LINKEDIN JOB ANALYSIS

PROJECT BY:

JATIN KUMAR SINGH

TASK:

- Scrape data from professional networking platform Linkedin using python library called Beautifulsoup (or similar) and collate information in the given format and make tables using the data

STEP 1 – WEB SCRAPPING USING PYTHON

- **LIBRARIES USED:**

❑ 1. PANDAS

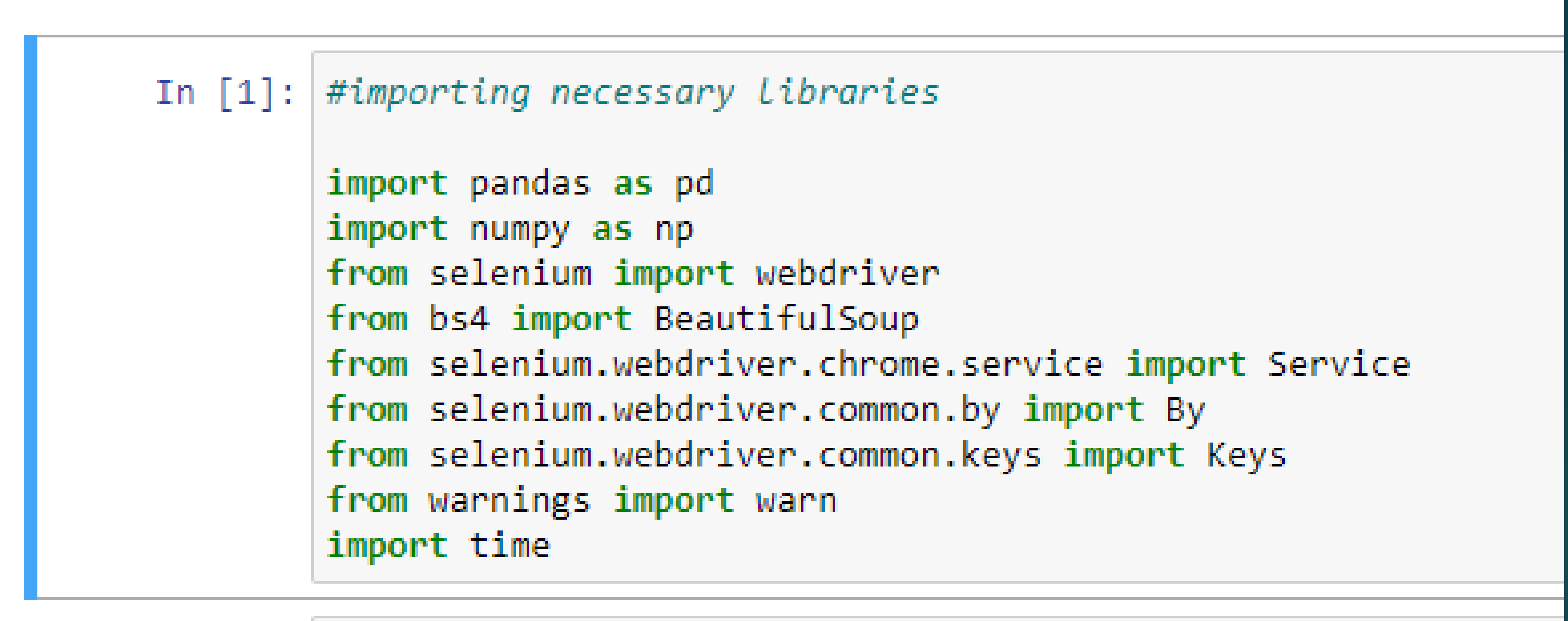
❑ 2. BEAUTIFULSOUP

❑ 3. SELENIUM

- **REFERENCES USED:**

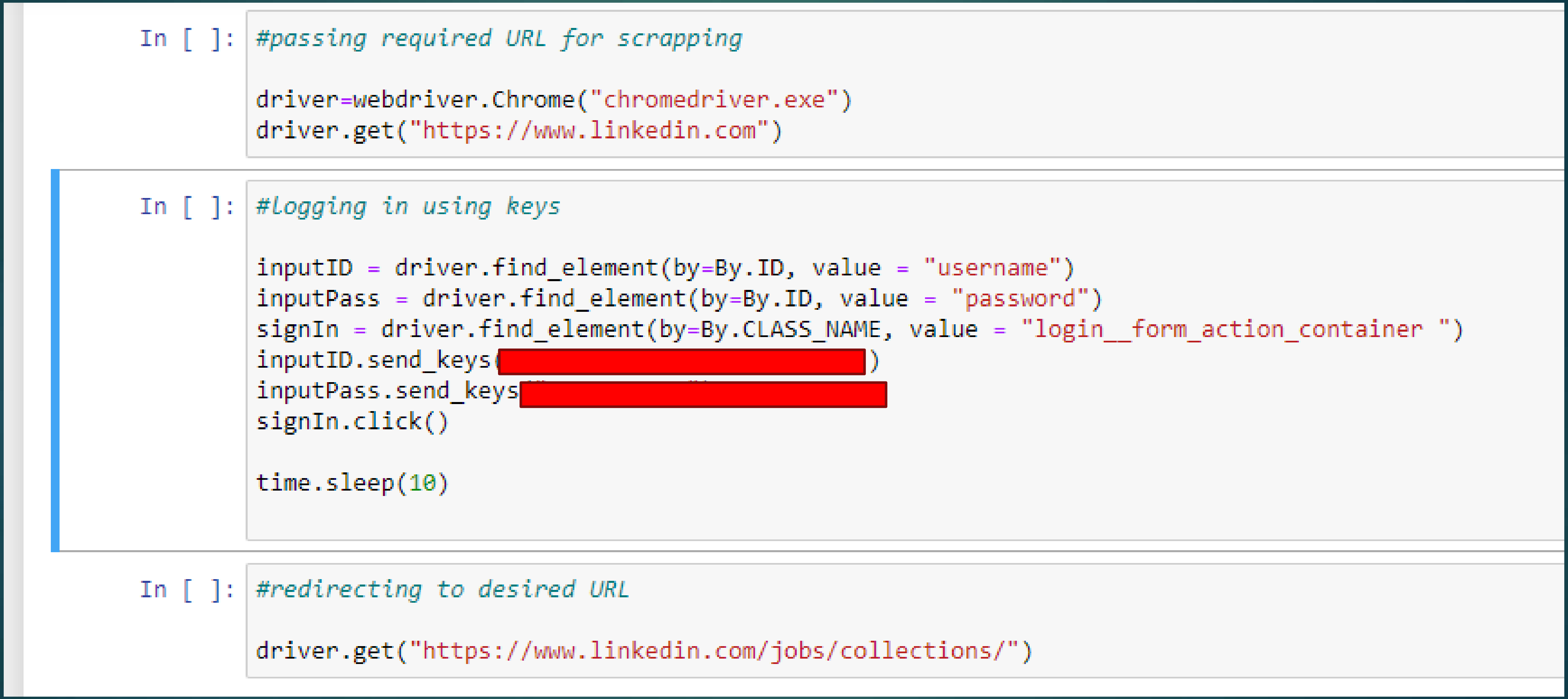
❑ 1. https://pypi.org/project/beautifulsoup4/

❑ 2. https://beautiful-soup-4.readthedocs.io/en/latest/



In [1]:

