>>>Import Statements<<<

In [1]:

```
#Python Imports
import os
import sys
import csv
import json
import time
import itertools
import numpy as np
import pandas as pd
from fuzzywuzzy import fuzz
from fuzzywuzzy import process
from selenium import webdriver
from IPython.display import Image
from selenium.webdriver.common.by import By
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.chrome.options import Options
from selenium.webdriver.chrome.service import Service
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC

########
chrome_options = Options()
#chrome_options.add_argument("--headless") # Ensure GUI is off
#chrome_options.add_argument("--no-sandbox")
browser = webdriver.Chrome(options=chrome_options)
browser.implicitly_wait(15) # seconds
```

In [2]:

```
#######
search_term_raw = 'mark cross'
brand_url = 'https://www.markcross.com/pages/location'
########

search_term = search_term_raw + " in "

zipcodes = pd.read_json('OnePager/top_500_zipcodes.json')['zip'].apply(lambc
region_zips = [
    '94110', '90210', '10001', '20001', '98101',
    '60601', '77002', '30303', '02108', '33131',
    '80202', '92101', '85004', '98104', '75201',
    '60611', '75205', '19104', '30363', '98109'
]


three_zips = region_zips
#['94123', '19104', '77494']
```

DuckDuckGo First Page (Browser)

In [3]:

```
duckduckGo = []
url = "https://duckduckgo.com/?q=ulla+johnson&va=u&t=he&ia=web"
```

```
browser.get(url)

for zipcode in three_zips:
    try:
        #Submit & Search
        search_form = browser.find_element(By.CLASS_NAME, 'js-search-input')
        search_form.clear()
        search_form.send_keys(search_term+zipcode)
        submit = browser.find_element(By.CLASS_NAME, 'search__button')
        time.sleep(1)
        submit.click() #Could add a 'time.sleep(1)' above and below
        time.sleep(1)

        #Collect Results
        results_box = browser.find_element(By.ID, 'links')
        results = results_box.find_elements(By.CLASS_NAME, 'nrn-react-div')

        #Scraping Through the Results
        for idx, store in enumerate(results, start=1):
            resultsInfo = {}
            resultsInfo['title'] = store.find_element(By.CLASS_NAME, 'ikg2IX
            resultsInfo['url'] = store.find_element(By.CLASS_NAME, 'LnpumSTh
            #print(resultsInfo['url'])
            resultsInfo['page_rank'] = idx
            resultsInfo['zipcode'] = zipcode
            resultsInfo['search_engine'] = 'duckduckgo'
            resultsInfo['search_term'] = search_term+zipcode
            duckduckGo.append(resultsInfo)

        print(zipcode, ' - ', len(results))

    except Exception as e:
        #print(e)
        print('ERROR', str(e)[0:100]+"...")
        browser.get(url)
```

```
94110  -  10
90210  -  10
ERROR Message: no such element: Unable to locate element: {"method":"css se
lector","selector":"[id="links"...
ERROR Message: no such element: Unable to locate element: {"method":"css se
lector","selector":"[id="links"...
98101  -  10
60601  -  10
77002  -  10
30303  -  10
02108  -  10
33131  -  10
80202  -  10
92101  -  10
85004  -  10
98104  -  10
75201  -  10
60611  -  10
75205  -  10
19104  -  10
30363  -  10
98109  -  10
```

In [4]:

```
pd.DataFrame(duckduckGo).head()
```

Out[4]:

	title	url	page_rank	zipcode	search_en
0	Location - Mark Cross	https://www.markcross.com/pages/location	1	94110	duckdu
1	Mark Cross reflects a rich heritage of America...	https://www.markcross.com/	2	94110	duckdu
2	Mark Cross, California (78 matches): Phone Num...	https://www.spokeo.com/Mark-Cross/California	3	94110	duckdu
3	Mark Cross Profiles | Facebook	https://www.facebook.com/public/Mark-Cross	4	94110	duckdu
4	Crossbody Bags - Mark Cross	https://www.markcross.com/collections/crossbod...	5	94110	duckdu

StartPage First Page (Browser)

In [5]:

```
startPage = []
url = 'https://www.startpage.com/en/'
browser.get(url)

for zipcode in three_zips:
    #Submit & Search
    browser.get(url)
    submit=browser.find_element(By.ID, 'search-btn')
    search_form = browser.find_element(By.ID, 'q')
    search_form.clear()
    search_form.send_keys(search_term+zipcode)
    submit.click()
    time.sleep(2)

    #Collect Results
    results_box = browser.find_element(By.CLASS_NAME, 'w-gl')
    results = results_box.find_elements(By.CLASS_NAME, 'w-gl__result')

    #Scraping Through the Results
    for idx, store in enumerate(results, start=1):
        resultsInfo = {}
        resultsInfo['title'] = store.find_element(By.TAG_NAME, 'h3').text
        resultsInfo['url'] = store.find_element(By.CLASS_NAME, 'result-link'
        resultsInfo['page_rank'] = idx
        resultsInfo['zipcode'] = zipcode
        resultsInfo['search_engine'] = 'startpage'
        resultsInfo['search_term'] = search_term+zipcode
        startPage.append(resultsInfo)

    print(zipcode, ' - ', len(results))
```

```
94110  -  10
90210  -  10
10001  -  10
20001  -  8
98101  -  10
60601  -  10
77002  -  10
30303  -  10
02108  -  10
33131  -  10
80202  -  10
92101  -  10
85004  -  10
98104  -  10
75201  -  10
60611  -  10
75205  -  10
19104  -  10
30363  -  10
98109  -  10
```

