

Revelio Data Processing - Private Firm

November 22, 2025

```
[1]: import pandas as pd
import swifter
pd.set_option('display.max_columns', 500)
pd.set_option('display.max_rows', 100)
```

1 Process reference table

```
[2]: files_position = []

import glob, os
os.chdir("D:/User.Data/Dan.Li/Revelio/umiami_202212_full_company_ref/")
for file in glob.glob("*.csv"):
    files_position.append(file)
files_position
```

```
[2]: ['company_ref_0000_part_00.csv',
'company_ref_0001_part_00.csv',
'company_ref_0002_part_00.csv',
'company_ref_0003_part_00.csv',
'company_ref_0004_part_00.csv',
'company_ref_0005_part_00.csv',
'company_ref_0006_part_00.csv',
'company_ref_0007_part_00.csv',
'company_ref_0008_part_00.csv',
'company_ref_0009_part_00.csv',
'company_ref_0010_part_00.csv',
'company_ref_0011_part_00.csv',
'company_ref_0012_part_00.csv',
'company_ref_0013_part_00.csv',
'company_ref_0014_part_00.csv',
'company_ref_0015_part_00.csv',
'company_ref_0016_part_00.csv',
'company_ref_0017_part_00.csv',
'company_ref_0018_part_00.csv',
'company_ref_0019_part_00.csv',
'company_ref_0020_part_00.csv',
'company_ref_0021_part_00.csv',
```

```
'company_ref_0022_part_00.csv',
'company_ref_0023_part_00.csv',
'company_ref_0024_part_00.csv',
'company_ref_0025_part_00.csv',
'company_ref_0026_part_00.csv',
'company_ref_0027_part_00.csv',
'company_ref_0028_part_00.csv',
'company_ref_0029_part_00.csv',
'company_ref_0030_part_00.csv',
'company_ref_0031_part_00.csv',
'company_ref_0032_part_00.csv',
'company_ref_0033_part_00.csv',
'company_ref_0034_part_00.csv',
'company_ref_0035_part_00.csv',
'company_ref_0036_part_00.csv',
'company_ref_0037_part_00.csv',
'company_ref_0038_part_00.csv',
'company_ref_0039_part_00.csv',
'company_ref_0040_part_00.csv',
'company_ref_0041_part_00.csv',
'company_ref_0042_part_00.csv']
```

```
[5]: postion_dir = "D:/User.Data/Dan.Li/Revelio/umiami_202212_full_company_ref/"

finished = -1
df1_sum = pd.DataFrame()
```

```
[ ]: for i in range(0, len(files_position)):

    if i <= finished:
        continue
    file_i = files_position[i]

    file_full = postion_dir + file_i
    filename = file_i.split(".")[0]

    print("Start to processs: ", file_full)
    df1 = pd.read_csv(file_full, encoding = "latin", on_bad_lines='skip',
↳lineterminator='\n',
                        dtype = {'rcid': str})
    df1["rcid"] = df1["rcid"].apply(lambda x:"rcid_" + str(x))

    df1_sum = df1_sum.append(df1)
    print("File finished: ", file_full)
    finished = i
```

Start to processs: D:/User.Data/Dan.Li/Revelio/umiami_202212_full_company_ref/c

```

company_ref_0000_part_00.csv
File finished: D:/User.Data/Dan.Li/Revelio/umiami_202212_full_company_ref/compa
ny_ref_0000_part_00.csv
Start to processs: D:/User.Data/Dan.Li/Revelio/umiami_202212_full_company_ref/c
ompany_ref_0001_part_00.csv
File finished: D:/User.Data/Dan.Li/Revelio/umiami_202212_full_company_ref/compa
ny_ref_0001_part_00.csv
Start to processs: D:/User.Data/Dan.Li/Revelio/umiami_202212_full_company_ref/c
ompany_ref_0002_part_00.csv
File finished: D:/User.Data/Dan.Li/Revelio/umiami_202212_full_company_ref/compa
ny_ref_0002_part_00.csv
Start to processs: D:/User.Data/Dan.Li/Revelio/umiami_202212_full_company_ref/c
ompany_ref_0003_part_00.csv
File finished: D:/User.Data/Dan.Li/Revelio/umiami_202212_full_company_ref/compa
ny_ref_0003_part_00.csv
Start to processs: D:/User.Data/Dan.Li/Revelio/umiami_202212_full_company_ref/c
ompany_ref_0004_part_00.csv
File finished: D:/User.Data/Dan.Li/Revelio/umiami_202212_full_company_ref/compa
ny_ref_0004_part_00.csv
Start to processs: D:/User.Data/Dan.Li/Revelio/umiami_202212_full_company_ref/c
ompany_ref_0005_part_00.csv
File finished: D:/User.Data/Dan.Li/Revelio/umiami_202212_full_company_ref/compa
ny_ref_0005_part_00.csv
Start to processs: D:/User.Data/Dan.Li/Revelio/umiami_202212_full_company_ref/c
ompany_ref_0006_part_00.csv
File finished: D:/User.Data/Dan.Li/Revelio/umiami_202212_full_company_ref/compa
ny_ref_0006_part_00.csv
Start to processs: D:/User.Data/Dan.Li/Revelio/umiami_202212_full_company_ref/c
ompany_ref_0007_part_00.csv

```

```
[16]: df1_sum = df1_sum.drop_duplicates()
```

```
[43]: comp_RL_matched_simple = df1_sum
```