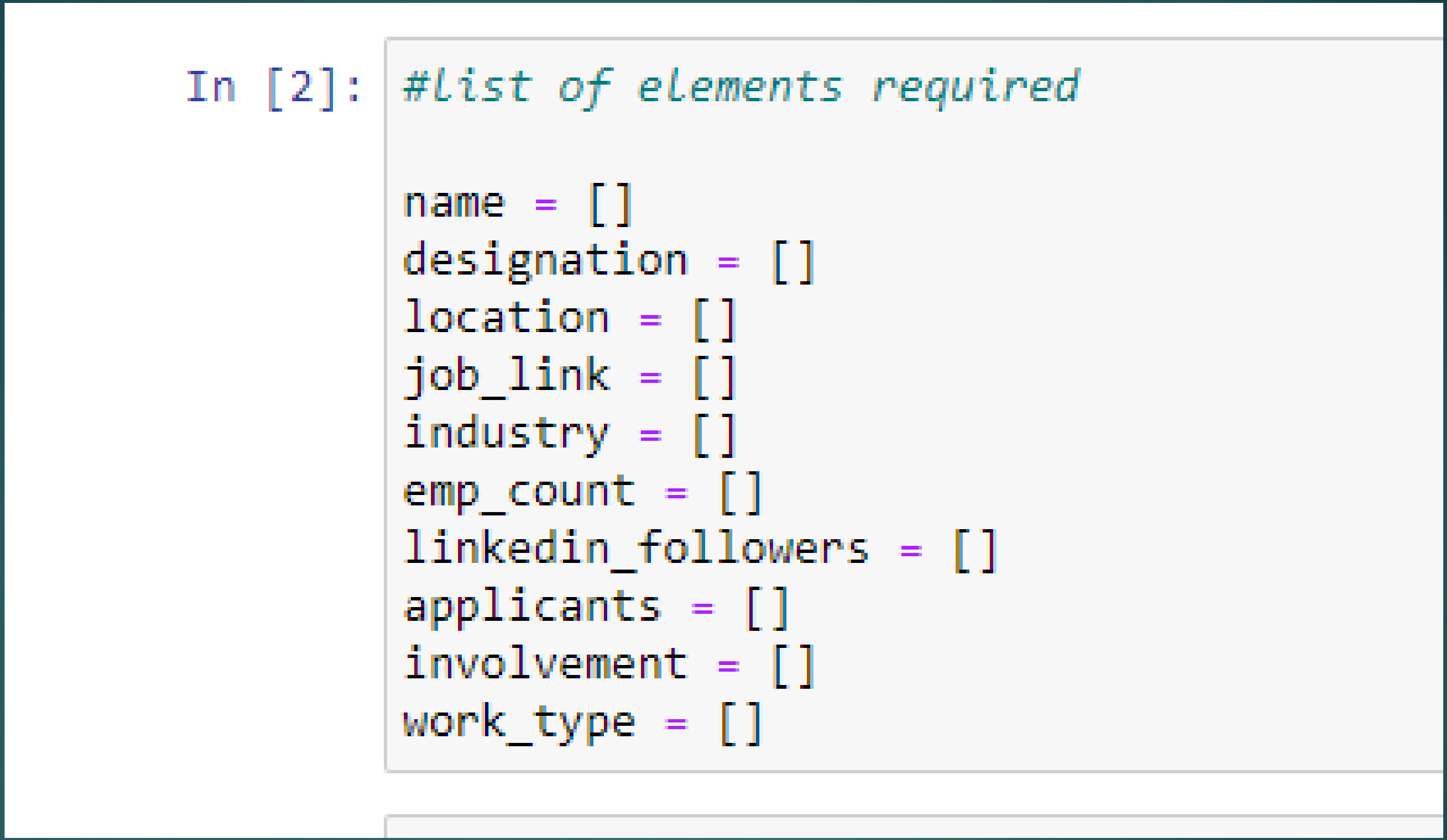
```
#importing necessary libraries

import pandas as pd
import numpy as np
from selenium import webdriver
from bs4 import BeautifulSoup
from selenium.webdriver.chrome.service import Service
from selenium.webdriver.common.by import By
from selenium.webdriver.common.keys import Keys
from warnings import warn
import time
```



```
In [ ]: #passing required URL for scrapping

        driver=webdriver.Chrome("chromedriver.exe")
        driver.get("https://www.linkedin.com")
```

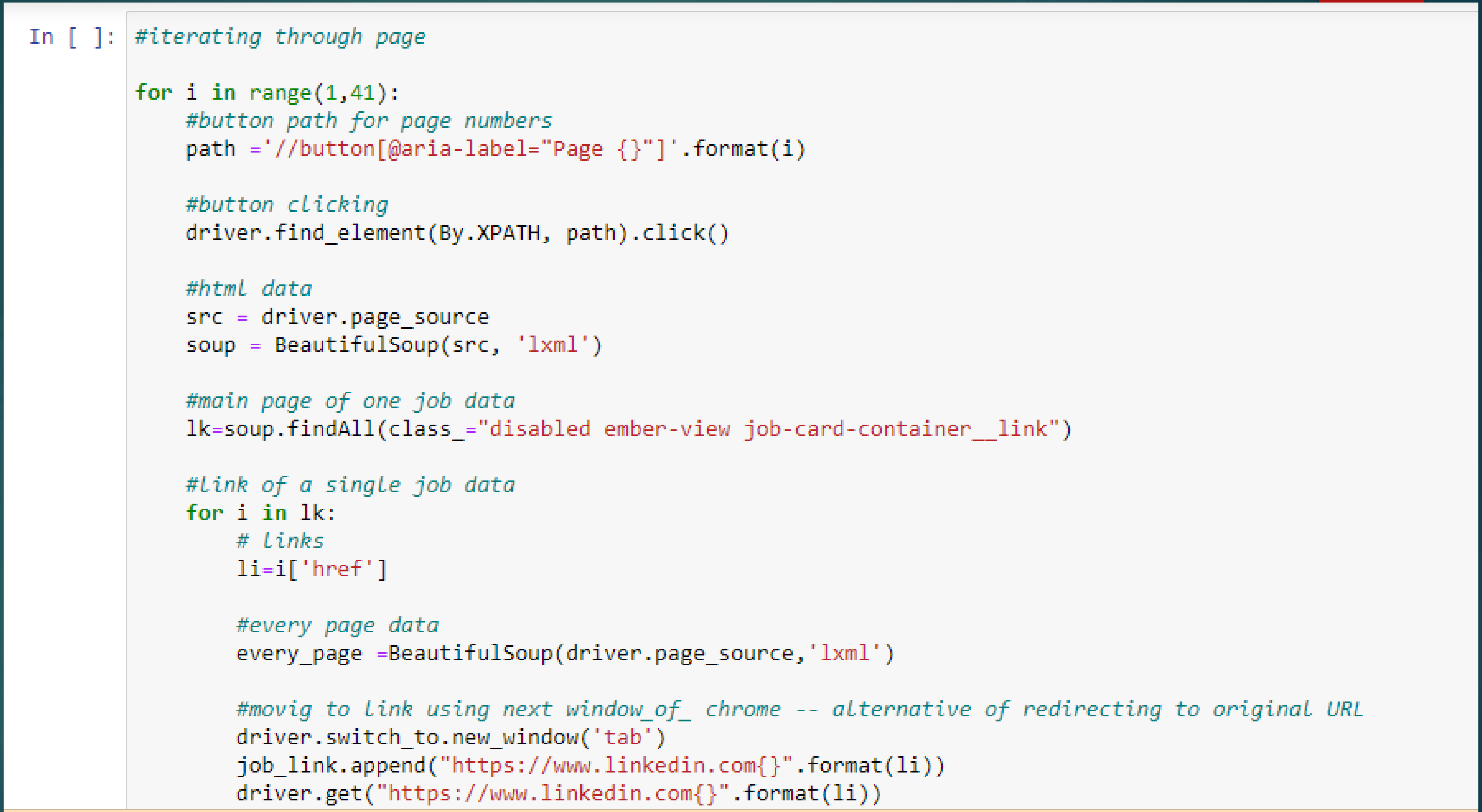
```
In [ ]: #logging in using keys

        inputID = driver.find_element(by=By.ID, value = "username")
        inputPass = driver.find_element(by=By.ID, value = "password")
        signIn = driver.find_element(by=By.CLASS_NAME, value = "login__form_action_container ")
        inputID.send_keys([redacted])
        inputPass.send_keys[redacted]
        signIn.click()

        time.sleep(10)
```

```
In [ ]: #redirecting to desired URL

        driver.get("https://www.linkedin.com/jobs/collections/")
```

