01 data preparation and exploration

April 5, 2022

[ ]:
```
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
from scipy.spatial import distance_matrix
np.random.seed(123)
```

get company data. filter by type of company:

- SOCIETA' DI CAPITALE|SU|SOCIETA' A RESPONSABILITA' LIMITATA CON UNICO SOCIO
- SOCIETA' DI CAPITALE|SR|SOCIETA' A RESPONSABILITA' LIMITATA
- SOCIETA' DI CAPITALE|SP|SOCIETA' PER AZIONI
- SOCIETA' DI CAPITALE|SD|SOCIETA' EUROPEA
- SOCIETA' DI CAPITALE|RS|SOCIETA' A RESPONSABILITA' LIMITATA SEMPLIFICATA
- SOCIETA' DI CAPITALE|RR|SOCIETA' A RESPONSABILITA' LIMITATA A CAPITALE RIDOTTO
- SOCIETA' DI CAPITALE|AU|SOCIETA' PER AZIONI CON SOCIO UNICO
- SOCIETA' DI CAPITALE|AA|SOCIETA' IN ACCOMANDITA PER AZIONI

[ ]:
```
# Data Acquisition
filename = r'../../_DataScience/____PHD_2021/_data/tidy/cmp.csv'
# cols_to_use = ['idCompany', 'name', 'cf', 'prov', 'sede_ul', 'ng2',␣
 ↪'stato_impresa',
#        'addetti_aaaa', 'addetti_indip', 'addetti_dip', 'capitale',
#        'capitale_valuta', 'imp_sedi_ee', 'imp_eefvg', 'is.sme', 'is.startup',
#        'is.fem', 'is.young', 'is.fore', 'yearsInBusiness']
cols_to_use = [ 'cf', 'prov', 'ng2', 'stato_impresa', 'yearsInBusiness']
companies = pd.read_csv(filename, dtype=str, usecols=cols_to_use)
companies['yearsInBusiness'] = companies['yearsInBusiness'].astype(float).
 ↪round(1)
company_types = ['SU','SR', 'SP','SD','RS','RR','AU','AA']
companies = companies[companies.prov.isin(['TS','GO','UD','PN'])]
companies = companies[companies.ng2.isin(company_types)]
companies = companies[companies.stato_impresa.isin(['ATTIVA'])]
companies.shape
```

[ ]:
```
(19339, 5)
```

[ ]:

```
filename = r'../../_DataScience/____PHD_2021/_data/tidy/bsd.csv'
# cols_to_use = ['cf', 'prov', 'year', 'totAssets', 'totIntang', 'accounts',␣
 ↪'totEquity',
#         'debts', 'prod', 'revenues', 'personnel', 'valCost', 'ammort',
#         'profLoss', 'valAdded', 'deprec', 'noi']
cols_to_use = ['cf',  'year','totAssets', 'totIntang', 'totEquity', 'noi']
bsd = pd.read_csv(filename, dtype=str, usecols=cols_to_use)
bsd = bsd[bsd.year == '2019']
bsd['totEquity'] = bsd['totEquity'].astype(float)
bsd['totAssets'] = bsd['totAssets'].astype(float)
bsd['noi'] = bsd['noi'].astype(float)
bsd['totIntang'] = bsd['totIntang'].astype(float)

bsd['rAssets'] = bsd.totAssets/bsd.totEquity
bsd['rNOI'] =     bsd.noi/bsd.totEquity
bsd['rIntang'] = bsd.totIntang/bsd.totEquity

cols_to_use = ['cf', 'rAssets', 'rNOI', 'rIntang']
bsd = bsd[ cols_to_use]

bsd.head(5)
```

[ ]:

```
             cf    rAssets      rNOI   rIntang
2   00002070324   3.112637  1.003312  0.003202
9   00007080369   1.683364  0.114665  0.044183
16  00007470933  26.512538  0.541276  4.157546
23  00009840315   1.380727  0.197934  0.025756
30  00012670303   1.328204  0.002143  0.003106
```