2 Position

```
[44]: files_position = []

import glob, os
os.chdir("D:/User.Data/Dan.Li/Revelio/umiami_202212_position/")
for file in glob.glob("*.csv"):
#     print(file)
    files_position.append(file)
files_position

```

```
[44]: ['position_0000_part_00.csv',
      'position_0000_part_01.csv',
```

'position_0000_part_02.csv',
'position_0000_part_03.csv',
'position_0000_part_04.csv',
'position_0000_part_05.csv',
'position_0001_part_00.csv',
'position_0001_part_01.csv',
'position_0001_part_02.csv',
'position_0001_part_03.csv',
'position_0001_part_04.csv',
'position_0001_part_05.csv',
'position_0002_part_00.csv',
'position_0002_part_01.csv',
'position_0002_part_02.csv',
'position_0002_part_03.csv',
'position_0002_part_04.csv',
'position_0002_part_05.csv',
'position_0003_part_00.csv',
'position_0003_part_01.csv',
'position_0003_part_02.csv',
'position_0003_part_03.csv',
'position_0003_part_04.csv',
'position_0003_part_05.csv',
'position_0004_part_00.csv',
'position_0004_part_01.csv',
'position_0004_part_02.csv',
'position_0004_part_03.csv',
'position_0004_part_04.csv',
'position_0004_part_05.csv',
'position_0005_part_00.csv',
'position_0005_part_01.csv',
'position_0005_part_02.csv',
'position_0005_part_03.csv',
'position_0005_part_04.csv',
'position_0005_part_05.csv',
'position_0006_part_00.csv',
'position_0006_part_01.csv',
'position_0006_part_02.csv',
'position_0006_part_03.csv',
'position_0006_part_04.csv',
'position_0006_part_05.csv',
'position_0007_part_00.csv',
'position_0007_part_01.csv',
'position_0007_part_02.csv',
'position_0007_part_03.csv',
'position_0007_part_04.csv',
'position_0007_part_05.csv',
'position_0008_part_00.csv',

'position_0008_part_01.csv',
'position_0008_part_02.csv',
'position_0008_part_03.csv',
'position_0008_part_04.csv',
'position_0008_part_05.csv',
'position_0009_part_00.csv',
'position_0009_part_01.csv',
'position_0009_part_02.csv',
'position_0009_part_03.csv',
'position_0009_part_04.csv',
'position_0009_part_05.csv',
'position_0010_part_00.csv',
'position_0010_part_01.csv',
'position_0010_part_02.csv',
'position_0010_part_03.csv',
'position_0010_part_04.csv',
'position_0010_part_05.csv',
'position_0011_part_00.csv',
'position_0011_part_01.csv',
'position_0011_part_02.csv',
'position_0011_part_03.csv',
'position_0011_part_04.csv',
'position_0011_part_05.csv',
'position_0012_part_00.csv',
'position_0012_part_01.csv',
'position_0012_part_02.csv',
'position_0012_part_03.csv',
'position_0012_part_04.csv',
'position_0012_part_05.csv',
'position_0013_part_00.csv',
'position_0013_part_01.csv',
'position_0013_part_02.csv',
'position_0013_part_03.csv',
'position_0013_part_04.csv',
'position_0013_part_05.csv',
'position_0014_part_00.csv',
'position_0014_part_01.csv',
'position_0014_part_02.csv',
'position_0014_part_03.csv',
'position_0014_part_04.csv',
'position_0014_part_05.csv',
'position_0015_part_00.csv',
'position_0015_part_01.csv',
'position_0015_part_02.csv',
'position_0015_part_03.csv',
'position_0015_part_04.csv',
'position_0015_part_05.csv',

'position_0016_part_00.csv',
'position_0016_part_01.csv',
'position_0016_part_02.csv',
'position_0016_part_03.csv',
'position_0016_part_04.csv',
'position_0016_part_05.csv',
'position_0017_part_00.csv',
'position_0017_part_01.csv',
'position_0017_part_02.csv',
'position_0017_part_03.csv',
'position_0017_part_04.csv',
'position_0017_part_05.csv',
'position_0018_part_00.csv',
'position_0018_part_01.csv',
'position_0018_part_02.csv',
'position_0018_part_03.csv',
'position_0018_part_04.csv',
'position_0018_part_05.csv',
'position_0019_part_00.csv',
'position_0019_part_01.csv',
'position_0019_part_02.csv',
'position_0019_part_03.csv',
'position_0019_part_04.csv',
'position_0019_part_05.csv',
'position_0020_part_00.csv',
'position_0020_part_01.csv',
'position_0020_part_02.csv',
'position_0020_part_03.csv',
'position_0020_part_04.csv',
'position_0020_part_05.csv',
'position_0021_part_00.csv',
'position_0021_part_01.csv',
'position_0021_part_02.csv',
'position_0021_part_03.csv',
'position_0021_part_04.csv',
'position_0021_part_05.csv',
'position_0022_part_00.csv',
'position_0022_part_01.csv',
'position_0022_part_02.csv',
'position_0022_part_03.csv',
'position_0022_part_04.csv',
'position_0022_part_05.csv',
'position_0023_part_00.csv',
'position_0023_part_01.csv',
'position_0023_part_02.csv',
'position_0023_part_03.csv',
'position_0023_part_04.csv',

```

'position_0023_part_05.csv',
'position_0024_part_00.csv',
'position_0024_part_01.csv',
'position_0024_part_02.csv',
'position_0024_part_03.csv',
'position_0024_part_04.csv',
'position_0024_part_05.csv',
'position_0025_part_00.csv',
'position_0025_part_01.csv',
'position_0025_part_02.csv',
'position_0025_part_03.csv',
'position_0025_part_04.csv',
'position_0025_part_05.csv',
'position_0026_part_00.csv',
'position_0026_part_01.csv',
'position_0026_part_02.csv',
'position_0026_part_03.csv',
'position_0026_part_04.csv',
'position_0026_part_05.csv',
'position_0027_part_00.csv',
'position_0027_part_01.csv',
'position_0027_part_02.csv',
'position_0027_part_03.csv',
'position_0027_part_04.csv',
'position_0027_part_05.csv']