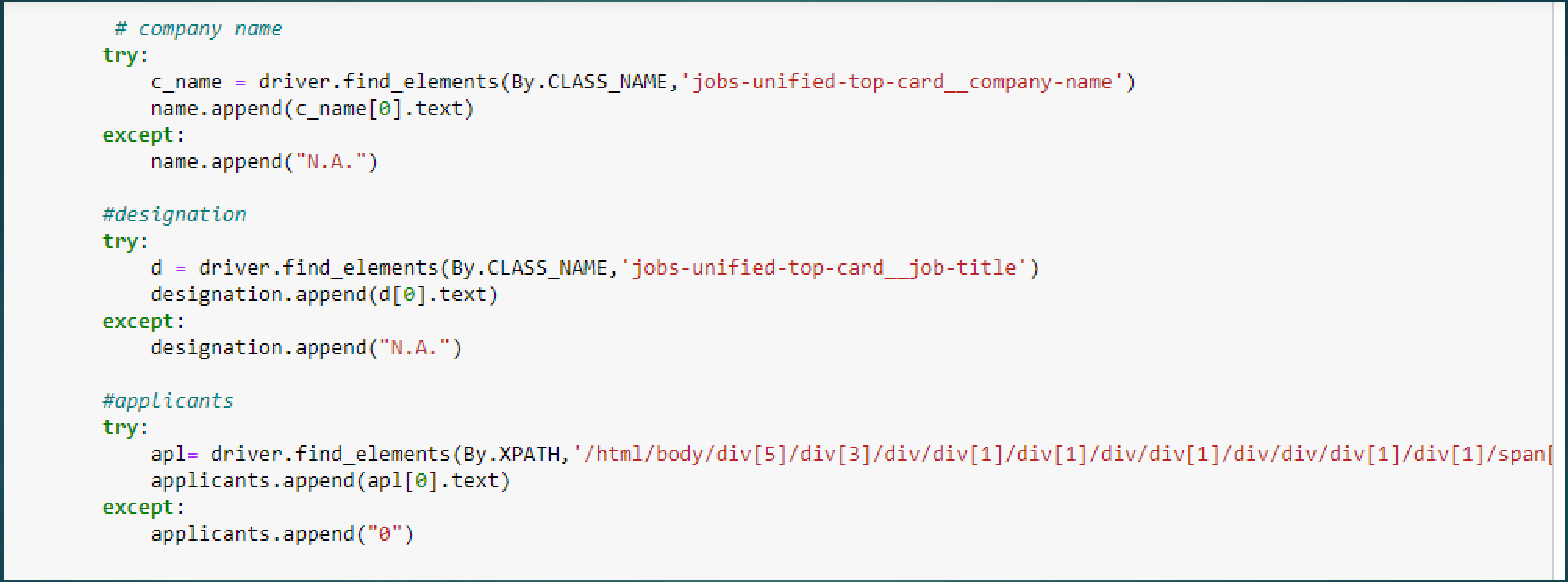


In [2]:

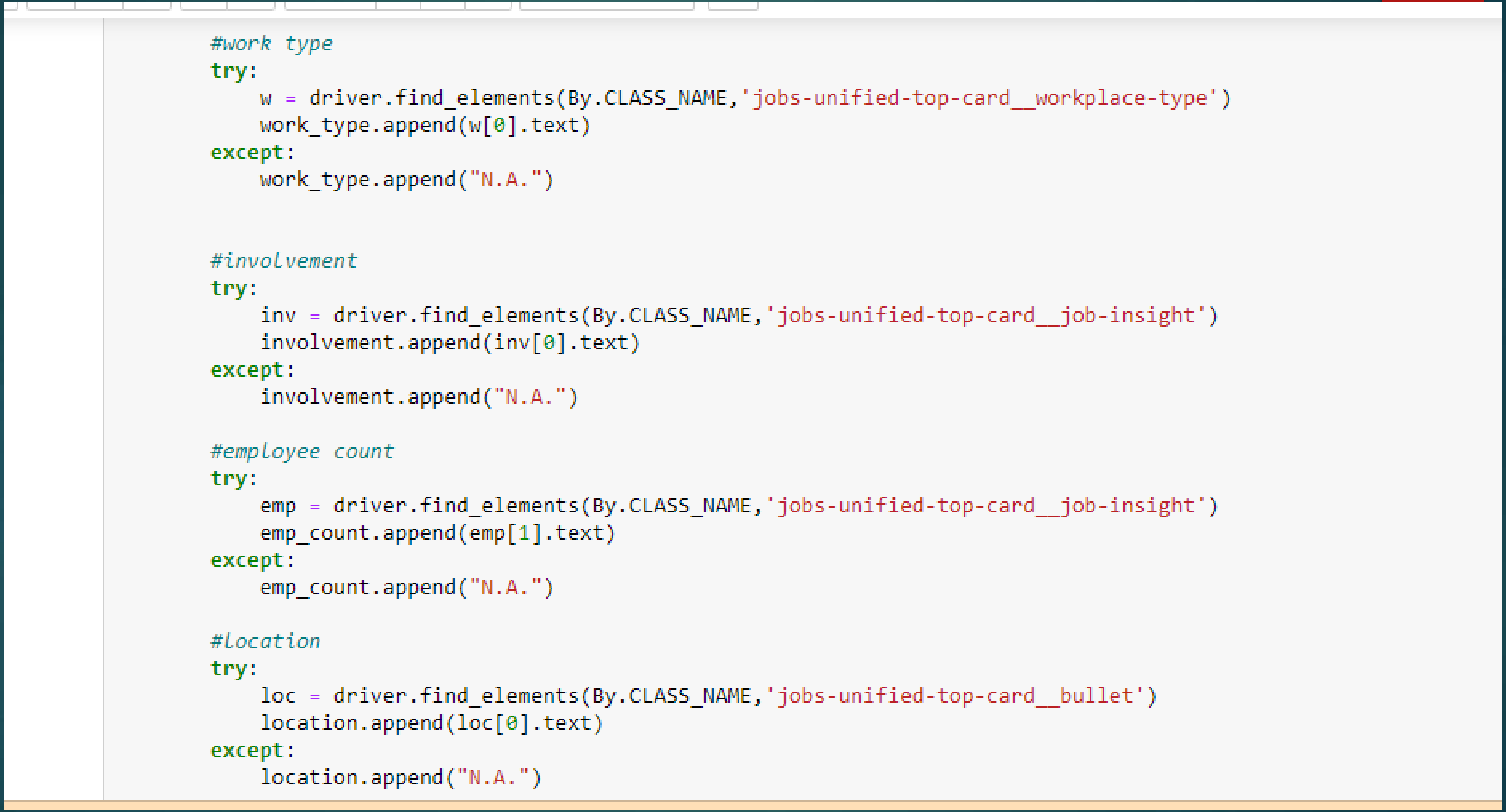
```
#list of elements required

name = []
designation = []
location = []
job_link = []
industry = []
emp_count = []
linkedin_followers = []
applicants = []
involvement = []
work_type = []
```



