Loading data

In [1]:

```
import sys
import pandas as pd
import numpy as np
df = pd.read_csv('/Users/xingzhangzhu/Downloads/rpsdata_rfs.csv')
```

In [43]:

```
features = 'absacc acc aeavol age agr baspread beta betasq bm bm_ia cash cashdeb
t cashpr cfp cfp_ia chatoia chcsho chempia chinv chmom chpmia chtx cinvest convi
nd currat depr divi divo dolvol dy ear egr ep gma grcapx grltnoa herf hire idiov
ol ill indmom invest lev lgr maxret mom12m mom1m mom36m mom6m ms mve mve_ia ninc
r operprof orgcap pchcapx_ia pchcurrat pchdepr pchgm_pchsale pchquick pchsale_pc
hinvt pchsale_pchrect pchsale_pchxsga pchsaleinv pctacc pricedelay ps quick rd r
d_mve rd_sale realestate retvol roaq roavol roeq roic rsup salecash saleinv sale
rec secured securedind sgr sin sp std_dolvol std_turn stdacc stdcf tang tb turn
 zerotrade'
keep94 = features.split(" ")
# add 'DATE' for grouping purposes
keep94.extend(['permno', 'DATE', 'RET'])
# create the same dataframe used in the paper
df_94 = df[keep94]
df_94.head()
```

Out[43]:

	absacc	acc	aeavol	age	agr	baspread	beta	betasq	bm	
0	0.113203	0.113203	1.001090	4	0.173800	0.014234	1.060413	1.124477	1.180962	0.
1	0.084369	0.084369	-0.613146	4	0.078283	0.022470	1.526010	2.328707	0.956692	-0.
2	NaN	NaN	NaN	1	NaN	0.511667	1.759512	3.095882	3.362003	1.
3	0.025227	0.025227	-0.491307	4	0.126033	0.020899	0.492884	0.242935	1.330341	0.
4	0.076974	0.076974	-0.256932	4	0.123875	0.016947	1.139157	1.297678	1.579284	0.

5 rows × 97 columns

Replace nan with cross-sectional median

In [44]:

```
# 数据清洗过程中发现，'orgcap', 'realestate', 'secured'
# 这三个因子在某些月份全是空值（因此无中位数），故弃之。
factors_to_drop = []
df_94_by_date = df_94.groupby('DATE')
for date, date_df in df_94_by_date:
        for index, value in enumerate(date_df.median()):
            if np.isnan(value):
                # 寻找任何median是NaN的index lable
                factors_to_drop.append(date_df.median().index[index])

print("Factors to drop:\n", set(factors_to_drop))
```

```
Factors to drop:
 {'realestate', 'orgcap', 'secured'}
```

In [45]:

```
# 查看这三个即将被drop的因子"非NaN"值的百分比
for factor in ['secured', 'orgcap', 'realestate']:
    print('{}: {:.2%}'.format(factor, (len(df_94[factor]) - df_94[factor].count
()) / len(df_94[factor])))
```

```
secured: 41.23%
orgcap: 27.96%
realestate: 57.71%
```

In [48]:

```
# Fill NaN with median
def fill_NaN_with_Median(df):
    return df.drop('DATE', 1).fillna(value=df.median())
```

In [49]:

```
# drop above factors
df_91 = df_94.drop(['secured', 'orgcap', 'realestate'], axis=1)
# group by 'DATE'
df_91_by_date = df_91.groupby('DATE')
# Fill NaN with median and return a multi-indexed dataframe
df_fiiled = df_91_by_date.apply(fill_NaN_with_Median).reset_index(level='DATE')
df_fiiled.head()
```

Out[49]:

	DATE	absacc	acc	aeavol	age	agr	baspread	beta	betasq	
0	19800131	0.113203	0.113203	1.001090	4	0.173800	0.014234	1.060413	1.124477	1.1
1	19800131	0.084369	0.084369	-0.613146	4	0.078283	0.022470	1.526010	2.328707	0.9
2	19800131	0.080049	0.076017	0.046328	1	0.136066	0.511667	1.759512	3.095882	3.3
3	19800131	0.025227	0.025227	-0.491307	4	0.126033	0.020899	0.492884	0.242935	1.3
4	19800131	0.076974	0.076974	-0.256932	4	0.123875	0.016947	1.139157	1.297678	1.5