In [6]:

```
pd.DataFrame(startPage).head()
```

Out[6]:

	title	url	page_rank	zipcode	search_eng
0	Mark Cross - The Flood Gallery	https://www.thefloodgallery.com/products/mark-...	1	94110	startpa
1	Mark Cross - University of California at Berke...	https://www.linkedin.com/in/crossfire	2	94110	startpa
2	MARK CROSS TO SHUT DOWN - Chicago Tribune	https://www.chicagotribune.com/news/ct-xpm-199...	3	94110	startpa
3	Location - Mark Cross	https://www.markcross.com/pages/location	4	94110	startpa
4	Mark Cross reflects a rich heritage of America...	https://www.markcross.com/	5	94110	startpa

Yahoo Search (Browser)

In [7]:

```
yahoo = []
url = "https://search.yahoo.com/search;_ylt=AwrEo_2emiRknKkNCTxDDWVH;_ylc=X1
browser.get(url)

for zipcode in three_zips:
    browser.get(url)
    #Submit & Search
    submit=browser.find_element(By.ID, 'sbq-submit')
    search_form = browser.find_element(By.ID, 'yschsp')
    search_form.clear()
    search_form.send_keys(search_term+zipcode)
    submit.click()
    time.sleep(2)

    #Collect Results
    results_box = browser.find_element(By.CLASS_NAME, 'searchCenterMiddle')
    results = results_box.find_elements(By.CLASS_NAME, 'algo')

    for idx, store in enumerate(results, start=1):
        resultsInfo = {}
        resultsInfo['title'] = store.find_element(By.CLASS_NAME, 'd-ib').get
        resultsInfo['url'] = store.find_element(By.CLASS_NAME, 'd-ib').get_a
        resultsInfo['page_rank'] = idx
        resultsInfo['zipcode'] = zipcode
        resultsInfo['search_engine'] = 'yahoo'
        resultsInfo['search_term'] = search_term+zipcode
        yahoo.append(resultsInfo)

    print(zipcode, ' - ', len(results))
```

```
94110  -  13
90210  -  12
10001  -  12
20001  -  12
98101  -  13
60601  -  13
77002  -  13
30303  -  13
02108  -  13
33131  -  13
80202  -  13
92101  -  13
85004  -  13
98104  -  13
75201  -  13
60611  -  13
75205  -  13
19104  -  13
30363  -  13
98109  -  13
```

In [8]:

```
pd.DataFrame(yahoo)
```

Out[8]:

	title	url	page_rank	zipcode	search
0	Location – Mark Cross	https://www.markcross.com/pages/location	1	94110	
1	MARK CROSS PA-C, NPI 1205964939 - Physician As...	https://npiprofile.com/npi/1205964939	2	94110	
2	Mark Cross | LinkedIn	https://www.linkedin.com/company/mark-cross	3	94110	
3	All Handbags – Mark Cross	https://www.markcross.com/collections/handbags	4	94110	
4	Mark Cross Profiles | Facebook	https://www.facebook.com/public/Mark-Cross	5	94110	
...	...	...	...	...	
252	Obituary for Mark J. Cross | Slattery Funeral ...	https://www.slatteryfuneralhome.com/obituary/M...	9	98109	
253	S. Mark Cross, Clinical Psychologist in Falls ...	https://mentaltherapy.io/psychologist/s-mark-c...	10	98109	
254	Mark Cross (748 matches): Phone Number, Email,...	https://www.spokeo.com/Mark-Cross	11	98109	
255	Mark A Kross, (970) 282-0890, Fort Collins — P...	https://clustrmaps.com/person/Kross-3dufon	12	98109	
256	Mark Cruz, Washington (28 matches): Phone Numb...	https://www.spokeo.com/Mark-Cruz/Washington	13	98109	

257 rows × 6 columns

Mojeek (Browser) - NEED HEAD?

In [9]:
```
mojeek = []
url = 'https://www.mojeek.com/'
browser.get(url)

for zipcode in three_zips:
    #Submit & Search
    browser.get(url)
    submit=browser.find_element(By.CLASS_NAME, 'search')
    search_form = browser.find_element(By.CLASS_NAME, 'js-search-input')
    search_form.clear()
    search_form.send_keys(search_term+zipcode)
    submit.click()
    time.sleep(2)

    try:
        #Collect Results
        results_box = browser.find_element(By.CLASS_NAME, 'results-standard'
        results = results_box.find_elements(By.TAG_NAME, 'li')

        #Scraping Through the Results
        for idx, store in enumerate(results, start=1):
            resultsInfo = {}
            resultsInfo['title'] = store.find_element(By.CLASS_NAME, 'title'
            resultsInfo['url'] = store.find_element(By.CLASS_NAME, 'title').
            resultsInfo['page_rank'] = idx
            resultsInfo['zipcode'] = zipcode
            resultsInfo['search_engine'] = 'mojeek'
            resultsInfo['search_term'] = search_term+zipcode

            mojeek.append(resultsInfo)
    except:
        print("no results/error")
        results = []

    print(zipcode, ' - ', len(results))
```

```
94110  –  10
90210  –  10
10001  –  10
20001  –  10
98101  –  10
60601  –  10
77002  –  10
30303  –  10
02108  –  10
33131  –  10
80202  –  10
92101  –  10
85004  –  10
98104  –  10
75201  –  10
60611  –  10
75205  –  10
19104  –  10
30363  –  10
98109  –  10
```