```

```

[45]: postion_dir = "D:/User.Data/Dan.Li/Revelio/umiami_202212_position/"
output_dir = "D:/User.Data/Dan.Li/Revelio/Processed_Data_PrivateFirms/
↳Position_cleaned/"
finished = -1

```

```

[ ]: for i in range(0, len(files_position)):

    if i <= finished:
        continue
    file_i = files_position[i]

    file_full = postion_dir + file_i
    filename = file_i.split(".")[0]
    output_full = output_dir + filename + ".pkl"

    print("Start to processs: ", file_full)
    df1 = pd.read_csv(file_full, encoding = "latin", on_bad_lines='skip',
↳lineterminator='\n',
                        dtype = {'rcid': str, 'position_id': str, 'user_id':str})
    df1 = df1[df1["rcid"]!="nan"]

```

```

df1["rcid"] = df1["rcid"].apply(lambda x:"rcid_" + str(x))
df1['position_id'] = df1['position_id'].apply(lambda x:"position_id_" +
↪str(x))
df1['user_id'] = df1['user_id'].apply(lambda x:"user_id_" + str(x))

df1 = df1.merge(comp_RL_matched_simple, left_on = ["rcid"], right_on =
↪["rcid"],how = "inner")

print(len(df1))
df1.to_pickle(output_full)
print("File finished: ", file_full)
finished = i

```

Start to processs:

D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0000_part_00.csv

C:\Users\dxl902\AppData\Local\conda\conda\envs\an_env\lib\site-
packages\IPython\core\interactiveshell.py:3457: DtypeWarning: Columns
(16,17,18,19,20,21,22,24,25,26,27,28,29,30,32,33,34,35,36,37,38,39,40,41,42,43)
have mixed types.Specify dtype option on import or set low_memory=False.
exec(code_obj, self.user_global_ns, self.user_ns)

0

File finished:

D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0000_part_00.csv

Start to processs:

D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0000_part_01.csv

C:\Users\dxl902\AppData\Local\conda\conda\envs\an_env\lib\site-
packages\IPython\core\interactiveshell.py:3457: DtypeWarning: Columns
(16,17,18,19,20,21,22,24,25,26,27,28,29,30,32,33,34,37,38,39,40,41,42,43) have
mixed types.Specify dtype option on import or set low_memory=False.
exec(code_obj, self.user_global_ns, self.user_ns)

515

File finished:

D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0000_part_01.csv

Start to processs:

D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0000_part_02.csv

C:\Users\dxl902\AppData\Local\conda\conda\envs\an_env\lib\site-
packages\IPython\core\interactiveshell.py:3457: DtypeWarning: Columns
(21,22,24,25,26,27,28,29,30,32,33,34,37,38,39,40,41) have mixed types.Specify
dtype option on import or set low_memory=False.
exec(code_obj, self.user_global_ns, self.user_ns)

15200

File finished:

D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0000_part_02.csv

```

Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0000_part_03.csv

C:\Users\dxl902\AppData\Local\conda\conda\envs\an_env\lib\site-
packages\IPython\core\interactiveshell.py:3457: DtypeWarning: Columns
(21,22,24,25,26,27,28,29,30,33,37,39,40) have mixed types.Specify dtype option
on import or set low_memory=False.
    exec(code_obj, self.user_global_ns, self.user_ns)

32324
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0000_part_03.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0000_part_04.csv

C:\Users\dxl902\AppData\Local\conda\conda\envs\an_env\lib\site-
packages\IPython\core\interactiveshell.py:3457: DtypeWarning: Columns
(21,22,24,25,26,27,28,29,30,33,37,38,39,40) have mixed types.Specify dtype
option on import or set low_memory=False.
    exec(code_obj, self.user_global_ns, self.user_ns)

18556
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0000_part_04.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0000_part_05.csv
6470
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0000_part_05.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0001_part_00.csv

C:\Users\dxl902\AppData\Local\conda\conda\envs\an_env\lib\site-
packages\IPython\core\interactiveshell.py:3457: DtypeWarning: Columns
(16,17,18,19,20,21,22,24,25,26,27,28,29,30,32,33,34,35,36,37,38,39,40,41) have
mixed types.Specify dtype option on import or set low_memory=False.
    exec(code_obj, self.user_global_ns, self.user_ns)

1
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0001_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0001_part_01.csv

C:\Users\dxl902\AppData\Local\conda\conda\envs\an_env\lib\site-
packages\IPython\core\interactiveshell.py:3457: DtypeWarning: Columns
(18,19,20,21,22,24,25,26,27,28,29,30,32,33,34,37,38,39,40,41,42,43) have mixed
types.Specify dtype option on import or set low_memory=False.
    exec(code_obj, self.user_global_ns, self.user_ns)

494
File finished:

```

```

D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0001_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0001_part_02.csv
14854
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0001_part_02.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0001_part_03.csv

C:\Users\dxl902\AppData\Local\conda\conda\envs\an_env\lib\site-
packages\IPython\core\interactiveshell.py:3457: DtypeWarning: Columns
(21,22,24,25,26,27,28,29,30,33,39,40) have mixed types.Specify dtype option on
import or set low_memory=False.
    exec(code_obj, self.user_global_ns, self.user_ns)

31898
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0001_part_03.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0001_part_04.csv
18432
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0001_part_04.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0001_part_05.csv
6274
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0001_part_05.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0002_part_00.csv

C:\Users\dxl902\AppData\Local\conda\conda\envs\an_env\lib\site-
packages\IPython\core\interactiveshell.py:3457: DtypeWarning: Columns
(16,17,18,19,20,21,22,24,25,26,27,28,29,30,32,33,34,35,36,37,38,39,41) have
mixed types.Specify dtype option on import or set low_memory=False.
    exec(code_obj, self.user_global_ns, self.user_ns)

0
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0002_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0002_part_01.csv
461
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0002_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0002_part_02.csv
14896
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0002_part_02.csv

```

Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0002_part_03.csv
31985
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0002_part_03.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0002_part_04.csv
18276
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0002_part_04.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0002_part_05.csv
6343
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0002_part_05.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0003_part_00.csv
4
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0003_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0003_part_01.csv
476
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0003_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0003_part_02.csv
14804
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0003_part_02.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0003_part_03.csv
32496
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0003_part_03.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0003_part_04.csv
18530
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0003_part_04.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0003_part_05.csv
6450
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0003_part_05.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0004_part_00.csv

```

C:\Users\dxl902\AppData\Local\conda\conda\envs\an_env\lib\site-
packages\IPython\core\interactiveshell.py:3457: DtypeWarning: Columns
(0,1,16,17,18,19,20,21,22,26,27,28,29,30,31,32,34,35,36,37,38,39,40,41,42,43)
have mixed types.Specify dtype option on import or set low_memory=False.
    exec(code_obj, self.user_global_ns, self.user_ns)

4
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0004_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0004_part_01.csv
451
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0004_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0004_part_02.csv
15109
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0004_part_02.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0004_part_03.csv
31813
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0004_part_03.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0004_part_04.csv
18205
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0004_part_04.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0004_part_05.csv
6320
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0004_part_05.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0005_part_00.csv
1
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0005_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0005_part_01.csv

C:\Users\dxl902\AppData\Local\conda\conda\envs\an_env\lib\site-
packages\IPython\core\interactiveshell.py:3457: DtypeWarning: Columns
(16,17,18,19,20,21,22,24,25,26,27,28,29,30,33,37,38,39,40,41,42,43) have mixed
types.Specify dtype option on import or set low_memory=False.
    exec(code_obj, self.user_global_ns, self.user_ns)

496

```

```

File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0005_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0005_part_02.csv
15036
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0005_part_02.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0005_part_03.csv
31796
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0005_part_03.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0005_part_04.csv
18297
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0005_part_04.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0005_part_05.csv
6510
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0005_part_05.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0006_part_00.csv
0
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0006_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0006_part_01.csv

C:\Users\dxl902\AppData\Local\conda\conda\envs\an_env\lib\site-
packages\IPython\core\interactiveshell.py:3457: DtypeWarning: Columns
(17,18,19,20,21,22,24,25,26,27,28,29,30,32,33,34,37,38,39,40,41,42,43) have
mixed types.Specify dtype option on import or set low_memory=False.
    exec(code_obj, self.user_global_ns, self.user_ns)

492
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0006_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0006_part_02.csv
15112
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0006_part_02.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0006_part_03.csv
32341
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0006_part_03.csv

```

Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0006_part_04.csv
18414
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0006_part_04.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0006_part_05.csv
6421
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0006_part_05.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0007_part_00.csv
2
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0007_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0007_part_01.csv
474
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0007_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0007_part_02.csv
15015
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0007_part_02.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0007_part_03.csv
31965
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0007_part_03.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0007_part_04.csv
18339
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0007_part_04.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0007_part_05.csv
6355
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0007_part_05.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0008_part_00.csv
7
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0008_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0008_part_01.csv
523

```

File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0008_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0008_part_02.csv
15310
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0008_part_02.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0008_part_03.csv
32159
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0008_part_03.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0008_part_04.csv
18391
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0008_part_04.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0008_part_05.csv

C:\Users\dxl902\AppData\Local\conda\conda\envs\an_env\lib\site-
packages\IPython\core\interactiveshell.py:3457: DtypeWarning: Columns
(21,22,24,25,26,27,28,29,30,32,33,34,37,38,39,40) have mixed types.Specify dtype
option on import or set low_memory=False.
    exec(code_obj, self.user_global_ns, self.user_ns)

6384
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0008_part_05.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0009_part_00.csv
1
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0009_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0009_part_01.csv
524
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0009_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0009_part_02.csv
15175
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0009_part_02.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0009_part_03.csv
32106
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0009_part_03.csv

```

```

Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0009_part_04.csv
18244
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0009_part_04.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0009_part_05.csv
6315
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0009_part_05.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_position/position_0010_part_00.csv

```

3 Aggregate Position File

```

[26]: files_position = []

import glob, os
os.chdir(r"D:\User.Data\Dan.
↳Li\Revelio\Processed_Data_PrivateFirms\Position_cleaned")
for file in glob.glob("*.pkl"):
#     print(file)
    files_position.append(file)
files_position[0:5]

```

```

[26]: ['position_0000_part_00.pkl',
'position_0000_part_01.pkl',
'position_0000_part_02.pkl',
'position_0000_part_03.pkl',
'position_0000_part_04.pkl']

