In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
```

In [2]:
```
#Load major_groups.csv

df = pd.read_csv("data/major_groups.csv", encoding = 'utf-8')

#Load public firms.csv
firms = pd.read_csv("data/public_firms.csv", encoding = 'utf-8')
```

In [3]:
```
df

#Industry chosen is Apparel and Accessory Stores.
#The corresponding code is 56 and the Index is 48
```

Out[3]:

	major_group	description
0	1	Agricultural Production Crops
1	2	Agriculture production livestock and animal sp...
2	7	Agricultural Services
3	8	Forestry
4	9	Fishing hunting and trapping
...	...	...
78	94	Administration Of Human Resource Programs
79	95	Administration Of Environmental Quality And Ho...
80	96	Administration Of Economic Programs
81	97	National Security And International Affairs
82	99	Nonclassifiable Establishments

83 rows × 2 columns

In [4]:
```
#Part 1

# Filter rows where the first two digits of 'sic' are 56. There are only 4 c
df_56 = firms[firms['sic'] // 100 == 56]

#Summary of the dataset
df_56.describe()
```

```
print(len(df_56))
```

1367

In [ ]:
```
df_56
```

Out[ ]:

	gvkey	fyear	location	conm	ipodate	sic	prcc_c	ch	
5923	2484	1995	USA	BURLINGTON COAT FACTORY INVS	NaN	5651	10.2500	14.520	1
5924	2484	1996	USA	BURLINGTON COAT FACTORY INVS	NaN	5651	13.0000	73.560	2
5925	2484	1997	USA	BURLINGTON COAT FACTORY INVS	NaN	5651	16.4370	157.394	5
5926	2484	1998	USA	BURLINGTON COAT FACTORY INVS	NaN	5651	16.3125	106.952	4
5927	2484	1999	USA	BURLINGTON COAT FACTORY INVS	NaN	5651	13.8750	127.818	6
...	...	...	...	...	...	...	...	...	
203507	187575	2016	USA	TILLY'S INC	2012/05/04	5600	13.1900	78.994	1
203508	187575	2017	USA	TILLY'S INC	2012/05/04	5600	14.7600	53.202	1
203509	187575	2018	USA	TILLY'S INC	2012/05/04	5600	10.8600	68.160	2
203510	187575	2019	USA	TILLY'S INC	2012/05/04	5600	12.2500	70.137	2
203511	187575	2020	USA	TILLY'S INC	2012/05/04	5600	8.1600	76.184	-

1367 rows × 12 columns

In [5]:
```
#Part 1

#number of unique firm year
nfirmyear = len(df_56['fyear'].unique())
```

```
print(f"The number of unique firm year is: {nfirmyear}")

#number of unique company names
ncompany = len(df_56['conm'].unique())
print(f"The number of unique company names is: {ncompany}")
```

```
The number of unique firm year is: 27
The number of unique company names is: 105
```

In [7]:
```
#Part 1

#Group the companies by their company names
df_56_cnt = df_56.groupby('conm').nunique()
df_56_cnt
```

Out[7]:

conm	gvkey	fyear	location	ipodate	sic	prcc_c	ch	ni	asset	sale	roa
ABERCROMBIE & FITCH -CL A	1	26	1	1	1	25	26	26	26	26	26
AMERN EAGLE OUTFITTERS INC	1	26	1	1	1	26	26	26	26	26	26
ANN INC	1	21	1	1	1	21	21	21	21	21	21
ARO LIQUIDATION INC	1	16	1	0	1	14	16	16	16	16	16
ASCENA RETAIL GROUP INC	1	26	1	0	1	26	26	26	26	26	26
...	...	...	...	...	...	...	...	...	...	...	...
VICTORIAS SECRET AND CO	1	2	1	1	1	0	2	2	2	2	2
WALKING CO HOLDINGS INC	1	13	1	1	1	11	13	13	13	13	13
WET SEAL INC	1	20	1	0	1	20	20	20	20	20	20
WHITE HOUSE INC-REDH	1	2	1	0	1	0	2	2	2	2	2
ZUMIEZ INC	1	18	1	1	1	16	18	18	18	18	18

105 rows × 11 columns

In [8]:
```
#Part 1

#number of companies with full record
full_record = df_56_cnt[df_56_cnt['fyear'] == 27]
len(full_record)
print(f"The number of companies with full record is: {len(full_record)}")
full_record
```

```
The number of companies with full record is: 11
```

Out[8]:

conm	gvkey	fyear	location	ipodate	sic	prcc_c	ch	ni	asset	sale	roa
BATH & BODY WORKS INC	1	27	1	0	1	27	27	27	27	27	27
BUCKLE INC	1	27	1	1	1	27	27	27	27	27	27
CATO CORP - CL A	1	27	1	0	1	27	27	27	27	27	27
DESTINATION XL GROUP INC	1	27	1	1	1	26	25	27	27	27	27
FOOT LOCKER INC	1	27	1	0	1	27	27	24	27	27	27
GAP INC	1	27	1	0	1	27	26	26	27	27	27
GENESCO INC	1	27	1	0	1	27	27	27	27	27	27
NORDSTROM INC	1	27	1	0	1	27	27	27	27	27	27
ROSS STORES INC	1	27	1	0	1	27	27	27	27	27	27
TJX COS INC (THE)	1	27	1	0	1	27	27	27	27	27	27
URBAN OUTFITTERS INC	1	27	1	1	1	27	27	27	27	27	27

In [9]:
```
df_56['fyear'].unique()

full_record
```

Out[9]:

conm	gvkey	fyear	location	ipodate	sic	prcc_c	ch	ni	asset	sale	roa
BATH & BODY WORKS INC	1	27	1	0	1	27	27	27	27	27	27
BUCKLE INC	1	27	1	1	1	27	27	27	27	27	27
CATO CORP - CL A	1	27	1	0	1	27	27	27	27	27	27
DESTINATION XL GROUP INC	1	27	1	1	1	26	25	27	27	27	27
FOOT LOCKER INC	1	27	1	0	1	27	27	24	27	27	27
GAP INC	1	27	1	0	1	27	26	26	27	27	27
GENESCO INC	1	27	1	0	1	27	27	27	27	27	27
NORDSTROM INC	1	27	1	0	1	27	27	27	27	27	27
ROSS STORES INC	1	27	1	0	1	27	27	27	27	27	27
TJX COS INC (THE)	1	27	1	0	1	27	27	27	27	27	27
URBAN OUTFITTERS INC	1	27	1	1	1	27	27	27	27	27	27

In [10]:
```
#Part 1B Question 1 and 2

# stock price column name: prcc_c
# sales column name: sale
# geographical distribution column name: location)

#top 10 firm with highest stock price in 2020
top_10_stock_price_2020 = df_56[df_56['fyear'] ==2020].nlargest(10, 'prcc_c'
top_10_stock_price_2020
print(f"top 10 stock price is: {top_10_stock_price_2020['conm'].unique()}")


#top 10 firm with highest sales in all of history

df_56_sum = df_56.groupby('conm')['sale'].sum()
df_56_sum = df_56_sum.reset_index() #converts series into dataframe
```

```
top_10_sales = df_56_sum.nlargest(10,"sale")
print(f"top 10 sales is: {top_10_sales['conm'].unique()}")
```

```
top 10 stock price is: ['BURLINGTON STORES INC' 'ROSS STORES INC' 'TJX COS I
NC (THE)'
 'CHILDRENS PLACE INC' 'CITI TRENDS INC' 'BOOT BARN HOLDINGS INC'
 'FOOT LOCKER INC' 'SHOE CARNIVAL INC' 'BATH & BODY WORKS INC'
 'ZUMIEZ INC']
top 10 sales is: ['TJX COS INC (THE)' 'GAP INC' 'BATH & BODY WORKS INC' 'NOR
DSTROM INC'
 'ROSS STORES INC' 'FOOT LOCKER INC' 'ABERCROMBIE & FITCH  -CL A'
 'ASCENA RETAIL GROUP INC' 'AMERN EAGLE OUTFITTERS INC'
 'DESIGNER BRANDS INC']
```

In [11]:
```
top_10_stock_price_2020
```

Out[11]:

	gvkey	fyear	location	conm	ipodate	sic	prcc_c	ch
68733	18675	2020	USA	BURLINGTON STORES INC	2013/10/02	5600	261.55	1380.276
33774	9248	2020	USA	ROSS STORES INC	NaN	5651	122.81	4819.293
43543	11672	2020	USA	TJX COS INC (THE)	NaN	5651	68.29	10469.570
141569	65430	2020	USA	CHILDRENS PLACE INC	1997/09/19	5600	50.10	63.548
184708	163051	2020	USA	CITI TRENDS INC	2005/05/18	5651	49.68	123.177
76291	21898	2020	USA	BOOT BARN HOLDINGS INC	2014/10/30	5661	43.36	73.148
43244	11584	2020	USA	FOOT LOCKER INC	NaN	5661	40.44	1680.000
94181	27938	2020	USA	SHOE CARNIVAL INC	1993/03/16	5661	39.18	106.532
23067	6733	2020	USA	BATH & BODY WORKS INC	NaN	5600	37.19	3903.000
184613	162988	2020	USA	ZUMIEZ INC	2005/05/06	5651	36.78	73.622

In [12]:
```
top_10_sales
```

Out[12]:

	conm	sale
92	TJX COS INC (THE)	531354.915
54	GAP INC	362527.300
8	BATH & BODY WORKS INC	274942.175
74	NORDSTROM INC	248159.506
80	ROSS STORES INC	188529.105
47	FOOT LOCKER INC	167706.000
0	ABERCROMBIE & FITCH -CL A	67874.646
4	ASCENA RETAIL GROUP INC	65366.513
1	AMERN EAGLE OUTFITTERS INC	63138.850
33	DESIGNER BRANDS INC	57096.129

In [13]:
```
#Part 1B Question 3