In [10]:
```
pd.DataFrame(mojeek).head()
```

Out[10]:

	title	url	page_rank	zipcode	search_e
0	Media Advisory: Man killed crossing the street...	https://walksf.org/news/for-reporters/press-re...	1	94110	n
1	1183 Hampshire St, San Francisco, CA 94110 - M...	https://www.coldwellbankerhomes.com/ca/san-fra...	2	94110	n
2	2391 Mission St San Francisco, CA 94110 - Reta...	https://www.showcase.com/2391-mission-st-san-f...	3	94110	n
3	Need body shop in south SF bay (Shameless cros...	https://www.thehondaforums.com/threads/need-bo...	4	94110	n
4	Teen with Strabismus filed under , strabismus,...	https://www.seevividly.com/picture/655/Teen_wi...	5	94110	n

Bing (Browser)

In [11]:
```
bing = []
url = "https://www.bing.com/search?q=ulla+johnson+19104&form=QBLH&sp=-1&ghc=
browser.get(url)

for zipcode in three_zips:
    browser.get(url)

    #Submit & Search
    submit=browser.find_element(By.ID, 'sb_go_par')
    search_form = browser.find_element(By.ID, 'sb_form_q')
    search_form.clear()
    search_form.send_keys(search_term+zipcode)
    submit.click()
    time.sleep(2)

    #PAGE 1
    results_box = browser.find_element(By.ID, 'b_results')
    results = results_box.find_elements(By.CLASS_NAME, 'b_algo')

    #Scraping Through the Results
    for idx, store in enumerate(results, start=1):
        resultsInfo = {}
        resultsInfo['title'] = store.find_element(By.TAG_NAME, 'a').get_attr
        resultsInfo['url'] =store.find_element(By.TAG_NAME, 'a').get_attribu
        #print(resultsInfo['url'])
        resultsInfo['page_rank'] = idx
        resultsInfo['zipcode'] = zipcode
        resultsInfo['search_engine'] = 'bing'
        resultsInfo['search_term'] = search_term+zipcode
        bing.append(resultsInfo)

    print(zipcode, ' - ', len(results))
```

```
94110  -  5
90210  -  10
10001  -  13
20001  -  14
98101  -  10
60601  -  10
77002  -  10
30303  -  10
02108  -  10
33131  -  10
80202  -  10
92101  -  10
85004  -  10
98104  -  10
75201  -  10
60611  -  10
75205  -  10
19104  -  10
30363  -  10
98109  -  10
```

In [12]:

```
pd.DataFrame(bing)
```

Out[12]:

	title	url	page_rank	zipcode	search_engine
0		https://www.markcross.com/pages/location	1	94110	bing
1		https://www.linkedin.com/company/mark-cross	2	94110	bing
2		https://www.whitepages.com/name/Mark-Cross	3	94110	bing
3		https://www.spokeo.com/Mark-Cross	4	94110	bing
4		https://www.linkedin.com/in/mark-cross-23aaa6a	5	94110	bing
...	...	...	...	...	...
197		https://www.spokeo.com/Mark-Cross/Washington	6	98109	bing
198		https://www.realtor.com/realestateagents/58ae8...	7	98109	bing
199		https://crosscountrymortgage.com/Seattle-WA-5531/	8	98109	bing
200		https://www.mylife.com/mark-cross/	9	98109	bing
201		https://www.linkedin.com/in/mark-cross-23aaa6a	10	98109	bing

202 rows × 6 columns

Yellow Pages (Shopping Specific)

In [13]:

```
yellowPages = []
url = 'https://www.yellowpages.com/'
browser.get(url)

for zipcode in three_zips:
    #Submit & Search
    submit=browser.find_element(By.TAG_NAME, 'button')
    search_form = browser.find_element(By.ID, 'query')
    location_form = browser.find_element(By.ID, 'location')
    search_form.clear()
    search_form.send_keys(search_term_raw)
    location_form.clear()
    location_form.send_keys(zipcode)
    submit.click()
    time.sleep(2)

    try:
        #Collect Results
        results_box = browser.find_element(By.CLASS_NAME, 'organic')
        results = results_box.find_elements(By.CLASS_NAME, 'result')
        #Scraping Through the Results
        for idx, store in enumerate(results, start=1):
            resultsInfo = {}
            resultsInfo['title'] = store.find_element(By.CLASS_NAME, 'busine
            resultsInfo['url'] = store.find_element(By.CLASS_NAME, 'business
            resultsInfo['page_rank'] = idx
            resultsInfo['zipcode'] = zipcode
            resultsInfo['search_engine'] = 'yellow_pages'
            resultsInfo['search_term'] = search_term_raw+ " " + zipcode
            yellowPages.append(resultsInfo)
        print(zipcode, ' - ', len(results))

    except Exception as e:
        print('ERROR', str(e)[0:75]+"...")

    browser.get(url)
```

```
94110  -  10
90210  -  30
10001  -  20
20001  -  5
98101  -  10
60601  -  11
77002  -  9
30303  -  7
02108  -  17
33131  -  4
80202  -  16
92101  -  8
85004  -  18
98104  -  10
75201  -  10
60611  -  11
75205  -  11
19104  -  7
30363  -  7
98109  -  10
```