5 rows × 94 columns

In [161]:

```
def shift_stock(df_stock):
    global n
    n += 1
    # num_stock = len(df['permno'].unique()) == 18700
    b = ("{:.2%}".format(n / 18700))
    sys.stdout.write('\r' + b)
    sys.stdout.flush()
    for k, v in frequency.items():
        if v == 'Annual':
            df_stock[k] = df_stock[k].shift(6)
        elif v == 'Quarterly':
            df_stock[k] = df_stock[k].shift(4)
        elif v == 'Monthly':
            df_stock[k] = df_stock[k].shift(1)
    return df_stock
```

Shift data based on frequency

In [198]:

```
table = """1 absacc Absolute accruals Bandyopadhyay, Huang & Wirjanto 2010, WP C
ompustat Annual
2 acc Working capital accruals Sloan 1996, TAR Compustat Annual
3 aeavol Abnormal earnings announcement volume Lerman, Livnat & Mendenhall 2007,
WP Compustat+CRSP Quarterly
4 age # years since rst Compustat coverage Jiang, Lee & Zhang 2005, RAS Compusta
t Annual
5 agr Asset growth Cooper, Gulen & Schill 2008, JF Compustat Annual
6 baspread Bid-ask spread Amihud & Mendelson 1989, JF CRSP Monthly
7 beta Beta Fama & MacBeth 1973, JPE CRSP Monthly
8 betasq Beta squared Fama & MacBeth 1973, JPE CRSP Monthly
9 bm Book-to-market Rosenberg, Reid & Lanstein 1985, JPM Compustat+CRSP Annual
10 bm_ia Industry-adjusted book to market Asness, Porter & Stevens 2000, WP Comp
ustat+CRSP Annual
11 cash Cash holdings Palazzo 2012, JFE Compustat Quarterly
12 cashdebt Cash ow to debt Ou & Penman 1989, JAE Compustat Annual
13 cashpr Cash productivity Chandrashekar & Rao 2009, WP Compustat Annual
14 cfp Cash ow to price ratio Desai, Rajgopal & Venkatachalam 2004, TAR Compusta
t Annual
15 cfp_ia Industry-adjusted cash ow to price ratio Asness, Porter & Stevens 200
0, WP Compustat Annual
16 chatoia Industry-adjusted change in asset turnover Soliman 2008, TAR Compusta
t Annual
17 chcsho Change in shares outstanding Ponti & Woodgate 2008, JF Compustat Annua
l
18 chempia Industry-adjusted change in employees Asness, Porter & Stevens 1994,
 WP Compustat Annual
19 chinv Change in inventory Thomas & Zhang 2002, RAS Compustat Annual
20 chmom Change in 6-month momentum Gettleman & Marks 2006, WP CRSP Monthly
21 chpmia Industry-adjusted change in prot margin Soliman 2008, TAR Compustat An
nual
22 chtx Change in tax expense Thomas & Zhang 2011, JAR Compustat Quarterly
23 cinvest Corporate investment Titman, Wei & Xie 2004, JFQA Compustat Quarterly
24 convind Convertible debt indicator Valta 2016, JFQA Compustat Annual
25 currat Current ratio Ou & Penman 1989, JAE Compustat Annual
26 depr Depreciation / PP&E Holthausen & Larcker 1992, JAE Compustat Annual
27 divi Dividend initiation Michaely, Thaler & Womack 1995, JF Compustat Annual
28 divo Dividend omission Michaely, Thaler & Womack 1995, JF Compustat Annual
29 dolvol Dollar trading volume Chordia, Subrahmanyam & Anshuman 2001, JFE CRSP
 Monthly
30 dy Dividend to price Litzenberger & Ramaswamy 1982, JF Compustat Annual
31 ear Earnings announcement return Kishore, Brandt, Santa-Clara & Venkatachalam
2008, WP Compustat+CRSP Quarterly
32 egr Growth in common shareholder equity Richardson, Sloan, Soliman & Tuna 200
5, JAE Compustat Annual
33 ep Earnings to price Basu 1977, JF Compustat Annual
34 gma Gross protability Novy-Marx 2013, JFE Compustat Annual
35 grcapx Growth in capital expenditures Anderson & Garcia-Feijoo 2006, JF Compu
stat Annual
36 grltnoa Growth in long term net operating assets Faireld, Whisenant & Yohn 20
03, TAR Compustat Annual
37 herf Industry sales concentration Hou & Robinson 2006, JF Compustat Annual
38 hire Employee growth rate Bazdresch, Belo & Lin 2014, JPE Compustat Annual
39 idiovol Idiosyncratic return volatility Ali, Hwang & Trombley 2003, JFE CRSP
 Monthly
40 ill Illiquidity Amihud 2002, JFM CRSP Monthly
41 indmom Industry momentum Moskowitz & Grinblatt 1999, JF CRSP Monthly
42 invest Capital expenditures and inventory Chen & Zhang 2010, JF Compustat Ann
ual
```

```
43 lev Leverage Bhandari 1988, JF Compustat Annual
44 lgr Growth in long-term debt Richardson, Sloan, Soliman & Tuna 2005, JAE Comp
ustat Annual