#How many firms in each location
top_10_location = df_56.groupby(['location','conm']).count()
top_10_location

unique_companies_per_location = df_56.groupby('location')['conm'].nunique()
unique_companies_per_location
```

Out[13]:
```
location
CAN      1
USA    104
Name: conm, dtype: int64
```

In [14]:
```
#Part 1B Question 4

pc = df_56.groupby('fyear').prcc_c.mean()

#Plot the chart
plt.plot(pc)
plt.title("Average Stock Price of Apparel and Accessory Stores Industry from
plt.xlabel("Year")
plt.ylabel("Average stock price")
plt.grid(True)
plt.show()
```

In [15]:
```
#Part 1B Question 5

#filter 2007 observations
df_2007 = df_56[df_56['fyear']==2007][['conm','prcc_c']].rename(columns={'pr
df_2007

#filter 2008 observations
df_2008 = df_56[df_56['fyear']==2008][['conm','prcc_c']].rename(columns={'pr
df_2008

#Join both the dataset
df_2007_2008 = df_2007.merge(df_2008, on = 'conm', how= 'inner')
df_2007_2008

#Remove all the null values
df_2007_2008 = df_2007_2008.dropna()
df_2007_2008

#Calculate the percentage change from 2007 to 2008 and sort to reflect the l
df_2007_2008['diff'] = ((df_2007_2008['prcc_c_2008'] - df_2007_2008['prcc_c_

df_2007_2008.sort_values(by = "diff")
```

Out[15]:

	conm	prcc_c_2007	prcc_c_2008	diff

44	EDDIE BAUER HOLDINGS INC	6.35	0.51	-91.968504
12	DESTINATION XL GROUP INC	5.18	0.52	-89.961390
22	PACIFIC SUNWEAR CALIF INC	14.11	1.59	-88.731396
36	TWEEN BRANDS INC	26.48	4.32	-83.685801
26	TALBOTS INC	11.82	2.39	-79.780034
1	CACHE INC	9.34	2.02	-78.372591
13	ANN INC	25.56	5.77	-77.425665
18	STEIN MART INC	4.74	1.13	-76.160338
39	BAKERS FOOTWEAR GROUP INC	2.30	0.57	-75.217391
34	J CREW GROUP INC	48.21	12.20	-74.694047
30	ABERCROMBIE & FITCH -CL A	79.97	23.07	-71.151682
41	ZUMIEZ INC	24.36	7.45	-69.417077
8	NORDSTROM INC	36.73	13.31	-63.762592
40	RTW RETAILWINDS INC	6.38	2.32	-63.636364
37	CHARLOTTE RUSSE HOLDING INC	16.15	6.49	-59.814241
32	COLDWATER CREEK INC	6.69	2.85	-57.399103
6	GENESCO INC	37.80	16.92	-55.238095
27	AMERN EAGLE OUTFITTERS INC	20.77	9.36	-54.935002
3	CHARMING SHOPPES INC	5.41	2.44	-54.898336
21	DESTINATION MATERNITY CORP	17.40	7.85	-54.885057
24	CHICOS FAS INC	9.03	4.18	-53.709856
16	CHRISTOPHER & BANKS CORP	11.45	5.60	-51.091703
17	TAILORED BRANDS INC	26.98	13.54	-49.814678
7	BATH & BODY WORKS INC	18.93	10.04	-46.962493
10	FOOT LOCKER INC	13.66	7.34	-46.266471
25	URBAN OUTFITTERS INC	27.26	14.98	-45.047689
31	STAGE STORES INC	14.80	8.25	-44.256757
38	ARO LIQUIDATION INC	26.50	16.10	-39.245283
5	GAP INC	21.28	13.39	-37.077068
19	BUCKLE INC	33.00	21.82	-33.878788

43	DSW INC-OLD	18.76	12.46	-33.582090
28	COLLECTIVE BRANDS INC	17.39	11.72	-32.604945
23	SHOE CARNIVAL INC	14.11	9.55	-32.317505
15	DESIGNER BRANDS INC	5.09	3.47	-31.827112
11	TJX COS INC (THE)	28.73	20.57	-28.402367
35	DELIAS INC	2.71	2.20	-18.819188
33	CHILDRENS PLACE INC	25.93	21.68	-16.390282
4	ASCENA RETAIL GROUP INC	12.51	10.74	-14.148681
42	CITI TRENDS INC	15.44	14.72	-4.663212
2	CATO CORP -CL A	15.66	15.10	-3.575990
9	ROSS STORES INC	25.57	29.73	16.269065
14	WET SEAL INC	2.33	2.97	27.467811
29	HOT TOPIC INC	5.82	9.27	59.278351
20	FINISH LINE INC -CL A	2.42	5.60	131.404959

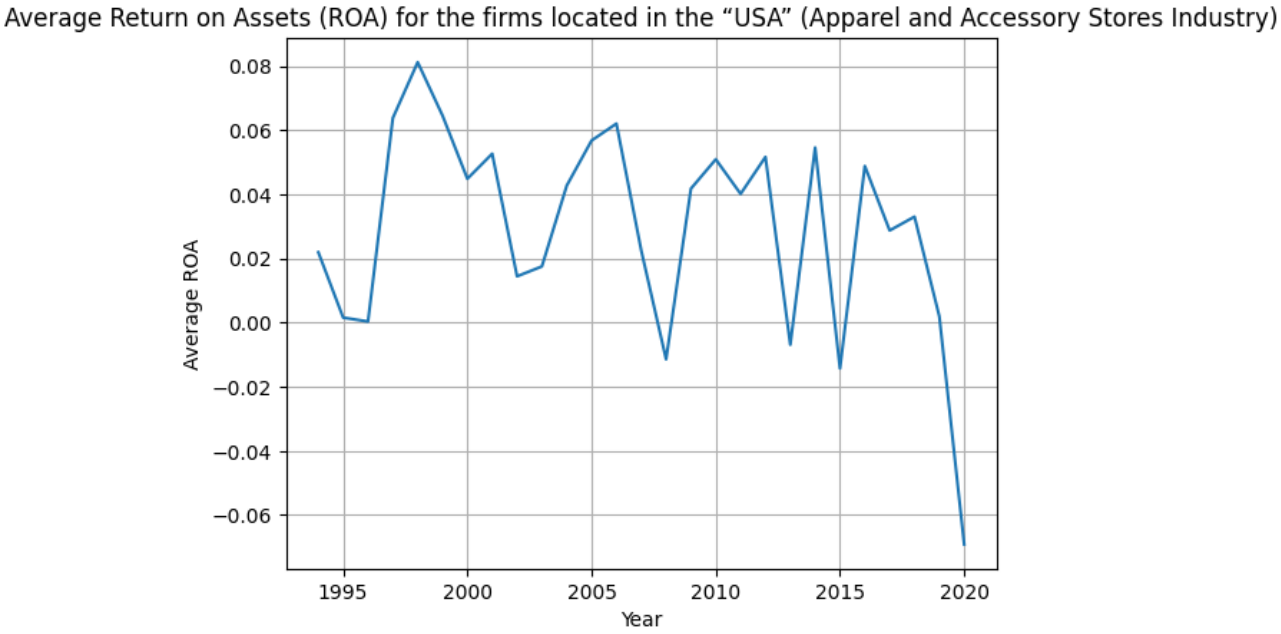
In [16]:
```
#Part 1B Question 6

#Drop the missing values from roa column
df_roa = df_56.dropna(subset=['roa'])

#Only select USA location for plotting
df_roa_USA = df_roa[df_roa['location'] == 'USA']

#Get the mean for plotting
df_roa_plot = df_roa_USA.groupby('fyear').roa.mean()

#Plot the chart
plt.plot(df_roa_plot)
plt.title("Average Return on Assets (ROA) for the firms located in the “USA”
plt.xlabel("Year")
plt.ylabel("Average ROA")
plt.grid(True)
plt.show()
```



In [19]:
```
#Part 2C

# 1. Convert all words to lowercase.
# 2. Remove punctuations.
# 3. Remove stop words based on the list of English stop words in NLTK

import string
from nltk.corpus import stopwords
from collections import Counter
from sklearn.feature_extraction.text import TfidfVectorizer

translator = str.maketrans('', '', string.punctuation)
sw = stopwords.words('english')
```

In [233…
```
df = pd.read_csv("data/2020_10K_item1_full.csv", encoding = 'utf-8')
df = df.drop_duplicates()
df
```

Out[233...

	cik	year	name	item_1_text	gvkey
0	1041588	2020	ACCESS-POWER INC	fixed expenses are previosuly documented in an...	66119
1	315374	2020	HURCO COMPANIES INC	General Hurco Companies, Inc. is an internatio...	5788
2	1622996	2020	ACRO BIOMEDICAL CO., LTD.	We have been engaged in the business of develo...	27584
3	1191334	2020	Chun Can Capital Group	CORPORATE HISTORY Chun Can Capital Group (form...	153614
12	1593204	2020	Adaiah Distribution Inc	General Adaiah Distribution Inc. was incorpora...	23706
...	...	...	...	...	...
5476	740664	2020	R F INDUSTRIES LTD	General RF Industries, Ltd. (together with sub...	2829
5477	1074828	2020	KNOW LABS, INC.	BACKGROUND AND CAPITAL STRUCTURE Know Labs, In...	166430
5478	40570	2020	GEE Group Inc.	General GEE Group Inc. (the Company , us , ...	5050
5479	1341726	2020	GULFSLOPE ENERGY, INC.	General GulfSlope Energy, Inc. is an independe...	175595
5480	72633	2020	NORTH EUROPEAN OIL ROYALTY TRUST	(a) General Development of Business. North Eur...	7959

5152 rows × 5 columns

In [234...