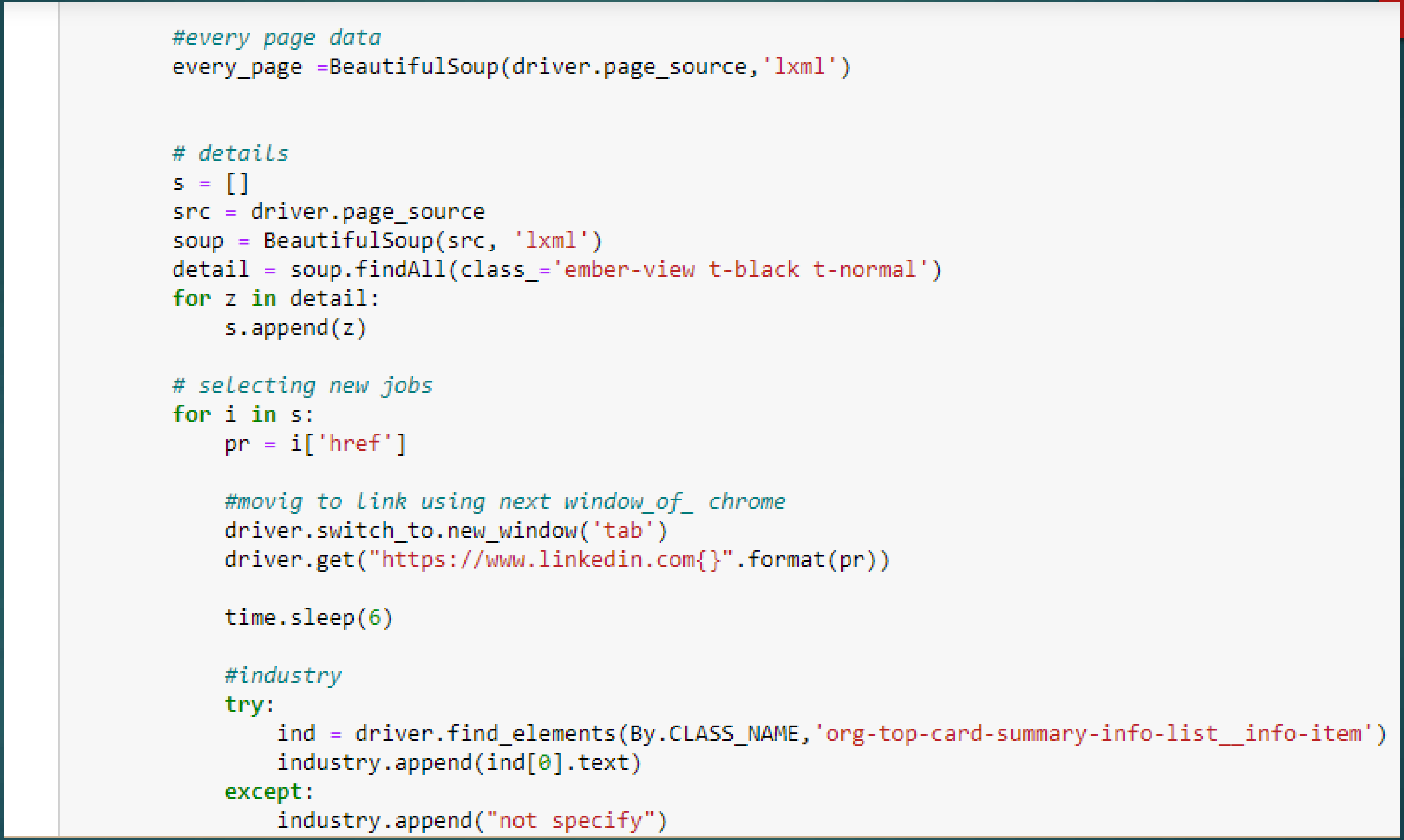
```python
#iterating through page

for i in range(1,41):
    #button path for page numbers
    path ='//button[@aria-label="Page {}"]'.format(i)

    #button clicking
    driver.find_element(By.XPATH, path).click()

    #html data
    src = driver.page_source
    soup = BeautifulSoup(src, 'lxml')

    #main page of one job data
    lk=soup.findAll(class_="disabled ember-view job-card-container__link")

    #link of a single job data
    for i in lk:
        # links
        li=i['href']

        #every page data
        every_page =BeautifulSoup(driver.page_source,'lxml')

        #movig to link using next window_of_ chrome -- alternative of redirecting to original URL
        driver.switch_to.new_window('tab')
        job_link.append("https://www.linkedin.com{}".format(li))
        driver.get("https://www.linkedin.com{}".format(li))
```



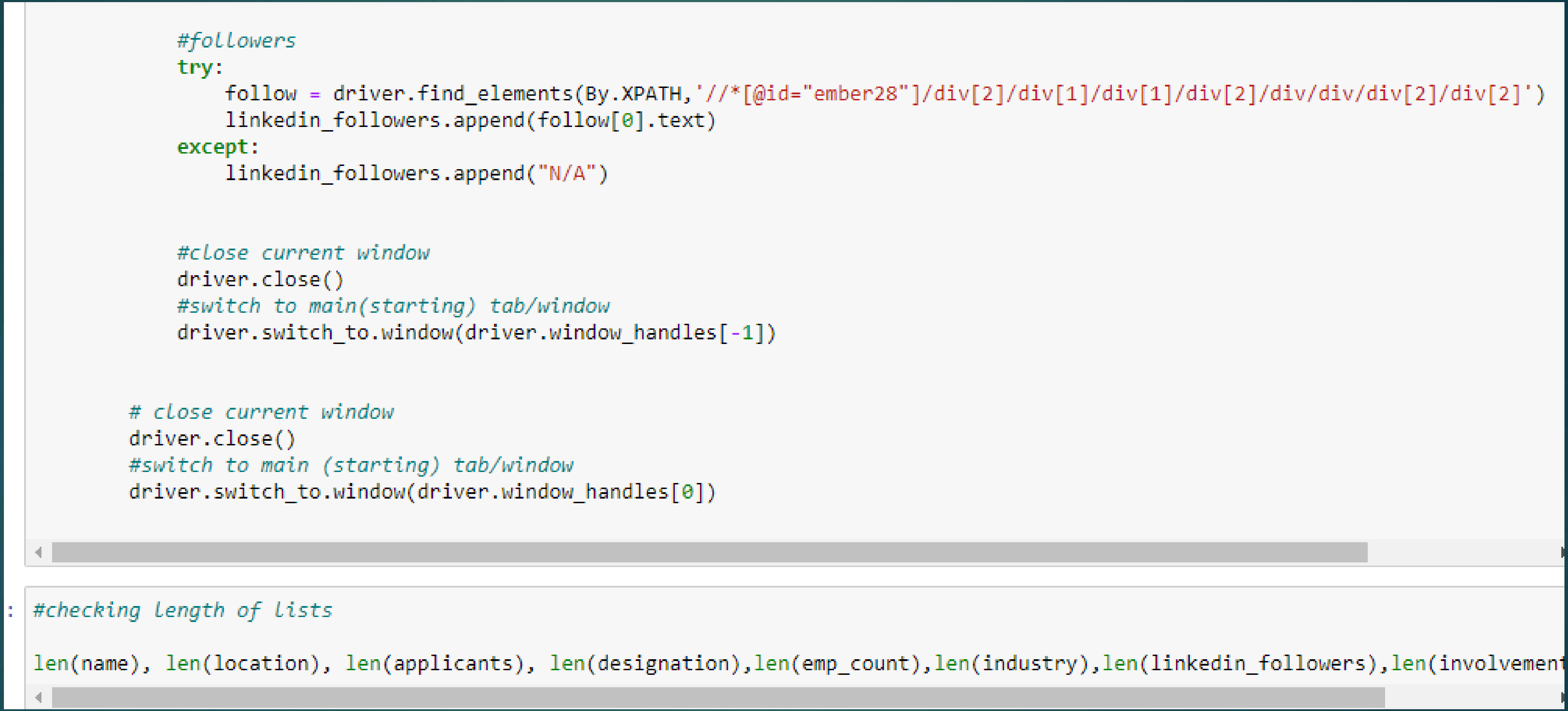
```
 # company name
try:
    c_name = driver.find_elements(By.CLASS_NAME,'jobs-unified-top-card__company-name')
    name.append(c_name[0].text)
except:
    name.append("N.A.")

#designation
try:
    d = driver.find_elements(By.CLASS_NAME,'jobs-unified-top-card__job-title')
    designation.append(d[0].text)
except:
    designation.append("N.A.")

#applicants
try:
    apl= driver.find_elements(By.XPATH,'/html/body/div[5]/div[3]/div/div[1]/div[1]/div/div[1]/div/div/div[1]/div[1]/span[
    applicants.append(apl[0].text)
except:
    applicants.append("0")
```



```python
        #work type
        try:
            w = driver.find_elements(By.CLASS_NAME,'jobs-unified-top-card__workplace-type')
            work_type.append(w[0].text)
        except:
            work_type.append("N.A.")


        #involvement
        try:
            inv = driver.find_elements(By.CLASS_NAME,'jobs-unified-top-card__job-insight')
            involvement.append(inv[0].text)
        except:
            involvement.append("N.A.")

        #employee count
        try:
            emp = driver.find_elements(By.CLASS_NAME,'jobs-unified-top-card__job-insight')
            emp_count.append(emp[1].text)
        except:
            emp_count.append("N.A.")

        #location
        try:
            loc = driver.find_elements(By.CLASS_NAME,'jobs-unified-top-card__bullet')
            location.append(loc[0].text)
        except:
            location.append("N.A.")
```