In [14]:

```
pd.DataFrame(yellowPages)
```

Out[14]:

	title	url	page_rank	zipcode	search_engine	se
0	Syn, Mark N	https://www.yellowpages.com/san-francisco-ca/m...	1	94110	yellow_pages	
1	Phillips, Mark A, PA	https://www.yellowpages.com/emeryville-ca/mip/...	2	94110	yellow_pages	
2	Mark Ryan Fine Art	https://www.yellowpages.com/oakland-ca/mip/mar...	3	94110	yellow_pages	
3	Traves, Mark W	https://www.yellowpages.com/san-mateo-ca/mip/t...	4	94110	yellow_pages	
4	Mark Medders	https://www.yellowpages.com/hayward-ca/mip/mar...	5	94110	yellow_pages	
...	...	...	...	...	...	
226	Mark L Bowers, PA	https://www.yellowpages.com/renton-wa/mip/mark...	6	98109	yellow_pages	
227	Mark A Aytch, PA-C	https://www.yellowpages.com/kent-wa/mip/mark-a...	7	98109	yellow_pages	
228	Mark M Mashita, Other	https://www.yellowpages.com/everett-wa/mip/mar...	8	98109	yellow_pages	
229	Mark R. Arrant, PA	https://www.yellowpages.com/tacoma-wa/mip/mark...	9	98109	yellow_pages	
230	Mark Walther, PA-C	https://www.yellowpages.com/tacoma-wa/mip/mark...	10	98109	yellow_pages	

231 rows × 6 columns

Store Locator (Shopping Specific)

In [15]:

```
resultsList = []
url = brand_url

for i, zipcode in enumerate(three_zips, start=1):
    try:
        browser.get(url)
        time.sleep(2)

        query_entry=browser.find_element(By.CLASS_NAME, 'stockist-query-entr
        input_field = query_entry.find_element(By.TAG_NAME, 'input')
        submit = query_entry.find_element(By.CLASS_NAME, 'stockist-search-bu

        input_field.clear()
        input_field.send_keys(zipcode)
        time.sleep(5)
        submit.click()
        time.sleep(8)

        search_results = browser.find_element(By.CLASS_NAME, 'stockist-resul
        res=search_results.find_elements(By.CLASS_NAME,'stockist-result')

        print(zipcode, " results:", len(res), ' -', i)

        if len(res) != 0:
            for idx, store in enumerate(res):
                storeInfo = {}
                storeInfo['title'] = store.find_element(By.CLASS_NAME, 'stoc
                storeInfo['page_rank'] = idx+1
                storeInfo['zipcode'] = zipcode
                address = [line.get_attribute("textContent") for line in
                          store.find_element(By.CLASS_NAME, 'stockist-result
                storeInfo['address'] = ", ".join(address)
                storeInfo['search_term'] = search_term_raw
                storeInfo['search_engine'] = 'store_locator'
                storeInfo['url'] = store.find_element(By.CLASS_NAME, 'stocki
                resultsList.append(storeInfo)
        else:
            storeInfo = {}
            storeInfo['title'] = "Stockist Store Locator " + search_term_raw
            storeInfo['page_rank'] = idx+1
            storeInfo['zipcode'] = zipcode
            storeInfo['search_term'] = 'no_store_found'
            storeInfo['search_engine'] = 'store_locator'
            storeInfo['url'] = url
            resultsList.append(storeInfo)
            print(storeInfo['title'])

    except Exception as e:
        print("ERROR", zipcode, i, str(e)[0:75]+"...")
        time.sleep(2)

print("===Done===")
```

```
94110  results: 1  - 1
90210  results: 0  - 2
Stockist Store Locator mark cross
10001  results: 5  - 3
20001  results: 4  - 4
98101  results: 1  - 5
60601  results: 0  - 6
Stockist Store Locator mark cross
77002  results: 3  - 7
30303  results: 1  - 8
02108  results: 5  - 9
33131  results: 2  - 10
80202  results: 0  - 11
Stockist Store Locator mark cross
92101  results: 4  - 12
85004  results: 0  - 13
Stockist Store Locator mark cross
98104  results: 1  - 14
75201  results: 3  - 15
60611  results: 0  - 16
Stockist Store Locator mark cross
75205  results: 3  - 17
19104  results: 4  - 18
30363  results: 1  - 19
98109  results: 1  - 20
===Done===
```

In [16]:

```
pd.DataFrame(resultsList)
```

Out[16]:

	title	page_rank	zipcode	address	search_term	search_engine	
0	Elyse Walker	1	94110	1234 Adam St., St. Helena, California 94574, ...	mark cross	store_locator	ht
1	Stockist Store Locator mark cross	1	90210	NaN	no_store_found	store_locator	https://wv
2	Jonathan Cohen	1	10001	833 Madison Avenue, New York, New York 10021,...	mark cross	store_locator	ht
3	Elyse Walker	2	10001	926 Madison Avenue, New York, New York 10021,...	mark cross	store_locator	ht
4	Five Story	3	10001	1020 Madison Avenue, New York, New York 10075...	mark cross	store_locator	ht
5	Julianne	4	10001	274 Main St, Port Washington, New York 11050, NY	mark cross	store_locator	ht
6	VRSNL	5	10001	18 Newbury Street, Boston, Massachusetts 2116...	mark cross	store_locator	ht
7	Elyse Walker	1	20001	926 Madison Avenue, New York, New York 10021,...	mark cross	store_locator	ht
8	Five Story	2	20001	1020 Madison Avenue, New York, New York 10075...	mark cross	store_locator	ht
9	Jonathan Cohen	3	20001	833 Madison Avenue, New York, New York 10021,...	mark cross	store_locator	ht
10	Julianne	4	20001	274 Main St, Port Washington, New York 11050, NY	mark cross	store_locator	ht
11	Amazon	1	98101	Seattle	mark cross	store_locator	ht
12	Stockist Store	1	60601	NaN	no_store_found	store_locator	https://wv

	title	page_rank	zipcode	address	search_term	search_engine	
	Locator mark cross						
13	Forty Five Ten	1	77002	1615 Main Street, Dallas, Texas 75201, United...	mark cross	store_locator	ht
14	Forty Five Ten	2	77002	60 Highland Park Village, Dallas, Texas 75205...	mark cross	store_locator	ht
15	Market Highland Park Village	3	77002	26 Highland Park Village, Dallas, Texas 75205...	mark cross	store_locator	ht
16	Capitol	1	30303	4010 Sharon Rd, Charlotte, North Carolina 282...	mark cross	store_locator	ht
17	VRSNL	1	02108	18 Newbury Street, Boston, Massachusetts 2116...	mark cross	store_locator	ht
18	Julianne	2	02108	274 Main St, Port Washington, New York 11050, NY	mark cross	store_locator	ht
19	Elyse Walker	3	02108	926 Madison Avenue, New York, New York 10021,...	mark cross	store_locator	ht
20	Five Story	4	02108	1020 Madison Avenue, New York, New York 10075...	mark cross	store_locator	ht
21	Jonathan Cohen	5	02108	833 Madison Avenue, New York, New York 10021,...	mark cross	store_locator	ht
22	Marissa Collections	1	33131	340 Royal Poinciana Way M337, Palm Beach, Flor...	mark cross	store_locator	ht
23	Marissa Collections	2	33131	1167 3rd Street, Naples, Florida 34102, Unite...	mark cross	store_locator	ht
24	Stockist Store Locator mark cross	2	80202	NaN	no_store_found	store_locator	https://wv

	title	page_rank	zipcode	address	search_term	search_engine	
25	Elyse Walker	1	92101	3444 Via Lido, Newport Beach, California 9266...	mark cross	store_locator	ht
26	Capitol	2	92101	Brentwood Country Mart, 225 26th St Suite 38A,...	mark cross	store_locator	ht
27	Elyse Walker	3	92101	15306 Antioch Street, Pacific Palisades, Calif...	mark cross	store_locator	ht
28	Elyse Walker	4	92101	4719 Commons Way, Suite J, Calabasas, Californ...	mark cross	store_locator	ht
29	Stockist Store Locator mark cross	4	85004	NaN	no_store_found	store_locator	https://wv
30	Amazon	1	98104	Seattle	mark cross	store_locator	ht
31	Forty Five Ten	1	75201	1615 Main Street, Dallas, Texas 75201, United...	mark cross	store_locator	ht
32	Forty Five Ten	2	75201	60 Highland Park Village, Dallas, Texas 75205...	mark cross	store_locator	ht
33	Market Highland Park Village	3	75201	26 Highland Park Village, Dallas, Texas 75205...	mark cross	store_locator	ht
34	Stockist Store Locator mark cross	3	60611	NaN	no_store_found	store_locator	https://wv
35	Forty Five Ten	1	75205	60 Highland Park Village, Dallas, Texas 75205...	mark cross	store_locator	ht
36	Market Highland Park Village	2	75205	26 Highland Park Village, Dallas, Texas 75205...	mark cross	store_locator	ht
37	Forty Five Ten	3	75205	1615 Main Street, Dallas,	mark cross	store_locator	ht

	title	page_rank	zipcode	address	search_term	search_engine	
				Texas 75201, United...			
38	Elyse Walker	1	19104	926 Madison Avenue, New York, New York 10021,...	mark cross	store_locator	ht
39	Five Story	2	19104	1020 Madison Avenue, New York, New York 10075...	mark cross	store_locator	ht
40	Jonathan Cohen	3	19104	833 Madison Avenue, New York, New York 10021,...	mark cross	store_locator	ht
41	Julianne	4	19104	274 Main St, Port Washington, New York 11050, NY	mark cross	store_locator	ht
42	Capitol	1	30363	4010 Sharon Rd, Charlotte, North Carolina 282...	mark cross	store_locator	ht
43	Amazon	1	98109	Seattle	mark cross	store_locator	ht

Google Search (Browser)

In [17]:
```
google = []
url = 'https://www.google.com/search?q=google'
browser.get(url)

for zipcode in three_zips:
    try:
        base_url = 'https://www.google.com/search?q=' + search_term_raw.repl
        browser.get(base_url)

        for i in range(0,2):
            browser.execute_script("window.scrollTo(0,document.body.scrollHe
            time.sleep(3)
            #print('scrolling...')
            try:
                more = browser.find_element(By.CLASS_NAME, 'RVQdVd')
                more.click()
                #print('load more click!')
            except:
                pass
                #print('pass', current_combo, ' - ', i)

        print("~done scrolling~")
        results = browser.find_elements(By.CLASS_NAME, 'yuRUbf')
        print(search_term_raw + ' - ', str(len(results)))

        for idx, blueLink in enumerate(results, 1):
            resultInfo = {}
            resultInfo['title'] = blueLink.find_element(By.TAG_NAME, 'a').fi
            resultInfo['url'] = blueLink.find_element(By.TAG_NAME, 'a').get_
            resultInfo['page_rank'] = idx
            resultInfo['zipcode'] = zipcode
            resultInfo['search_engine'] = 'google_search'
            resultInfo['search_term'] = str(search_term_raw.replace(" ", "+"
            try:
                #resultInfo['link_website'] = blueLink.find_element(By.TAG_N
                resultInfo['title'] = resultInfo['title'] + " " + blueLink.f
            except:
                pass
            google.append(resultInfo)

        browser.quit()
        browser = webdriver.Chrome(options=chrome_options) #can add 'sleep(2
    except Exception as e:
        #print(e)
        print('ERROR', str(e)[0:100]+"...")
        browser.quit()
        browser = webdriver.Chrome(options=chrome_options) #can add 'sleep(2
~done scrolling~
mark cross -  31
~done scrolling~
mark cross -  51
~done scrolling~
mark cross -  31
~done scrolling~
mark cross -  19
~done scrolling~
mark cross -  49
~done scrolling~
mark cross -  31
~done scrolling~
mark cross -  31
~done scrolling~
mark cross -  32
~done scrolling~
mark cross -  31
~done scrolling~
mark cross -  29
~done scrolling~
mark cross -  29
~done scrolling~
mark cross -  49
~done scrolling~
mark cross -  29
~done scrolling~
mark cross -  39
~done scrolling~
mark cross -  29
~done scrolling~
mark cross -  19
~done scrolling~
mark cross -  29
~done scrolling~
mark cross -  9
~done scrolling~
mark cross -  51
~done scrolling~
mark cross -  19
```

```
In [18]: pd.DataFrame(google)
```