```

```

[27]: output_dir = "D:/User.Data/Dan.Li/Revelio/Processed_Data_PrivateFirms/
↳Position_cleaned/"
finished = -1
df_positions = pd.DataFrame()

```

```

[28]: for i in range(0, len(files_position)):

    if i <= finished:
        continue
    output_full = output_dir + files_position[i]
    df1 = pd.read_pickle(output_full)
#     df1["rcid"] = df1["rcid_y"]

    df1 = df1[["user_id", "position_id", "company_cleaned",
               "region", "country", "state", "msa",
               "startdate", "enddate", "jobtitle_raw",

```

```

        "salary", "seniority",
        "rn", "rcid", "onet_code", "factset_entity_id"]

df_positions = df_positions.append(df1)
print("File finished: ", files_position[i])
finished = i

```

```

File finished: position_0000_part_00.pkl
File finished: position_0000_part_01.pkl
File finished: position_0000_part_02.pkl
File finished: position_0000_part_03.pkl
File finished: position_0000_part_04.pkl
File finished: position_0000_part_05.pkl
File finished: position_0001_part_00.pkl
File finished: position_0001_part_01.pkl
File finished: position_0001_part_02.pkl
File finished: position_0001_part_03.pkl
File finished: position_0001_part_04.pkl
File finished: position_0001_part_05.pkl
File finished: position_0002_part_00.pkl
File finished: position_0002_part_01.pkl
File finished: position_0002_part_02.pkl
File finished: position_0002_part_03.pkl
File finished: position_0002_part_04.pkl
File finished: position_0002_part_05.pkl
File finished: position_0003_part_00.pkl
File finished: position_0003_part_01.pkl
File finished: position_0003_part_02.pkl
File finished: position_0003_part_03.pkl
File finished: position_0003_part_04.pkl
File finished: position_0003_part_05.pkl
File finished: position_0004_part_00.pkl
File finished: position_0004_part_01.pkl
File finished: position_0004_part_02.pkl
File finished: position_0004_part_03.pkl
File finished: position_0004_part_04.pkl
File finished: position_0004_part_05.pkl
File finished: position_0005_part_00.pkl
File finished: position_0005_part_01.pkl
File finished: position_0005_part_02.pkl
File finished: position_0005_part_03.pkl
File finished: position_0005_part_04.pkl
File finished: position_0005_part_05.pkl
File finished: position_0006_part_00.pkl
File finished: position_0006_part_01.pkl
File finished: position_0006_part_02.pkl
File finished: position_0006_part_03.pkl
File finished: position_0006_part_04.pkl

```

File finished: position_0006_part_05.pkl
File finished: position_0007_part_00.pkl
File finished: position_0007_part_01.pkl
File finished: position_0007_part_02.pkl
File finished: position_0007_part_03.pkl
File finished: position_0007_part_04.pkl
File finished: position_0007_part_05.pkl
File finished: position_0008_part_00.pkl
File finished: position_0008_part_01.pkl
File finished: position_0008_part_02.pkl
File finished: position_0008_part_03.pkl
File finished: position_0008_part_04.pkl
File finished: position_0008_part_05.pkl
File finished: position_0009_part_00.pkl
File finished: position_0009_part_01.pkl
File finished: position_0009_part_02.pkl
File finished: position_0009_part_03.pkl
File finished: position_0009_part_04.pkl
File finished: position_0009_part_05.pkl
File finished: position_0010_part_00.pkl
File finished: position_0010_part_01.pkl
File finished: position_0010_part_02.pkl
File finished: position_0010_part_03.pkl
File finished: position_0010_part_04.pkl
File finished: position_0010_part_05.pkl
File finished: position_0011_part_00.pkl
File finished: position_0011_part_01.pkl
File finished: position_0011_part_02.pkl
File finished: position_0011_part_03.pkl
File finished: position_0011_part_04.pkl
File finished: position_0011_part_05.pkl
File finished: position_0012_part_00.pkl
File finished: position_0012_part_01.pkl
File finished: position_0012_part_02.pkl
File finished: position_0012_part_03.pkl
File finished: position_0012_part_04.pkl
File finished: position_0012_part_05.pkl
File finished: position_0013_part_00.pkl
File finished: position_0013_part_01.pkl
File finished: position_0013_part_02.pkl
File finished: position_0013_part_03.pkl
File finished: position_0013_part_04.pkl
File finished: position_0013_part_05.pkl
File finished: position_0014_part_00.pkl
File finished: position_0014_part_01.pkl
File finished: position_0014_part_02.pkl
File finished: position_0014_part_03.pkl
File finished: position_0014_part_04.pkl