[ ]:

```
companies = companies.merge(bsd, on='cf')

companies.head(5)
```

[ ]:

```
            cf prov ng2 stato_impresa  yearsInBusiness    rAssets      rNOI  \
0  00002070324   TS  SR        ATTIVA             53.0   3.112637  1.003312
1  00007470933   PN  SR        ATTIVA             59.2  26.512538  0.541276
2  00009840315   GO  SR        ATTIVA             58.5   1.380727  0.197934
3  00012670303   UD  SP        ATTIVA             51.1   1.328204  0.002143
4  00018160309   UD  SR        ATTIVA             54.6   6.935805  0.115856

    rIntang
0  0.003202
1  4.157546
2  0.025756
3  0.003106
4  0.231687
```

```
[ ]: cols_to_use = ['cf', 'yearsInBusiness','rAssets','rIntang', 'rNOI']
     companies=companies[cols_to_use]
     companies.head(5)
```

```
[ ]:             cf  yearsInBusiness    rAssets   rIntang      rNOI
     0  00002070324             53.0   3.112637  0.003202  1.003312
     1  00007470933             59.2  26.512538  4.157546  0.541276
     2  00009840315             58.5   1.380727  0.025756  0.197934
     3  00012670303             51.1   1.328204  0.003106  0.002143
     4  00018160309             54.6   6.935805  0.231687  0.115856
```

```
[ ]: filename = r'../../_DataScience/____PHD_2021/_data/tidy/rating.csv'
     #cols_to_use = ['cf', 'final_rank', 'evaluation_date', 'is_consolidated',␣
      ↪'rating010','year']
     cols_to_use = ['cf', 'rating010','year']
     rating = pd.read_csv(filename, dtype=str, usecols=cols_to_use)
     rating = rating[rating.year == '2019']
     rating['rating010'] = rating['rating010'].astype(float)
     rating.head(5)
```

```
[ ]:              cf  rating010  year
     2   00008980328        1.0  2019
     6   00019410307        1.0  2019
     10  00037070323        1.0  2019
     14  00039970314        1.0  2019
     18  00041170317        1.0  2019
```

```
[ ]: companies = companies.merge(rating, on='cf')
     companies.head(5)
```

```
[ ]:             cf  yearsInBusiness    rAssets   rIntang      rNOI  rating010  \
     0  00002070324             53.0   3.112637  0.003202  1.003312        9.0
     1  00007470933             59.2  26.512538  4.157546  0.541276        5.0
     2  00009840315             58.5   1.380727  0.025756  0.197934        9.0
     3  00012670303             51.1   1.328204  0.003106  0.002143        6.0
     4  00018160309             54.6   6.935805  0.231687  0.115856        2.0

        year
     0  2019
     1  2019
     2  2019
     3  2019
     4  2019
```

```
[ ]: cols_to_use = ['cf', 'yearsInBusiness','rAssets','rIntang', 'rNOI', 'rating010']
     companies=companies[cols_to_use]
```

```
companies.head(5)
```

```
[ ]:             cf  yearsInBusiness    rAssets   rIntang      rNOI  rating010
     0  00002070324             53.0   3.112637  0.003202  1.003312        9.0
     1  00007470933             59.2  26.512538  4.157546  0.541276        5.0
     2  00009840315             58.5   1.380727  0.025756  0.197934        9.0
     3  00012670303             51.1   1.328204  0.003106  0.002143        6.0
     4  00018160309             54.6   6.935805  0.231687  0.115856        2.0
```

```
[ ]: filename = r'../../_DataScience/____PHD_2021/_data/tidy/nace.csv'
     #cols_to_use = ['cf', 'idCompany', 'id_localiz', 'loc_n', 'code_type',␣
      ↪'division','code']
     cols_to_use = ['cf', 'loc_n', 'code_type', 'division']
     nace = pd.read_csv(filename, dtype=str, usecols=cols_to_use)
     nace = nace[ nace.code_type == "I"]
     nace = nace[ nace.loc_n == "0"]
     nace.drop_duplicates(subset=['cf'], keep='first', inplace=True,␣
      ↪ignore_index=True)
     cols_to_use = ['cf', 'division']
     nace=nace[cols_to_use]
     nace.columns = ['cf', 'NACE_division']
     nace.head(5)
```

```
[ ]:             cf NACE_division
     0  00002070324            52
     1  00007470933            25
     2  00008120313            47
     3  00008900938            11
     4  00012650933            69
```

```
[ ]: companies = companies.merge(nace, on='cf')
     companies.head(5)
```

```
[ ]:             cf  yearsInBusiness    rAssets   rIntang      rNOI  rating010  \
     0  00002070324             53.0   3.112637  0.003202  1.003312        9.0
     1  00007470933             59.2  26.512538  4.157546  0.541276        5.0
     2  00018160309             54.6   6.935805  0.231687  0.115856        2.0
     3  00030810311             47.0   4.114307  0.005739 -0.078252        4.0
     4  00039490313             42.9   1.828177  0.088395  0.206470        8.0

       NACE_division
     0            52
     1            25
     2            28
     3            47
     4            20
```