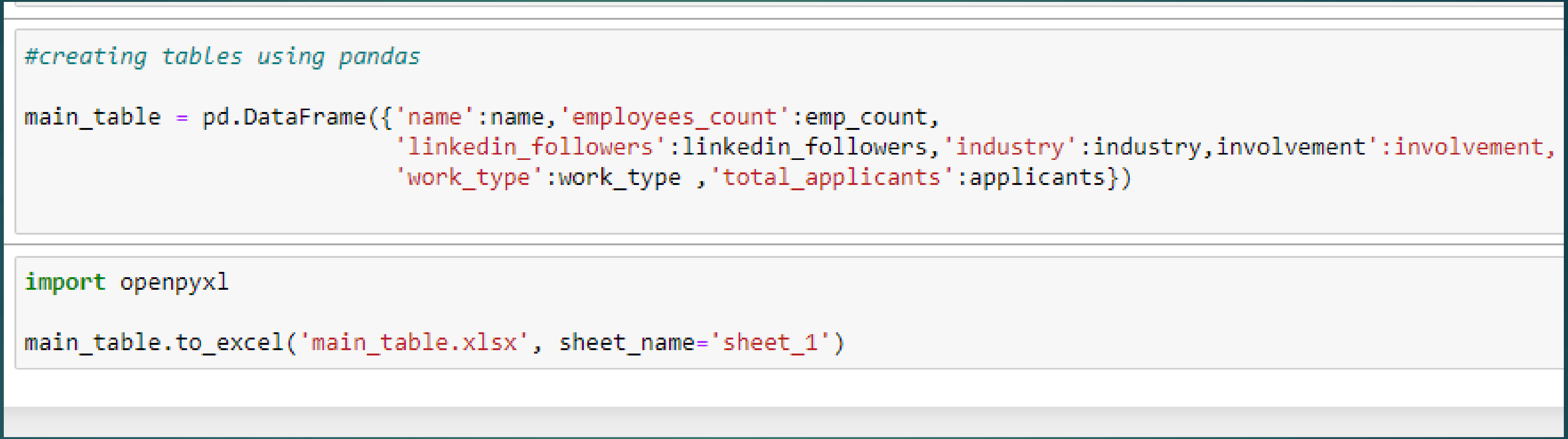
```
#every page data
every_page =BeautifulSoup(driver.page_source,'lxml')


# details
s = []
src = driver.page_source
soup = BeautifulSoup(src, 'lxml')
detail = soup.findAll(class_='ember-view t-black t-normal')
for z in detail:
    s.append(z)

# selecting new jobs
for i in s:
    pr = i['href']

    #movig to link using next window_of_ chrome
    driver.switch_to.new_window('tab')
    driver.get("https://www.linkedin.com{}".format(pr))

    time.sleep(6)

    #industry
    try:
        ind = driver.find_elements(By.CLASS_NAME,'org-top-card-summary-info-list__info-item')
        industry.append(ind[0].text)
    except:
        industry.append("not specify")
```



```
            #followers
            try:
                follow = driver.find_elements(By.XPATH,'//*[@id="ember28"]/div[2]/div[1]/div[1]/div[2]/div/div/div[2]/div[2]')
                linkedin_followers.append(follow[0].text)
            except:
                linkedin_followers.append("N/A")


            #close current window
            driver.close()
            #switch to main(starting) tab/window
            driver.switch_to.window(driver.window_handles[-1])


        # close current window
        driver.close()
        #switch to main (starting) tab/window
        driver.switch_to.window(driver.window_handles[0])
```

```
#checking length of lists

len(name), len(location), len(applicants), len(designation),len(emp_count),len(industry),len(linkedin_followers),len(involvement
```

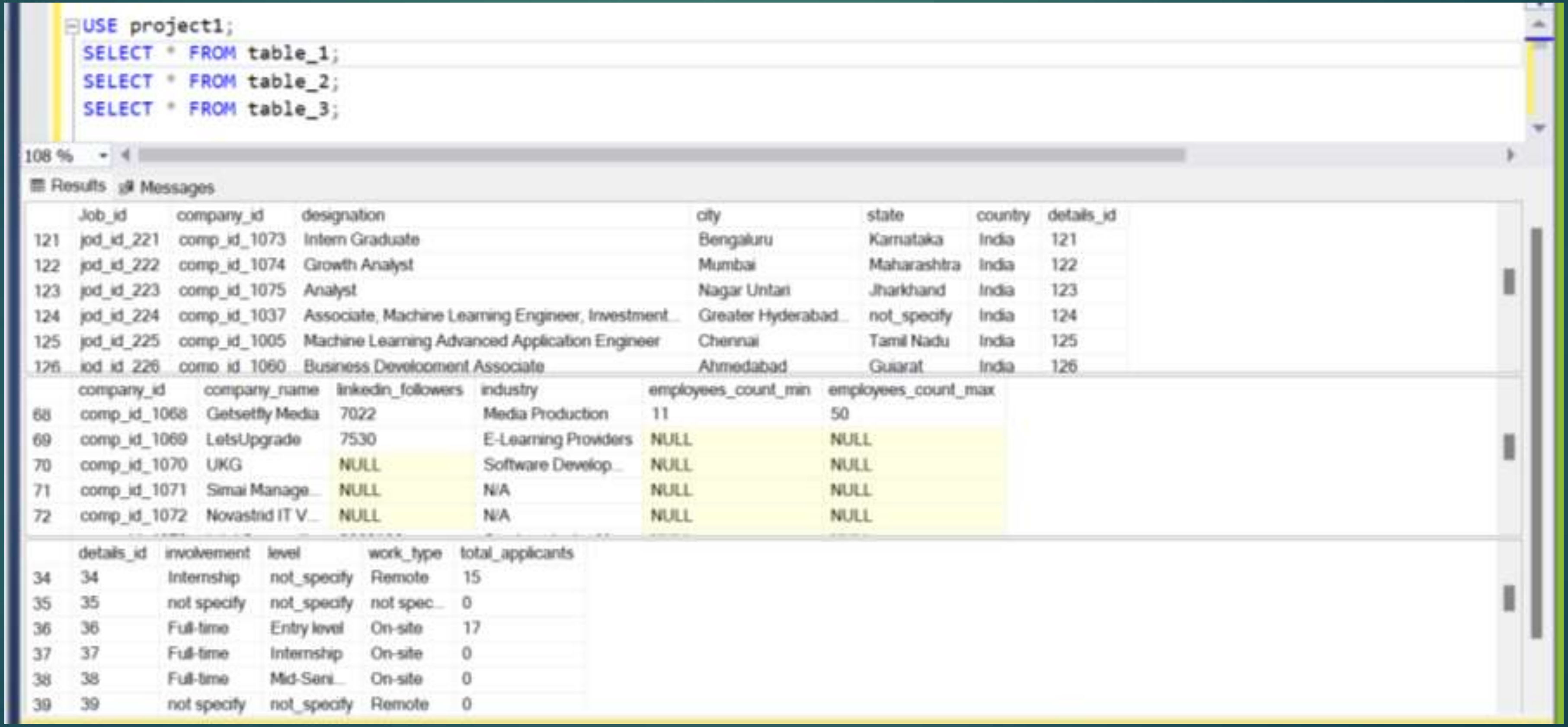
FINALLY MAKING A TABLE FROM LISTS USING PANDAS



```
#creating tables using pandas

main_table = pd.DataFrame({'name':name,'employees_count':emp_count,
                           'linkedin_followers':linkedin_followers,'industry':industry,involvement':involvement,
                           'work_type':work_type ,'total_applicants':applicants})
```

```
import openpyxl

main_table.to_excel('main_table.xlsx', sheet_name='sheet_1')
```

