | USA | /funding-round/3ca6d41f3553a6f8d62209874e87d83e | venture |
| **936** | /organization/aboutone | AboutOne | operating | USA | /funding-round/ac4512dff659967d6aa0237f8c5cd6e5 | venture |
| **948** | /organization/abra | Abra | operating | USA | /funding-round/cd7d853628a80a27c1aadcff92826550 | venture |
| **956** | /organization/absolutdata | AbsolutData | operating | USA | /funding-round/1a448e0b75b346e473edb8f7e44a4ca3 | venture |
| **1010** | /organization/acadiasoft | AcadiaSoft | operating | USA | /funding-round/44a3010dd331ee5f89feefc70925f3ff | venture |
| **1011** | /organization/acadiasoft | AcadiaSoft | operating | USA | /funding-round/94de8f62876c3ae085029fdd15cd5650 | venture |
| **1045** | /organization/accelerated-vision-group | Accelerated Vision Group | operating | USA | /funding-round/4c703de9c312a0cc14f8307fb0383096 | venture |
| **1046** | /organization/accelerated-vision-group | Accelerated Vision Group | operating | USA | /funding-round/efc17c623b56a27ee73dca0f0155def3 | venture |
| **1072** | /organization/accelops | AccelOps | operating | USA | /funding-round/76abbf3b54bd6ad3abc7b503adecfb42 | venture |
| **1073** | /organization/accelops | AccelOps | operating | USA | /funding-round/c521b592ec7c69178447aa7242d90995 | venture |
| **1087** | /organization/accept-software | Accept Software | acquired | USA | /funding-round/03a2db742f933b0532e2f433f39a2b21 | venture |
| **1088** | /organization/accept-software | Accept Software | acquired | USA | /funding-round/2d1ef4ff3a49c29fa18f4362d394229d | venture |
| **1089** | /organization/accept-software | Accept Software | acquired | USA | /funding-round/330381d6fb6a2be7e13e18c5e89dad7b | venture |
| **1090** | /organization/accept-software | Accept Software | acquired | USA | /funding-round/703bc67685438d1344a7a3b2d432518f | venture |
| **...** | ... | ... | ... | ... | ... | ... |
| **87539** | /organization/zebit | Zebit | closed | GBR | /funding-round/50fe985400785190c9fe4ba317aa8223 | venture |
| **87558** | /organization/zecter | Zecter | acquired | USA | /funding-round/47c308302fd912249b19a00e1fdd1cc2 | venture |
| **87560** | /organization/zecter | Zecter | acquired | USA | /funding-round/d7f4567eae495b093fccfaefa953a054 | venture |
| **87689** | /organization/zenpayroll | Gusto | operating | USA | /funding-round/3cc998e7270da32bc897f7e2381a0931 | venture |
| **87690** | /organization/zenpayroll | Gusto | operating | USA | /funding-round/9ec1c859afcff414a15853077f2b3db7 | venture |
| **87719** | /organization/zentrick | Zentrick | operating | USA | /funding-round/ae0c94ecdac4d40a26415b003c730e06 | venture |
| **87734** | /organization/zeo | Zeo | closed | USA | /funding-round/45b22e40f0c237f2867bab4ef34ae1c0 | venture |

