1.2 Data Collection - Collect company data from compustat

First, we setup our connection to the wrds database.

In [1]:
```
import wrds_tools
import parameters

wrds = wrds_tools.WrdsConnection(wrds_username=parameters.wrds_username)
```

```
/home/julian/PycharmProjects/corporate_disruptions/venv/lib/python3.6/site-pa
ckages/psycopg2/__init__.py:144: UserWarning: The psycopg2 wheel package will
be renamed from release 2.8; in order to keep installing from binary please u
se "pip install psycopg2-binary" instead. For details see: <http://initd.org/
psycopg/docs/install.html#binary-install-from-pypi>.
  """)

Loading library list...
Done
```

Setup observation period and grab the basic info we need.

In [2]:
```
from datetime import date

wrds.set_selection_period(start_date=date(year=1999, month=1, day=1),
                          end_date=date(year=2018, month=12, day=31))
wrds.build_sp500()
```

Add names and industry classification system GICS

In [3]:
```
wrds.add_names()
wrds.dataset['name'] = wrds.dataset['name'].str.title()

wrds.add_industry_classifiers(get_gics=True)

wrds.head(3)
```

Out[3]:

	gvkey	name	SIC	NAICS	GICS_group	GICS_industry	GICS_sector	GICS_subindustry
0	001013	Adc Telecommunications Inc	3661	334210	4520	452010	45	45201020
1	001045	American Airlines Group Inc	4512	481111	2030	203020	20	20302010
2	001075	Pinnacle West Capital Corp	4911	2211	5510	551010	55	55101010

1.2.1 Filter by industry

We use the GICS industry classification system to filter. For an overview, see https://en.wikipedia.org/wiki/Global_Industry_Classification_Standard (https://en.wikipedia.org/wiki/Global_Industry_Classification_Standard) . Industry Group 2550 is Retailing.

In [4]:
```
wrds.filter_by_industry(industry_code='2550', classification_system='GICS_grou
```

In [5]:
```
print('Number of observations: ', len(wrds.dataset),
      '\n\n----------------\n')
print(wrds.dataset['name'])
wrds.dataset.head()
```

```
Number of observations:  49

----------------

0                Best Buy Co Inc
1                  Officemax Inc
2         Circuit City Stores Inc
3                    Target Corp
4             Dillards Inc  -Cl A
5             Dollar General Corp
6            Family Dollar Stores
7                     Macy'S Inc
8                        Gap Inc
9                Genuine Parts Co
10                Home Depot Inc
11            Sears Holdings Corp
12                   L Brands Inc
13           Lowe'S Companies Inc
14       May Department Stores Co
15                  Nordstrom Inc
16                Penney (J C) Co
17     Pep Boys-Manny Moe & Jack
18                Autonation Inc
19                Ross Stores Inc
20             Sears Roebuck & Co
21                 Rs Legacy Corp
22                  Toys R Us Inc
23                Foot Locker Inc
24              Tjx Companies Inc
25                   Big Lots Inc
26                   Tiffany & Co
27              Office Depot Inc
28            Signet Jewelers Ltd
29                    Staples Inc
30                   Autozone Inc
31                    Kohl'S Corp
32          Bed Bath & Beyond Inc
33        O'Reilly Automotive Inc
34                   Petsmart Inc
35           Urban Outfitters Inc
36              Tractor Supply Co
37                Dollar Tree Inc
38    Abercrombie & Fitch  -Cl A
39                     Carmax Inc
40                 Amazon.Com Inc
41           Booking Holdings Inc
42              Expedia Group Inc
43                  Gamestop Corp
44         Advance Auto Parts Inc
45                    Netflix Inc
46                       Lkq Corp
47                Ulta Beauty Inc
48                Tripadvisor Inc
Name: name, dtype: object
```

Out[5]:

	gvkey	name	SIC	NAICS	GICS_group	GICS_industry	GICS_sector	GICS_subindustry
0	002184	Best Buy Co Inc	5731	443142	2550	255040	25	25504020

1.2.2 Filter out Internet & Direct Marketing Retail

Internet & Direct Marketing Retail is the GICS Sub-Industry 25502020.

In [6]:
```
wrds.dataset[wrds.dataset['GICS_subindustry'] == '25502020']
```

Out[6]:

	gvkey	name	SIC	NAICS	GICS_group	GICS_industry	GICS_sector	GICS_subindustry
40	064768	Amazon.Com Inc	5961	454111	2550	255020	25	25502020
41	119314	Booking Holdings Inc	7370	519130	2550	255020	25	25502020
42	126296	Expedia Group Inc	4700	561510	2550	255020	25	25502020
45	147579	Netflix Inc	7841	532230	2550	255020	25	25502020
48	199356	Tripadvisor Inc	7370	519130	2550	255020	25	25502020

In [7]:
```
wrds.dataset = wrds.dataset[wrds.dataset['GICS_subindustry'] != '25502020']
wrds.dataset = wrds.dataset.reset_index(drop=True)
wrds.dataset.head(3)
```

Out[7]:

	gvkey	name	SIC	NAICS	GICS_group	GICS_industry	GICS_sector	GICS_subindustry
0	002184	Best Buy Co Inc	5731	443142	2550	255040	25	25504020
1	002290	Officemax Inc	5110	424120	2550	255040	25	25504040
2	003054	Circuit City Stores Inc	5731	443112	2550	255040	25	25504020

1.3 Add executives data

Set observation period.

In [8]:
```
wrds.set_observation_period(start_date=date(year=2009, month=1, day=1),
                            end_date=date(year=2018, month=12, day=31))
```

In [9]:
```
wrds.add_executives()

wrds.head()
```

Out[9]:

	gvkey	name	SIC	NAICS	GICS_group	GICS_industry	GICS_sector	GICS_subindustry	execid	yea
0	002184	Best Buy Co Inc	5731	443142	2550	255040	25	25504020	06175	200
1	002184	Best Buy Co Inc	5731	443142	2550	255040	25	25504020	28397	200
2	002184	Best Buy Co Inc	5731	443142	2550	255040	25	25504020	28397	201
3	002184	Best Buy Co Inc	5731	443142	2550	255040	25	25504020	28397	201
4	002184	Best Buy Co Inc	5731	443142	2550	255040	25	25504020	28397	201

In [10]:
```
len(wrds.dataset.execid.unique())
```

Out[10]: 521

1.4 Output dataframe as .feather

The .feather format allows us to further manipulate the dataframe in R without having to specify column types again.

In [11]:
```
selection = wrds.return_dataframe()
```

We transform the year column to datatype 'object' (character strings) because feather cannot handle pandas 'Int64' datatype yet.

In [12]:
```
selection['year'] = selection['year'].astype('object')
```

In [13]:
```
from datetime import date

selection.to_feather('sample {}.feather'.format(str(date.today())))
```