Out[18]:

	title	url	page_rank	zipcode	s
0	Mark Cross The Flood Gallery	https://www.thefloodgallery.com/products/mark-...	1	94110	9
1	Mark Cross - University of California at Berke...	https://www.linkedin.com/in/crossfire	2	94110	9
2	MARK CROSS TO SHUT DOWN Chicago Tribune	https://www.chicagotribune.com/news/ct-xpm-199...	3	94110	9
3	Location Mark Cross	https://www.markcross.com/pages/location	4	94110	9
4	Mark Cross reflects a rich heritage of America...	https://www.markcross.com/	5	94110	9
...	...	...	...	...	
632	Seattle, Seattle, WA, 98109 - Restaurant For S...	https://www.loopnet.com/Listing/18255017/Seatt...	15	98109	9
633	Boneless Ribeye Mishima Reserve	https://www.mishimareserve.com/our-products/bo...	16	98109	9
634	285 8th Ave N Seattle WA 98109 Commercial Brok...	https://www.commercialmls.com/Search/ListingDe...	17	98109	9
635	Systems Immunogenetics of Biodefense in the Co...	https://www.galelab.org/collaborative-cross	18	98109	9
636	Find a Doctor | Swedish Health Services Swedis...	https://schedule.swedish.org/	19	98109	9

637 rows × 6 columns

Google Shopping (Shopping Specific)

In [19]:

```
def get_storeInfo(store, zipcode_here, current_rank_here, term):
    data = {}
    data['title'] = store.find_element(By.CLASS_NAME, 'MxVeme').text
    data['page_rank'] = current_rank_here
    data['zipcode'] = zipcode_here
    data['search_engine'] = 'google_shopping'
    data['search_term'] = term
    data['url'] = store.find_element(By.CLASS_NAME, 'k7eIUb').find_element(B
    data['address'] = store.find_element(By.CLASS_NAME, 'lSS0Af').text
    return data
```

In [20]:

```
google_shopping = []
url_base = 'https://www.google.com/search?q=*&tbm=shop'
browser.get(url)

for zipcode in three_zips:
    try:
        current_combo = url_base.replace("*", search_term_raw.replace(" ", "
        term_str = search_term_raw.replace(" ", "+") + "+in+"+str(zipcode)
        browser.get(current_combo) #Get the link
        morePlaces = True

        results = browser.find_element(By.XPATH, '//div[@jscontroller="lcX38
        stores = results.find_elements(By.CLASS_NAME, 'FFnM0')
        print("LEN:", len(stores), zipcode)

        while morePlaces == True:
            #for length in range(len(stores)-3): #How many times to click th
            try:
                button = results.find_element(By.CLASS_NAME, 't6JUTe')
                button.click()
                time.sleep(1)
            except:
                pass #print("no more 'more places' button")
                morePlaces = False

        for idx, store in enumerate(stores):
            google_shopping.append(get_storeInfo(store, zipcode, idx, term_s
        #time.sleep(1)

    except Exception as e:
        print('ERROR', str(e)[0:75]+"...")

    browser.quit()
    browser = webdriver.Chrome(options=chrome_options)
```

```
LEN: 28 94110
ERROR Message: no such element: Unable to locate element: {"method":"xpat
h","sele...
LEN: 16 10001
ERROR Message: no such element: Unable to locate element: {"method":"xpat
h","sele...
LEN: 45 98101
LEN: 24 60601
LEN: 11 77002
LEN: 3 30303
LEN: 16 02108
LEN: 7 33131
LEN: 15 80202
LEN: 9 92101
LEN: 6 85004
LEN: 12 98104
LEN: 15 75201
LEN: 26 60611
LEN: 15 75205
LEN: 40 19104
LEN: 15 30363
LEN: 5 98109
```

In [21]:

```python
pd.DataFrame(google_shopping)
```

Out[21]:

	title	page_rank	zipcode	search_engine	search_term	
0	Nordstrom	0	94110	google_shopping	mark+cross+in+94110	https://maps. dad
1	Bloomingdale's	1	94110	google_shopping	mark+cross+in+94110	https://maps. dad
2	Neiman Marcus	2	94110	google_shopping	mark+cross+in+94110	https://maps. da
3	Neiman Marcus	3	94110	google_shopping	mark+cross+in+94110	https://maps. da
4	Bloomingdale's	4	94110	google_shopping	mark+cross+in+94110	https://maps. dac
...	...	...	...	...	...	
303	Nordstrom	0	98109	google_shopping	mark+cross+in+98109	https://maps. da
304	Grainger Industrial Supply	1	98109	google_shopping	mark+cross+in+98109	https://maps. dac
305	Hallmark	2	98109	google_shopping	mark+cross+in+98109	https://maps. dad
306	Tractor Supply Company	3	98109	google_shopping	mark+cross+in+98109	https://maps. c
307	Dightmans Bible Book Center	4	98109	google_shopping	mark+cross+in+98109	https://maps. dad

308 rows × 7 columns

Brave Search (Browser)

In [22]:

```
'''
braveSearch = []
url = 'https://search.brave.com/'
browser.get(url)

for zipcode in three_zips:
    #Submit & Search
    browser.get(url)
    submit=browser.find_element(By.ID, 'submit-button')
    search_form = browser.find_element(By.ID, 'searchbox')
    search_form.clear()
    search_form.send_keys(search_term+zipcode)
    submit.click()
    time.sleep(2)

    #Collect Results
    results_box = browser.find_element(By.ID, 'results')
    results = results_box.find_elements(By.CLASS_NAME, 'fdb')

    #Scraping Through the Results
    for idx, store in enumerate(results, start=1):
        resultsInfo = {}
        resultsInfo['title'] = store.find_element(By.CLASS_NAME, 'snippet-ti
        resultsInfo['url'] = store.find_element(By.CLASS_NAME, 'result-heade
        resultsInfo['page_rank'] = idx
        resultsInfo['zipcode'] = zipcode
        resultsInfo['search_engine'] = 'brave_search'
        resultsInfo['search_term'] = search_term+zipcode
        braveSearch.append(resultsInfo)

    print(zipcode, ' - ', len(results))
    browser.quit()

    chrome_options = Options()
    #chrome_options.add_argument("--headless") # Ensure GUI is off
    #chrome_options.add_argument("--no-sandbox")
    browser = webdriver.Chrome(options=chrome_options)
    browser.implicitly_wait(15) # seconds

    browser = webdriver.Chrome(options=chrome_options)

    time.sleep(2)
'''
```

Out[22]:

```
'\nbraveSearch = []\nurl = \'https://search.brave.com/\'\nbrowser.get(url)
\n\nfor zipcode in three_zips:\n    #Submit & Search\n    browser.get(url)
\n    submit=browser.find_element(By.ID, \'submit-button\')\n    search_for
m = browser.find_element(By.ID, \'searchbox\')\n    search_form.clear()\n
search_form.send_keys(search_term+zipcode)\n    submit.click()\n    time.sl
eep(2)\n    \n    #Collect Results\n    results_box = browser.find_element
(By.ID, \'results\')\n    results = results_box.find_elements(By.CLASS_NAM
E, \'fdb\')\n\n    #Scraping Through the Results\n    for idx, store in enu
merate(results, start=1):\n        resultsInfo = {}\n        resultsInfo
[\'title\'] = store.find_element(By.CLASS_NAME, \'snippet-title\').get_attr
ibute("textContent").strip()\n        resultsInfo[\'url\'] = store.find_ele
ment(By.CLASS_NAME, \'result-header\').get_attribute(\'href\')\n        res
ultsInfo[\'page_rank\'] = idx\n        resultsInfo[\'zipcode\'] = zipcode\n
resultsInfo[\'search_engine\'] = \'brave_search\'\n        resultsInfo[\'se
arch_term\'] = search_term+zipcode\n        braveSearch.append(resultsInfo)
\n        \n    print(zipcode, \' - \', len(results))\n    browser.quit()\n
\n    chrome_options = Options()\n    #chrome_options.add_argument("--headl
ess") # Ensure GUI is off\n    #chrome_options.add_argument("--no-sandbox")
\n    browser = webdriver.Chrome(options=chrome_options)\n    browser.impli
citly_wait(15) # seconds\n    \n    browser = webdriver.Chrome(options=chro
me_options)\n    \n    time.sleep(2)\n'
```

In [23]:

```
#pd.DataFrame(braveSearch).head()
```

>>> Combine Dataframes <<<

- 'search_term_raw'
- Shopping Specific gives store name, while Browser gives web page name

In [24]:

```
bing_df = pd.DataFrame(bing)
yahoo_df = pd.DataFrame(yahoo)
mojeek_df = pd.DataFrame(mojeek)
google_df = pd.DataFrame(google)
startPage_df = pd.DataFrame(startPage)
duckduckGo_df = pd.DataFrame(duckduckGo)
#braveSearch_df = pd.DataFrame(braveSearch)
yellowPages_df = pd.DataFrame(yellowPages)
google_shopping_df = pd.DataFrame(google_shopping)
results_df = pd.DataFrame(resultsList)

combined_df = pd.concat([bing_df, yahoo_df, mojeek_df, startPage_df, duckduc
                         #braveSearch_df,
                         yellowPages_df, google_df, google_shopping_df, resu

combined_df['search_term_raw'] = pd.Series(search_term_raw, index=combined_c

df_dict = combined_df.to_dict(orient='records')
#df_dict
```

In [25]:

```
combined_df
```

Out[25]:

	title	url	page_rank	zipcode	search_engine	s
0		https://www.markcross.com/pages/location	1	94110	bing	m
1		https://www.linkedin.com/company/mark-cross	2	94110	bing	m
2		https://www.whitepages.com/name/Mark-Cross	3	94110	bing	m
3		https://www.spokeo.com/Mark-Cross	4	94110	bing	m
4		https://www.linkedin.com/in/mark-cross-23aaa6a	5	94110	bing	m
...	...	...	...	...	...	
39	Five Story	https://www.google.com/maps/dir/?api=1&destina...	2	19104	store_locator	
40	Jonathan Cohen	https://www.google.com/maps/dir/?api=1&destina...	3	19104	store_locator	
41	Julianne	https://www.google.com/maps/dir/?api=1&destina...	4	19104	store_locator	
42	Capitol	https://www.google.com/maps/dir/?api=1&destina...	1	30363	store_locator	
43	Amazon	https://www.google.com/maps/dir/?api=1&destina...	1	98109	store_locator	

2257 rows × 8 columns

Is this result a store/retailer?