45 maxret Maximum daily return Bali, Cakici & Whitelaw 2011, JFE CRSP Monthly
46 mom12m 12-month momentum Jegadeesh 1990, JF CRSP Monthly
47 mom1m 1-month momentum Jegadeesh & Titman 1993, JF CRSP Monthly
48 mom36m 36-month momentum Jegadeesh & Titman 1993, JF CRSP Monthly
49 mom6m 6-month momentum Jegadeesh & Titman 1993, JF CRSP Monthly
50 ms Financial statement score Mohanram 2005, RAS Compustat Quarterly
51 mve_ia Size Banz 1981, JFE CRSP Monthly
52 mve ia Industry-adjusted size Asness, Porter & Stevens 2000, WP Compustat Ann
ual
53 nincr Number of earnings increases Barth, Elliott & Finn 1999, JAR Compustat
 Quarterly
54 operprof Operating protability Fama & French 2015, JFE Compustat Annual
55 orgcap Organizational capital Eisfeldt & Papanikolaou 2013, JF Compustat Annu
al
56 pchcapx_ia Industry adjusted % change in capital expenditures Abarbanell & Bu
shee 1998, TAR Compustat Annual
57 pchcurrat % change in current ratio Ou & Penman 1989, JAE Compustat Annual
58 pchdepr % change in depreciation Holthausen & Larcker 1992, JAE Compustat Ann
ual
59 pchgm_pchsale % change in gross margin - % change in sales Abarbanell & Bushe
e 1998, TAR Compustat Annual
60 pchquick % change in quick ratio Ou & Penman 1989, JAE Compustat Annual
61 pchsale_pchinvt % change in sales - % change in inventory Abarbanell & Bushee
1998, TAR Compustat Annual
62 pchsale_pchrect % change in sales - % change in A/R Abarbanell & Bushee 1998,
TAR Compustat Annual
63 pchsale_pchxsga % change in sales - % change in SG&A Abarbanell & Bushee 199
8, TAR Compustat Annual
64 pchsaleinv % change sales-to-inventory Ou & Penman 1989, JAE Compustat Annual
65 pctacc Percent accruals Hafzalla, Lundholm & Van Winkle 2011, TAR Compustat A
nnual
66 pricedelay Price delay Hou & Moskowitz 2005, RFS CRSP Monthly
67 ps Financial statements score Piotroski 2000, JAR Compustat Annual
68 quick Quick ratio Ou & Penman 1989, JAE Compustat Annual
69 rd R&D increase Eberhart, Maxwell & Siddique 2004, JF Compustat Annual
70 rd_mve R&D to market capitalization Guo, Lev & Shi 2006, JBFA Compustat Annua
l
71 rd_sale R&D to sales Guo, Lev & Shi 2006, JBFA Compustat Annual
72 realestate Real estate holdings Tuzel 2010, RFS Compustat Annual
73 retvol Return volatility Ang, Hodrick, Xing & Zhang 2006, JF CRSP Monthly
74 roaq Return on assets Balakrishnan, Bartov & Faurel 2010, JAE Compustat Quart
erly
75 roavol Earnings volatility Francis, LaFond, Olsson & Schipper 2004, TAR Compu
stat Quarterly
76 roeq Return on equity Hou, Xue & Zhang 2015, RFS Compustat Quarterly
77 roic Return on invested capital Brown & Rowe 2007, WP Compustat Annual
78 rsup Revenue surprise Kama 2009, JBFA Compustat Quarterly
79 salecash Sales to cash Ou & Penman 1989, JAE Compustat Annual
80 saleinv Sales to inventory Ou & Penman 1989, JAE Compustat Annual
81 salerec Sales to receivables Ou & Penman 1989, JAE Compustat Annual
82 secured Secured debt Valta 2016, JFQA Compustat Annual
83 securedind Secured debt indicator Valta 2016, JFQA Compustat Annual
84 sgr Sales growth Lakonishok, Shleifer & Vishny 1994, JF Compustat Annual
85 sin Sin stocks Hong & Kacperczyk 2009, JFE Compustat Annual
86 sp Sales to price Barbee, Mukherji, & Raines 1996, FAJ Compustat Annual
87 std_dolvol Volatility of liquidity (dollar trading volume) Chordia, Subrahman
yam & Anshuman 2001, JFE CRSP Monthly
88 std_turn Volatility of liquidity (share turnover) Chordia, Subrahmanyam, &Ans
```

```
human 2001, JFE CRSP Monthly
89 stdacc Accrual volatility Bandyopadhyay, Huang & Wirjanto 2010, WP Compustat
 Quarterly
90 stdcf Cash ow volatility Huang 2009, JEF Compustat Quarterly
91 tang Debt capacity/rm tangibility Almeida & Campello 2007, RFS Compustat Annu
al
92 tb Tax income to book income Lev & Nissim 2004, TAR Compustat Annual
93 turn Share turnover Datar, Naik & Radclie 1998, JFM CRSP Monthly
94 zerotrade Zero trading days Liu 2006, JFE CRSP Monthly"""

frequency = {}

for line in table.split('\n'):
    if line.split()[1] not in ['realestate', 'orgcap', 'secured']:
        frequency[line.split()[1]] = line.split()[-1]

len(frequency)
```

Out[198]:

```
91
```

In [70]:

```
df_fiiled_by_stock = df_fiiled.groupby('permno')
```

In [172]:

```
df_list = []
n = 0
c = 0
for stock_code, stock_df in df_fiiled_by_stock:
    df_list.append(shift_stock(stock_df.copy()))
```

```
100.00%
```

In [196]:

```
# concatenate into a finalized dataframe
shift = pd.concat(df_list)
shift.head()
```

Out[196]:

	DATE	absacc	acc	aeavol	age	agr	baspread	beta	betasq	bm	.
351579	19870331	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	.
355995	19870430	NaN	NaN	NaN	NaN	NaN	0.029304	-0.060573	0.003669	NaN	.
360391	19870529	NaN	NaN	NaN	NaN	NaN	0.042022	-0.159291	0.025374	NaN	.
364797	19870630	NaN	NaN	NaN	NaN	NaN	0.037871	-0.172907	0.029897	NaN	.
369236	19870731	NaN	NaN	2.484184	NaN	NaN	0.055419	-0.149683	0.022405	NaN	.

5 rows × 93 columns

Test if the shifted & filled dataframe is what we wanted

In [195]:

```
# 'baspread' is updated monthly
# the data in row 32072 was shifted 1 month back to row 28319 as expected
print(shift.loc[[28319, 32072]][['baspread', 'std_dolvol']])
print()
print(df.loc[[28319, 32072]][['baspread', 'std_dolvol']])
```

```
       baspread  std_dolvol
28319       NaN         NaN
32072  0.016031    0.799338

       baspread  std_dolvol
28319  0.016031         NaN
32072  0.017000         NaN
```

In [188]:

```
# test median fill nan
# In the month of 1980/09/30, the median value of 'std_dolvol' is 0.79933817045
# this number replaced the nan value as expected in row 32072
df_91_by_date.get_group(19800930)['std_dolvol'].median()
```

Out[188]:

```
0.79933817045
```

In [ ]:

```

```