# Data Cleaning

In [116]:

```
import pandas as pd
import numpy as np
df_comp=pd.read_csv('C:\\Users\\lenovo\\Desktop\\Spark Foundation EDA\\compani
es.csv',encoding= "ISO-8859-1")
df_map=pd.read_csv('C:\\Users\\lenovo\\Desktop\\Spark Foundation EDA\\mapping.
csv',encoding= "ISO-8859-1")
df_rou2=pd.read_csv('C:\\Users\\lenovo\\Desktop\\Spark Foundation EDA\\rounds
2.csv',encoding = "ISO-8859-1")
```

Taking a look at the companies data set

In [117]:

```
df_comp.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 66368 entries, 0 to 66367
Data columns (total 11 columns):
Unnamed: 0       66368 non-null int64
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: int64(1), object(10)
memory usage: 5.6+ MB
```

In [118]:
```
df_comp.head()
```

Out[118]:

|   | Unnamed: 0 | permalink | name | homepage_url | category_list | statu |
|---|---|---|---|---|---|---|
| 0 | 0 | /Organization/-Fame | #fame | http://livfame.com | Media | operatir |
| 1 | 1 | /Organization/-Qounter | :Qounter | http://www.qounter.com | Application Platforms\|Real Time\|Social Network... | operatir |
| 2 | 2 | /Organization/-The-One-Of-Them-Inc- | (THE) ONE of THEM,Inc. | http://oneofthem.jp | Apps\|Games\|Mobile | operatir |
| 3 | 3 | /Organization/0-6-Com | 0-6.com | http://www.0-6.com | Curated Web | operatir |
| 4 | 4 | /Organization/004-Technologies | 004 Technologies | http://004gmbh.de/en/004-interact | Software | operatir |

Removing the first row column

In [119]:
```
df_comp.drop('Unnamed: 0',axis=1, inplace=True)
```

In [120]:
```
df_comp.head(5)
```

Out[120]:

|   | permalink | name | homepage_url | category_list | status | country_ |
|---|---|---|---|---|---|---|
| 0 | /Organization/-Fame | #fame | http://livfame.com | Media | operating |  |
| 1 | /Organization/-Qounter | :Qounter | http://www.qounter.com | Application Platforms\|Real Time\|Social Network... | operating |  |
| 2 | /Organization/-The-One-Of-Them-Inc- | (THE) ONE of THEM,Inc. | http://oneofthem.jp | Apps\|Games\|Mobile | operating |  |
| 3 | /Organization/0-6-Com | 0-6.com | http://www.0-6.com | Curated Web | operating |  |
| 4 | /Organization/004-Technologies | 004 Technologies | http://004gmbh.de/en/004-interact | Software | operating |  |

In [121]:

```
df_rou2.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 [122]:

```
df_rou2.head()
```

Out[122]:

| | company_permalink | funding_round_permalink | funding_round_type | funding_ro |
|---|---|---|---|---|
| 0 | /organization/-fame | /funding-round/9a01d05418af9f794eebff7ace91f638 | venture | |
| 1 | /ORGANIZATION/-QOUNTER | /funding-round/22dacff496eb7acb2b901dec1dfe5633 | venture | |
| 2 | /organization/-qounter | /funding-round/b44fbb94153f6cdef13083530bb48030 | seed | |
| 3 | /ORGANIZATION/-THE-ONE-OF-THEM-INC- | /funding-round/650b8f704416801069bb178a1418776b | venture | |
| 4 | /organization/0-6-com | /funding-round/5727accaeaa57461bd22a9bdd945382d | venture | |

Making the Primary keys uniform in both the data sets.

In [123]:

```
df_comp['permalink']=df_comp['permalink'].str.lower()
df_rou2['company_permalink']=df_rou2['company_permalink'].str.lower()
```

Checking that the meta data of all the companies in round 2 is available with us.

In [124]:

```
len(df_comp['permalink'])
```

Out[124]:

```
66368
```

In [125]:

```
len(df_rou2['company_permalink'])
```

Out[125]:

```
114949
```

In [126]:

```
len(df_rou2['company_permalink'].unique())
```

Out[126]:

```
66370
```

In [127]:
```
df_rou2.loc[~df_rou2['company_permalink'].isin(df_comp['permalink'])]
```

Out[127]:

| | company_permalink | funding_round_permalink | funding_round_type |
|---|---|---|---|
| 77 | /organization/10â°north | /funding-round/b41ff7de932f8b6e5bbeed3966c0ed6a | equity_crowdfunding |
| 729 | /organization/51wofang-æ□ å¿§æ□□æ□¿ | /funding-round/346b9180d276a74e0fbb2825e66c6f5b | venture |
| 2670 | /organization/adslinkedâ□¢ | /funding-round/449ae54bb63c768c232955ca6911dee4 | seed |
| 3166 | /organization/aesthetic-everythingâ®-social-ne... | /funding-round/62593455f1a69857ed05d5734cc04132 | equity_crowdfunding |
| 3291 | /organization/affluent-attachã©-club-2 | /funding-round/626678bdf1654bc4df9b1b34647a4df1 | seed |
| ... | ... | ... | ... |
| 110545 | /organization/whodatâ□□s-spaces | /funding-round/d5d6db3d1e6c54d71a63b3aa0c9278e6 | seed |
| 113839 | /organization/zengame-ç¦□æ¸¸ç§□æ□□ | /funding-round/6ba28fb4f3eadf5a9c6c81bc5dde6cdf | seed |
| 114946 | /organization/ã□eron | /funding-round/59f4dce44723b794f21ded3daed6e4fe | venture |
| 114947 | /organization/ã□asys-2 | /funding-round/35f09d0794651719b02bbfd859ba9ff5 | seed |
| 114948 | /organization/ä°novatiff-reklam-ve-tanä±tä±m-h... | /funding-round/af942869878d2cd788ef5189b435ebc4 | grant |

74 rows × 6 columns

This is a problem we usally face due to problem while encoding and decoding data.

Since we want the meta data of all the companies performing a left join to retain all the data of the company data set.

In [128]:
```
master_frame=pd.merge(df_comp, df_rou2, how='left', left_on='permalink', right
_on='company_permalink')
```

In [129]:

```
master_frame.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 114943 entries, 0 to 114942
Data columns (total 16 columns):
permalink                  114943 non-null object
name                       114942 non-null object
homepage_url               108809 non-null object
category_list              111535 non-null object
status                     114943 non-null object
country_code               106271 non-null object
state_code                 104003 non-null object
region                     104782 non-null object
city                       104785 non-null object
founded_at                 94422 non-null object
company_permalink          114875 non-null object
funding_round_permalink    114875 non-null object
funding_round_type         114875 non-null object
funding_round_code         31132 non-null object
funded_at                  114875 non-null object
raised_amount_usd          94915 non-null float64
dtypes: float64(1), object(15)
memory usage: 14.9+ MB
```

In [130]:

```
master_dataframe.head()
```

Out[130]:

| | permalink | name | homepage_url | category_list | status | country_code | s |
|---|---|---|---|---|---|---|---|
| 0 | /organization/-fame | #fame | http://livfame.com | Media | operating | IND | |
| 1 | /organization/-qounter | :Qounter | http://www.qounter.com | Application Platforms\|Real Time\|Social Network... | operating | USA | |
| 2 | /organization/-qounter | :Qounter | http://www.qounter.com | Application Platforms\|Real Time\|Social Network... | operating | USA | |
| 3 | /organization/-the-one-of-them-inc- | (THE) ONE of THEM,Inc. | http://oneofthem.jp | Apps\|Games\|Mobile | operating | NaN | |
| 4 | /organization/0-6-com | 0-6.com | http://www.0-6.com | Curated Web | operating | CHN | |

```
In [131]: master_frame.isnull().sum()
```

```
Out[131]: permalink                   0
          name                        1
          homepage_url             6134
          category_list            3408
          status                      0
          country_code             8672
          state_code              10940
          region                  10161
          city                    10158
          founded_at              20521
          company_permalink          68
          funding_round_permalink    68
          funding_round_type         68
          funding_round_code      83811
          funded_at                  68
          raised_amount_usd       20028
          dtype: int64
```

Since we have a lot of misisng values