[ ]:

```
filename = r'../../_DataScience/____PHD_2021/_data/tidy/empl_stock.csv'
#cols_to_use = ['cf', 'name', 'rea', 'prov', 'StockProv', 'StockAll',␣
 ↪'date_stock']
cols_to_use = ['cf', 'StockAll']
emp_stock = pd.read_csv(filename, dtype=str, usecols=cols_to_use)
emp_stock.columns = ['cf', 'staff_count']
emp_stock.head(5)

companies = companies.merge(emp_stock, on='cf')
companies.head(5)
```

[ ]:

```
            cf  yearsInBusiness    rAssets   rIntang      rNOI  rating010  \
0  00002070324             53.0   3.112637  0.003202  1.003312        9.0
1  00002070324             53.0   3.112637  0.003202  1.003312        9.0
2  00007470933             59.2  26.512538  4.157546  0.541276        5.0
3  00018160309             54.6   6.935805  0.231687  0.115856        2.0
4  00018160309             54.6   6.935805  0.231687  0.115856        2.0

  NACE_division staff_count
0            52          16
1            52           0
2            25          24
3            28           1
4            28          12
```

[ ]:

```
filename = r'../../_DataScience/____PHD_2021/_data/tidy/empl_flows.csv'
cols_to_use = ['cf', 'year', 'turnover','balance']
empl_flows = pd.read_csv(filename, dtype=str, usecols=cols_to_use)
empl_flows = empl_flows[ empl_flows.year == '2014']
empl_flows['cf'] = empl_flows['cf'].str.strip()
empl_flows.columns = ['cf', 'year', 'staff_turnover', 'staff_variation']
```

[ ]:

```
companies = companies.merge(empl_flows, on='cf')
```

[ ]:

```
coords = pd.read_csv(r'./maps/FVG/companies.csv', dtype='str')
coords.columns = ['ind', 'cf', 'company','unit', 'lat', 'lon']
coords['lat'] = coords['lat'].astype(float)
coords['lon'] = coords['lon'].astype(float)
coords = coords[ coords.unit == 'SEDE']
coords = coords[ [ 'cf',  'lat', 'lon'] ]
coords.shape
```

[ ]:

```
(16624, 3)
```

[ ]:

```
companies = companies.merge(coords, on='cf')
```

[ ]:

```python
companies = companies.sample(frac=1)
```

[ ]:

```python
companies.reset_index(inplace=True)
```

[ ]:

```python
companies = companies[ [ 'index','yearsInBusiness', 'rAssets', 'rIntang',␣
 ↪'rNOI', 'rating010',
       'NACE_division', 'staff_count', 'staff_turnover', 'staff_variation',␣
 ↪'lat', 'lon'] ]
```

[ ]:

```python
companies.to_csv(r'./data/data.csv', index=False)
```