- Via store terms + Stockist threshold

In [26]:

```
#####
threshold = 70

def inStockistValue(str1, str2):
    wratio = fuzz.WRatio(str1, str2)
    token_set = fuzz.token_set_ratio(str1, str2)
    return (wratio+token_set)/2
'''
def isNameOfStore(potentialstore_str):
    isStore = True
    #INSERT CODE HERE
    if isStore == True:
        return 0
    else:
        return 1
'''
#####

s1 = "Le Specs | Designer | NET-A-PORTER"
s2 = "Le Specs Stockist"
s_list = ["Liquor Store Katy - Alcohol Delivery Spec's Beer & Wine", "Specs

def extractTop(search_result, choices):
    val_list = []
    for choice in choices:
        val_list.append((choice, inStockistValue(search_result, choice)))
    max_ratio_item = max(val_list, key=lambda x: x[1])
    return max_ratio_item

print(inStockistValue(s1, s2))
print(extractTop(s2, s_list))
```

```
75.0
('Specs Beer & Wine', 54.5)
```

In [27]:

```
unique_zips = combined_df['zipcode'].unique().tolist()

updated_dict = []

for zcode in unique_zips:
    locator_stores_list = combined_df[(combined_df['search_engine'] == 'stor
    locator_stores_list.append(search_term_raw + " Stockist Store Locator")

    print(zcode, '\n', locator_stores_list)
    loop_dict = combined_df[(combined_df['search_engine'] != 'store_locator'

    for search_result in loop_dict:
        search_result['extractTop_name'] = extractTop(search_result['title']
        search_result['extractTop_value'] = extractTop(search_result['title'
        if search_result['extractTop_value'] >= threshold:
            search_result['is_stockist_store'] = 1
        else:
            search_result['is_stockist_store'] = 0
    updated_dict += loop_dict
```

```
94110
 ['Elyse Walker', 'mark cross Stockist Store Locator']
90210
 ['Stockist Store Locator mark cross', 'mark cross Stockist Store Locator']
10001
 ['Jonathan Cohen ', 'Elyse Walker', 'Five Story', 'Julianne ', 'VRSNL', 'mark cross Stockist Store Locator']
20001
 ['Elyse Walker', 'Five Story', 'Jonathan Cohen ', 'Julianne ', 'mark cross Stockist Store Locator']
98101
 ['Amazon', 'mark cross Stockist Store Locator']
60601
 ['Stockist Store Locator mark cross', 'mark cross Stockist Store Locator']
77002
 ['Forty Five Ten', 'Forty Five Ten', 'Market Highland Park Village', 'mark cross Stockist Store Locator']
30303
 ['Capitol', 'mark cross Stockist Store Locator']
02108
 ['VRSNL', 'Julianne ', 'Elyse Walker', 'Five Story', 'Jonathan Cohen ', 'mark cross Stockist Store Locator']
33131
 ['Marissa Collections', 'Marissa Collections', 'mark cross Stockist Store Locator']
80202
 ['Stockist Store Locator mark cross', 'mark cross Stockist Store Locator']
92101
 ['Elyse Walker', 'Capitol', 'Elyse Walker', 'Elyse Walker', 'mark cross Stockist Store Locator']
85004
 ['Stockist Store Locator mark cross', 'mark cross Stockist Store Locator']
98104
 ['Amazon', 'mark cross Stockist Store Locator']
75201
 ['Forty Five Ten', 'Forty Five Ten', 'Market Highland Park Village', 'mark cross Stockist Store Locator']
60611
 ['Stockist Store Locator mark cross', 'mark cross Stockist Store Locator']
75205
 ['Forty Five Ten', 'Market Highland Park Village', 'Forty Five Ten', 'mark cross Stockist Store Locator']
19104
 ['Elyse Walker', 'Five Story', 'Jonathan Cohen ', 'Julianne ', 'mark cross Stockist Store Locator']
30363
 ['Capitol', 'mark cross Stockist Store Locator']
98109
 ['Amazon', 'mark cross Stockist Store Locator']
```

In [28]:

```
new_df = pd.DataFrame(updated_dict)
df = pd.concat([new_df, results_df])
df
```

Out[28]:

	title	url	page_rank	zipcode	search_engine	s
0		https://www.markcross.com/pages/location	1	94110	bing	m
1		https://www.linkedin.com/company/mark-cross	2	94110	bing	m
2		https://www.whitepages.com/name/Mark-Cross	3	94110	bing	m
3		https://www.spokeo.com/Mark-Cross	4	94110	bing	m
4		https://www.linkedin.com/in/mark-cross-23aaa6a	5	94110	bing	m
...	...	...	...	...	...	
39	Five Story	https://www.google.com/maps/dir/?api=1&destina...	2	19104	store_locator	
40	Jonathan Cohen	https://www.google.com/maps/dir/?api=1&destina...	3	19104	store_locator	
41	Julianne	https://www.google.com/maps/dir/?api=1&destina...	4	19104	store_locator	
42	Capitol	https://www.google.com/maps/dir/?api=1&destina...	1	30363	store_locator	
43	Amazon	https://www.google.com/maps/dir/?api=1&destina...	1	98109	store_locator	

2257 rows × 11 columns

Is this result a store name?

- Need metric threshold to say it is

In [29]:

```
df.to_csv("mark_cross_mega.csv")
```