```
In [132]: master_frame.isnull().sum()*100/len(master_frame['permalink'])
```

```
Out[132]: permalink                   0.000000
          name                        0.000870
          homepage_url                5.336558
          category_list               2.964948
          status                      0.000000
          country_code                7.544609
          state_code                  9.517761
          region                      8.840034
          city                        8.837424
          founded_at                 17.853197
          company_permalink           0.059160
          funding_round_permalink     0.059160
          funding_round_type          0.059160
          funding_round_code         72.915271
          funded_at                   0.059160
          raised_amount_usd          17.424289
          dtype: float64
```

All these rows would not help in our Analysis.

```
In [133]: master_frame.drop('homepage_url',axis=1,inplace=True)
          master_frame.drop('funding_round_code',axis=1,inplace=True)
          master_frame.drop('founded_at',axis=1,inplace=True)
          master_frame.drop('state_code',axis=1,inplace=True)
```

In [134]:
```
master_frame.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 114943 entries, 0 to 114942
Data columns (total 12 columns):
permalink                  114943 non-null object
name                       114942 non-null object
category_list              111535 non-null object
status                     114943 non-null object
country_code               106271 non-null object
region                     104782 non-null object
city                       104785 non-null object
company_permalink          114875 non-null object
funding_round_permalink    114875 non-null object
funding_round_type         114875 non-null object
funded_at                  114875 non-null object
raised_amount_usd          94915 non-null float64
dtypes: float64(1), object(11)
memory usage: 11.4+ MB
```

The column raised_amount_usd of atmost significance to us.

In [136]:
```
from scipy.stats import kurtosis
master_frame['raised_amount_usd'].kurtosis()
```

Out[136]:
```
19222.12972387389
```

This is as far from a normal distribution as anything could be. Since teh values only account for 17% of the value, rather than replacing with the mean let's just drop the values.

In [137]:
```
master_frame.drop(master_frame[master_frame['raised_amount_usd'].isnull()].index, inplace = True)
```

In [138]:
```
master_frame.isnull().sum()
```

Out[138]:
```
permalink                     0
name                          1
category_list              1038
status                        0
country_code               5830
region                     7027
city                       7024
company_permalink             0
funding_round_permalink       0
funding_round_type            0
funded_at                     0
raised_amount_usd             0
dtype: int64
```

# Investment Type Analysis

In [139]:
```
master_frame['funding_round_type'].unique()
```

Out[139]:
```
array(['venture', 'seed', 'undisclosed', 'convertible_note',
       'private_equity', 'debt_financing', 'angel', 'grant',
       'equity_crowdfunding', 'post_ipo_equity', 'post_ipo_debt',
       'product_crowdfunding', 'secondary_market',
       'non_equity_assistance'], dtype=object)
```

In [140]:
```
master_frame.groupby('funding_round_type').mean().sort_values(by='raised_amount_usd',ascending=False)
```

Out[140]:

| | raised_amount_usd |
|---|---|
| **funding_round_type** | |
| post_ipo_debt | 1.687046e+08 |
| post_ipo_equity | 8.218249e+07 |
| secondary_market | 7.964963e+07 |
| private_equity | 7.334146e+07 |
| undisclosed | 1.925276e+07 |
| debt_financing | 1.704353e+07 |
| venture | 1.174943e+07 |
| grant | 4.312660e+06 |
| convertible_note | 1.457327e+06 |
| product_crowdfunding | 1.363131e+06 |
| angel | 9.588918e+05 |
| seed | 7.198925e+05 |
| equity_crowdfunding | 5.391133e+05 |
| non_equity_assistance | 4.112031e+05 |

Since The money that is supposed to be invested in should be between 5million-15million USD, the best appropriate funding type is venture. Hypotheisis testing wont work since we dont have a normal distribution.

In [141]:
```
df_venture= master_frame[master_frame["funding_round_type"]=="venture"]
```

In [142]:
```
df_venture.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 50219 entries, 0 to 114935
Data columns (total 12 columns):
permalink                  50219 non-null object
name                       50219 non-null object
category_list              49719 non-null object
status                     50219 non-null object
country_code               48105 non-null object
region                     47509 non-null object
city                       47509 non-null object
company_permalink          50219 non-null object
funding_round_permalink    50219 non-null object
funding_round_type         50219 non-null object