```
#code from lecture
def clean_text(text):
    # lower case
    clean_text = text.lower()

    # remove punctuation
    clean_text = clean_text.translate(translator)

    # remove stopwords
    clean_words = [w for w in clean_text.split() if w not in sw]

    return ' '.join(clean_words)

df['item_1_clean'] = df['item_1_text'].apply(clean_text)
```

In [235...

```
#Part 2D
```

```
#D-1
#get unique company key
df_56_com = df_56['gvkey'].unique()
df_56_com
```

Out[235…
```
array([  2484,   2595,   2818,   2938,   3087,   3122,   3824,   4072,
         4218,   4469,   4892,   4990,   5109,   6400,   6733,   7272,
         7922,   8390,   8504,   9248,   9305,  10969,  11584,  11672,
        11866,  12221,  13292,  13338,  13339,  13381,  13842,  14083,
        15346,  16580,  17911,  18675,  20895,  21828,  21898,  22612,
        23973,  24171,  24334,  24545,  24621,  25020,  25108,  25167,
        25186,  25234,  25353,  25357,  27773,  27936,  27937,  27938,
        27981,  29105,  29150,  29264,  29854,  30059,  30435,  30699,
        31829,  32194,  39310,  61397,  61445,  61868,  62667,  63621,
        63643,  63874,  64304,  64820,  65430,  65484,  65592,  66526,
       116104, 119474, 120716, 122778, 124176, 125275, 147661, 148224,
       148372, 149057, 154754, 160571, 162988, 163051, 163601, 164058,
       164103, 179342, 180227, 183570, 184323, 185190, 185227, 187041,
       187575])
```

In [236…
```
#use the company key above to filter the main dataframe
df_56_word = df[df['gvkey'].isin(df_56_com)]
df_56_word
```

Out[236…

	cik	year	name	item_1_text	gvkey	item_1_clean
1278	877422	2020	SpartanNash Co	Overview SpartanNash Company (together with it...	63874	overview spartannash company together subsidia...
2680	919012	2020	AMERICAN EAGLE OUTFITTERS INC	General American Eagle Outfitters, Inc. (the...	30059	general american eagle outfitters inc company ...
2868	1579298	2020	Burlington Stores, Inc.	Overview We are a nationally recognized retail...	18675	overview nationally recognized retailer highqu...
2997	897429	2020	CHICO'S FAS, INC.	Overview Founded in 1983, Chico s FAS, Inc.1, ...	27981	overview founded 1983 chico fas inc1 leading o...
3110	1318008	2020	Zumiez Inc	Zumiez Inc., including its wholly-owned subsid...	162988	zumiez inc including whollyowned

						subsidiaries ...
3174	883943	2020	CHRISTOPHER & BANKS CORP	Overview Christopher & Banks Corporation is a ...	25108	overview christopher banks corporation nationa...
3216	1483510	2020	EXPRESS, INC.	In this section, Express , we , us , the C...	184323	section express us company refer express inc c...
3225	39911	2020	GAP INC	General The Gap, Inc. (Gap Inc., the Company,...	4990	general gap inc gap inc company incorporated s...
3281	1041859	2020	Childrens Place, Inc.	As used in this Annual Report on Form 10-K, re...	65430	used annual report form 10k references company...
3306	813298	2020	DESTINATION XL GROUP, INC.	Destination XL Group, Inc., together with its ...	13381	destination xl group inc together subsidiaries...
3355	72333	2020	NORDSTROM INC	DESCRIPTION OF BUSINESS Founded in 1901 as a r...	7922	description business founded 1901 retail shoe ...
3653	18255	2020	CATO CORP	General The Company, founded in 1946, operated...	2818	general company founded 1946 operated 1281 fas...
3672	850209	2020	FOOT LOCKER, INC.	Overview We are a medical device company that ...	11584	overview medical device company develops comme...
3874	701985	2020	L Brands, Inc.	General L Brands, Inc. (we or the Company)...	6733	general l brands inc company operates highly c...
4008	745732	2020	ROSS STORES, INC.	Ross Stores, Inc. and its subsidiaries (we	9248	ross stores inc subsidiaries company

				o...		operate t...
4009	1018840	2020	ABERCROMBIE & FITCH CO /DE/	GENERAL Abercrombie & Fitch Co. (A&F), a com...	63643	general abercrombie fitch co af company incorp...
4015	912615	2020	URBAN OUTFITTERS INC	General We are a leading lifestyle products an...	29150	general leading lifestyle products services co...
4024	895447	2020	SHOE CARNIVAL INC	Our Company Shoe Carnival, Inc. is one of the ...	27938	company shoe carnival inc one nation largest f...
4037	1524025	2020	TILLY'S, INC.	Tillys is a leading destination specialty reta...	187575	tillys leading destination specialty retailer ...
4051	18498	2020	GENESCO INC	General Genesco Inc. (Genesco , Company , w...	5109	general genesco inc genesco company us incorpo...
4055	885245	2020	BUCKLE INC	The Buckle, Inc. (the Company) is a retailer...	25234	buckle inc company retailer medium betterprice...
4103	884217	2020	TAILORED BRANDS INC	Corporate Overview Nanning Tongji Hospital, In...	25167	corporate overview nanning tongji hospital inc...
4311	1399935	2020	Francesca's Holdings CORP	General Francesca s Holdings Corporation was i...	187041	general francesca holdings corporation incorpo...
4315	1319947	2020	Designer Brands Inc.	Overview Designer Brands Inc., originally foun...	24171	overview designer brands inc originally founde...
4316	1319947	2020	Designer	Overview Designer Brands Inc., originally	163601	overview designer brands inc

			Brands Inc.	foun...		originally founde...
4422	1318484	2020	Citi Trends Inc	Overview and History We are a value-priced ret...	163051	overview history valuepriced retailer fashion ...
4489	1610250	2020	Boot Barn Holdings, Inc.	Unless otherwise provided in this report, all ...	21898	unless otherwise provided report references re...
4634	1211351	2020	RTW Retailwinds, Inc.	Overview RTW Retailwinds, Inc., formerly known...	160571	overview rtw retailwinds inc formerly known ne...
4652	1687932	2020	J.Jill, Inc.	In this Annual Report, unless otherwise indica...	30435	annual report unless otherwise indicated conte...
4660	884940	2020	STEIN MART INC	OVERVIEW Headquartered in Jacksonville, Florid...	25186	overview headquartered jacksonville florida st...

In [237…
```
#D-2
#code from lecture
from collections import Counter

#word count function
def get_keywords_wc(text):
    c = Counter(text.split())
    words = []
    for pair in c.most_common(10):
        words.append(pair[0])
    return ' '.join(words)
```

In [238…
```
#get word count using get_keywords_wc
df_56_word['keyword_clean_wc'] = df_56_word.item_1_clean.apply(get_keywords_
df_56_word
```

