Companies file analysis -
Companies file has 66368 entries and 10 columns with various features

1. permalink, home_url and name are three company specific identifiers and name column has 1 missing value, with home_url with 5038 missing values.

2. category_list gives the industry category the company belongs to.

3. country_code, state_code, region_code and city are geography specific identifiers and have some missing values.

4. status has also some missing values along with founded_at.

5. We will deal with each column one-by-one

6. Lets get some initial analysis of rounds file.

Rounds file analysis
Rounds file has 114949 entries and 6 columns with various features

1. company_permalink seems to be similar to that of the permalink column in companies file

2. fundaing_round_permalink is another name identifier for the funding round

3. funding round type gives us the idea of the type of funding done.

4. funding_round_code is another feature with codes about the funding round.
    1. It has 83809 missing values

5. Raised_amount_usd gives the funding in USD.
    1. Max amount raised is 21.27 billion USD
    2. Min is 0 and average funding is 10 million USD
    3. The average is inflated due the high outliers.

Trivia -

* rounds2 has 90247 unique permalink
* companies has 66368 unique permalink entries
* We convert them to lower case

* There seems to be some weird characters in the company_permalink in rounds file.
* Lets check the csv file to see if these entries are actually the same as in dataframe

```
29597                               /organization/e-cãbica
31863          /organization/energystone-games-çµç³æ¸¸æ
45176                  /organization/huizuche-com-æ ç§ÿè½¦
58473                /organization/magnet-tech-ç£ç³ç§æ
101036    /organization/tipcat-interactive-æ²èÿä¿¡æ¯ç...
109969               /organization/weiche-tech-åè½¦ç§æ
113839                   /organization/zengame-ç¦
æ¸¸ç§æ

Name: company_permalink, dtype: object
* The csv file seems to have the above 7 entries in correct format.
* This means there is an issue in decoding the file in python
* The main reason that python is not able to decode is because of
various languages this data
  is collected from various countries.



Final outcome -

Uncommon permalink in rounds to that in companies: Series([], Name:
company_permalink, dtype: object)
rounds 66368
companies 66368
```