Part 2: Data cleaning

```
#
#     > We have cleaned the files now we have to treat the missing
values we found in rest of the columns.
#     > This time we will import from the cleaned csv files


# # Missing value treatment
```

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

```
#     > create a master data frame for ease of analysis
#     > we can use pd.merge to merge on company_permalink column
#     > after merging we can drop any one of the permalink column as
they are redundant
```

Percent missing value in merged dataframe

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

# 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.

Dropping columns

```
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
```

After dropping
Missing columns include category_list, country_code and raised_amount_usd.
We can not simply delete these columns as category_list will be used for merging with the mapping file.
country_code and raised_amount_usd are useful from business perspective.
We have to carefully tread through the raised_amount_usd column as it has about 17% missing values

Raised_amount_column

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

The mean amount of funding is 10 million USD. The median is about 1.7 million USD.
The highest amount invested is about 21.7 billion USD The data is highly skewed and has very large outliers. This clearly inflate the mean.
This suggests we have no other option but to delete the missing values in raised_amount_usd as we can not impute them with mean or median

After deleting the null entries in raised_amount and country_codes ,

imputing Tell_it_in name in names row.
And deleting null rows in category_list we get clean data which we save to .csv files

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 88529 entries, 0 to 114947
Data columns (total 9 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   permalink                88529 non-null  object 
 1   name                     88529 non-null  object 
 2   category_list            88529 non-null  object 
 3   status                   88529 non-null  object 
 4   country_code             88529 non-null  object 
 5   funding_round_permalink  88529 non-null  object 
 6   funding_round_type       88529 non-null  object 
 7   funded_at                88529 non-null  object 
 8   raised_amount_usd        88529 non-null  float64
dtypes: float64(1), object(8)
```

We have treated all the missing values. Now we have 88529 out of 114948 entries left after clean-up.
We have about 78% of our initial data. Which is low but as the data has ~89K entries, we can do some solid analysis to them.
Now we can put the cleaned master data to a csv file