```
/var/folders/_b/v51rk4qj7hxcq50ftvsf39540000gn/T/ipykernel_52426/2379778571.
py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/
stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df_56_word['keyword_clean_wc'] = df_56_word.item_1_clean.apply(get_keyword
s_wc)
```

Out[238...

	cik	year	name	item_1_text	gvkey	item_1_clean	keyword_
1278	877422	2020	SpartanNash Co	Overview SpartanNash Company (together with it...	63874	overview spartannash company together subsidia...	company retail d
2680	919012	2020	AMERICAN EAGLE OUTFITTERS INC	General American Eagle Outfitters, Inc. (the...	30059	general american eagle outfitters inc company ...	stores ser 1 officer
2868	1579298	2020	Burlington Stores, Inc.	Overview We are a nationally recognized retail...	18675	overview nationally recognized retailer highqu...	st merchar believ
2997	897429	2020	CHICO'S FAS, INC.	Overview Founded in 1983, Chico s FAS, Inc.1, ...	27981	overview founded 1983 chico fas inc1 leading o...	me brands fiscal st
3110	1318008	2020	Zumiez Inc	Zumiez Inc., including its wholly-owned subsid...	162988	zumiez inc including whollyowned subsidiaries ...	stores be merchar
3174	883943	2020	CHRISTOPHER & BANKS CORP	Overview Christopher & Banks Corporation is a ...	25108	overview christopher banks corporation nationa...	merchand custo cust
3216	1483510	2020	EXPRESS, INC.	In this section, Express , we , us , the C...	184323	section express us company refer express inc c...	me express s
				General The Gap, Inc. (Gap		general gap inc gap inc	presider

3225	39911	2020	GAP INC	Inc., the Company,...	4990	company incorporated s...	execu bra
3281	1041859	2020	Childrens Place, Inc.	As used in this Annual Report on Form 10-K, re...	65430	used annual report form 10k references company...	child plac compar
3306	813298	2020	DESTINATION XL GROUP, INC.	Destination XL Group, Inc., together with its ...	13381	destination xl group inc together subsidiaries...	store fiscal casu
3355	72333	2020	NORDSTROM INC	DESCRIPTION OF BUSINESS Founded in 1901 as a r...	7922	description business founded 1901 retail shoe ...	merchand
3653	18255	2020	CATO CORP	General The Company, founded in 1946, operated...	2818	general company founded 1946 operated 1281 fas...	merchand sales cred
3672	850209	2020	FOOT LOCKER, INC.	Overview We are a medical device company that ...	11584	overview medical device company develops comme...	system devic medi
3874	701985	2020	L Brands, Inc.	General L Brands, Inc. (we or the Company)...	6733	general l brands inc company operates highly c...	secu stores b prod
4008	745732	2020	ROSS STORES, INC.	Ross Stores, Inc. and its subsidiaries (we o...	9248	ross stores inc subsidiaries company operate t...	stores me discour de
4009	1018840	2020	ABERCROMBIE & FITCH CO /DE/	GENERAL Abercrombie & Fitch Co. (A&F), a com...	63643	general abercrombie fitch co af company incorp...	comp customer merc
4015	912615	2020	URBAN OUTFITTERS INC	General We are a leading lifestyle	29150	general leading lifestyle products	stores me segment s

				products an...		services co...	
4024	895447	2020	SHOE CARNIVAL INC	Our Company Shoe Carnival, Inc. is one of the ...	27938	company shoe carnival inc one nation largest f...	st cus custom
4037	1524025	2020	TILLY'S, INC.	Tillys is a leading destination specialty reta...	187575	tillys leading destination specialty retailer ...	stores merchar be
4051	18498	2020	GENESCO INC	General Genesco Inc. (Genesco , Company , w...	5109	general genesco inc genesco company us incorpo...	footw 2 operatio
4055	885245	2020	BUCKLE INC	The Buckle, Inc. (the Company) is a retailer...	25234	buckle inc company retailer medium betterprice...	comp st merchar
4103	884217	2020	TAILORED BRANDS INC	Corporate Overview Nanning Tongji Hospital, In...	25167	corporate overview nanning tongji hospital inc...	nth comp inc stat
4311	1399935	2020	Francesca's Holdings CORP	General Francesca s Holdings Corporation was i...	187041	general francesca holdings corporation incorpo...	me boutiqu
4315	1319947	2020	Designer Brands Inc.	Overview Designer Brands Inc., originally foun...	24171	overview designer brands inc originally founde...	custom bran retail
4316	1319947	2020	Designer Brands Inc.	Overview Designer Brands Inc., originally foun...	163601	overview designer brands inc originally founde...	custom bran retail
4422	1318484	2020	Citi Trends Inc	Overview and History We are a value-priced ret...	163051	overview history valuepriced retailer fashion ...	st me fashi

4489	1610250	2020	Boot Barn Holdings, Inc.	Unless otherwise provided in this report, all ...	21898	unless otherwise provided report references re...	magr gasificatic
4634	1211351	2020	RTW Retailwinds, Inc.	Overview RTW Retailwinds, Inc., formerly known...	160571	overview rtw retailwinds inc formerly known ne...	compa store me yea
4652	1687932	2020	J.Jill, Inc.	In this Annual Report, unless otherwise indica...	30435	annual report unless otherwise indicated conte...	custon year fisc
4660	884940	2020	STEIN MART INC	OVERVIEW Headquartered in Jacksonville, Florid...	25186	overview headquartered jacksonville florida st...	stores me 2020 s cu

In [239…

```
#code from lecture
def get_keywords_tfidf(document_list):
    '''
    Input: A list of documents (text)
    Output: The corresponding top 10 keywords for each document based on tf-
    '''

    # Step 1: Create the TF-IDF vectorizer
    vectorizer = TfidfVectorizer()

    # Step 2: Calculate the TF-IDF matrix
    tfidf_matrix = vectorizer.fit_transform(document_list)

    # Step 3: Get feature names (words)
    feature_names = vectorizer.get_feature_names_out()

    # Step 4: Extract top 10 keywords for each text
    top_keywords = []
    for i in range(len(document_list)):

        if i %100 == 0:
            print(f'Processing the {i}/{len(document_list)} document.')

        feature_index = tfidf_matrix[i, :].nonzero()[1]
        tfidf_scores = zip(feature_index, [tfidf_matrix[i, x] for x in featu
```

```
        sorted_tfidf_scores = sorted(tfidf_scores, key=lambda x: x[1], rever
        top_keywords.append(' '.join([feature_names[i] for i, _ in sorted_tf

    return top_keywords
```

In [241…

```
#get tf- idf using get_keywords_tfidf
from sklearn.feature_extraction.text import TfidfVectorizer
keywords = get_keywords_tfidf(df_56_word.item_1_clean.tolist())
df_56_word['tfidf'] = keywords
df_56_word
```

```
Processing the 0/30 document.
/var/folders/_b/v51rk4qj7hxcq50ftvsf39540000gn/T/ipykernel_52426/3610273905.
py:4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/
stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df_56_word['tfidf'] = keywords
```

Out[241…

	cik	year	name	item_1_text	gvkey	item_1_clean	keyword_
1278	877422	2020	SpartanNash Co	Overview SpartanNash Company (together with it...	63874	overview spartannash company together subsidia...	company retail d
2680	919012	2020	AMERICAN EAGLE OUTFITTERS INC	General American Eagle Outfitters, Inc. (the...	30059	general american eagle outfitters inc company ...	stores ser 1 officer
2868	1579298	2020	Burlington Stores, Inc.	Overview We are a nationally recognized retail...	18675	overview nationally recognized retailer highqu...	st merchar believ
2997	897429	2020	CHICO'S FAS, INC.	Overview Founded in 1983, Chico s FAS, Inc.1, ...	27981	overview founded 1983 chico fas inc1 leading o...	me brands fiscal st
3110	1318008	2020	Zumiez Inc	Zumiez Inc., including its wholly-owned subsid...	162988	zumiez inc including whollyowned subsidiaries ...	stores be merchar
			CHRISTOPHER	Overview Christopher &		overview christopher	merchanc

3174	883943	2020	& BANKS CORP	Banks Corporation is a ...	25108	banks corporation nationa...	custo cust
3216	1483510	2020	EXPRESS, INC.	In this section, Express , we , us , the C...	184323	section express us company refer express inc c...	me express s
3225	39911	2020	GAP INC	General The Gap, Inc. (Gap Inc., the Company,...	4990	general gap inc gap inc company incorporated s...	presider execu bra
3281	1041859	2020	Childrens Place, Inc.	As used in this Annual Report on Form 10-K, re...	65430	used annual report form 10k references company...	child plac compar
3306	813298	2020	DESTINATION XL GROUP, INC.	Destination XL Group, Inc., together with its ...	13381	destination xl group inc together subsidiaries...	store fiscal casu
3355	72333	2020	NORDSTROM INC	DESCRIPTION OF BUSINESS Founded in 1901 as a r...	7922	description business founded 1901 retail shoe ...	merchand
3653	18255	2020	CATO CORP	General The Company, founded in 1946, operated...	2818	general company founded 1946 operated 1281 fas...	merchand sales cred
3672	850209	2020	FOOT LOCKER, INC.	Overview We are a medical device company that ...	11584	overview medical device company develops comme...	system devic medi
3874	701985	2020	L Brands, Inc.	General L Brands, Inc. (we or the Company)...	6733	general l brands inc company operates highly c...	secu stores b prod
4008	745732	2020	ROSS STORES, INC.	Ross Stores, Inc. and its subsidiaries (	9248	ross stores inc subsidiaries company	stores me discour

				we o...		operate t...	de
4009	1018840	2020	ABERCROMBIE & FITCH CO /DE/	GENERAL Abercrombie & Fitch Co. (A&F), a com...	63643	general abercrombie fitch co af company incorp...	comp customer merc
4015	912615	2020	URBAN OUTFITTERS INC	General We are a leading lifestyle products an...	29150	general leading lifestyle products services co...	stores me segment s
4024	895447	2020	SHOE CARNIVAL INC	Our Company Shoe Carnival, Inc. is one of the ...	27938	company shoe carnival inc one nation largest f...	st cus custom
4037	1524025	2020	TILLY'S, INC.	Tillys is a leading destination specialty reta...	187575	tillys leading destination specialty retailer ...	stores merchar be
4051	18498	2020	GENESCO INC	General Genesco Inc. (Genesco , Company , w...	5109	general genesco inc genesco company us incorpo...	footw 2 operatio
4055	885245	2020	BUCKLE INC	The Buckle, Inc. (the Company) is a retailer...	25234	buckle inc company retailer medium betterprice...	comp st merchar
4103	884217	2020	TAILORED BRANDS INC	Corporate Overview Nanning Tongji Hospital, In...	25167	corporate overview nanning tongji hospital inc...	nth comp inc stat
4311	1399935	2020	Francesca's Holdings CORP	General Francesca s Holdings Corporation was i...	187041	general francesca holdings corporation incorpo...	me boutiqu
4315	1319947	2020	Designer Brands Inc.	Overview Designer Brands Inc., originally	24171	overview designer brands inc originally	custom bran retail

				foun...		founde...	
4316	1319947	2020	Designer Brands Inc.	Overview Designer Brands Inc., originally foun...	163601	overview designer brands inc originally founde...	custom bran retail
4422	1318484	2020	Citi Trends Inc	Overview and History We are a value-priced ret...	163051	overview history valuepriced retailer fashion ...	st me fashi
4489	1610250	2020	Boot Barn Holdings, Inc.	Unless otherwise provided in this report, all ...	21898	unless otherwise provided report references re...	magr gasificati
4634	1211351	2020	RTW Retailwinds, Inc.	Overview RTW Retailwinds, Inc., formerly known...	160571	overview rtw retailwinds inc formerly known ne...	compa store me yea
4652	1687932	2020	J.Jill, Inc.	In this Annual Report, unless otherwise indica...	30435	annual report unless otherwise indicated conte...	custon year fisc
4660	884940	2020	STEIN MART INC	OVERVIEW Headquartered in Jacksonville, Florid...	25186	overview headquartered jacksonville florida st...	stores me 2020 s cu

In [242…