STEP 2 – USING MS SQL FOR TABLE CREATION

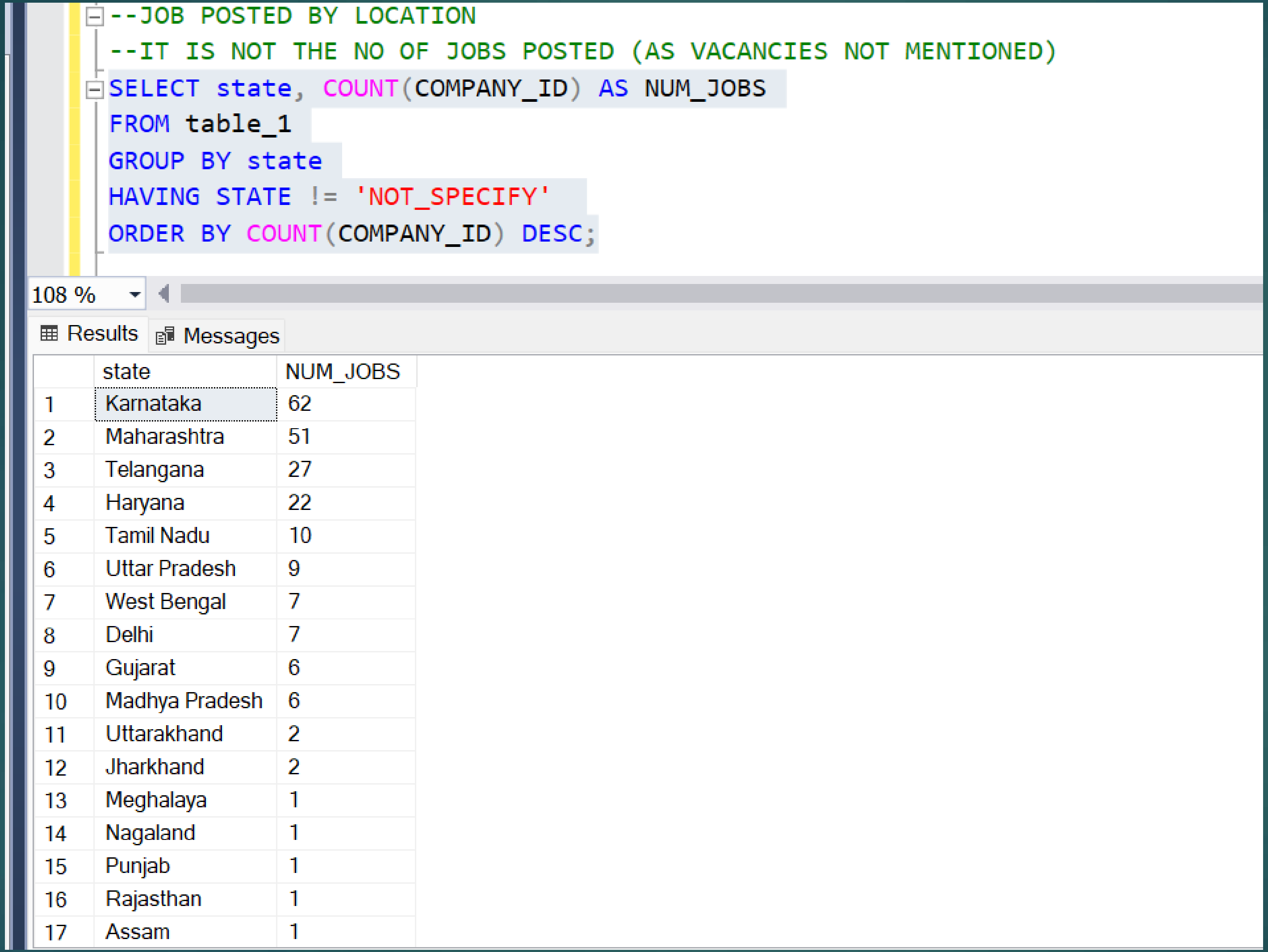


```sql
USE project1;
SELECT * FROM table_1;
SELECT * FROM table_2;
SELECT * FROM table_3;
```

	Job_id	company_id	designation	city	state	country	details_id
121	jod_id_221	comp_id_1073	Intern Graduate	Bengaluru	Karnataka	India	121
122	jod_id_222	comp_id_1074	Growth Analyst	Mumbai	Maharashtra	India	122
123	jod_id_223	comp_id_1075	Analyst	Nagar Untari	Jharkhand	India	123
124	jod_id_224	comp_id_1037	Associate, Machine Learning Engineer, Investment...	Greater Hyderabad...	not_specify	India	124
125	jod_id_225	comp_id_1005	Machine Learning Advanced Application Engineer	Chennai	Tamil Nadu	India	125
126	jod_id_226	comp_id_1060	Business Development Associate	Ahmedabad	Gujarat	India	126

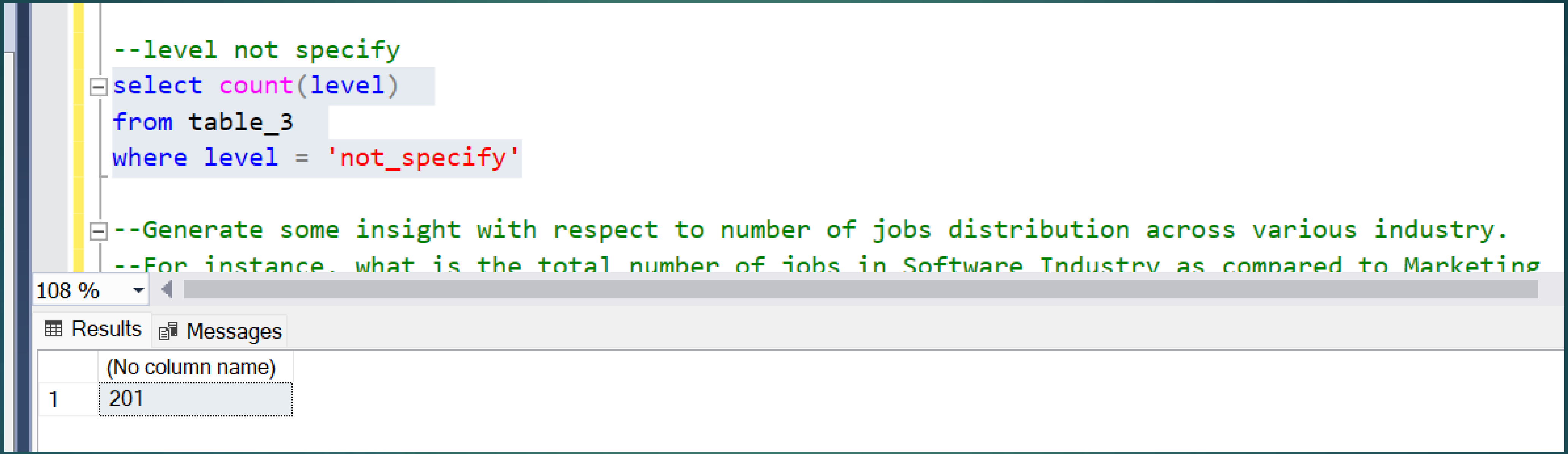
	company_id	company_name	linkedin_followers	industry	employees_count_min	employees_count_max
68	comp_id_1068	Getsetfly Media	7022	Media Production	11	50
69	comp_id_1069	LetsUpgrade	7530	E-Learning Providers	NULL	NULL
70	comp_id_1070	UKG	NULL	Software Develop...	NULL	NULL
71	comp_id_1071	Simai Manage...	NULL	N/A	NULL	NULL
72	comp_id_1072	Novastrid IT V...	NULL	N/A	NULL	NULL

	details_id	involvement	level	work_type	total_applicants
34	34	Internship	not_specify	Remote	15
35	35	not specify	not_specify	not spec...	0
36	36	Full-time	Entry level	On-site	17
37	37	Full-time	Internship	On-site	0
38	38	Full-time	Mid-Seni...	On-site	0
39	39	not specify	not_specify	Remote	0