File finished: position_0014_part_05.pkl
File finished: position_0015_part_00.pkl
File finished: position_0015_part_01.pkl
File finished: position_0015_part_02.pkl
File finished: position_0015_part_03.pkl
File finished: position_0015_part_04.pkl
File finished: position_0015_part_05.pkl
File finished: position_0016_part_00.pkl
File finished: position_0016_part_01.pkl
File finished: position_0016_part_02.pkl
File finished: position_0016_part_03.pkl
File finished: position_0016_part_04.pkl
File finished: position_0016_part_05.pkl
File finished: position_0017_part_00.pkl
File finished: position_0017_part_01.pkl
File finished: position_0017_part_02.pkl
File finished: position_0017_part_03.pkl
File finished: position_0017_part_04.pkl
File finished: position_0017_part_05.pkl
File finished: position_0018_part_00.pkl
File finished: position_0018_part_01.pkl
File finished: position_0018_part_02.pkl
File finished: position_0018_part_03.pkl
File finished: position_0018_part_04.pkl
File finished: position_0018_part_05.pkl
File finished: position_0019_part_00.pkl
File finished: position_0019_part_01.pkl
File finished: position_0019_part_02.pkl
File finished: position_0019_part_03.pkl
File finished: position_0019_part_04.pkl
File finished: position_0019_part_05.pkl
File finished: position_0020_part_00.pkl
File finished: position_0020_part_01.pkl
File finished: position_0020_part_02.pkl
File finished: position_0020_part_03.pkl
File finished: position_0020_part_04.pkl
File finished: position_0020_part_05.pkl
File finished: position_0021_part_00.pkl
File finished: position_0021_part_01.pkl
File finished: position_0021_part_02.pkl
File finished: position_0021_part_03.pkl
File finished: position_0021_part_04.pkl
File finished: position_0021_part_05.pkl
File finished: position_0022_part_00.pkl
File finished: position_0022_part_01.pkl
File finished: position_0022_part_02.pkl
File finished: position_0022_part_03.pkl
File finished: position_0022_part_04.pkl

File finished: position_0022_part_05.pkl
File finished: position_0023_part_00.pkl
File finished: position_0023_part_01.pkl
File finished: position_0023_part_02.pkl
File finished: position_0023_part_03.pkl
File finished: position_0023_part_04.pkl
File finished: position_0023_part_05.pkl
File finished: position_0024_part_00.pkl
File finished: position_0024_part_01.pkl
File finished: position_0024_part_02.pkl
File finished: position_0024_part_03.pkl
File finished: position_0024_part_04.pkl
File finished: position_0024_part_05.pkl
File finished: position_0025_part_00.pkl
File finished: position_0025_part_01.pkl
File finished: position_0025_part_02.pkl
File finished: position_0025_part_03.pkl
File finished: position_0025_part_04.pkl
File finished: position_0025_part_05.pkl
File finished: position_0026_part_00.pkl
File finished: position_0026_part_01.pkl
File finished: position_0026_part_02.pkl
File finished: position_0026_part_03.pkl
File finished: position_0026_part_04.pkl
File finished: position_0026_part_05.pkl
File finished: position_0027_part_00.pkl
File finished: position_0027_part_01.pkl
File finished: position_0027_part_02.pkl
File finished: position_0027_part_03.pkl
File finished: position_0027_part_04.pkl
File finished: position_0027_part_05.pkl

```
[31]: os.chdir(r"D:\User.Data\Dan.Li\Revelio")  
      df_positions.to_pickle("Processed_Data_PrivateFirms/df_positions.pkl")
```

```
[32]: df_positions["d_comp_user"] = 1
```

```
[33]: df_positions_simple = df_positions[["user_id", "d_comp_user"]]  
      df_positions_simple = df_positions_simple.drop_duplicates("user_id")
```

```
[34]: df_positions_simple.to_pickle("Processed_Data_PrivateFirms/comp_userid.pkl")
```

4 User

```
[1]: import pandas as pd
import os
os.chdir("D:/User.Data/Dan.Li/Revelio/")
userid_keep = pd.read_pickle("Processed_Data_PrivateFirms/comp_userid.pkl")
```

```
[5]: user_ids_to_keep = userid_keep['user_id'].tolist()
```

```
[3]: files_position = []

import glob, os
os.chdir("D:/User.Data/Dan.Li/Revelio/umiami_202212_user/")
for file in glob.glob("*.csv"):
    #     print(file)
    files_position.append(file)
files_position
```

```
[3]: ['user_0000_part_00.csv',
      'user_0000_part_01.csv',
      'user_0001_part_00.csv',
      'user_0001_part_01.csv',
      'user_0002_part_00.csv',
      'user_0002_part_01.csv',
      'user_0003_part_00.csv',
      'user_0003_part_01.csv',
      'user_0004_part_00.csv',
      'user_0004_part_01.csv',
      'user_0005_part_00.csv',
      'user_0005_part_01.csv',
      'user_0006_part_00.csv',
      'user_0006_part_01.csv',
      'user_0007_part_00.csv',
      'user_0007_part_01.csv',
      'user_0008_part_00.csv',
      'user_0008_part_01.csv',
      'user_0009_part_00.csv',
      'user_0009_part_01.csv',
      'user_0010_part_00.csv',
      'user_0010_part_01.csv',
      'user_0011_part_00.csv',
      'user_0011_part_01.csv',
      'user_0012_part_00.csv',
      'user_0012_part_01.csv',
      'user_0013_part_00.csv',
      'user_0013_part_01.csv',
      'user_0014_part_00.csv',
```

```

'user_0014_part_01.csv',
'user_0015_part_00.csv',
'user_0015_part_01.csv',
'user_0016_part_00.csv',
'user_0016_part_01.csv',
'user_0017_part_00.csv',
'user_0017_part_01.csv',
'user_0018_part_00.csv',
'user_0018_part_01.csv',
'user_0019_part_00.csv',
'user_0019_part_01.csv',
'user_0020_part_00.csv',
'user_0020_part_01.csv',
'user_0021_part_00.csv',
'user_0021_part_01.csv',
'user_0022_part_00.csv',
'user_0022_part_01.csv',
'user_0023_part_00.csv',
'user_0023_part_01.csv',
'user_0024_part_00.csv',
'user_0024_part_01.csv',
'user_0025_part_00.csv',
'user_0025_part_01.csv',
'user_0026_part_00.csv',
'user_0026_part_01.csv',
'user_0027_part_00.csv',
'user_0027_part_01.csv']

```

```

[6]: user_dir = "D:/User.Data/Dan.Li/Revelio/umiami_202212_user/"
output_dir = "D:/User.Data/Dan.Li/Revelio/Processed_Data_PrivateFirms/
↳User_cleaned/"
finished = -1

```

```

[7]: for i in range(0, len(files_position)):

    if i <= finished:
        continue
    file_i = files_position[i]

    file_full = user_dir + file_i
    filename = file_i.split(".")[0]
    output_full = output_dir + filename + ".pkl"

    print("Start to processs: ", file_full)

    df1 = pd.read_csv(file_full, encoding = "latin", on_bad_lines='skip',
↳lineterminator='\n',

```

```

dtype = {'user_id':str})
df1['user_id'] = df1['user_id'].apply(lambda x:"user_id_" + str(x))
df1 = df1[df1['user_id'].isin(user_ids_to_keep)]

df1.to_pickle(output_full)
print("File finished: ", file_full)
finished = i

```

Start to processs:

D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0000_part_00.csv

C:\Users\dxl902\AppData\Local\conda\conda\envs\an_env\lib\site-packages\IPython\core\interactiveshell.py:3457: DtypeWarning: Columns (4,5,6,7,8,9,10,11,12,20,21) have mixed types.Specify dtype option on import or set low_memory=False.

exec(code_obj, self.user_global_ns, self.user_ns)

File finished:

D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0000_part_00.csv

Start to processs:

D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0000_part_01.csv

C:\Users\dxl902\AppData\Local\conda\conda\envs\an_env\lib\site-packages\IPython\core\interactiveshell.py:3457: DtypeWarning: Columns (4,5,6,20) have mixed types.Specify dtype option on import or set low_memory=False.

exec(code_obj, self.user_global_ns, self.user_ns)

File finished:

D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0000_part_01.csv

Start to processs:

D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0001_part_00.csv

File finished:

D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0001_part_00.csv

Start to processs:

D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0001_part_01.csv

C:\Users\dxl902\AppData\Local\conda\conda\envs\an_env\lib\site-packages\IPython\core\interactiveshell.py:3457: DtypeWarning: Columns (4,5,6,7,8,9,10,12,20) have mixed types.Specify dtype option on import or set low_memory=False.

exec(code_obj, self.user_global_ns, self.user_ns)

File finished:

D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0001_part_01.csv

Start to processs:

D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0002_part_00.csv

File finished:

D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0002_part_00.csv

Start to processs:

D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0002_part_01.csv

```

C:\Users\dxl902\AppData\Local\conda\conda\envs\an_env\lib\site-
packages\IPython\core\interactiveshell.py:3457: DtypeWarning: Columns
(4,5,6,12,20) have mixed types.Specify dtype option on import or set
low_memory=False.
    exec(code_obj, self.user_global_ns, self.user_ns)

File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0002_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0003_part_00.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0003_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0003_part_01.csv

C:\Users\dxl902\AppData\Local\conda\conda\envs\an_env\lib\site-
packages\IPython\core\interactiveshell.py:3457: DtypeWarning: Columns
(4,5,12,20) have mixed types.Specify dtype option on import or set
low_memory=False.
    exec(code_obj, self.user_global_ns, self.user_ns)

File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0003_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0004_part_00.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0004_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0004_part_01.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0004_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0005_part_00.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0005_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0005_part_01.csv

C:\Users\dxl902\AppData\Local\conda\conda\envs\an_env\lib\site-
packages\IPython\core\interactiveshell.py:3457: DtypeWarning: Columns
(4,5,6,7,8,12,20) have mixed types.Specify dtype option on import or set
low_memory=False.
    exec(code_obj, self.user_global_ns, self.user_ns)

File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0005_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0006_part_00.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0006_part_00.csv

```

```

Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0006_part_01.csv

C:\Users\dxl902\AppData\Local\conda\conda\envs\an_env\lib\site-
packages\IPython\core\interactiveshell.py:3457: DtypeWarning: Columns
(4,5,6,7,12,20) have mixed types.Specify dtype option on import or set
low_memory=False.
    exec(code_obj, self.user_global_ns, self.user_ns)

File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0006_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0007_part_00.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0007_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0007_part_01.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0007_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0008_part_00.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0008_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0008_part_01.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0008_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0009_part_00.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0009_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0009_part_01.csv

C:\Users\dxl902\AppData\Local\conda\conda\envs\an_env\lib\site-
packages\IPython\core\interactiveshell.py:3457: DtypeWarning: Columns
(4,5,6,7,8,9,10,11,12,20) have mixed types.Specify dtype option on import or set
low_memory=False.
    exec(code_obj, self.user_global_ns, self.user_ns)

File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0009_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0010_part_00.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0010_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0010_part_01.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0010_part_01.csv

```

```

Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0011_part_00.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0011_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0011_part_01.csv

C:\Users\dxl902\AppData\Local\conda\conda\envs\an_env\lib\site-
packages\IPython\core\interactiveshell.py:3457: DtypeWarning: Columns
(4,5,6,7,8,9,12,20) have mixed types.Specify dtype option on import or set
low_memory=False.
    exec(code_obj, self.user_global_ns, self.user_ns)

File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0011_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0012_part_00.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0012_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0012_part_01.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0012_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0013_part_00.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0013_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0013_part_01.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0013_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0014_part_00.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0014_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0014_part_01.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0014_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0015_part_00.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0015_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0015_part_01.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0015_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0016_part_00.csv

```

File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0016_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0016_part_01.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0016_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0017_part_00.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0017_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0017_part_01.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0017_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0018_part_00.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0018_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0018_part_01.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0018_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0019_part_00.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0019_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0019_part_01.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0019_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0020_part_00.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0020_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0020_part_01.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0020_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0021_part_00.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0021_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0021_part_01.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0021_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0022_part_00.csv

```

File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0022_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0022_part_01.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0022_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0023_part_00.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0023_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0023_part_01.csv

C:\Users\dxl902\AppData\Local\conda\conda\envs\an_env\lib\site-
packages\IPython\core\interactiveshell.py:3457: DtypeWarning: Columns (4,12,20)
have mixed types.Specify dtype option on import or set low_memory=False.
    exec(code_obj, self.user_global_ns, self.user_ns)

File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0023_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0024_part_00.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0024_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0024_part_01.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0024_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0025_part_00.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0025_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0025_part_01.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0025_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0026_part_00.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0026_part_00.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0026_part_01.csv
File finished:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0026_part_01.csv
Start to processs:
D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0027_part_00.csv

C:\Users\dxl902\AppData\Local\conda\conda\envs\an_env\lib\site-
packages\IPython\core\interactiveshell.py:3457: DtypeWarning: Columns

```

(4,5,6,7,8,9,12,20,21) have mixed types.Specify dtype option on import or set low_memory=False.

```
exec(code_obj, self.user_global_ns, self.user_ns)
```

File finished:

D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0027_part_00.csv

Start to processs:

D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0027_part_01.csv

File finished:

D:/User.Data/Dan.Li/Revelio/umiami_202212_user/user_0027_part_01.csv

5 Aggregated User File

```
[9]: files = []

import glob, os
os.chdir(r"D:\User.Data\Dan.
↳Li\Revelio\Processed_Data_PrivateFirms\User_cleaned")
for file in glob.glob("*.pkl"):
#     print(file)
    files.append(file)
files[0:5]
```

```
[9]: ['user_0000_part_00.pkl',
      'user_0000_part_01.pkl',
      'user_0001_part_00.pkl',
      'user_0001_part_01.pkl',
      'user_0002_part_00.pkl']
```

```
[10]: output_dir = "D:/User.Data/Dan.Li/Revelio/Processed_Data_PrivateFirms/
↳User_cleaned/"
finished = -1
df_users = pd.DataFrame()
```

```
[11]: for i in range(0, len(files)):

    if i <= finished:
        continue
    output_full = output_dir + files[i]
    df1 = pd.read_pickle(output_full)

    df1 = df1.drop(columns=['native_prob', 'multiple_prob', 'prestige',
↳'highest_degree', 'updated_dt', 'f_prob', 'm_prob', 'white_prob',
↳'black_prob', "api_prob", "hispanic_prob",
        "user_location"])
    df_users = df_users.append(df1)
    print("File finished: ", files[i])
```

```
finished = i
```

```
File finished: user_0000_part_00.pkl
File finished: user_0000_part_01.pkl
File finished: user_0001_part_00.pkl
File finished: user_0001_part_01.pkl
File finished: user_0002_part_00.pkl
File finished: user_0002_part_01.pkl
File finished: user_0003_part_00.pkl
File finished: user_0003_part_01.pkl
File finished: user_0004_part_00.pkl
File finished: user_0004_part_01.pkl
File finished: user_0005_part_00.pkl
File finished: user_0005_part_01.pkl
File finished: user_0006_part_00.pkl
File finished: user_0006_part_01.pkl
File finished: user_0007_part_00.pkl
File finished: user_0007_part_01.pkl
File finished: user_0008_part_00.pkl
File finished: user_0008_part_01.pkl
File finished: user_0009_part_00.pkl
File finished: user_0009_part_01.pkl
File finished: user_0010_part_00.pkl
File finished: user_0010_part_01.pkl
File finished: user_0011_part_00.pkl
File finished: user_0011_part_01.pkl
File finished: user_0012_part_00.pkl
File finished: user_0012_part_01.pkl
File finished: user_0013_part_00.pkl
File finished: user_0013_part_01.pkl
File finished: user_0014_part_00.pkl
File finished: user_0014_part_01.pkl
File finished: user_0015_part_00.pkl
File finished: user_0015_part_01.pkl
File finished: user_0016_part_00.pkl
File finished: user_0016_part_01.pkl
File finished: user_0017_part_00.pkl
File finished: user_0017_part_01.pkl
File finished: user_0018_part_00.pkl
File finished: user_0018_part_01.pkl
File finished: user_0019_part_00.pkl
File finished: user_0019_part_01.pkl
File finished: user_0020_part_00.pkl
File finished: user_0020_part_01.pkl
File finished: user_0021_part_00.pkl
File finished: user_0021_part_01.pkl
File finished: user_0022_part_00.pkl
File finished: user_0022_part_01.pkl
```

```
File finished: user_0023_part_00.pkl  
File finished: user_0023_part_01.pkl  
File finished: user_0024_part_00.pkl  
File finished: user_0024_part_01.pkl  
File finished: user_0025_part_00.pkl  
File finished: user_0025_part_01.pkl  
File finished: user_0026_part_00.pkl  
File finished: user_0026_part_01.pkl  
File finished: user_0027_part_00.pkl  
File finished: user_0027_part_01.pkl
```

```
[13]: df_users = df_users.drop_duplicates("user_id")
```

```
[14]: os.chdir(r"D:\User.Data\Dan.Li\Revelio")  
df_users.to_pickle("Processed_Data_PrivateFirms/df_users_clean.pkl")
```