|  | permalink | name | status | country_code | funding_round_permalink | funding |
|---|---|---|---|---|---|---|
| **87735** | /organization/zeo | Zeo | closed | USA | /funding-round/9b567c9830a4501758d99ba6529e8ac0 | venture |
| **87744** | /organization/zephyr-health | Zephyr Health | operating | USA | /funding-round/4edc7d9233a1a58643bff77b87332038 | venture |
| **87745** | /organization/zephyr-health | Zephyr Health | operating | USA | /funding-round/6bbf6cac4cf2565afa4cf8625dadb834 | venture |
| **87746** | /organization/zephyr-health | Zephyr Health | operating | USA | /funding-round/734a64f4ffd197a3539c9bc6ff7af9b5 | venture |
| **87759** | /organization/zeptor | Zeptor | operating | USA | /funding-round/eb853f91865c697754fb1f94e9b795b3 | venture |
| **87823** | /organization/zestfinance | ZestFinance | operating | USA | /funding-round/33cd4d4fa967d1fec848e082260e23a9 | venture |
| **87824** | /organization/zestfinance | ZestFinance | operating | USA | /funding-round/3814934c1697ee07d2dd53b6fcc32cbc | venture |
| **87826** | /organization/zestfinance | ZestFinance | operating | USA | /funding-round/ade89e5c3e55d6f2ec0ddcd20ee085eb | venture |
| **87827** | /organization/zestfinance | ZestFinance | operating | USA | /funding-round/d0a50c9928ba4b9fcb94120c6bc22bd8 | venture |
| **87923** | /organization/ziegler | Ziegler | ipo | USA | /funding-round/c4933a2fe9a4d9b0a2ec19fb4cfe1083 | venture |
| **88084** | /organization/ziprealty | ZipRealty | acquired | USA | /funding-round/1b57e619d3474963a31605197172cb06 | venture |
| **88193** | /organization/zonarsystems | Zonar Systems | operating | USA | /funding-round/f0126dbea5d6075d8d4a1c2d106d9eca | venture |
| **88231** | /organization/zoomdata | Zoomdata | operating | USA | /funding-round/0095bec234eec6448bc49570045bd89b | venture |
| **88232** | /organization/zoomdata | Zoomdata | operating | USA | /funding-round/639c4b4cae7be6e0746b0fbe07e78bc0 | venture |
| **88243** | /organization/zoominfo | ZoomInfo | operating | USA | /funding-round/8cdc750a5e5793323af50ca23dee162e | venture |
| **88255** | /organization/zoomsafer | ZoomSafer | acquired | USA | /funding-round/9fa41d0600b98191504e0113859783b4 | venture |
| **88256** | /organization/zoomsafer | ZoomSafer | acquired | USA | /funding-round/f35c1ec421d78b31ae2a9e2c893e86c6 | venture |
| **88269** | /organization/zoopla | Zoopla | ipo | GBR | /funding-round/0ec759962079a8997eb1632d6c1a769b | venture |
| **88270** | /organization/zoopla | Zoopla | ipo | GBR | /funding-round/98da1f441a55c9a9629a256828923e38 | venture |
| **88291** | /organization/zopa | Zopa | operating | GBR | /funding-round/2a55d435c3433d8f903526c050c19361 | venture |
| **88292** | /organization/zopa | Zopa | operating | GBR | /funding-round/4b0740cb83da8d2af9d221e5455f8923 | venture |
| **88293** | /organization/zopa | Zopa | operating | GBR | /funding-round/54dbfbd899caf7d1d4b2b7676065f303 | venture |
| **88294** | /organization/zopa | Zopa | operating | GBR | /funding-round/720b9f244c1f4d4fed63361d3bb0aa22 | venture |

2616 rows × 9 columns

Notice that values such as 'analytics', 'business analytics', 'finance', 'nanatechnology' etc. are not present in the mapping file.

Let's have a look at the values which are present in the mapping file but not in the main dataframe df.

```
In [80]:
```

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

Out[80]:

| | category_list | Automotive & Sports | Blanks | Cleantech / Semiconductors | Entertainment | Health | Manufacturing | News, Search and Messaging | Others | Social, Financ Analyti Adverti |
|---|---|---|---|---|---|---|---|---|---|---|
| **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 |
| **59** | big data a0lytics | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **64** | biomass power generation | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| **69** | boating industry | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **77** | building owners | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| **79** | business a0lytics | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **85** | business travelers | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| **89** | can0bis | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| **91** | career ma0gement | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **94** | casual games | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| **97** | charities | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **103** | chi0 internet | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| **104** | civil engineers | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| **110** | cloud-based music | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| **114** | cloud ma0gement | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| **119** | collectibles | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| **145** | contact ma0gement | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **198** | digital rights ma0gement | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| **199** | digital sig0ge | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| **200** | direct advertising | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **210** | document ma0gement | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| | educatio0l | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |

|  | category_list | Automotive & Sports | Blanks | Cleantech / Semiconductors | Entertainment | Health | Manufacturing | News, Search and Messaging | Others | Social, Financ Analyti Adverti |
|---|---|---|---|---|---|---|---|---|---|---|
| 223 | games | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 224 | edutainment | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 232 | email newsletters | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 240 | energy ma0gement | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 241 | energy storage | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 605 | social business | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 610 | social games | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 612 | social media ma0gement | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 614 | social media platforms | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 616 | social news | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 617 | social recruiting | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 618 | social television | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 619 | social travel | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 625 | speech recognition | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 630 | stock exchanges | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 632 | subscription service | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 633 | supply chain ma0gement | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 634 | surveys | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 637 | task ma0gement | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 638 | taxis | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 639 | tea | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 647 | tourism | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 655 | universities | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 656 | university students | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 657 | unmanned air systems | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 658 | usability | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 659 | user experience design | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 662 | vending and concessions | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 664 | veteri0ry | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 669 | video streaming | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 670 | virtual | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |

|  | category_list | Automotive & Sports | Blanks | Cleantech / Semiconductors | Entertainment | Health | Manufacturing | News, Search and Messaging | Others | Social, Financ Analyti Adverti |
|---|---|---|---|---|---|---|---|---|---|---|
|  | workforces |  |  |  |  |  |  |  |  |  |
| 672 | waste management | 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

In [81]:

```
# 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
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
                      ...
658                             usability
659                user experience design
660                        user interface
661                             utilities
662               vending and concessions
663                       venture capital
664                            veterinary
665                                 video
666                    video conferencing
667                           video games
668                       video on demand
669                       video streaming
670                    virtual workforces
671                                  voip
672                      waste management
673                                 watch
674                                 water
675                    water purification
676                             wearables
677                          web browsers
678                            web design
679                       web development
680                           web hosting
681                             web tools
682                              weddings
683                             wholesale
```

```
684                     wine and spirits
685                             wireless
686                                women
687                         young adults
Name: category_list, Length: 687, dtype: object
```

This looks fine now. Let's now merge the two dataframes.

In [84]:

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

Out[84]:

| | permalink | name | status | country_code | funding_round_permalink | funding_round_type | func |
|---|---|---|---|---|---|---|---|
| **0** | /organization/-fame | #fame | operating | IND | /funding-round/9a01d05418af9f794eebff7ace91f638 | venture | 05-0<br>2015 |
| **1** | /organization/90min | 90min | operating | GBR | /funding-round/21a2cbf6f2fb2a1c2a61e04bf930dfe6 | venture | 06-1<br>2015 |
| **2** | /organization/90min | 90min | operating | GBR | /funding-round/bd626ed022f5c66574b1afe234f3c90d | venture | 07-0<br>2013 |
| **3** | /organization/90min | 90min | operating | GBR | /funding-round/fd4b15e8c97ee2ffc0acccdbe1a98810 | venture | 26-0<br>2014 |
| **4** | /organization/all-def-digital | All Def Digital | operating | USA | /funding-round/452a2342fe720285c3b92e9bd927d9ba | venture | 06-0<br>2014 |

In [85]:

```
# 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[85]:

| | permalink | name | status | country_code | funding_round_permalink | funding_round_type | func |
|---|---|---|---|---|---|---|---|
| **0** | /organization/-fame | #fame | operating | IND | /funding-round/9a01d05418af9f794eebff7ace91f638 | venture | 05-0<br>2015 |
| **1** | /organization/90min | 90min | operating | GBR | /funding-round/21a2cbf6f2fb2a1c2a61e04bf930dfe6 | venture | 06-1<br>2015 |
| **2** | /organization/90min | 90min | operating | GBR | /funding-round/bd626ed022f5c66574b1afe234f3c90d | venture | 07-0<br>2013 |
| **3** | /organization/90min | 90min | operating | GBR | /funding-round/fd4b15e8c97ee2ffc0acccdbe1a98810 | venture | 26-0<br>2014 |
| **4** | /organization/all-def-digital | All Def Digital | operating | USA | /funding-round/452a2342fe720285c3b92e9bd927d9ba | venture | 06-0<br>2014 |

In [86]:

```
# look at the column types and names
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
```

### Converting the 'wide' dataframe to 'long'

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 [87]:

```
# store the value and id variables in two separate arrays

# store the value variables in one Series
value_vars = df.columns[9:18]

# take the setdiff() to get the rest of the variables
id_vars = np.setdiff1d(df.columns, value_vars)

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

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

['country_code' 'funded_at' 'funding_round_permalink' 'funding_round_type'
 'main_category' 'name' 'permalink' 'raised_amount_usd' 'status']
```

In [88]:

```
# convert into long
long_df = pd.melt(df,
        id_vars=list(id_vars),
        value_vars=list(value_vars))

long_df.head()
```

Out[88]:

|  | country_code | funded_at | funding_round_permalink | funding_round_type | main_category | name | perma |
|---|---|---|---|---|---|---|---|

|  | country_code | funded_at | funding_round_permalink | funding_round_type | main_category | name | perma |
|---|---|---|---|---|---|---|---|
| 0 | IND | 05-01-2015 | /funding-round/9a01d05418af9f794ccbff7acc91f638 | venture | media | #fame | /organ |
| 1 | GBR | 06-10-2015 | /funding-round/21a2cbf6f2fb2a1c2a61e04bf930dfe6 | venture | media | 90min | /organi |
| 2 | GBR | 07-05-2013 | /funding-round/bd626ed022f5c66574b1afe234f3c90d | venture | media | 90min | /organi |
| 3 | GBR | 26-03-2014 | /funding-round/fd4b15e8c97ee2ffc0acccdbe1a98810 | venture | media | 90min | /organi |
| 4 | USA | 06-08-2014 | /funding-round/452a2342fe720285c3b92e9bd927d9ba | venture | media | All Def Digital | /organi def-dig |

In [90]:

```
# remove rows having value=0
long_df = long_df[long_df['value']==1]
long_df = long_df.drop('value', axis=1)
```

In [91]:

```
# look at the new df
long_df.head()
len(long_df)
```

Out[91]:

```
38788
```

In [98]:

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

In [99]:

```
# info
long_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 38788 entries, 25828 to 349075
Data columns (total 10 columns):
country_code               38788 non-null object
funded_at                  38788 non-null object
funding_round_permalink    38788 non-null object
funding_round_type         38788 non-null object
main_category              38788 non-null object
name                       38788 non-null object
permalink                  38788 non-null object
raised_amount_usd          38788 non-null float64
status                     38788 non-null object
sector                     38788 non-null object
dtypes: float64(1), object(9)
memory usage: 3.3+ 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 and the amount of investment in the three countries.

In [100]:

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

# first, let's also filter for investment range between 5 and 15m
df = long_df[(long_df['raised_amount_usd'] >= 5000000) & (long_df['raised_amount_usd'] <= 15000000)
]
```

In [101]:

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

Out[101]:

| | | count | sum |
|---|---|---|---|
| **country_code** | **sector** | | |
| **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 |
| **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 |
| **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 |

This will be much more easy to understand using a plot.

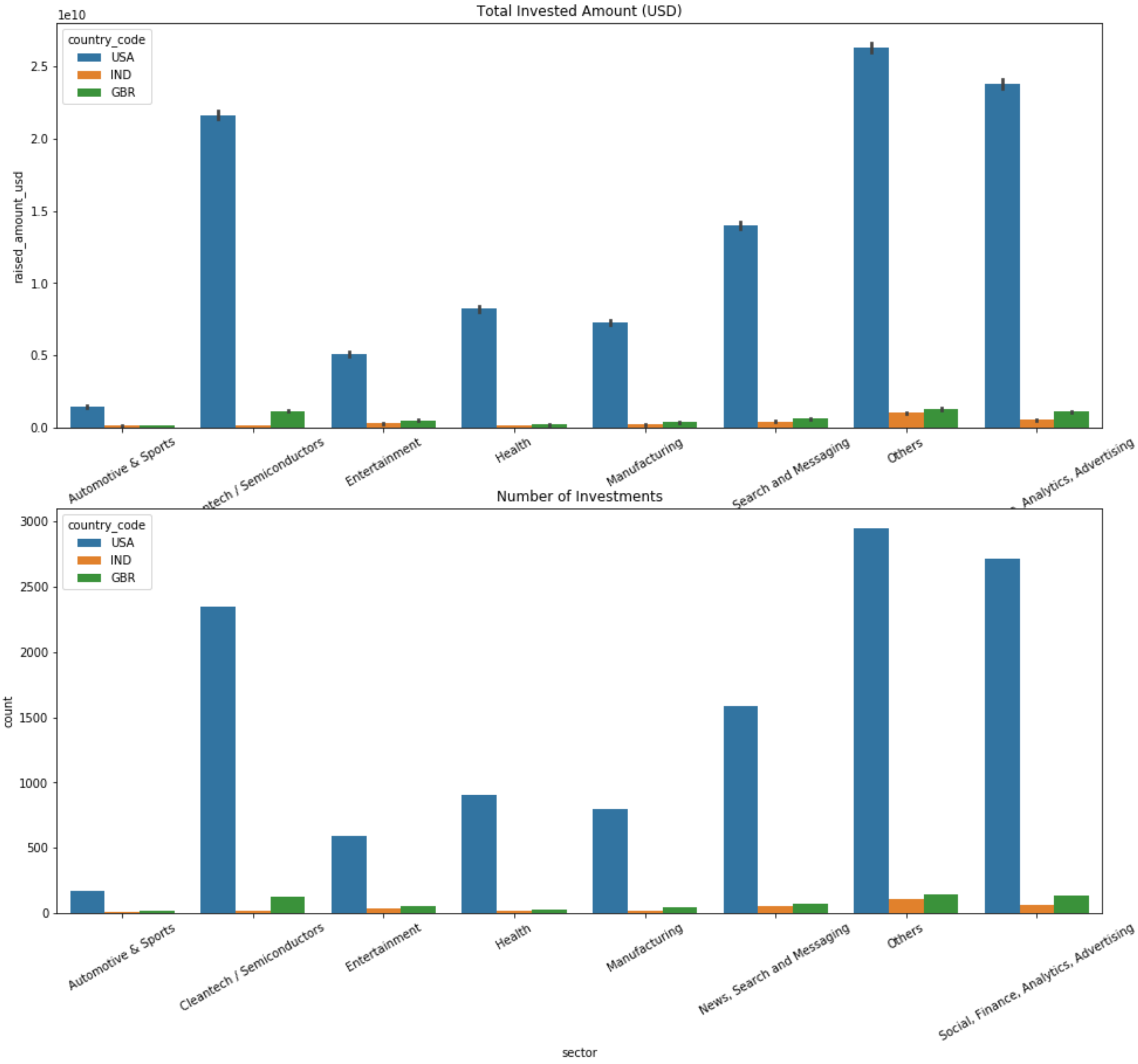
In [102]:

```
# plotting sector-wise count and sum of investments in the three countries
plt.figure(figsize=(16, 14))

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)')

plt.subplot(2, 1, 2)
q = sns.countplot(x='sector', hue='country_code', data=df)
q.set_xticklabels(q.get_xticklabels(),rotation=30)
plt.title('Number of Investments')


plt.show()
```



Thus, the top country in terms of the number of investments (and the total amount invested) is the USA. The sectors 'Others', 'Social, Finance, Analytics and Advertising' and 'Cleantech/Semiconductors' are the most heavily invested ones.

In case you don't want to consider 'Others' as a sector, 'News, Search and Messaging' is the next best sector.