funded_at                  50219 non-null object
raised_amount_usd          50219 non-null float64
dtypes: float64(1), object(11)
memory usage: 5.0+ MB
```

In [143]:
```
df_venture.head()
```

Out[143]:

| | permalink | name | category_list | status | country_code | region | |
|---|---|---|---|---|---|---|---|
| 0 | /organization/-fame | #fame | Media | operating | IND | Mumbai | Mur |
| 3 | /organization/-the-one-of-them-inc- | (THE) ONE of THEM,Inc. | Apps\|Games\|Mobile | operating | NaN | NaN | |
| 4 | /organization/0-6-com | 0-6.com | Curated Web | operating | CHN | Beijing | Be |
| 8 | /organization/0ndine-biomedical-inc | Ondine Biomedical Inc. | Biotechnology | operating | CAN | Vancouver | Vanco |
| 10 | /organization/0xdata | H2O.ai | Analytics | operating | USA | SF Bay Area | Mour |

# Country

WE'll decide what Country the investment is to made based on the past capital invsetment trend of the Country.

In [144]:
```
df_venture.groupby('country_code').sum().sort_values(by='raised_amount_usd', ascending=False)
```

Out[144]:

| | raised_amount_usd |
|---|---|
| **country_code** | |
| USA | 4.225108e+11 |
| CHN | 3.983542e+10 |
| GBR | 2.024563e+10 |
| IND | 1.439186e+10 |
| CAN | 9.583332e+09 |
| ... | ... |
| MCO | 6.570000e+05 |
| SAU | 5.000000e+05 |
| CMR | 3.595610e+05 |
| GTM | 3.000000e+05 |
| MMR | 2.000000e+05 |

97 rows × 1 columns

We can see that the top three countries with the maximum invesment in the past.

In [145]:
```
df_country=pd.DataFrame(master_frame[master_frame['country_code'].isin(['USA','CHN','GBR'])])
```

In [146]:
```
df_country.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 68990 entries, 2 to 114935
Data columns (total 12 columns):
permalink                  68990 non-null object
name                       68989 non-null object
category_list              68588 non-null object
status                     68990 non-null object
country_code               68990 non-null object
region                     68465 non-null object
city                       68465 non-null object
company_permalink          68990 non-null object
funding_round_permalink    68990 non-null object
funding_round_type         68990 non-null object
funded_at                  68990 non-null object
raised_amount_usd          68990 non-null float64
dtypes: float64(1), object(11)
memory usage: 6.8+ MB
```

In [147]:
```
df_country.head()
```

Out[147]:

| | permalink | name | category_list | status | country_code | region | city | compa |
|---|---|---|---|---|---|---|---|---|
| 2 | /organization/-qounter | :Qounter | Application Platforms\|Real Time\|Social Network... | operating | USA | DE - Other | Delaware City | /organi |
| 4 | /organization/0-6-com | 0-6.com | Curated Web | operating | CHN | Beijing | Beijing | /organ |
| 9 | /organization/0xdata | H2O.ai | Analytics | operating | USA | SF Bay Area | Mountain View | /orga |
| 10 | /organization/0xdata | H2O.ai | Analytics | operating | USA | SF Bay Area | Mountain View | /orga |
| 11 | /organization/0xdata | H2O.ai | Analytics | operating | USA | SF Bay Area | Mountain View | /orga |

# Sector Analysis

In [148]:
```
df_map.info()
```

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

In [150]:
```
df_map.head(0)
```

Out[150]:

| category_list | Automotive & Sports | Blanks | Cleantech / Semiconductors | Entertainment | Health | Manufacturing | Mes |
|---|---|---|---|---|---|---|---|

In [51]:
```
df_map = pd.melt(df_map, id_vars =['category_list'], value_vars =['Automotive
 & Sports',
                                                              'Cleantech / Sem
iconductors','Entertainment',
                                                             'Health','Manufac
turing','News, Search and Messaging','Others',
                                                             'Social, Finance,
Analytics, Advertising'])
```

In [52]:
```
df_map.head(5)
```

Out[52]:

| | category_list | variable | value |
|---|---|---|---|
| 0 | NaN | Automotive & Sports | 0 |
| 1 | 3D | Automotive & Sports | 0 |
| 2 | 3D Printing | Automotive & Sports | 0 |
| 3 | 3D Technology | Automotive & Sports | 0 |
| 4 | Accounting | Automotive & Sports | 0 |

Null values are of no use here, hence dropping them.

In [54]:
```
df_map.dropna(inplace=True)
```

In [55]:
```
df_map.head(10)
```

Out[55]:

| | category_list | variable | value |
|---|---|---|---|
| 1 | 3D | Automotive & Sports | 0 |
| 2 | 3D Printing | Automotive & Sports | 0 |
| 3 | 3D Technology | Automotive & Sports | 0 |
| 4 | Accounting | Automotive & Sports | 0 |
| 5 | Active Lifestyle | Automotive & Sports | 0 |
| 6 | Ad Targeting | Automotive & Sports | 0 |
| 7 | Advanced Materials | Automotive & Sports | 0 |
| 8 | Adventure Travel | Automotive & Sports | 1 |
| 9 | Advertising | Automotive & Sports | 0 |
| 10 | Advertising Exchanges | Automotive & Sports | 0 |

In [57]:
```
df_map = df_map[df_map.value == 1]