```
#D-3 word cloud
#code from lecture
from wordcloud import WordCloud
text1 = ' '.join(df_56_word['keyword_clean_wc'].tolist())

#word cloud for word_count
wordcloud1 = WordCloud(width=800, height=400, background_color='white').gene
plt.figure(figsize=(10,5))
plt.imshow(wordcloud1)
plt.savefig('keyword_wc.png') # save as PNG file
```

```
plt.axis('off')

plt.show()
```

In [243…

```
#word cloud for tf-idf
text2 = ' '.join(df_56_word['tfidf'].tolist())
wordcloud2 = WordCloud(width=800, height=400, background_color='white').gene

plt.figure(figsize=(10,5))
plt.imshow(wordcloud2)
plt.savefig('tfidf.png') # save as PNG file
plt.axis('off')

plt.show()
```

In [244…
```
#Part 2E rain a word2vec model and analyze word similarities
#code from lecture
from gensim.models import Word2Vec
sent = [row.split() for row in df['item_1_clean']]

#train and save model
model = Word2Vec(sent, min_count=1, vector_size=50, workers=3, window=3, sg
model.save("word2vec.model")
model = Word2Vec.load("word2vec.model")
```

In [245…
```
from DocumentSimilarity import DocumentSimilarity
d = DocumentSimilarity(model = model, gvkeys=df_56_word['gvkey'], conm = df_
                       keywordslist = df_56_word['tfidf'])
```

In [246…
```
#top5 similar word
model.wv.most_similar(positive='footwear', topn=5)
```

Out[246…
```
[('apparel', 0.9550002217292786),
 ('outerwear', 0.8890790343284607),
 ('clothing', 0.8635386228561401),
 ('shoes', 0.8574903011322021),
 ('sportswear', 0.8462600708007812)]
```

In [248…
```
model.wv.most_similar(positive='store', topn=5)
```

Out[248…
```
[('stores', 0.8738605380058289),
 ('restaurant', 0.8545236587524414),
 ('inrestaurant', 0.8387573957443237),
 ('instore', 0.8363032937049866),
 ('popup', 0.8332494497299194)]
```

In [249…
```
model.wv.most_similar(positive='brand', topn=5)
```

Out[249…
```
[('brands', 0.9274031519889832),
 ('widelyrecognized', 0.8536772727966309),
 ('lee®', 0.8514106273651123),
 ('guess', 0.8509753346443176),
 ('bergio', 0.841139554977417)]
```

In [253…
```
unique_company =df_56_word[['gvkey','name']].drop_duplicates()
unique_company
```

Out[253…

	gvkey	name
1278	63874	SpartanNash Co
2680	30059	AMERICAN EAGLE OUTFITTERS INC
2868	18675	Burlington Stores, Inc.
2997	27981	CHICO'S FAS, INC.
3110	162988	Zumiez Inc
3174	25108	CHRISTOPHER & BANKS CORP
3216	184323	EXPRESS, INC.
3225	4990	GAP INC
3281	65430	Childrens Place, Inc.
3306	13381	DESTINATION XL GROUP, INC.
3355	7922	NORDSTROM INC
3653	2818	CATO CORP
3672	11584	FOOT LOCKER, INC.
3874	6733	L Brands, Inc.
4008	9248	ROSS STORES, INC.
4009	63643	ABERCROMBIE & FITCH CO /DE/
4015	29150	URBAN OUTFITTERS INC
4024	27938	SHOE CARNIVAL INC
4037	187575	TILLY'S, INC.
4051	5109	GENESCO INC
4055	25234	BUCKLE INC
4103	25167	TAILORED BRANDS INC

4311	187041	Francesca's Holdings CORP
4315	24171	Designer Brands Inc.
4316	163601	Designer Brands Inc.
4422	163051	Citi Trends Inc
4489	21898	Boot Barn Holdings, Inc.
4634	160571	RTW Retailwinds, Inc.
4652	30435	J.Jill, Inc.
4660	25186	STEIN MART INC

In [223…

```
#inspired by chatgpt - calcuate market share by fyear
df_56['Market_Share'] = df_56.groupby('fyear')['sale'].transform(lambda x: x
df_56[df_56['gvkey'] == 4990]
```

```
/var/folders/_b/v51rk4qj7hxcq50ftvsf39540000gn/T/ipykernel_52426/2195832630.
py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/
stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df_56['Market_Share'] = df_56.groupby('fyear')['sale'].transform(lambda x:
x / x.sum())
```

Out[223…

	gvkey	fyear	location	conm	ipodate	sic	prcc_c	ch	ni	
15929	4990	1994	USA	GAP INC	NaN	5651	30.500	414.487	320.240	20
15930	4990	1995	USA	GAP INC	NaN	5651	42.000	579.566	354.039	23
15931	4990	1996	USA	GAP INC	NaN	5651	30.125	485.644	452.859	26
15932	4990	1997	USA	GAP INC	NaN	5651	35.437	913.169	533.901	33
15933	4990	1998	USA	GAP INC	NaN	5651	56.125	565.253	824.539	39
15934	4990	1999	USA	GAP INC	NaN	5651	46.000	450.352	1127.065	51
15935	4990	2000	USA	GAP INC	NaN	5651	25.500	408.794	877.497	70

15936	4990	2001	USA	GAP INC	NaN	5651	13.940	1035.749	-7.764	76
15937	4990	2002	USA	GAP INC	NaN	5651	15.520	3388.514	477.457	99
15938	4990	2003	USA	GAP INC	NaN	5651	23.210	3334.000	1030.000	103
15939	4990	2004	USA	GAP INC	NaN	5651	21.120	2245.000	1150.000	100
15940	4990	2005	USA	GAP INC	NaN	5651	17.640	2035.000	1113.000	88
15941	4990	2006	USA	GAP INC	NaN	5651	19.500	2030.000	778.000	85
15942	4990	2007	USA	GAP INC	NaN	5651	21.280	1724.000	833.000	78
15943	4990	2008	USA	GAP INC	NaN	5651	13.390	1715.000	967.000	75
15944	4990	2009	USA	GAP INC	NaN	5651	20.950	2348.000	1102.000	79
15945	4990	2010	USA	GAP INC	NaN	5651	22.140	1561.000	1204.000	70
15946	4990	2011	USA	GAP INC	NaN	5651	18.550	1885.000	833.000	74
15947	4990	2012	USA	GAP INC	NaN	5651	31.040	1460.000	1135.000	74
15948	4990	2013	USA	GAP INC	NaN	5651	39.080	1510.000	1280.000	78
15949	4990	2014	USA	GAP INC	NaN	5651	42.110	1515.000	1262.000	76
15950	4990	2015	USA	GAP INC	NaN	5651	24.700	1370.000	920.000	74
15951	4990	2016	USA	GAP INC	NaN	5651	22.440	1783.000	676.000	76
15952	4990	2017	USA	GAP INC	NaN	5651	34.060	1783.000	848.000	79
15953	4990	2018	USA	GAP INC	NaN	5651	25.760	1081.000	1003.000	80
15954	4990	2019	USA	GAP INC	NaN	5651	17.680	1364.000	351.000	136

15955	4990	2020	USA	GAP INC	NaN	5651	20.190	1988.000	-665.000	137

In [254…
```
firm = 4990
ls = d.most_similar(firm = firm, topn = 5)
data = []
data.append(firm)
```

In [256…
```
for l in ls:
    data.append(l[0])
data
```

Out[256… `[4990, 30059, 27938, 63643, 25234, 5109]`

In [257…
```
similarity = df_56[df_56['gvkey'].isin(data)]
```

In [260…
```
similarity
```

Out[260…

	gvkey	fyear	location	conm	ipodate	sic	prcc_c	ch	
15929	4990	1994	USA	GAP INC	NaN	5651	30.500	414.487	3
15930	4990	1995	USA	GAP INC	NaN	5651	42.000	579.566	3
15931	4990	1996	USA	GAP INC	NaN	5651	30.125	485.644	4
15932	4990	1997	USA	GAP INC	NaN	5651	35.437	913.169	5
15933	4990	1998	USA	GAP INC	NaN	5651	56.125	565.253	8
...	...	...	...	...	...	...	...	...	
132224	63643	2016	USA	ABERCROMBIE & FITCH -CL A	1996/09/25	5651	12.000	547.189	
132225	63643	2017	USA	ABERCROMBIE & FITCH -CL A	1996/09/25	5651	17.430	675.558	
132226	63643	2018	USA	ABERCROMBIE & FITCH -CL A	1996/09/25	5651	20.050	723.135	
132227	63643	2019	USA	ABERCROMBIE & FITCH -CL A	1996/09/25	5651	17.290	671.267	
132228	63643	2020	USA	ABERCROMBIE & FITCH -CL A	1996/09/25	5651	20.360	1104.862	-

159 rows × 13 columns

In [263…

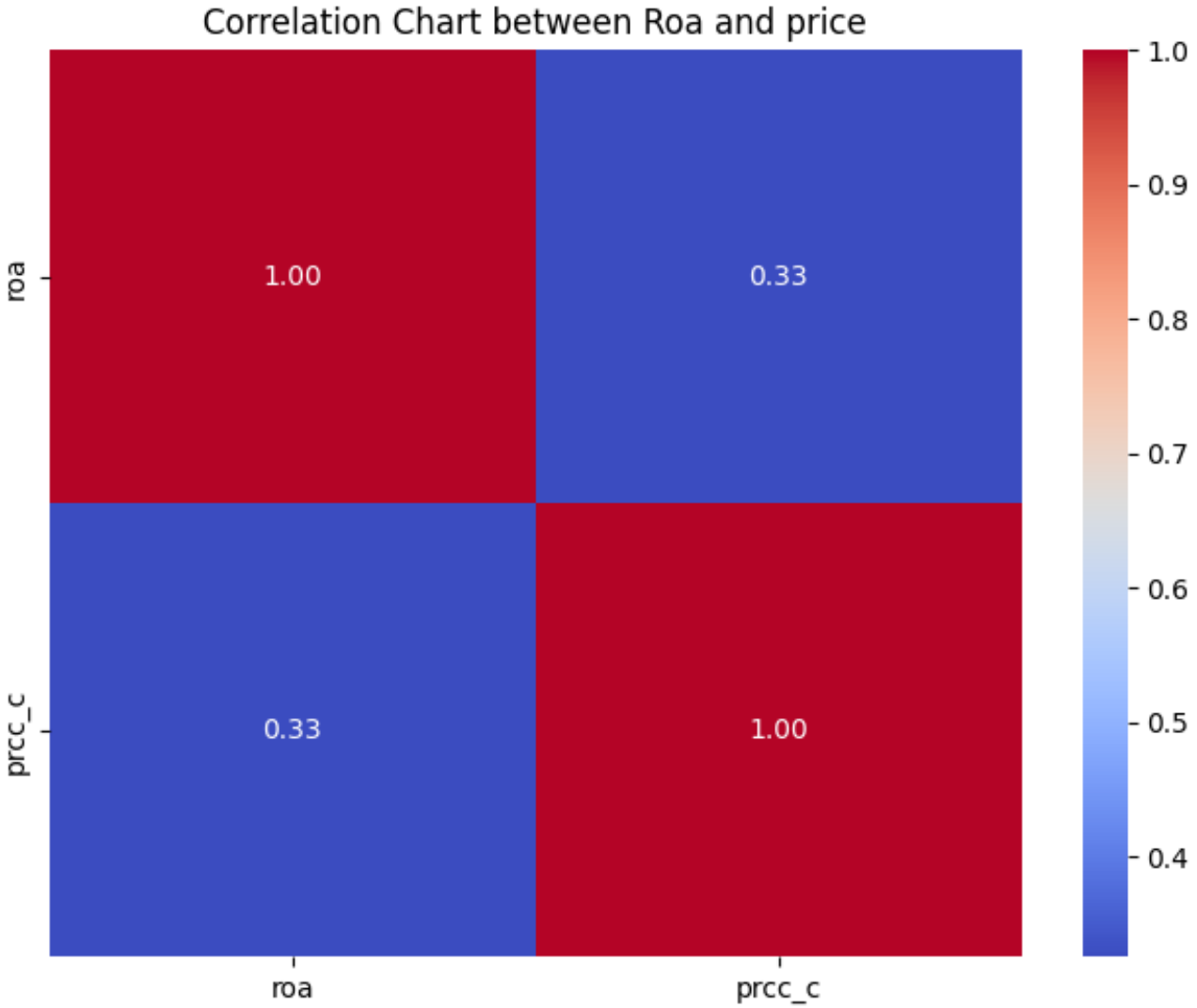
```
import seaborn as sns

