CitySearch Web Scraping

Implementation (8/20/24)

Importing libraries

In [ ]:

```python
import pandas as pd
import time
import re

from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.common.exceptions import TimeoutException
```

Navigating to the main page of CitySearch

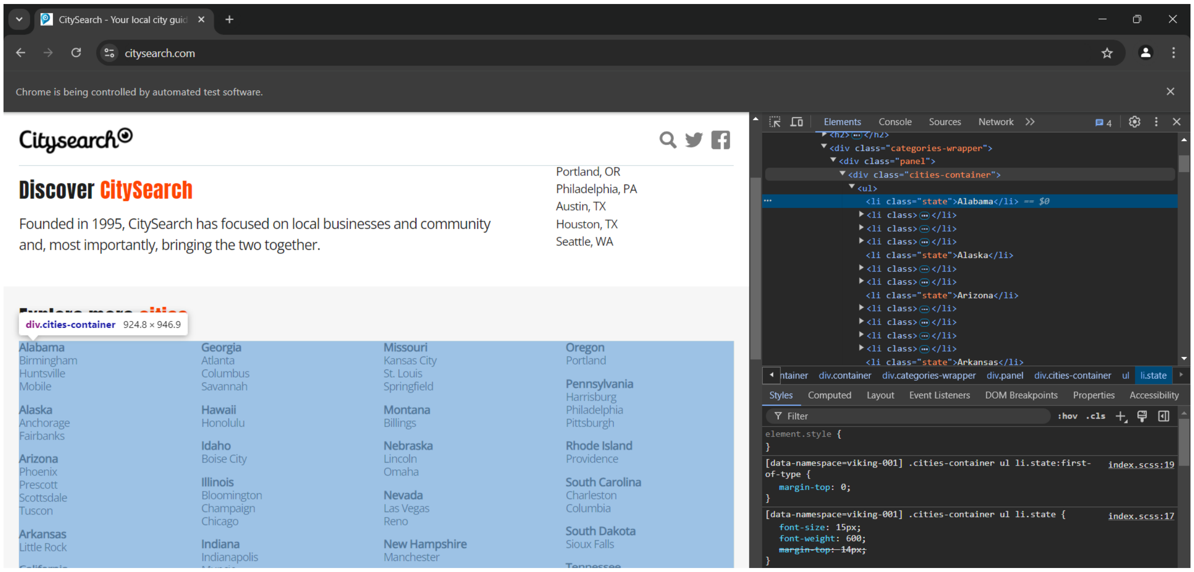
In [ ]:

```python
driver = webdriver.Chrome()
driver.get("https://www.citysearch.com/")
```

Extracting the links to individual cities

In [ ]:

```python
container = driver.find_element(By.CSS_SELECTOR, "div.cities-container
cities = container.find_elements(By.CSS_SELECTOR, "li:not([class*='sta

city_links = [city.get_attribute("href") for city in cities]
state, city = re.search("(?<=\.com\/).*", city_links[0]).group().split
```



Discover CitySearch

Founded in 1995, CitySearch has focused on local businesses and community and, most importantly, bringing the two together.

Portland, OR
Philadelphia, PA
Austin, TX
Houston, TX
Seattle, WA

Explore more cities

Alabama
Birmingham
Huntsville
Mobile

Alaska
Anchorage
Fairbanks

Arizona
Phoenix
Prescott
Scottsdale
Tuscon

Arkansas
Little Rock

Georgia
Atlanta
Columbus
Savannah

Hawaii
Honolulu

Idaho
Boise City

Illinois
Bloomington
Champaign
Chicago

Indiana
Indianapolis

Missouri
Kansas City
St. Louis
Springfield

Montana
Billings

Nebraska
Lincoln
Omaha

Nevada
Las Vegas
Reno

New Hampshire
Manchester

Oregon
Portland

Pennsylvania
Harrisburg
Philadelphia
Pittsburgh

Rhode Island
Providence

South Carolina
Charleston
Columbia

South Dakota
Sioux Falls

```
<h2>…</h2>
<div class="categories-wrapper">
  <div class="panel">
    <div class="cities-container">
      <ul>
        <li class="state">Alabama</li> == $0
        <li class>…</li>
        <li class>…</li>
        <li class>…</li>
        <li class="state">Alaska</li>
        <li class>…</li>
        <li class>…</li>
        <li class="state">Arizona</li>
        <li class>…</li>
        <li class>…</li>
        <li class>…</li>
        <li class>…</li>
        <li class="state">Arkansas</li>
```

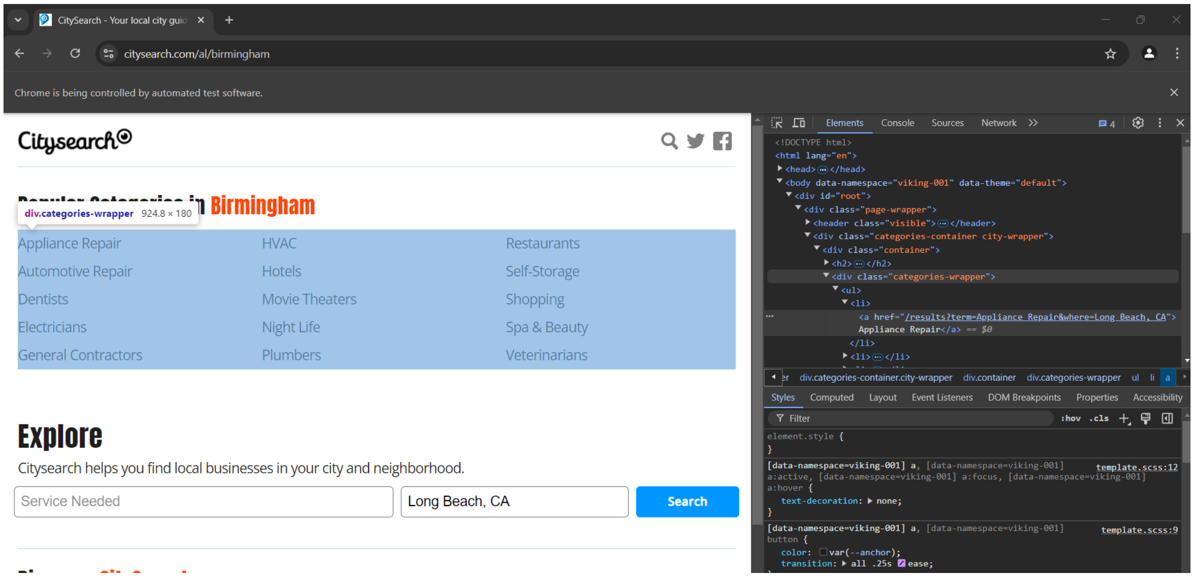
```
element.style {
}
[data-namespace=viking-001] .cities-container ul li.state:first-    index.scss:19
of-type {
  margin-top: 0;
}
[data-namespace=viking-001] .cities-container ul li.state {         index.scss:17
  font-size: 15px;
  font-weight: 600;
  ~~margin-top: 14px;~~
}
```

Navigating to a city link and gathering popular jobs

If we have keywords of specific industries we're interested in, I can iterate over them instead of iterating over popular industries. Also, if we have a list of states or cities we're interested in, I can also iterate over those.

In [ ]:

```
driver.get(city_links[0]) # as an example will be going through the jo
try:
    elem = WebDriverWait(driver, 10).until(EC.presence_of_element_loca
except TimeoutException:
    print("Timed out waiting for page to load")

popular_jobs = driver.find_elements(By.CSS_SELECTOR, 'div.categories-w
popular_jobs_links = [job.get_attribute("href") for job in popular_job
```


CitySearch - Your local city guid
citysearch.com/al/birmingham
Chrome is being controlled by automated test software.
Citysearch
div.categories-wrapper 924.8 × 180
Elements
Console
Sources
Network
4
er
div.categories-container.city-wrapper
div.container
div.categories-wrapper
ul
li
a
Styles
Computed
Layout
Event Listeners
DOM Breakpoints
Properties
Accessibility
Filter
:hov
.cls
Search

Popular Categories in Birmingham

Appliance Repair
Automotive Repair
Dentists
Electricians
General Contractors

HVAC
Hotels
Movie Theaters
Night Life
Plumbers

Restaurants
Self-Storage
Shopping
Spa & Beauty
Veterinarians

Explore

Citysearch helps you find local businesses in your city and neighborhood.

Service Needed | Long Beach, CA

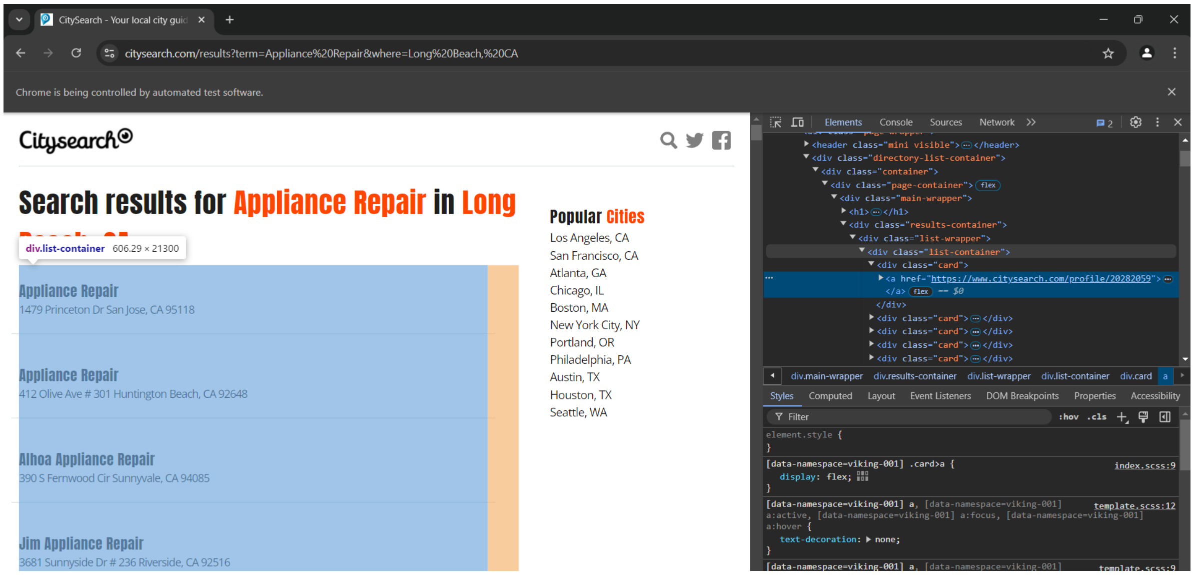
```
<!DOCTYPE html>
<html lang="en">
<head>…</head>
<body data-namespace="viking-001" data-theme="default">
  <div id="root">
    <div class="page-wrapper">
      <header class="visible">…</header>
      <div class="categories-container city-wrapper">
        <div class="container">
          <h2>…</h2>
          <div class="categories-wrapper">
            <ul>
              <li>
                <a href="/results?term=Appliance Repair&where=Long Beach, CA">
                Appliance Repair</a> == $0
              </li>
              <li>…</li>
```

```
element.style {
}
[data-namespace=viking-001] a, [data-namespace=viking-001]        template.scss:12
a:active, [data-namespace=viking-001] a:focus, [data-namespace=viking-001]
a:hover {
  text-decoration: ▸ none;
}
[data-namespace=viking-001] a, [data-namespace=viking-001]        template.scss:9
button {
  color: var(--anchor);
  transition: ▸ all .25s ease;
```

Navigating to first popular job and extracting links to jobs

In [ ]:

```python
driver.get(popular_jobs_links[0])

try:
    elem = WebDriverWait(driver, 10).until(EC.presence_of_element_loca
except TimeoutException:
    print("Timed out waiting for page to load")

job_cards = driver.find_elements(By.CSS_SELECTOR, "div.list-container
job_cards_links = [job.get_attribute("href") for job in job_cards]
```



Chrome is being controlled by automated test software.

Search results for Appliance Repair in Long

div.list-container 606.29 × 21300

Appliance Repair
1479 Princeton Dr San Jose, CA 95118

Appliance Repair
412 Olive Ave # 301 Huntington Beach, CA 92648

Alhoa Appliance Repair
390 S Fernwood Cir Sunnyvale, CA 94085

Jim Appliance Repair
3681 Sunnyside Dr # 236 Riverside, CA 92516

Popular Cities
Los Angeles, CA
San Francisco, CA
Atlanta, GA
Chicago, IL
Boston, MA
New York City, NY
Portland, OR
Philadelphia, PA
Austin, TX
Houston, TX
Seattle, WA

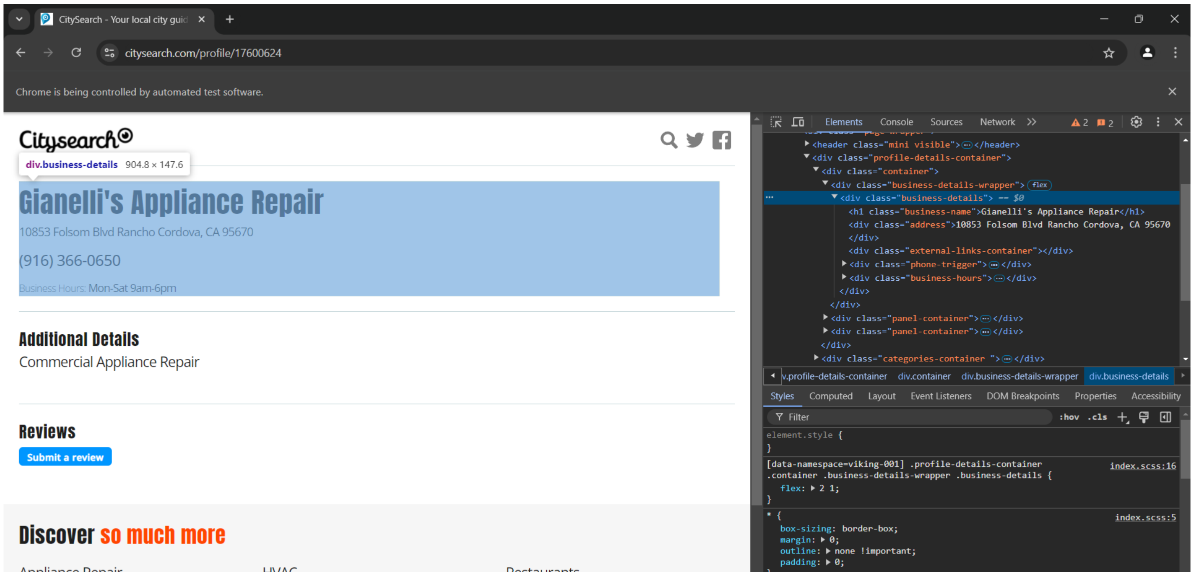
```html
<header class="mini visible">…</header>
<div class="directory-list-container">
  <div class="container">
    <div class="page-container"> flex
      <div class="main-wrapper">
        <h1>…</h1>
        <div class="results-container">
          <div class="list-wrapper">
            <div class="list-container">
              <div class="card">
                <a href="https://www.citysearch.com/profile/20282059">…
                </a> flex == $0
              </div>
              <div class="card">…</div>
              <div class="card">…</div>
              <div class="card">…</div>
              <div class="card">…</div>
```

```css
element.style {
}
[data-namespace=viking-001] .card>a {                    index.scss:9
  display: flex;
}
[data-namespace=viking-001] a, [data-namespace=viking-001]   template.scss:12
a:active, [data-namespace=viking-001] a:focus, [data-namespace=viking-001]
a:hover {
  text-decoration: ▶ none;
}
[data-namespace=viking-001] a, [data-namespace=viking-001]    template.scss:9
```

Scraping job description

In [ ]:

```python
business_list = []

for i in range(0, 5): #5 should be job_cards_link's length when implem

    driver.get(job_cards_links[i])

    try:
        elem = WebDriverWait(driver, 10).until(EC.presence_of_element_
    except TimeoutException:
        print("Timed out waiting for page to load")

    # not sure if all business have all their contact info so creating
    business_details = driver.find_elements(By.CSS_SELECTOR, 'div.busi

    business_details_dict = {
        entry.get_attribute("class"): entry.text
        for entry in business_details
    }

    business_list.append(business_details_dict)
    time.sleep(3)
```



Chrome is being controlled by automated test software.

Gianelli's Appliance Repair

10853 Folsom Blvd Rancho Cordova, CA 95670

(916) 366-0650

Business Hours: Mon-Sat 9am-6pm

Additional Details

Commercial Appliance Repair

Reviews

Submit a review

Discover so much more

Appliance Repair HVAC Restaurants

```html
▶<header class="mini visible">⋯</header>
▼<div class="profile-details-container">
  ▼<div class="container">
    ▼<div class="business-details-wrapper"> flex
      ▼<div class="business-details"> == $0
          <h1 class="business-name">Gianelli's Appliance Repair</h1>
          <div class="address">10853 Folsom Blvd Rancho Cordova, CA 95670
          </div>
          <div class="external-links-container"></div>
        ▶<div class="phone-trigger">⋯</div>
        ▶<div class="business-hours">⋯</div>
        </div>
      </div>
    ▶<div class="panel-container">⋯</div>
    ▶<div class="panel-container">⋯</div>
    </div>
  ▶<div class="categories-container ">⋯</div>
```

```css
element.style {
}
[data-namespace=viking-001] .profile-details-container          index.scss:16
.container .business-details-wrapper .business-details {
   flex: ▶ 2 1;
}
* {                                                              index.scss:5
   box-sizing: border-box;
   margin: ▶ 0;
   outline: ▶ none !important;
   padding: ▶ 0;
```

Converting to dataframe, renaming columns and exporting to csv

In [ ]:

```
df = pd.DataFrame.from_dict(business_list)

df.rename(columns={
    "business-name": "business name",
    "external-links-container": "external link",
    "phone-trigger": "phone number",
    "business-hours": "business hours"
}, inplace=True)

df.to_csv(f'./{state}_{city}.csv', index=False) # example al_birmingha
```

In [ ]:

```
driver.quit()
```

Possible Improvements and Changes

Depending on the company's needs, I can store these values elsewhere instead of a CSV. Possibly in MongoDB or an SQL database.

When scraping large amounts of data, the current script may run into memory issues. During full implementation, I'll refactor the script into something more modular or OOP. Selenium has something called Page Object Model (POM). I'm not too familiar with POM but I am more than willing to try!

Implementation (8/26/24)

Importing Libraries

In [ ]:

```
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.common.exceptions import TimeoutException
import pandas as pd
import time
import re
```

Key for converting state name to acronym

In [ ]:

```python
us_state_to_abbrev = {
    "Alabama": "AL",
    "Alaska": "AK",
    "Arizona": "AZ",
    "Arkansas": "AR",
    "California": "CA",
    "Colorado": "CO",
    "Connecticut": "CT",
    "Delaware": "DE",
    "Florida": "FL",
    "Georgia": "GA",
    "Hawaii": "HI",
    "Idaho": "ID",
    "Illinois": "IL",
    "Indiana": "IN",
    "Iowa": "IA",
    "Kansas": "KS",
    "Kentucky": "KY",
    "Louisiana": "LA",
    "Maine": "ME",
    "Maryland": "MD",
    "Massachusetts": "MA",
    "Michigan": "MI",
    "Minnesota": "MN",
    "Mississippi": "MS",
    "Missouri": "MO",
    "Montana": "MT",
    "Nebraska": "NE",
    "Nevada": "NV",
    "New Hampshire": "NH",
    "New Jersey": "NJ",
    "New Mexico": "NM",
    "New York": "NY",
    "North Carolina": "NC",
    "North Dakota": "ND",
    "Ohio": "OH",
    "Oklahoma": "OK",
    "Oregon": "OR",
    "Pennsylvania": "PA",
    "Rhode Island": "RI",
    "South Carolina": "SC",
    "South Dakota": "SD",
    "Tennessee": "TN",
    "Texas": "TX",
    "Utah": "UT",
    "Vermont": "VT",
    "Virginia": "VA",
    "Washington": "WA",
    "West Virginia": "WV",
    "Wisconsin": "WI",
    "Wyoming": "WY",
    "District of Columbia": "DC",
    "American Samoa": "AS",
    "Guam": "GU",
    "Northern Mariana Islands": "MP",
    "Puerto Rico": "PR",
    "United States Minor Outlying Islands": "UM",
    "U.S. Virgin Islands": "VI",
}

```

Importing .xlxs file and filling in merged cells with previous value

In [ ]:

```
df = pd.read_excel("google_maps_keywords.xlsx")
df.loc[:, ["Country", "State"]] = df.loc[:, ["Country", "State"]].ffil
```

Opening CitySearch

In [ ]:

```
driver = webdriver.Chrome()
driver.get("https://www.citysearch.com/")
```

Finding container for links of all cities and saving it to a variable

In [ ]:

```
container = driver.find_element(By.CSS_SELECTOR, "div.cities-container
cities = container.find_elements(By.CSS_SELECTOR, "li:not([class*='sta
city_links = [city.get_attribute("href") for city in cities]
state, city = re.search("(?<=\.com\/).*", city_links[0]).group().split

```

Grouping dataframe by country and state

During implementation, I would look over all the countries and state. This would be outer most loop. Until implementation, I have hard coded the script to only loop over the cities in the US.

In [ ]:

```
grouped_df = df.groupby("Country")
grouped_countries = grouped_df.get_group("United States") #todo during
grouped_states = grouped_countries.groupby("State")

states = grouped_states.groups.keys()
```

Looping for the states, cities, and industries in that order.

Some cities didn't have any job postings for some industries, so I've added a TryExcept block to cover for those situations. Currently, it's only looping for the first 5 job postings, but during implementation, the loop would continue until it's extracted everything.

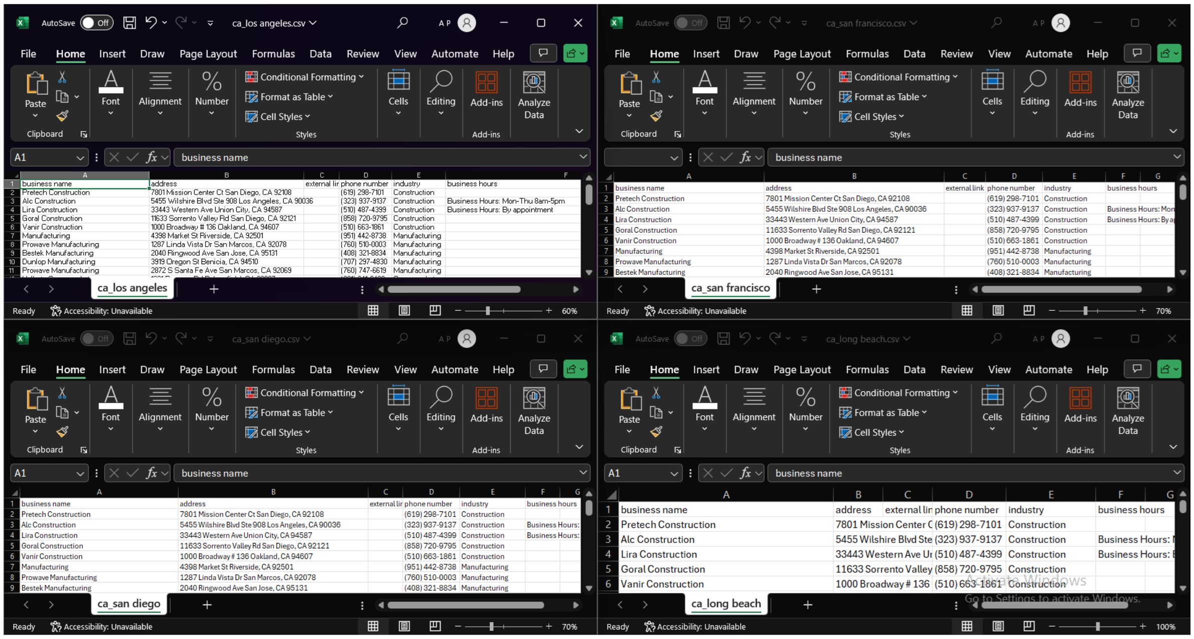
In [ ]:

```
for state in states:
    state_cities = grouped_states.get_group(state)["City"]
    state_industries = grouped_states.get_group(state)["Industry"]
    state_abbrev = us_state_to_abbrev[state]

    for city in state_cities:
        business_list = []

        if (pd.isnull(city)): continue

        for industry in state_industries:
            if (pd.isnull(industry)): continue

            url = f"https://www.citysearch.com/results?term={industry.
            print("-------------------------", url, "-----------------
            driver.get(url)

            # extracting all links to jobs in this category
            try:
                elem = WebDriverWait(driver, 10).until(
                    EC.presence_of_element_located((By.CSS_SELECTOR, "
            except TimeoutException:
                print("Timed out waiting for page to load: Most likely

            job_cards = driver.find_elements(By.CSS_SELECTOR, "div.lis
            job_cards_links = [job.get_attribute("href") for job in jo

            # visiting each job link for the current industry and scra

            if (len(job_cards_links) == 0): continue

            # for i in range(len(job_cards_links)): # todo uncomment t
            for i in range(5):
                print(job_cards_links[i])
                driver.get(job_cards_links[i])

                try:
                    elem = WebDriverWait(driver, 10).until(
                        EC.presence_of_element_located((By.CSS_SELECTO
                except TimeoutException:
                    print("Timed out waiting for page to load")

                # not sure if all business have all their contact info
                business_details = driver.find_elements(By.CSS_SELECTO

                business_details_dict = {
                    entry.get_attribute("class"): entry.text
                    for entry in business_details
                }
                business_details_dict["industry"] = industry

                business_list.append(business_details_dict)
                time.sleep(3)

        df = pd.DataFrame.from_dict(business_list)
        df.rename(columns={
            "business-name": "business name",
            "external-links-container": "external link",
            "phone-trigger": "phone number",
            "business-hours": "business hours"
        }, inplace=True)

        df.to_csv(f'./{state_abbrev.lower()}_{city.lower()}.csv', inde
```

In [ ]:

```
driver.quit()
```

Results


AutoSave Off
ca_los angeles.csv
File Home Insert Draw Page Layout Formulas Data Review View Automate Help
Paste Clipboard Font Alignment Number Conditional Formatting Format as Table Cell Styles Styles Cells Editing Add-ins Analyze Data
A1
Ready
Accessibility: Unavailable
60%
ca_san francisco.csv
70%
ca_san diego.csv
ca_long beach.csv
100%
Activate Windows
Go to Settings to activate Windows.

ca_los angeles

	A	B	C	D	E	F
1	business name	address	external lir	phone number	industry	business hours
2	Pretech Construction	7801 Mission Center Ct San Diego, CA 92108		(619) 298-7101	Construction	
3	Alc Construction	5455 Wilshire Blvd Ste 908 Los Angeles, CA 90036		(323) 937-9137	Construction	Business Hours: Mon-Thu 8am-5pm
4	Lira Construction	33443 Western Ave Union City, CA 94587		(510) 487-4399	Construction	Business Hours: By appointment
5	Goral Construction	11633 Sorrento Valley Rd San Diego, CA 92121		(858) 720-9795	Construction	
6	Vanir Construction	1000 Broadway # 136 Oakland, CA 94607		(510) 663-1861	Construction	
7	Manufacturing	4398 Market St Riverside, CA 92501		(951) 442-8738	Manufacturing	
8	Prowave Manufacturing	1287 Linda Vista Dr San Marcos, CA 92078		(760) 510-0003	Manufacturing	
9	Bestek Manufacturing	2040 Ringwood Ave San Jose, CA 95131		(408) 321-8834	Manufacturing	
10	Dunlop Manufacturing	3919 Oregon St Benicia, CA 94510		(707) 297-4830	Manufacturing	
11	Prowave Manufacturing	2872 S Santa Fe Ave San Marcos, CA 92069		(760) 747-6619	Manufacturing	

ca_san francisco

	A	B	C	D	E	F
1	business name	address	external link	phone number	industry	business hours
2	Pretech Construction	7801 Mission Center Ct San Diego, CA 92108		(619) 298-7101	Construction	
3	Alc Construction	5455 Wilshire Blvd Ste 908 Los Angeles, CA 90036		(323) 937-9137	Construction	Business Hours: Mon
4	Lira Construction	33443 Western Ave Union City, CA 94587		(510) 487-4399	Construction	Business Hours: By a
5	Goral Construction	11633 Sorrento Valley Rd San Diego, CA 92121		(858) 720-9795	Construction	
6	Vanir Construction	1000 Broadway # 136 Oakland, CA 94607		(510) 663-1861	Construction	
7	Manufacturing	4398 Market St Riverside, CA 92501		(951) 442-8738	Manufacturing	
8	Prowave Manufacturing	1287 Linda Vista Dr San Marcos, CA 92078		(760) 510-0003	Manufacturing	
9	Bestek Manufacturing	2040 Ringwood Ave San Jose, CA 95131		(408) 321-8834	Manufacturing	

ca_san diego

	A	B	C	D	E	F
1	business name	address	external lir	phone number	industry	business hours
2	Pretech Construction	7801 Mission Center Ct San Diego, CA 92108		(619) 298-7101	Construction	
3	Alc Construction	5455 Wilshire Blvd Ste 908 Los Angeles, CA 90036		(323) 937-9137	Construction	Business Hours:
4	Lira Construction	33443 Western Ave Union City, CA 94587		(510) 487-4399	Construction	Business Hours:
5	Goral Construction	11633 Sorrento Valley Rd San Diego, CA 92121		(858) 720-9795	Construction	
6	Vanir Construction	1000 Broadway # 136 Oakland, CA 94607		(510) 663-1861	Construction	
7	Manufacturing	4398 Market St Riverside, CA 92501		(951) 442-8738	Manufacturing	
8	Prowave Manufacturing	1287 Linda Vista Dr San Marcos, CA 92078		(760) 510-0003	Manufacturing	
9	Bestek Manufacturing	2040 Ringwood Ave San Jose, CA 95131		(408) 321-8834	Manufacturing	

ca_long beach

	A	B	C	D	E	F
1	business name	address	external lir	phone number	industry	business hours
2	Pretech Construction	7801 Mission Center C		(619) 298-7101	Construction	
3	Alc Construction	5455 Wilshire Blvd Ste		(323) 937-9137	Construction	Business Hours: I
4	Lira Construction	33443 Western Ave Ur		(510) 487-4399	Construction	Business Hours: E
5	Goral Construction	11633 Sorrento Valley		(858) 720-9795	Construction	
6	Vanir Construction	1000 Broadway # 136		(510) 663-1861	Construction	

I think it's ready for full implementation!