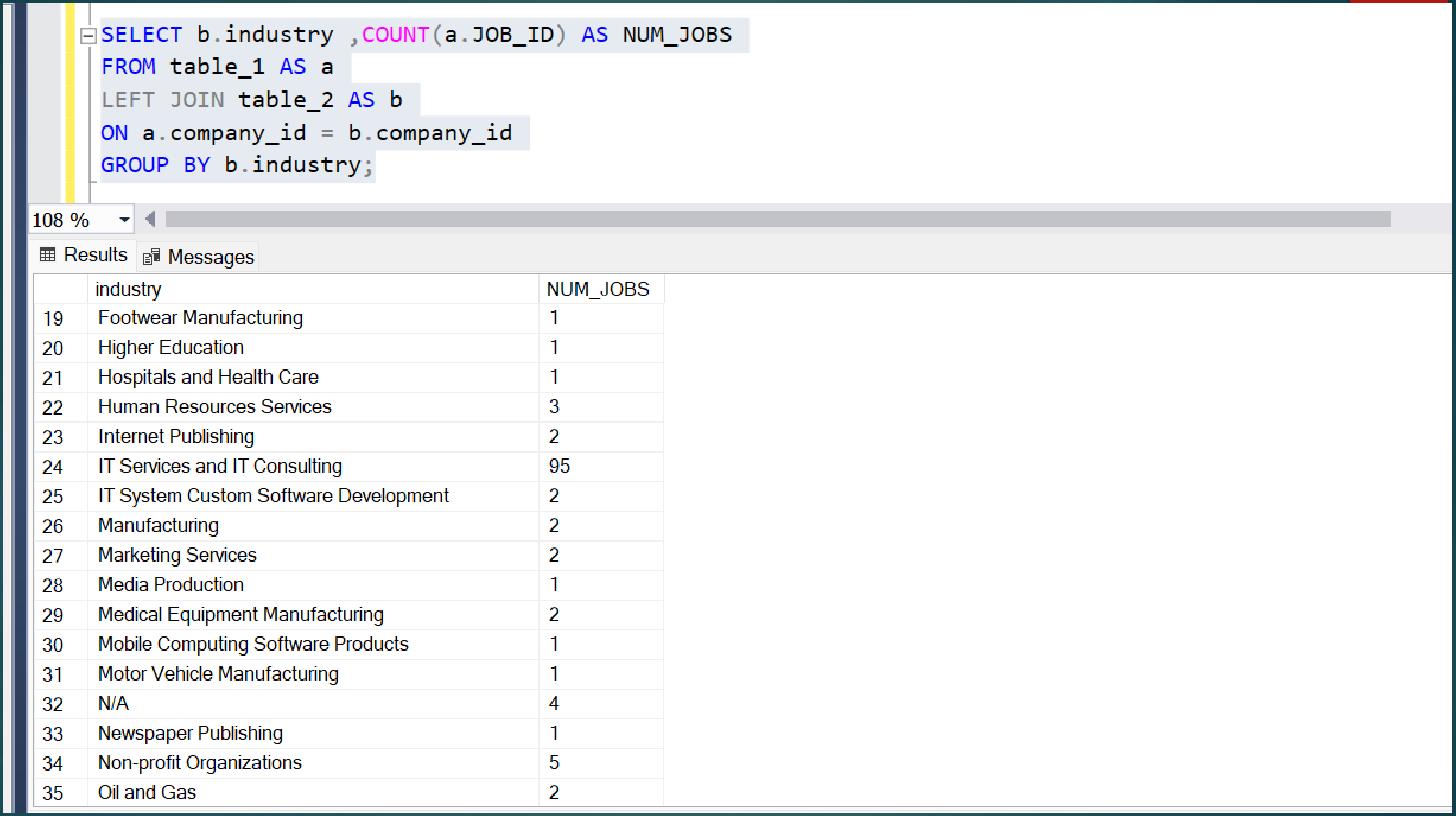
```sql
--JOB POSTED BY LOCATION
--IT IS NOT THE NO OF JOBS POSTED (AS VACANCIES NOT MENTIONED)
SELECT state, COUNT(COMPANY_ID) AS NUM_JOBS
FROM table_1
GROUP BY state
HAVING STATE != 'NOT_SPECIFY'
ORDER BY COUNT(COMPANY_ID) DESC;
```

	state	NUM_JOBS
1	Karnataka	62
2	Maharashtra	51
3	Telangana	27
4	Haryana	22
5	Tamil Nadu	10
6	Uttar Pradesh	9
7	West Bengal	7
8	Delhi	7
9	Gujarat	6
10	Madhya Pradesh	6
11	Uttarakhand	2
12	Jharkhand	2
13	Meghalaya	1
14	Nagaland	1
15	Punjab	1
16	Rajasthan	1
17	Assam	1



```
--level not specify
select count(level)
from table_3
where level = 'not_specify'

--Generate some insight with respect to number of jobs distribution across various industry.
--For instance, what is the total number of jobs in Software Industry as compared to Marketing
```

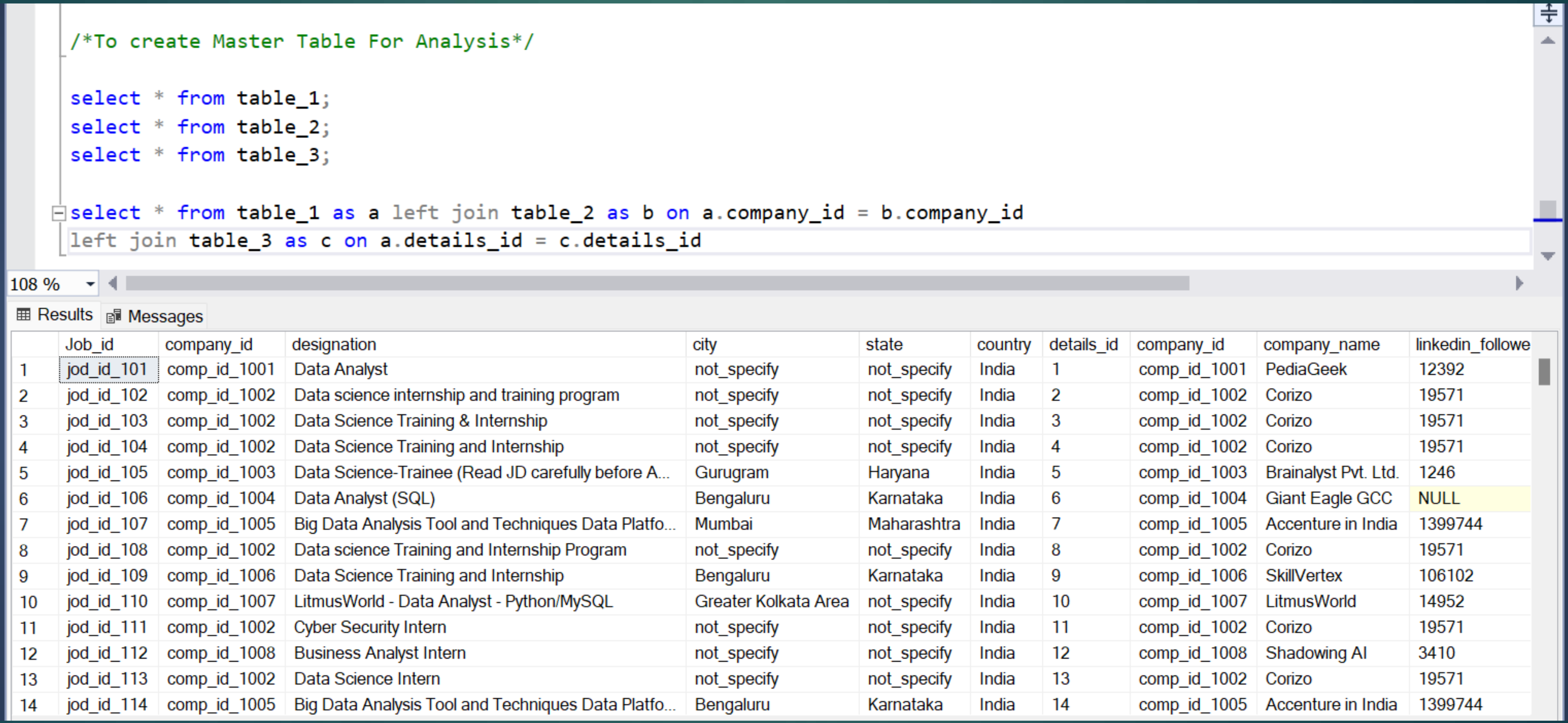
	(No column name)
1	201



```sql
SELECT b.industry ,COUNT(a.JOB_ID) AS NUM_JOBS
FROM table_1 AS a
LEFT JOIN table_2 AS b
ON a.company_id = b.company_id
GROUP BY b.industry;
```