# Assuming you have a DataFrame df with columns "Roa" and "Proc"
# Replace df with your actual DataFrame name and column names as needed

# Calculate the correlation matrix
correlation_matrix = similarity[['roa', 'prcc_c']].corr()

# Create a heatmap to visualize the correlation
plt.figure(figsize=(8, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")

# Set the title
plt.title('Correlation Chart between Roa and price')
plt.savefig('cor_price_roa.png')
# Display the plot
plt.show()
```

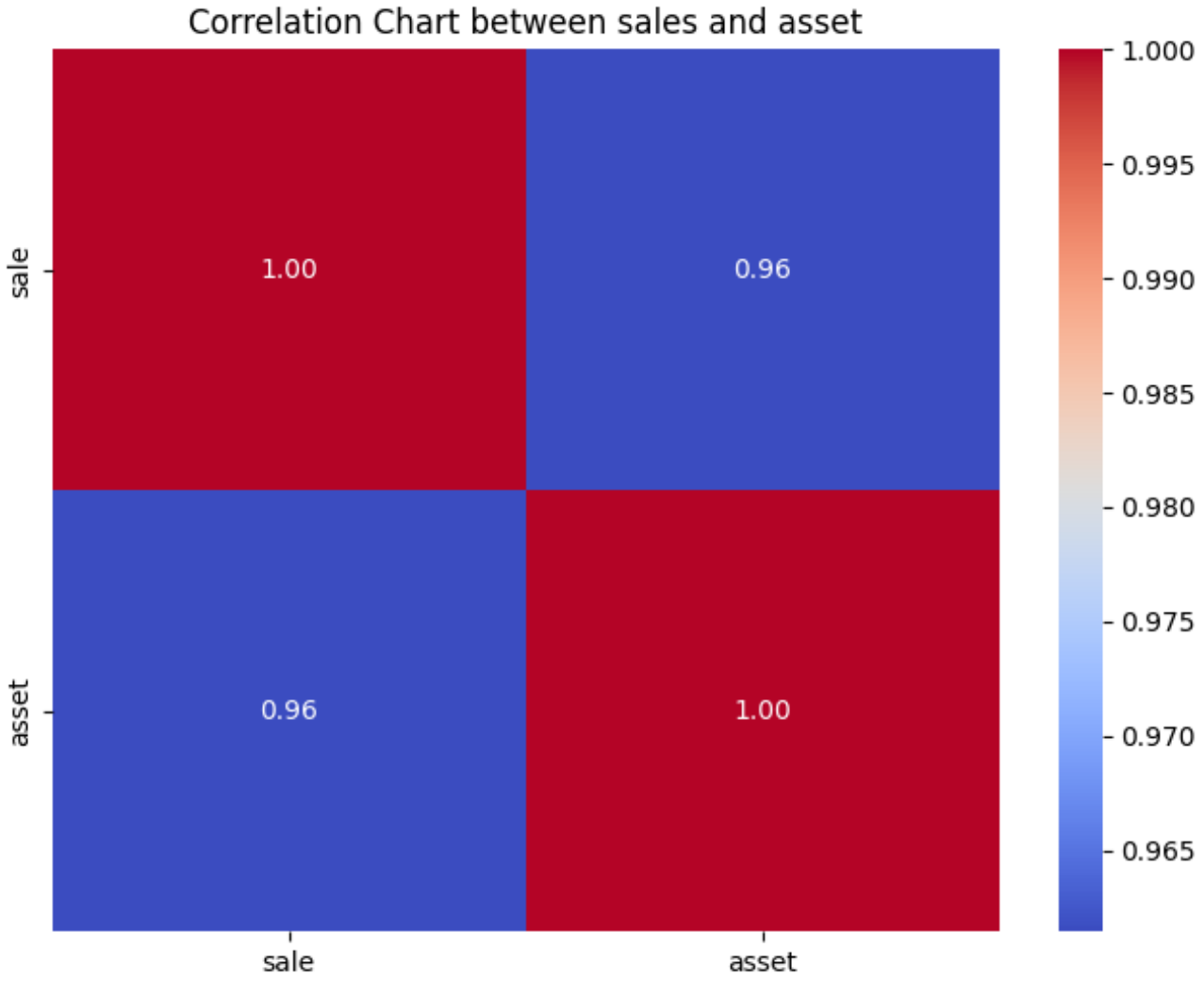


In [264…

```
correlation_matrix = similarity[['sale','asset']].corr()
```

```
# Create a heatmap to visualize the correlation
plt.figure(figsize=(8, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")

# Set the title
plt.title('Correlation Chart between sales and asset')
plt.savefig('cor_asset_sales.png')
# Display the plot
plt.show()
```



In [226…

```
#inspired by chatgpt - calculate price change by stocks.
similarity['price_move'] = similarity.groupby('gvkey')['prcc_c'].pct_change(
similarity[similarity['gvkey']==63643]
```