```

In [58]:
```
df_map.head(5)
```

Out[58]:

| | category_list | variable | value |
|---|---|---|---|
| 8 | Adventure Travel | Automotive & Sports | 1 |
| 14 | Aerospace | Automotive & Sports | 1 |
| 45 | Auto | Automotive & Sports | 1 |
| 46 | Automated Kiosk | Automotive & Sports | 1 |
| 47 | Automotive | Automotive & Sports | 1 |

In [60]:
```
df_map.drop('value', axis=1,inplace=True)
df_map.rename(columns={'category_list':'primary_sector','variable':'main_secto
r'},inplace=True)
```

In [65]:
```
df_map.head(7)
```

Out[65]:

| | primary_sector | main_sector |
|---|---|---|
| 8 | Adventure Travel | Automotive & Sports |
| 14 | Aerospace | Automotive & Sports |
| 45 | Auto | Automotive & Sports |
| 46 | Automated Kiosk | Automotive & Sports |
| 47 | Automotive | Automotive & Sports |
| 57 | Bicycles | Automotive & Sports |
| 69 | Boating Industry | Automotive & Sports |

Since the category List has more than one values associated with a country so considering just the first one. Hence cleaning the data accordingly.

In [66]:
```
df_country['primary_sector'] = df_country['category_list'].str.split('|', n =
2, expand = True)[[0]]
```

In [67]:
```
df_country.head()
```

Out[67]:

| | permalink | name | category_list | status | country_code | region | city | compa |
|---|---|---|---|---|---|---|---|---|
| 2 | /organization/-qounter | :Qounter | Application Platforms\|Real Time\|Social Network... | operating | USA | DE - Other | Delaware City | /organi |
| 4 | /organization/0-6-com | 0-6.com | Curated Web | operating | CHN | Beijing | Beijing | /organ |
| 9 | /organization/0xdata | H2O.ai | Analytics | operating | USA | SF Bay Area | Mountain View | /orga |
| 10 | /organization/0xdata | H2O.ai | Analytics | operating | USA | SF Bay Area | Mountain View | /orga |
| 11 | /organization/0xdata | H2O.ai | Analytics | operating | USA | SF Bay Area | Mountain View | /orga |

Since we want all the information on the country data set, performing a left join on the tables.

In [68]:
```
df_country=pd.merge(df_country,df_map,how='left',on='primary_sector')
```

In [69]:
```
df_country.head()
```

Out[69]:

| | permalink | name | category_list | status | country_code | region | city | compar |
|---|---|---|---|---|---|---|---|---|
| 0 | /organization/-qounter | :Qounter | Application Platforms\|Real Time\|Social Network... | operating | USA | DE - Other | Delaware City | /organiz |
| 1 | /organization/0-6-com | 0-6.com | Curated Web | operating | CHN | Beijing | Beijing | /organiz |
| 2 | /organization/0xdata | H2O.ai | Analytics | operating | USA | SF Bay Area | Mountain View | /organ |
| 3 | /organization/0xdata | H2O.ai | Analytics | operating | USA | SF Bay Area | Mountain View | /organ |
| 4 | /organization/0xdata | H2O.ai | Analytics | operating | USA | SF Bay Area | Mountain View | /organ |

Now we have our df_country data frame with only the country we want to invest in with the type of funding and also the main 8 sectors to which the companies belong to. Applying the condition for the invesment amount and divind the country sets further on the basis of region

In [70]:
```
df_fund_USA=df_country[(df_country['country_code'] == 'USA') & (df_country.rai
sed_amount_usd > 5000000.0) & (df_country.raised_amount_usd < 15000000.0)]
```

In [71]:
```
df_fund_USA.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 11354 entries, 5 to 68985
Data columns (total 14 columns):
permalink                  11354 non-null object
name                       11354 non-null object
category_list              11271 non-null object
status                     11354 non-null object
country_code               11354 non-null object
region                     11343 non-null object
city                       11343 non-null object
company_permalink          11354 non-null object
funding_round_permalink    11354 non-null object
funding_round_type         11354 non-null object
funded_at                  11354 non-null object
raised_amount_usd          11354 non-null float64
primary_sector             11271 non-null object
main_sector                10422 non-null object
dtypes: float64(1), object(13)
memory usage: 1.3+ MB
```

In [72]:
```
df_fund_USA.head()
```

Out[72]:

| | permalink | name | category_list | status | country_code | region | |
|---|---|---|---|---|---|---|---|
| 5 | /organization/0xdata | H2O.ai | Analytics | operating | USA | SF Bay Area | M |
| 9 | /organization/1-800-publicrelations-inc- | 1-800-PublicRelations, Inc. | Internet Marketing\|Media\|Public Relations | operating | USA | New York City | N |
| 56 | /organization/128-technology | 128 Technology | Service Providers\|Technology | operating | USA | Boston | Bu |
| 61 | /organization/1366-technologies | 1366 Technologies | Manufacturing | operating | USA | Boston | |
| 62 | /organization/1366-technologies | 1366 Technologies | Manufacturing | operating | USA | Boston | |

In [75]:

```
df_fund_USA.sort_values(by='raised_amount_usd',ascending=False)
```

Out[75]:

| | permalink | name | category_list | status |
|---|---|---|---|---|
| 29722 | /organization/intermolecular | Intermolecular | Semiconductors | ipo |
| 56416 | /organization/spidercloud-wireless | SpiderCloud Wireless | Enterprise Software | operating |
| 58550 | /organization/synos-technology | Synos Technology | Manufacturing | acquired |
| 68372 | /organization/zenverge | Zenverge | Semiconductors | acquired |
| 34881 | /organization/luminal | Luminal | Cloud Computing\|Infrastructure\|Security\|Software | operating |
| ... | ... | ... | ... | ... |
| 53171 | /organization/setpoint-medical | SetPoint Medical | Biotechnology | operating |
| 40965 | /organization/noesis-energy | Noesis | Clean Energy\|Finance Technology\|FinTech | operating |
| 59035 | /organization/tarana-wireless | Tarana Wireless | Mobile\|Wireless | operating |
| 59073 | /organization/taris-biomedical | TARIS Biomedical | Biotechnology | operating |
| 32288 | /organization/knowledge-factor | Knowledge Factor | Software | operating |

11354 rows × 14 columns

In [78]:

```
df_fund_USA.groupby('main_sector').sum().sort_values(by='raised_amount_usd',as
cending=False)
```

Out[78]:

| | raised_amount_usd |
|---|---|
| main_sector | |
| Others | 2.387400e+10 |
| Cleantech / Semiconductors | 2.124917e+10 |
| Social, Finance, Analytics, Advertising | 1.466620e+10 |
| News, Search and Messaging | 1.179915e+10 |
| Health | 8.287338e+09 |
| Manufacturing | 6.954982e+09 |
| Entertainment | 4.408507e+09 |
| Automotive & Sports | 1.428994e+09 |

In [89]:

```
df_fund_USA.groupby('main_sector').mean().sort_values(by='raised_amount_usd',ascending=False)
```

Out[89]:

| main_sector | raised_amount_usd |
|---|---|
| Cleantech / Semiconductors | 9.069217e+06 |
| Health | 8.988436e+06 |
| Manufacturing | 8.974171e+06 |
| Others | 8.931537e+06 |
| Automotive & Sports | 8.766834e+06 |
| Social, Finance, Analytics, Advertising | 8.735079e+06 |
| News, Search and Messaging | 8.707860e+06 |
| Entertainment | 8.610366e+06 |

In [79]:

```
df_fund_CHN=df_country[(df_country['country_code'] == 'CHN') & (df_country.raised_amount_usd > 5000000.0) & (df_country.raised_amount_usd < 15000000.0)]
```

In [80]:

```
df_fund_CHN.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 428 entries, 29 to 68525
Data columns (total 14 columns):
permalink                  428 non-null object
name                       428 non-null object
category_list              425 non-null object
status                     428 non-null object
country_code               428 non-null object
region                     386 non-null object
city                       386 non-null object
company_permalink          428 non-null object
funding_round_permalink    428 non-null object
funding_round_type         428 non-null object
funded_at                  428 non-null object
raised_amount_usd          428 non-null float64
primary_sector             425 non-null object
main_sector                401 non-null object
dtypes: float64(1), object(13)
memory usage: 50.2+ KB
```

```
In [81]: df_fund_CHN.head()
```

Out[81]:

| | permalink | name | category_list | status | country_code | region | |
|---|---|---|---|---|---|---|---|
| 29 | /organization/1006-tv | 1006.tv | Games\|Media | operating | CHN | Beijing | |
| 54 | /organization/123feng-com | 123Feng.Com | NaN | operating | CHN | Hangzhou | H |
| 87 | /organization/19pay | 19pay | Finance\|FinTech | operating | CHN | Beijing | |
| 286 | /organization/3i-systems | 3i Systems | Semiconductors | closed | CHN | Guangdong | Gu |
| 381 | /organization/4s91-com | 4s91.com | Mobile | operating | CHN | Guangzhou | Gu |

```
In [82]: df_fund_CHN.sort_values(by='raised_amount_usd',ascending=False)
```

Out[82]:

| | permalink | name | category_list | status | country_code | re |
|---|---|---|---|---|---|---|
| 25704 | /organization/guokang-health-management | Guokang Health Management | Health and Wellness | operating | CHN | Sher |
| 59555 | /organization/tencho-technology | Tencho Technology | Enterprise Software | closed | CHN | Guang |
| 53424 | /organization/shenzhen-jucheng-enterprise-mana... | Shenzhen Jucheng Enterprise Management Consult... | Consulting | operating | CHN | Sher |
| 16186 | /organization/damai-cn | Damai.cn | E-Commerce | operating | CHN | C ( |
| 32782 | /organization/lamahui | Lamahui | E-Commerce\|E-Commerce Platforms\|Mobile Commerce | operating | CHN | |
| ... | ... | ... | ... | ... | ... | |
| 60539 | /organization/three-nod-group | 3Nod | Enterprise Software | operating | CHN | Sher |
| 58132 | /organization/suzhou-tianma-medical-group | Tianma Medical Group | Biotechnology | operating | CHN | Sha |
| 18369 | /organization/e-buy-china-business-consulting-... | E-Buy | Enterprise Software | operating | CHN | Sha |
| 6940 | /organization/beijing-kylin-network-informatio... | Kylin Network | Games | operating | CHN | B |
| 12143 | /organization/chinanetcenter | ChinaNetCenter | Enterprise Software | operating | CHN | B |

428 rows × 14 columns

In [87]:
```
df_fund_CHN.groupby('main_sector').sum().sort_values(by='raised_amount_usd',ascending=False)
```

Out[87]:

| | raised_amount_usd |
|---:|---:|
| **main_sector** | |
| Others | 1.167736e+09 |
| Social, Finance, Analytics, Advertising | 5.998333e+08 |
| News, Search and Messaging | 5.697095e+08 |
| Entertainment | 4.531943e+08 |
| Cleantech / Semiconductors | 3.577425e+08 |
| Manufacturing | 2.952911e+08 |
| Health | 2.189122e+08 |
| Automotive & Sports | 1.122786e+08 |

In [88]:
```
df_fund_CHN.groupby('main_sector').mean().sort_values(by='raised_amount_usd',ascending=False)
```

Out[88]:

| | raised_amount_usd |
|---:|---:|
| **main_sector** | |
| Health | 1.042439e+07 |
| Automotive & Sports | 1.020715e+07 |
| Social, Finance, Analytics, Advertising | 9.833334e+06 |
| News, Search and Messaging | 9.339501e+06 |
| Others | 9.267745e+06 |
| Manufacturing | 9.227848e+06 |
| Cleantech / Semiconductors | 9.172885e+06 |
| Entertainment | 9.063885e+06 |

In [90]:
```
df_fund_GBR=df_country[(df_country['country_code'] == 'GBR') & (df_country.raised_amount_usd > 5000000.0) & (df_country.raised_amount_usd < 15000000.0)]
```

In [91]:
```
df_fund_GBR.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 692 entries, 220 to 68970
Data columns (total 14 columns):
permalink                  692 non-null object
name                       692 non-null object
category_list              684 non-null object
status                     692 non-null object
country_code               692 non-null object
region                     657 non-null object
city                       657 non-null object
company_permalink          692 non-null object
funding_round_permalink    692 non-null object
funding_round_type         692 non-null object
funded_at                  692 non-null object
raised_amount_usd          692 non-null float64
primary_sector             684 non-null object
main_sector                637 non-null object
dtypes: float64(1), object(13)
memory usage: 81.1+ KB
```

In [92]:
```
df_fund_GBR.head()
```

Out[92]:

| | permalink | name | category_list | status | country_code |
|---|---|---|---|---|---|
| 220 | /organization/365scores | 365Scores | Android\|Apps\|iPhone\|Mobile\|Sports | operating | GBR |
| 431 | /organization/5app | 5app | Mobile\|Software\|Web Design\|Web Development | operating | GBR |
| 503 | /organization/7digital | 7digital | Content Creators\|Content Delivery\|Licensing\|Mu... | acquired | GBR |
| 504 | /organization/7digital | 7digital | Content Creators\|Content Delivery\|Licensing\|Mu... | acquired | GBR |
| 547 | /organization/90min | 90min | Media\|News\|Publishing\|Soccer\|Sports | operating | GBR |

In [93]:

```
df_fund_GBR.groupby('main_sector').sum().sort_values(by='raised_amount_usd',ascending=False)
```

Out[93]:

|  | raised_amount_usd |
|---:|---:|
| **main_sector** |  |
| Cleantech / Semiconductors | 1.372091e+09 |
| Others | 1.288714e+09 |
| Social, Finance, Analytics, Advertising | 8.427263e+08 |
| News, Search and Messaging | 7.107990e+08 |
| Manufacturing | 4.523687e+08 |
| Entertainment | 4.363067e+08 |
| Health | 2.710470e+08 |
| Automotive & Sports | 1.760204e+08 |

In [94]:

```
df_fund_GBR.groupby('main_sector').mean().sort_values(by='raised_amount_usd',ascending=False)
```

Out[94]:

|  | raised_amount_usd |
|---:|---:|
| **main_sector** |  |
| Automotive & Sports | 9.264231e+06 |
| Others | 8.766764e+06 |
| Health | 8.743453e+06 |
| Cleantech / Semiconductors | 8.739434e+06 |
| Entertainment | 8.726134e+06 |
| Manufacturing | 8.699398e+06 |
| News, Search and Messaging | 8.668280e+06 |
| Social, Finance, Analytics, Advertising | 8.512386e+06 |

In [ ]:

```
```