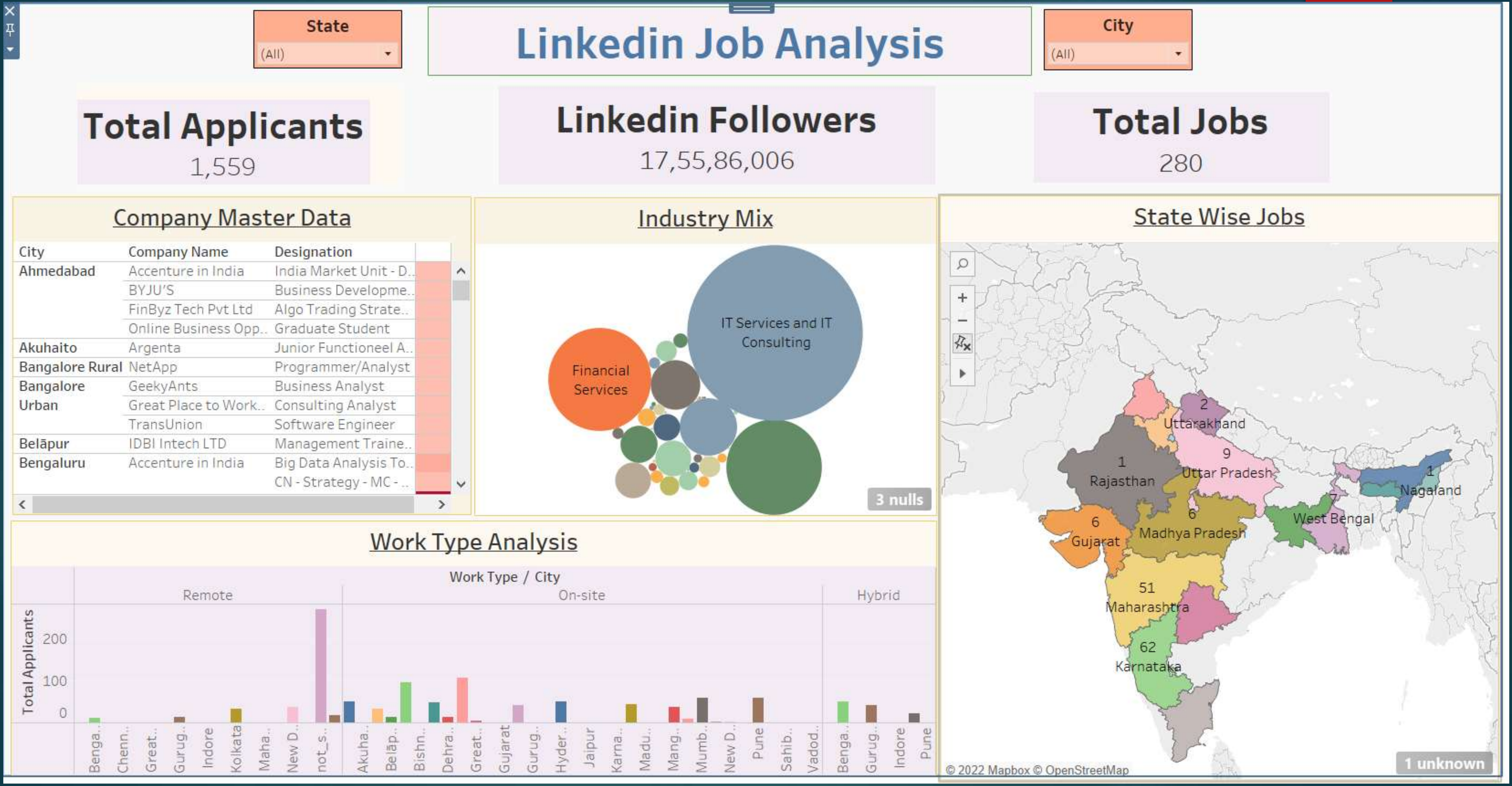
	industry	NUM_JOBS
19	Footwear Manufacturing	1
20	Higher Education	1
21	Hospitals and Health Care	1
22	Human Resources Services	3
23	Internet Publishing	2
24	IT Services and IT Consulting	95
25	IT System Custom Software Development	2
26	Manufacturing	2
27	Marketing Services	2
28	Media Production	1
29	Medical Equipment Manufacturing	2
30	Mobile Computing Software Products	1
31	Motor Vehicle Manufacturing	1
32	N/A	4
33	Newspaper Publishing	1
34	Non-profit Organizations	5
35	Oil and Gas	2

CREATING MASTER TABLE FOR TABLEAU



```
/*To create Master Table For Analysis*/

select * from table_1;
select * from table_2;
select * from table_3;

select * from table_1 as a left join table_2 as b on a.company_id = b.company_id
left join table_3 as c on a.details_id = c.details_id
```

	Job_id	company_id	designation	city	state	country	details_id	company_id	company_name	linkedin_followe
1	jod_id_101	comp_id_1001	Data Analyst	not_specify	not_specify	India	1	comp_id_1001	PediaGeek	12392
2	jod_id_102	comp_id_1002	Data science internship and training program	not_specify	not_specify	India	2	comp_id_1002	Corizo	19571
3	jod_id_103	comp_id_1002	Data Science Training & Internship	not_specify	not_specify	India	3	comp_id_1002	Corizo	19571
4	jod_id_104	comp_id_1002	Data Science Training and Internship	not_specify	not_specify	India	4	comp_id_1002	Corizo	19571
5	jod_id_105	comp_id_1003	Data Science-Trainee (Read JD carefully before A...	Gurugram	Haryana	India	5	comp_id_1003	Brainalyst Pvt. Ltd.	1246
6	jod_id_106	comp_id_1004	Data Analyst (SQL)	Bengaluru	Karnataka	India	6	comp_id_1004	Giant Eagle GCC	NULL
7	jod_id_107	comp_id_1005	Big Data Analysis Tool and Techniques Data Platfo...	Mumbai	Maharashtra	India	7	comp_id_1005	Accenture in India	1399744
8	jod_id_108	comp_id_1002	Data science Training and Internship Program	not_specify	not_specify	India	8	comp_id_1002	Corizo	19571
9	jod_id_109	comp_id_1006	Data Science Training and Internship	Bengaluru	Karnataka	India	9	comp_id_1006	SkillVertex	106102
10	jod_id_110	comp_id_1007	LitmusWorld - Data Analyst - Python/MySQL	Greater Kolkata Area	not_specify	India	10	comp_id_1007	LitmusWorld	14952
11	jod_id_111	comp_id_1002	Cyber Security Intern	not_specify	not_specify	India	11	comp_id_1002	Corizo	19571
12	jod_id_112	comp_id_1008	Business Analyst Intern	not_specify	not_specify	India	12	comp_id_1008	Shadowing AI	3410
13	jod_id_113	comp_id_1002	Data Science Intern	not_specify	not_specify	India	13	comp_id_1002	Corizo	19571
14	jod_id_114	comp_id_1005	Big Data Analysis Tool and Techniques Data Platfo...	Bengaluru	Karnataka	India	14	comp_id_1005	Accenture in India	1399744



State
(All)

Linkedin Job Analysis

City
(All)

Total Applicants
1,559

Linkedin Followers
17,55,86,006

Total Jobs
280

Company Master Data

City	Company Name	Designation
Ahmedabad	Accenture in India	India Market Unit - D..
	BYJU'S	Business Developme..
	FinByz Tech Pvt Ltd	Algo Trading Strate..
	Online Business Opp..	Graduate Student
Akuhaito	Argenta	Junior Functioneel A..
Bangalore Rural	NetApp	Programmer/Analyst
Bangalore Urban	GeekyAnts	Business Analyst
	Great Place to Work..	Consulting Analyst
	TransUnion	Software Engineer
Belāpur	IDBI Intech LTD	Management Traine..
Bengaluru	Accenture in India	Big Data Analysis To..
		CN - Strategy - MC - ..

Industry Mix

State Wise Jobs

Work Type Analysis