```
/var/folders/_b/v51rk4qj7hxcq50ftvsf39540000gn/T/ipykernel_52426/4264984879.
py:1: FutureWarning: The default fill_method='ffill' in SeriesGroupBy.pct_ch
ange is deprecated and will be removed in a future version. Call ffill befor
e calling pct_change to retain current behavior and silence this warning.
  similarity['price_move'] = similarity.groupby('gvkey')['prcc_c'].pct_chang
e()*100
/var/folders/_b/v51rk4qj7hxcq50ftvsf39540000gn/T/ipykernel_52426/4264984879.
py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/
stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  similarity['price_move'] = similarity.groupby('gvkey')['prcc_c'].pct_chang
e()*100
```

Out[226...

	gvkey	fyear	location	conm	ipodate	sic	prcc_c	ch
132203	63643	1995	USA	ABERCROMBIE & FITCH -CL A	1996/09/25	5651	NaN	0.874
132204	63643	1996	USA	ABERCROMBIE & FITCH -CL A	1996/09/25	5651	16.5000	1.945
132205	63643	1997	USA	ABERCROMBIE & FITCH -CL A	1996/09/25	5651	31.2500	42.667
132206	63643	1998	USA	ABERCROMBIE & FITCH -CL A	1996/09/25	5651	70.7500	163.564
132207	63643	1999	USA	ABERCROMBIE & FITCH -CL A	1996/09/25	5651	26.6875	147.908
132208	63643	2000	USA	ABERCROMBIE & FITCH -CL A	1996/09/25	5651	20.0000	137.581
132209	63643	2001	USA	ABERCROMBIE & FITCH -CL A	1996/09/25	5651	26.5300	167.664
132210	63643	2002	USA	ABERCROMBIE & FITCH -CL A	1996/09/25	5651	20.4600	391.035
132211	63643	2003	USA	ABERCROMBIE & FITCH -CL A	1996/09/25	5651	24.7100	56.373
132212	63643	2004	USA	ABERCROMBIE & FITCH -CL A	1996/09/25	5651	46.9500	350.368
132213	63643	2005	USA	ABERCROMBIE & FITCH -CL A	1996/09/25	5651	65.1800	50.687
132214	63643	2006	USA	ABERCROMBIE & FITCH -CL A	1996/09/25	5651	69.6300	81.959
				ABERCROMBIE				

132215	63643	2007	USA	& FITCH -CL A	1996/09/25	5651	79.9700	118.044
132216	63643	2008	USA	ABERCROMBIE & FITCH -CL A	1996/09/25	5651	23.0700	522.122
132217	63643	2009	USA	ABERCROMBIE & FITCH -CL A	1996/09/25	5651	34.8500	680.113
132218	63643	2010	USA	ABERCROMBIE & FITCH -CL A	1996/09/25	5651	57.6300	826.353
132219	63643	2011	USA	ABERCROMBIE & FITCH -CL A	1996/09/25	5651	48.8400	583.495
132220	63643	2012	USA	ABERCROMBIE & FITCH -CL A	1996/09/25	5651	47.9700	643.505
132221	63643	2013	USA	ABERCROMBIE & FITCH -CL A	1996/09/25	5651	32.9100	600.116
132222	63643	2014	USA	ABERCROMBIE & FITCH -CL A	1996/09/25	5651	28.6400	530.192
132223	63643	2015	USA	ABERCROMBIE & FITCH -CL A	1996/09/25	5651	27.0000	588.578
132224	63643	2016	USA	ABERCROMBIE & FITCH -CL A	1996/09/25	5651	12.0000	547.189
132225	63643	2017	USA	ABERCROMBIE & FITCH -CL A	1996/09/25	5651	17.4300	675.558
132226	63643	2018	USA	ABERCROMBIE & FITCH -CL A	1996/09/25	5651	20.0500	723.135
132227	63643	2019	USA	ABERCROMBIE & FITCH -CL A	1996/09/25	5651	17.2900	671.267
132228	63643	2020	USA	ABERCROMBIE & FITCH -CL A	1996/09/25	5651	20.3600	1104.862

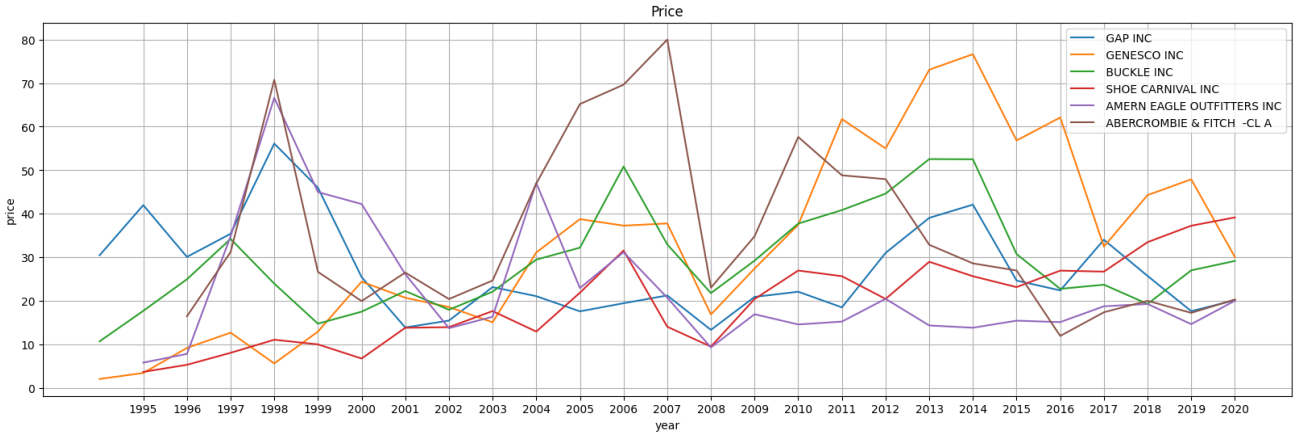
In [275…

```
plt.figure(figsize=(20, 6))

for gvkey, group_df in similarity.groupby('gvkey'):
    company_name = group_df['conm'].iloc[0]
    plt.plot(group_df['fyear'],group_df['prcc_c'], label = f' {company_name}

plt.xticks(group_df['fyear'])
plt.grid(True)
plt.xlabel('year')
plt.ylabel('price')
plt.title('Price')
```

```
plt.legend()
plt.savefig('price_linechart.png')
plt.show()
```

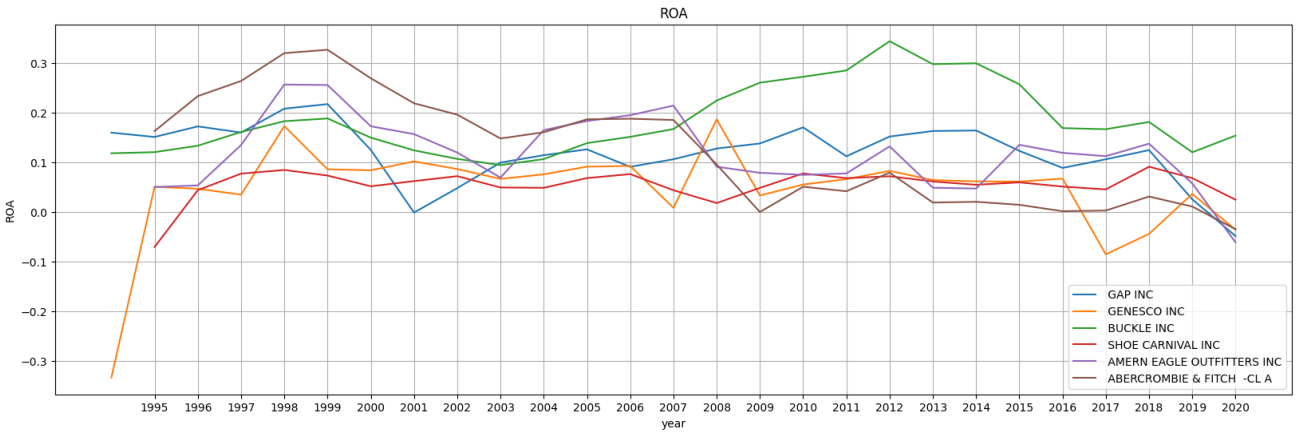


In [290…

```
plt.figure(figsize=(20, 6))

for gvkey, group_df in similarity.groupby('gvkey'):
    company_name = group_df['conm'].iloc[0]
    plt.plot(group_df['fyear'],group_df['roa'], label = f' {company_name}')

plt.xticks(group_df['fyear'])
plt.grid(True)
plt.xlabel('year')
plt.ylabel('ROA')
plt.title('ROA')
plt.legend()
plt.savefig('roa_linechart.png')
```

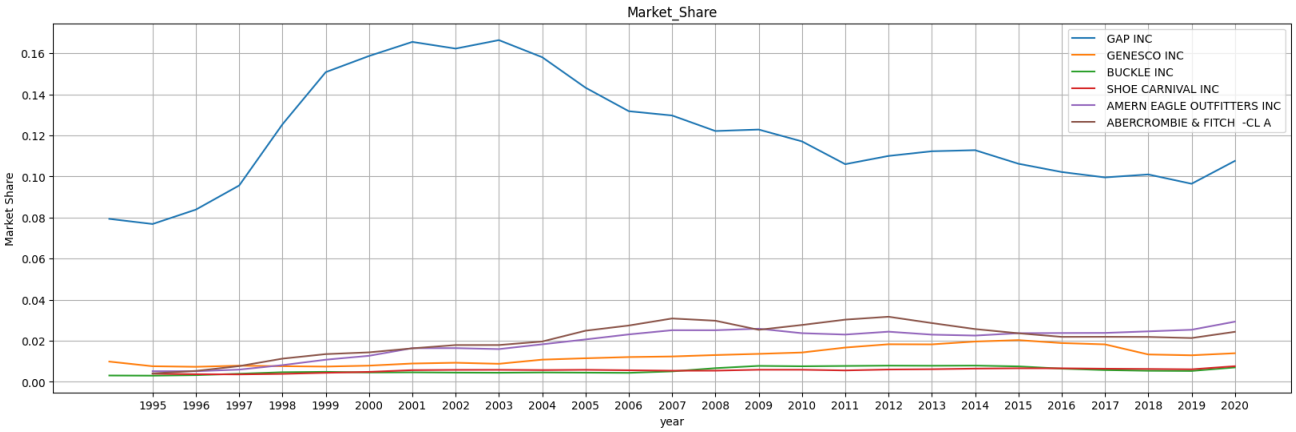


In [289…

```
plt.figure(figsize=(20, 6))

for gvkey, group_df in similarity.groupby('gvkey'):
    company_name = group_df['conm'].iloc[0]
    plt.plot(group_df['fyear'],group_df['Market_Share'], label = f' {company
```

```
plt.xticks(group_df['fyear'])
plt.grid(True)
plt.xlabel('year')
plt.ylabel('Market Share')
plt.legend()
plt.title('Market_Share')
plt.savefig('mktshare_linechart.png')
plt.show()
```

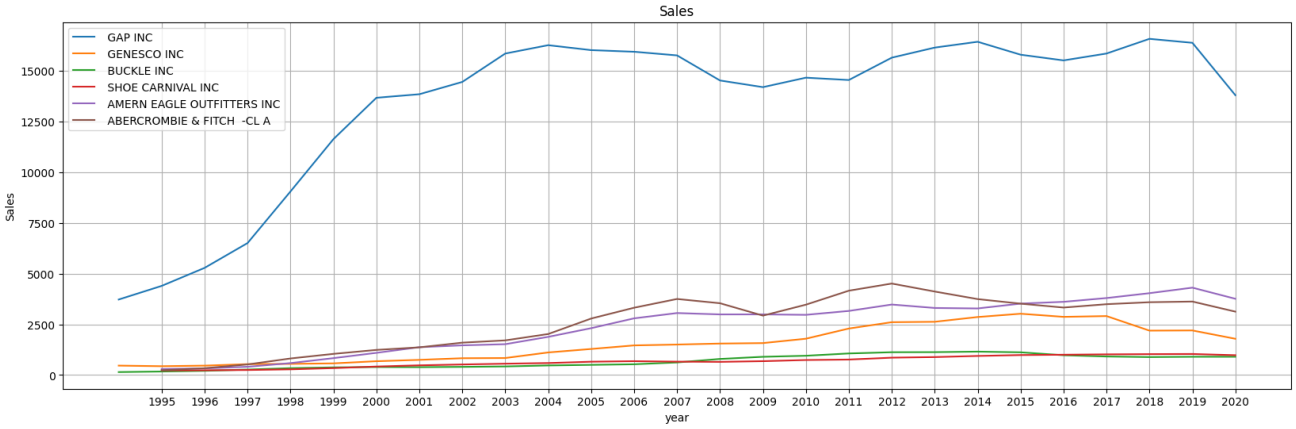


In [288…

```
plt.figure(figsize=(20, 6))

for gvkey, group_df in similarity.groupby('gvkey'):
    company_name = group_df['conm'].iloc[0]
    plt.plot(group_df['fyear'],group_df['sale'], label = f' {company_name}')

plt.xticks(group_df['fyear'])
plt.grid(True)
plt.xlabel('year')
plt.ylabel('Sales')
plt.title('Sales')
plt.legend()
plt.savefig('Sales_linechart.png')
plt.show()
```

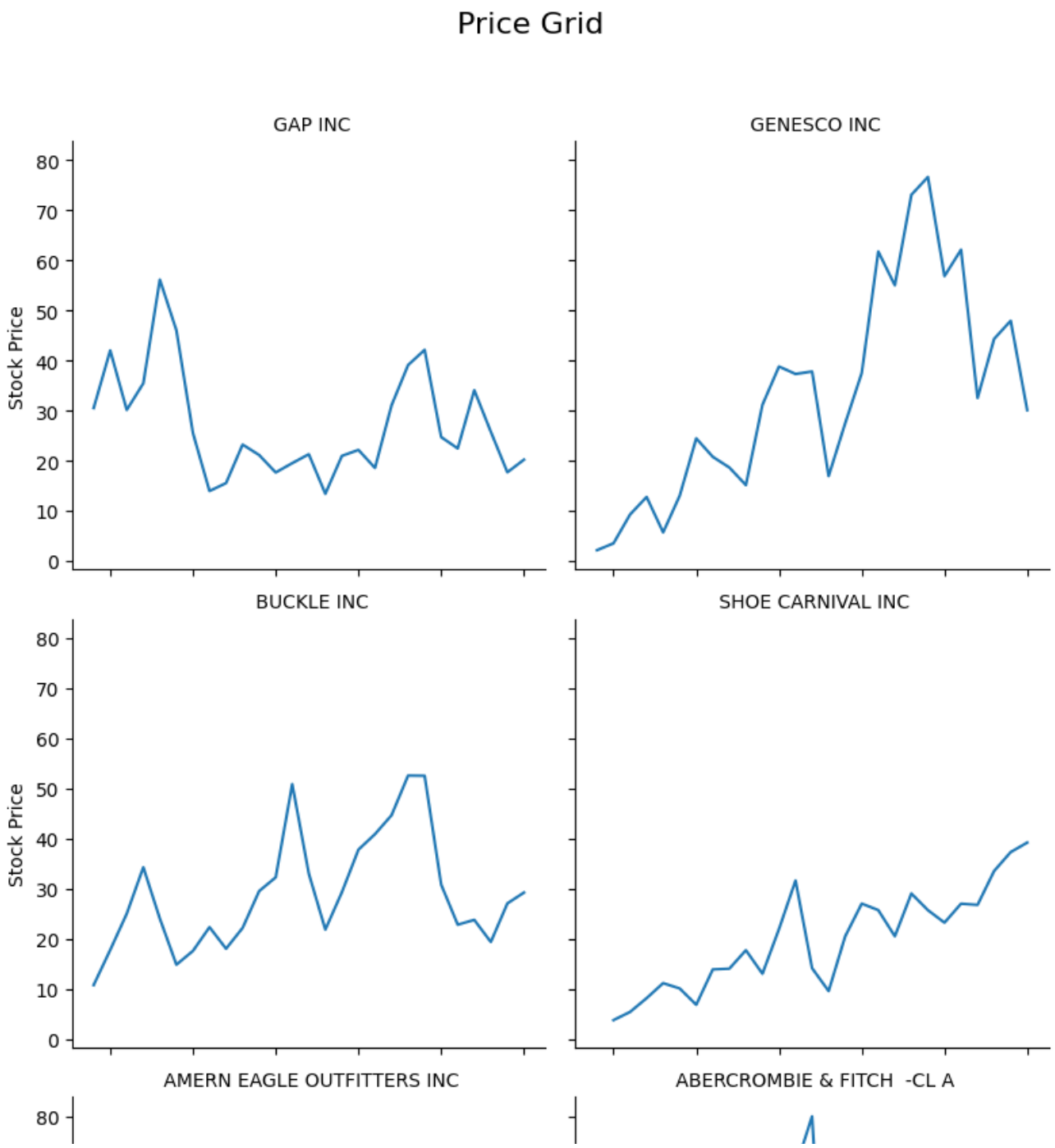


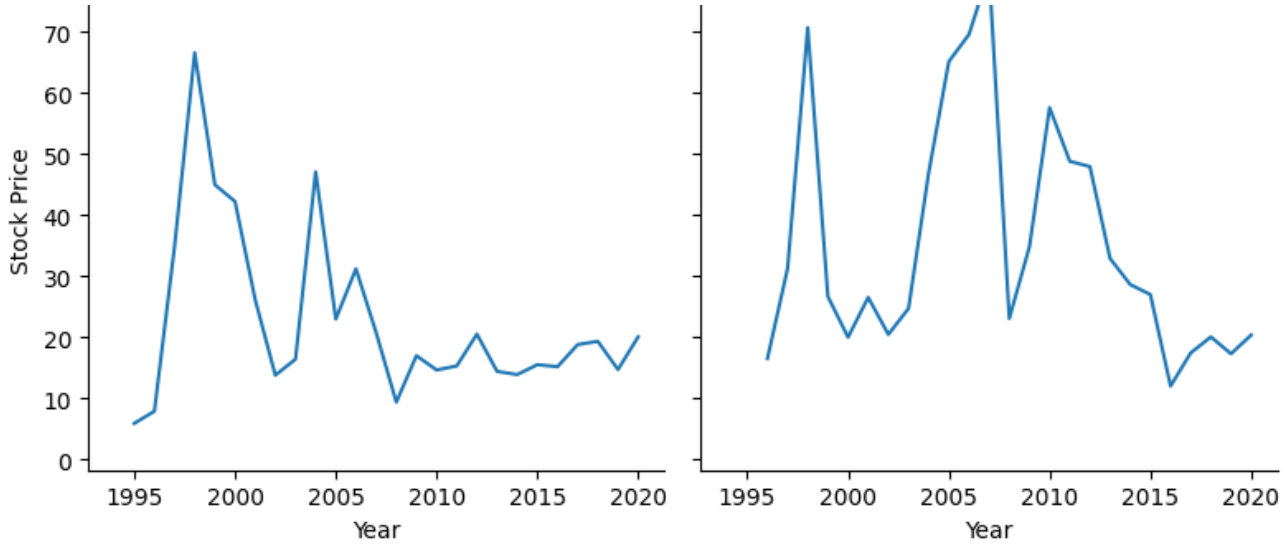
In [287…

```
g = sns.FacetGrid(data=similarity, col='conm', col_wrap=2, height=4)

# Map line plots to the FacetGrid
g.map(plt.plot, 'fyear', 'prcc_c')

# Set labels and titles
g.set_axis_labels('Year', 'Stock Price')
g.set_titles(col_template=' {col_name}')
g.fig.subplots_adjust(top=0.9)
g.fig.suptitle('Price Grid', fontsize=16)
plt.savefig('facetgrid.png')
```




Stock Price
70
60
50
40
30
20
10
0
1995
2000
2005
2010
2015
2020
Year
1995
2000
2005
2010
2015
2020
Year

In [ ]: