Project Report

on

Web Scraping and Data Visualization

(linked in)

Submitted

to

Gangaaram Technologies

Submitted

by

K Sai Sarath Reddy

Gangaaram Technologies Private Limited, First Floor,
6-1-184 Varadaraja Nagar, Tirupati,
Andhra Pradesh-517501.

TABLE OF CONTENT

1.ABSTRACT

This paper outlines a comprehensive workflow for web data scraping, processing, and visualization using various tools and technologies. We start with data scraping, employing Python's `Requests` library for retrieving web content and `Beautiful Soup` for parsing HTML to extract relevant data. `Selenium` is also discussed for handling dynamic web content.

After extraction, data is organized and cleaned using the `Pandas` library to create data frames. This includes managing missing data and transforming data types for preliminary analysis. The cleaned data is then exported to Excel sheets, ready for integration with analytical tools.

We detail the process of importing Excel data into Power BI, covering steps to connect, load, and prepare the data for visualization. Power BI's tools for data analysis and visualization, such as bar charts and line graphs, are demonstrated to extract actionable insights. Additionally, the creation of interactive dashboards in Power BI is explored, showcasing the design and implementation of user-friendly interfaces for presenting key insights.

The paper concludes by summarizing the workflow's significance, acknowledging limitations, and suggesting areas for future improvement. This workflow empowers data-driven decision-making by transforming unstructured web data into meaningful visual insights.

2.Introduction

In today’s data-driven world, the ability to extract and analyze vast amounts of web data has become increasingly vital for businesses, researchers, and various industries. Data scraping, the process of automatically extracting information from websites, enables the collection of large datasets that can provide valuable insights and drive informed decision-making.

The significance of data scraping spans numerous industries. In finance, it facilitates the monitoring of market trends and sentiment analysis. In e-commerce, it aids in price comparison and competitive analysis. Healthcare can benefit from scraping medical publications and databases for the latest research findings. Similarly, in marketing, data scraping helps track consumer behavior and social media trends.

This paper aims to present a comprehensive workflow for handling web data, covering the entire process from data extraction to visualization. The overall workflow includes the following steps:

2.1. **Data Exploring**: Understanding the structure of target websites and identifying the data to be extracted.

2.2. **Data Scraping**: Using tools like `Requests` and `Beautiful Soup` for extracting data from static web pages, and `Selenium` for dynamic content.

2.3. **Data Storing**: Organizing and storing the extracted data using data frames in `Pandas` for efficient manipulation.

2.4. **Data Analysis**: Cleaning and analyzing the data using `Pandas` to uncover patterns and insights.

2.5. **Data Visualization**: Creating visual representations of the data using Power BI to make the insights easily understandable and actionable.

3.Tools and Technologies Covered

3.1.Requests:

A Python library for sending HTTP requests and retrieving web content.

3.2.Beautiful Soup:

A Python library for parsing HTML and XML documents to extract data.

3.3.Selenium:

A tool for automating web browsers, useful for scraping dynamic web pages.

3.4.Pandas:

A powerful data manipulation library in Python, used for organizing, cleaning, and analyzing data.

3.5.Excel:

For storing data frames and facilitating data transfer to other tools.

3.6.Power BI:

A business analytics service for creating interactive visualizations and dashboards.

4.Data Scraping

we perform Data Scraping in Python using Beautiful Soup, requests and Selenium Libraries

Data scraping, also known as web scraping, is a technique used to extract large amounts of data from websites. This process involves sending requests to web servers, retrieving the content, and parsing it to collect specific pieces of information. As the internet grows, so does the volume of data available online, making data scraping an invaluable tool across various industries.

4.1.Significance in Various Industries

4.1.1. Finance: Financial analysts use data scraping to gather information on stock prices, market trends, and economic indicators from multiple sources. This real-time data is crucial for making informed investment decisions and conducting market research.

4.1.2. E-commerce: In the competitive world of online retail, businesses scrape data from competitors' websites to monitor pricing strategies, product availability, and customer reviews. This information helps in adjusting prices dynamically, optimizing inventory, and improving customer satisfaction.

4.1.3. Healthcare: Researchers and healthcare professionals scrape medical journals, research papers, and clinical trial databases to stay updated with the latest developments in medical science. This can accelerate the pace of medical research and improve patient care by providing timely access to critical information.

4.1.4. Marketing: Marketers scrape social media platforms, forums, and review sites to analyze consumer sentiment and track brand reputation. This data is vital for creating targeted marketing campaigns, understanding customer preferences, and enhancing product offerings.

4.1.1. Real Estate: Real estate companies use data scraping to collect information on property listings, market trends, and pricing from various real estate websites. This data helps in market analysis, valuation of properties, and identifying investment opportunities.

5.Beautiful Soup and Requests

Beautiful Soup and Requests are two essential Python libraries widely used scraping.They enable developers to efficiently retrieve and parse web content, transforming unstructured HTML data into a structured format suitable for analysis.

5.1.Requests

Requests is a simple and elegant HTTP library for Python, designed to send HTTP/1.1 requests with ease. It abstracts the complexities of making HTTP requests behind a simple API, allowing developers to focus on interacting with the data rather than managing the underlying network details.

5.2.Beautiful Soup

Beautiful Soup is a Python library for parsing HTML and XML documents. It creates a parse tree from page source code that can be used to extract data from HTML, making it easy to navigate and search through the web page's structure.

5.2.1.Key Features:

- **Parsing Flexibility:** Supports multiple parsers, including the built-in Python parser and lxml.
- **Navigable Structure:** Allows easy navigation and searching of the parse tree using tags, attributes, and text.
- **Robust Handling:** Deals with imperfect HTML documents, making it resilient to poorly formatted web pages.

5.3)Combining Requests and Beautiful Soup

The true power of web scraping is realized when combining Requests and Beautiful Soup. Requests handles the retrieval of the web page, while Beautiful Soup parses and extracts the desired information.

Example Workflow:

5.3.1. Send a Request: Use Requests to fetch the web page content.
5.3.2. Parse the HTML: Feed the content to Beautiful Soup for parsing.
5.3.3. Extract Data: Navigate through the parse tree to extract specific data.

6.Selenium

Explain what Selenium is and how it differs from Beautiful Soup. Discuss the scenarios where selenium is more appropriate, such as scraping dynamic content. Provide code examples to demonstrate how to set up Selenium, navigate web pages, and extract data.

6.1.The code we wrote for extracting data from linkedin.

```
from bs4 import BeautifulSoup
from selenium import webdriver
from selenium.webdriver import ActionChains
from selenium.webdriver.common.by import By
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.chrome.service import Service
from selenium.webdriver.chrome.options import Options
from selenium.webdriver.support.ui import WebDriverWait
import pandas as pd
import time
from openpyxl.workbook import Workbook
from openpyxl import load_workbook
import os
path=r"D:/internshipwork/webscraping/chrome-win64/chrome.exe"
driver=webdriver.Chrome()
driver.get("https://www.linkedin.com/jobs/search/currentJobId=3939432016&f_TPR=r604800&keywords=data%20analyst&origin=JOB_SEARCH_PAGE_JOB_FILTER&refresh=true&sortBy=R&spellCorrectionEnabled=true")
driver.maximize_window()
j=driver.find_element(By.XPATH,'/html/body/div[1]/header/nav/section/button')
time.sleep(2)
```

```
b=driver.find_element(By.XPATH,'//*[@id="job-search-bar-keywords"]')#//*[@id="job-search-bar-keywords"]
for i in range(1):
    b.send_keys(Keys.BACKSPACE)
b.send_keys('is')
k=driver.find_element(By.XPATH,'//*[@id="job-search-bar-location"]')#//*[@id="job-search-bar-location"]
for i in range(13):
    k.send_keys(Keys.BACKSPACE)
k.send_keys('India')
k.send_keys(Keys.ENTER)
unordered_list= driver.find_elements(By.CLASS_NAME,"jobs-search__results-list")
for ul in unordered_list:
    list_items = ul.find_elements(By.TAG_NAME, "li")


    for item in list_items:
        try:
            hiring_item= item.find_element(By.CLASS_NAME,"job-posting-benefits__text")


        except:
            print("")
 job_title = driver.find_elements(By.CLASS_NAME,"base-search-card__title")
 company_name = driver.find_elements(By.CLASS_NAME,"base-search-card__subtitle")
location = driver.find_elements(By.CLASS_NAME,"job-search-card__location")
hiring_information = driver.find_elements(By.CLASS_NAME,"job-posting-benefits__text")
posted_when = driver.find_elements(By.CLASS_NAME,"job-search-card__listdate--new")
```

```
        #a=str(i+2)
    #link=driver.find_element(By.DATA-TEST-PAGINATION-PAGE-BTN_NAME,"a")
job_title1=[]
company_name1=[]
location1=[]
hiring_information1=[]
posted_when1=[]
Seniority_level1=[]
Employment_type1=[]
Job_function1=[]
Industries1=[]

print(len(job_title))
a=len(job_title)
list_items = ul.find_elements(By.TAG_NAME, "li")



for item in list_items:
    print(1)
    print(item.text )

for item in list_items:
    try:
        job_title= item.find_element(By.CLASS_NAME,"base-search-card__title")
        job_title1.append(job_title.text)
    except:
        job_title1.append("---------")
```

```
for item in list_items:
    try:
        company_name= item.find_element(By.CLASS_NAME,"base-search-card__subtitle")
        company_name1.append(company_name.text)
    except:
        company_name1.append("---------")
 for item in list_items:
    try:
        location= item.find_element(By.CLASS_NAME,"job-search-card__location")
        location1.append(location.text)
    except:
        location1.append("---------")
for item in list_items:
    try:
        hiring_item= item.find_element(By.CLASS_NAME,"job-posting-benefits__text")
        hiring_information1.append(hiring_item.text)
    except:
        hiring_information1.append("---------")

for item in list_items:
    try:
        posted_item= item.find_element(By.CLASS_NAME,"job-search-card__listdate--new")
        posted_when1.append(posted_item.text)
    except:
        posted_when1.append("---------")
s=[]
```

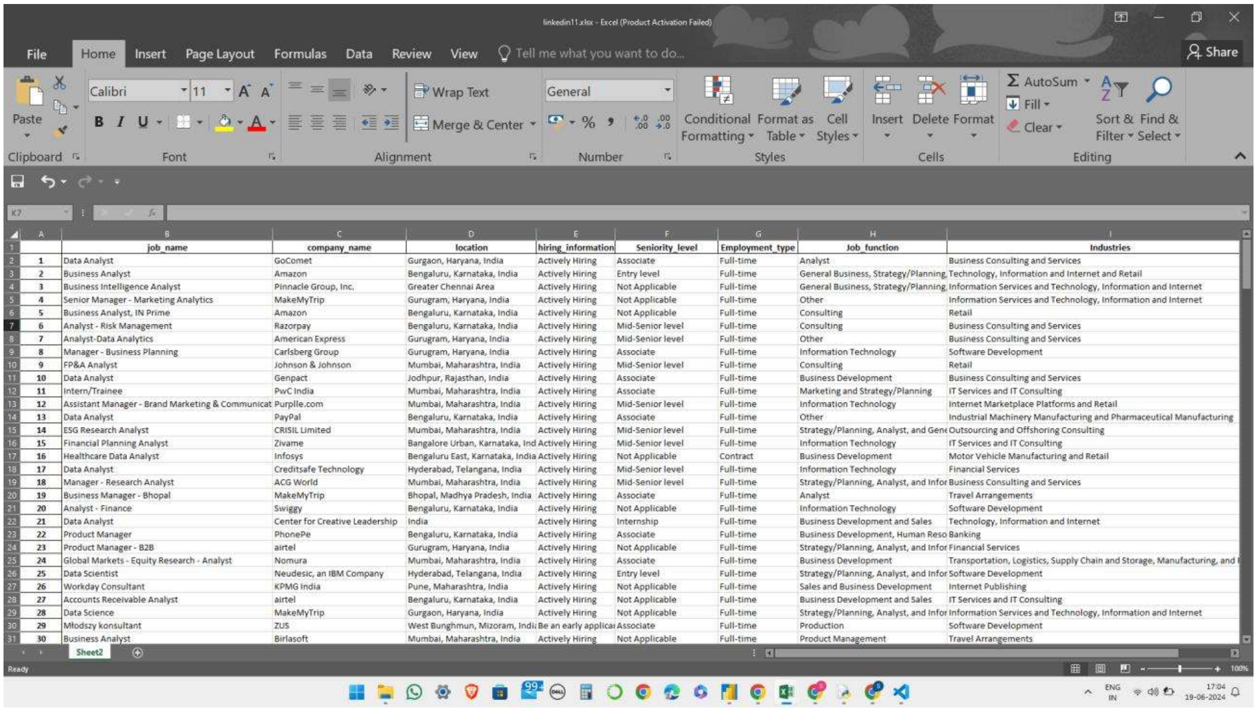
```
for item in list_items:
    l=item.find_element(By.TAG_NAME,'a')
    link=l.get_attribute("href")
    s.append(link)
a=0

for i in s:
    a=a+1
    print(a)
    driver.get(i)
    inside_items=driver.find_elements(By.CLASS_NAME,"description__job-criteria-item")
    #print(2)
    for i in inside_items:
        n=i.find_element(By.CSS_SELECTOR,"span")
        l=i.find_element(By.TAG_NAME,"h3")
        print(n.text)
         z=l.text

       if(z=="Seniority level"):
            Seniority_level1.append(n.text)
            print('a')
        elif(z=="Employment type"):
            Employment_type1.append(n.text)
            print('b')
        elif(z=="Job function"):
            Job_function1.append(n.text)
            print('c')
        else:
            Industries1.append(n.text)
            print('d')


    #time.sleep(0.1)
    time.sleep(4)


print(len(job_title1))
print(len(company_name1))
print(len(location1))
print(len(hiring_information1))
print(len(posted_when1))
print(len(Seniority_level1))
print(len(Employment_type1))
print(len(Job_function1))
print(len(Industries1))
```

7.Usage of Data Frames in Pandas

```
mydataset'= {job_title':job_title1, 'company_name':company_name1, 'location':location1,
'hiring_information':hiring_information1, 'posted_when':posted_when1,
'Employment_type':Employment_type1, 'Job_function':Job_function1,
'Seniority_level':Seniority_level1, 'Industries':Industries1}
myvar = pd.DataFrame(mydataset)
a=myvar
```

8.We use excel to Transfer Data Frames to Excel

```
with pd.ExcelWriter('linkedin.xlsx') as writer:
        myvar.to_excel(writer, sheet_name='sheet1')
```

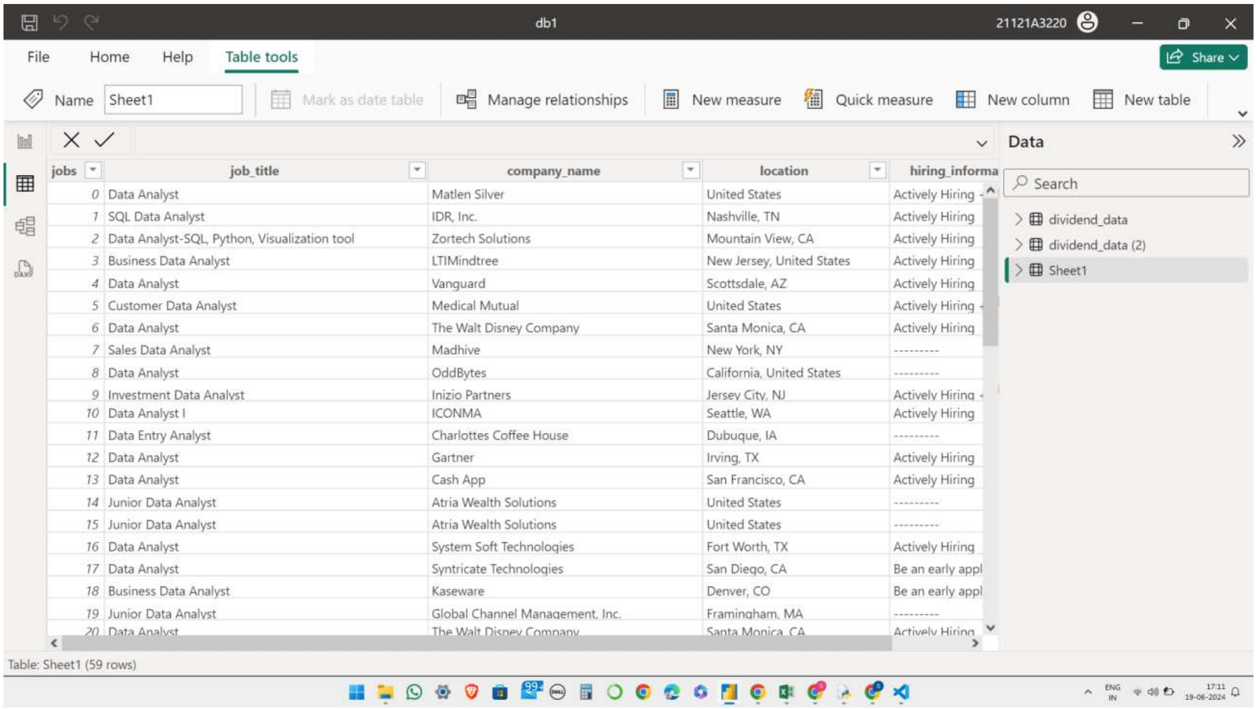


	job_name	company_name	location	hiring_information	Seniority_level	Employment_type	Job_function	Industries
1	Data Analyst	GoComet	Gurgaon, Haryana, India	Actively Hiring	Associate	Full-time	Analyst	Business Consulting and Services
2	Business Analyst	Amazon	Bengaluru, Karnataka, India	Actively Hiring	Entry level	Full-time	General Business, Strategy/Planning	Technology, Information and Internet and Retail
3	Business Intelligence Analyst	Pinnacle Group, Inc.	Greater Chennai Area	Actively Hiring	Not Applicable	Full-time	General Business, Strategy/Planning	Information Services and Technology, Information and Internet
4	Senior Manager - Marketing Analytics	MakeMyTrip	Gurugram, Haryana, India	Actively Hiring	Not Applicable	Full-time	Other	Information Services and Technology, Information and Internet
5	Business Analyst, IN Prime	Amazon	Bengaluru, Karnataka, India	Actively Hiring	Not Applicable	Full-time	Consulting	Retail
6	Analyst - Risk Management	Razorpay	Bengaluru, Karnataka, India	Actively Hiring	Mid-Senior level	Full-time	Consulting	Business Consulting and Services
7	Analyst-Data Analytics	American Express	Gurugram, Haryana, India	Actively Hiring	Mid-Senior level	Full-time	Other	Business Consulting and Services
8	Manager - Business Planning	Carlsberg Group	Gurugram, Haryana, India	Actively Hiring	Associate	Full-time	Information Technology	Software Development
9	FP&A Analyst	Johnson & Johnson	Mumbai, Maharashtra, India	Actively Hiring	Mid-Senior level	Full-time	Consulting	Retail
10	Data Analyst	Genpact	Jodhpur, Rajasthan, India	Actively Hiring	Associate	Full-time	Business Development	Business Consulting and Services
11	Intern/Trainee	PwC India	Mumbai, Maharashtra, India	Actively Hiring	Associate	Full-time	Marketing and Strategy/Planning	IT Services and IT Consulting
12	Assistant Manager - Brand Marketing & Communicat	Purplle.com	Mumbai, Maharashtra, India	Actively Hiring	Mid-Senior level	Full-time	Information Technology	Internet Marketplace Platforms and Retail
13	Data Analyst	PayPal	Bengaluru, Karnataka, India	Actively Hiring	Associate	Full-time	Other	Industrial Machinery Manufacturing and Pharmaceutical Manufacturing
14	ESG Research Analyst	CRISIL Limited	Mumbai, Maharashtra, India	Actively Hiring	Mid-Senior level	Full-time	Strategy/Planning, Analyst, and Gene	Outsourcing and Offshoring Consulting
15	Financial Planning Analyst	Zivame	Bangalore Urban, Karnataka, Ind	Actively Hiring	Mid-Senior level	Full-time	Information Technology	IT Services and IT Consulting
16	Healthcare Data Analyst	Infosys	Bengaluru East, Karnataka, India	Actively Hiring	Not Applicable	Contract	Business Development	Motor Vehicle Manufacturing and Retail
17	Data Analyst	Creditsafe Technology	Hyderabad, Telangana, India	Actively Hiring	Mid-Senior level	Full-time	Information Technology	Financial Services
18	Manager - Research Analyst	ACG World	Mumbai, Maharashtra, India	Actively Hiring	Mid-Senior level	Full-time	Strategy/Planning, Analyst, and Infor	Business Consulting and Services
19	Business Manager - Bhopal	MakeMyTrip	Bhopal, Madhya Pradesh, India	Actively Hiring	Associate	Full-time	Analyst	Travel Arrangements
20	Analyst - Finance	Swiggy	Bengaluru, Karnataka, India	Actively Hiring	Not Applicable	Full-time	Information Technology	Software Development
21	Data Analyst	Center for Creative Leadership	India	Actively Hiring	Internship	Full-time	Business Development and Sales	Technology, Information and Internet
22	Product Manager	PhonePe	Bengaluru, Karnataka, India	Actively Hiring	Associate	Full-time	Business Development, Human Reso	Banking
23	Product Manager - B2B	airtel	Gurugram, Haryana, India	Actively Hiring	Not Applicable	Full-time	Strategy/Planning, Analyst, and Infor	Financial Services
24	Global Markets - Equity Research - Analyst	Nomura	Mumbai, Maharashtra, India	Actively Hiring	Associate	Full-time	Business Development	Transportation, Logistics, Supply Chain and Storage, Manufacturing, and
25	Data Scientist	Neudesic, an IBM Company	Hyderabad, Telangana, India	Actively Hiring	Entry level	Full-time	Strategy/Planning, Analyst, and Infor	Software Development
26	Workday Consultant	KPMG India	Pune, Maharashtra, India	Actively Hiring	Not Applicable	Full-time	Sales and Business Development	Internet Publishing
27	Accounts Receivable Analyst	airtel	Bengaluru, Karnataka, India	Actively Hiring	Not Applicable	Full-time	Business Development and Sales	IT Services and IT Consulting
28	Data Science	MakeMyTrip	Gurgaon, Haryana, India	Actively Hiring	Not Applicable	Full-time	Strategy/Planning, Analyst, and Infor	Information Services and Technology, Information and Internet
29	Młodszy konsultant	ZUS	West Bunghmun, Mizoram, Indi:	Be an early applicai	Associate	Full-time	Production	Software Development
30	Business Analyst	Birlasoft	Mumbai, Maharashtra, India	Actively Hiring	Not Applicable	Full-time	Product Management	Travel Arrangements

Similarly we can retrieve the data from data frames. How much we wanted.

9.Loading excel Data into Power Bi

We can load data of excel sheet by browsing it.



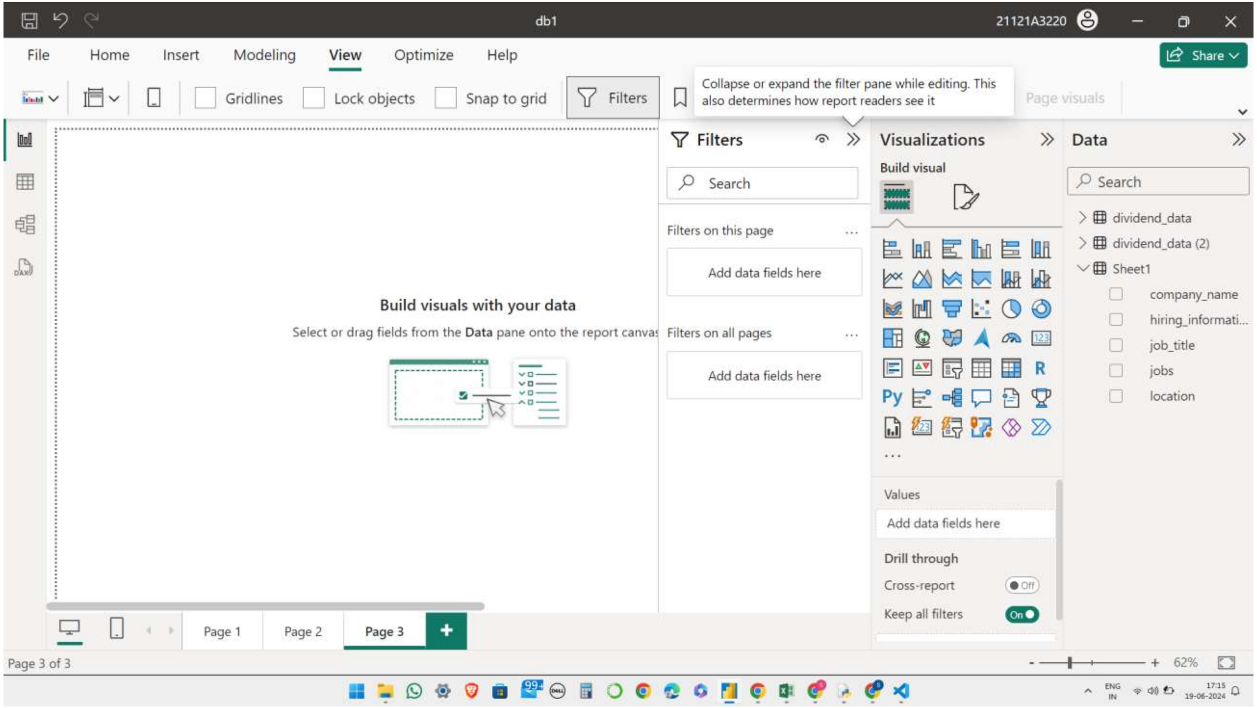
Name: Sheet1

jobs	job_title	company_name	location	hiring_informa
0	Data Analyst	Matlen Silver	United States	Actively Hiring -
1	SQL Data Analyst	IDR, Inc.	Nashville, TN	Actively Hiring
2	Data Analyst-SQL, Python, Visualization tool	Zortech Solutions	Mountain View, CA	Actively Hiring
3	Business Data Analyst	LTIMindtree	New Jersey, United States	Actively Hiring
4	Data Analyst	Vanguard	Scottsdale, AZ	Actively Hiring
5	Customer Data Analyst	Medical Mutual	United States	Actively Hiring -
6	Data Analyst	The Walt Disney Company	Santa Monica, CA	Actively Hiring
7	Sales Data Analyst	Madhive	New York, NY	---------
8	Data Analyst	OddBytes	California, United States	---------
9	Investment Data Analyst	Inizio Partners	Jersey City, NJ	Actively Hiring -
10	Data Analyst I	ICONMA	Seattle, WA	Actively Hiring
11	Data Entry Analyst	Charlottes Coffee House	Dubuque, IA	---------
12	Data Analyst	Gartner	Irving, TX	Actively Hiring
13	Data Analyst	Cash App	San Francisco, CA	Actively Hiring
14	Junior Data Analyst	Atria Wealth Solutions	United States	---------
15	Junior Data Analyst	Atria Wealth Solutions	United States	---------
16	Data Analyst	System Soft Technologies	Fort Worth, TX	Actively Hiring
17	Data Analyst	Syntricate Technologies	San Diego, CA	Be an early appl
18	Business Data Analyst	Kaseware	Denver, CO	Be an early appl
19	Junior Data Analyst	Global Channel Management, Inc.	Framingham, MA	---------
20	Data Analyst	The Walt Disney Company	Santa Monica, CA	Actively Hiring

Table: Sheet1 (59 rows)

10.Analysing of Data columns

We can analyse using columns of data in tables



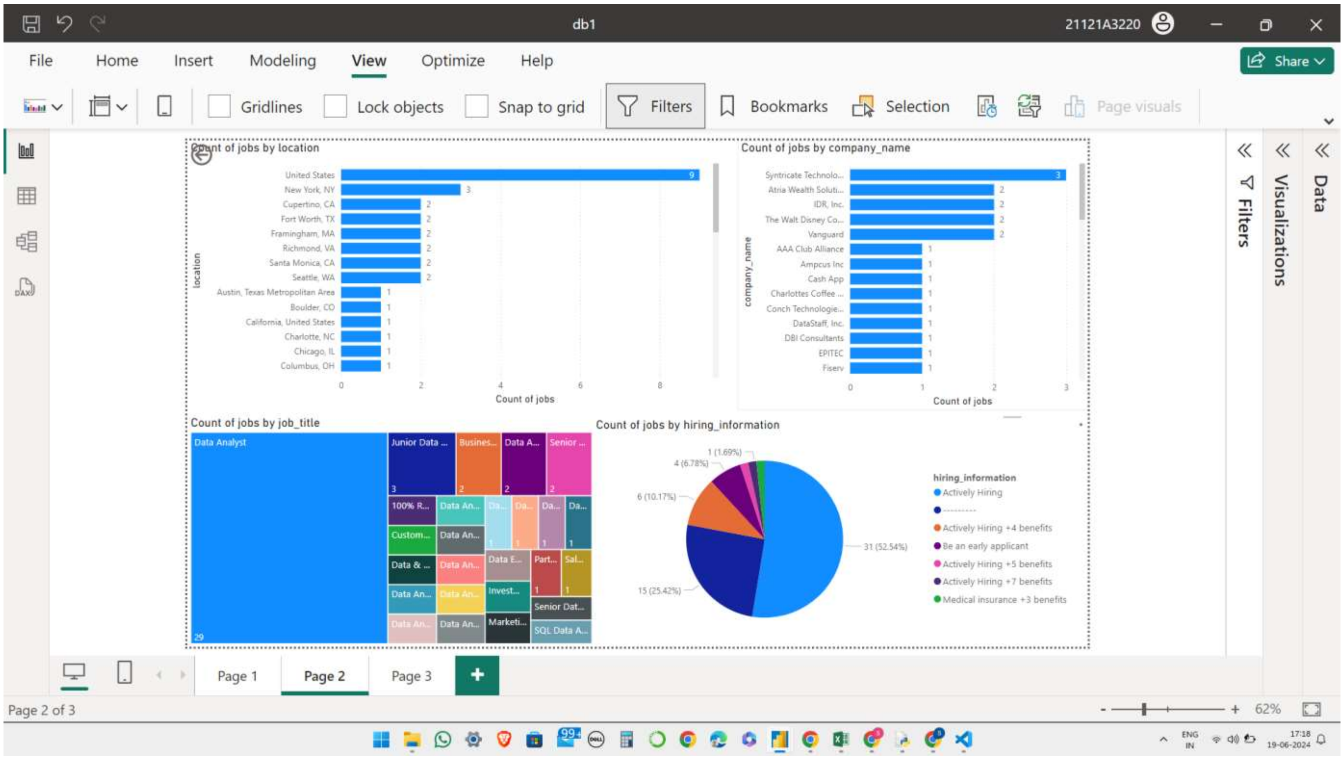
Collapse or expand the filter pane while editing. This also determines how report readers see it

Build visuals with your data

Select or drag fields from the Data pane onto the report canvas

11.Showing Visualisations of Data and Presentation of Data in Dash Boards

We can create dash boards out the data that has been collected.

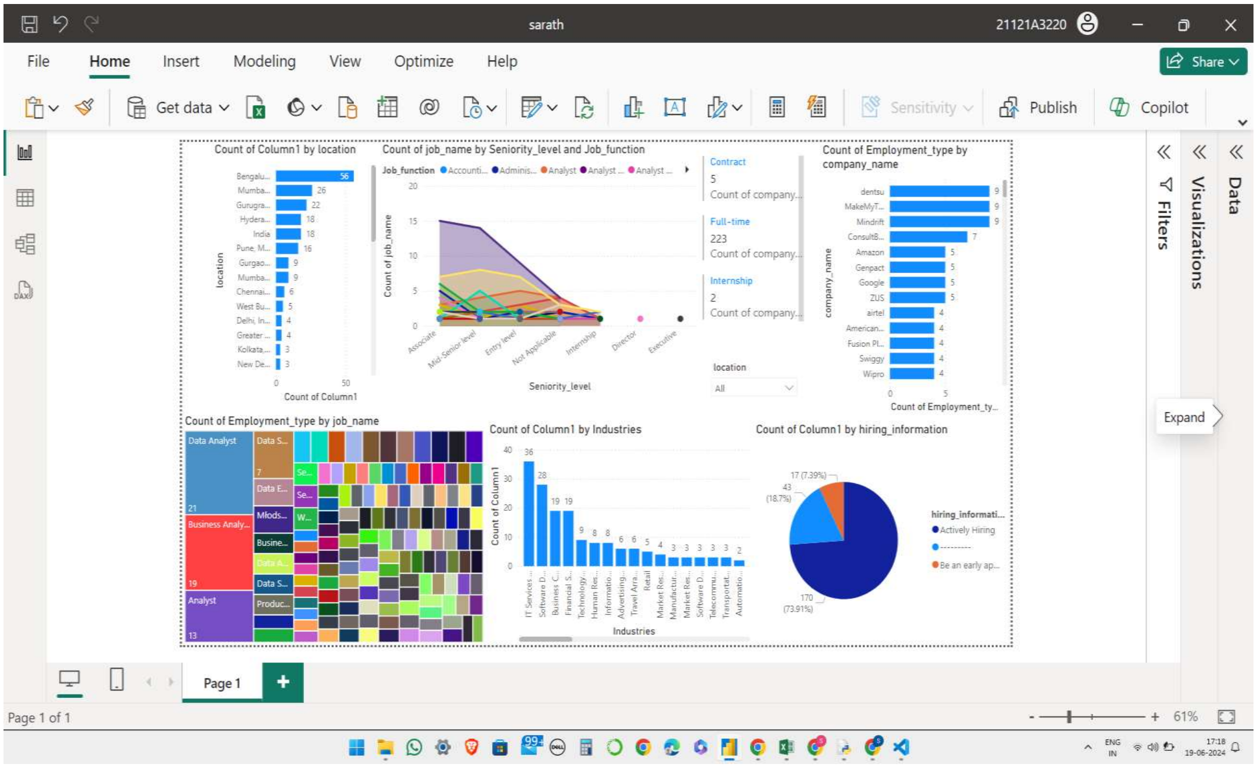


We have use 4 data visualisations her for data that had been analysed.

We performed the data visualisation and shown each data visualisation here.

12.Improvements after updating

We have updated information using selenium code again and iterated throgh each job and extracted more data presented second dash board.



We have use 8 data visualisations her for data that had been analysed.

We performed the data visualisation and shown each data visualisation here.

13.Conclusion

The integration of Selenium, Pandas, and Power BI creates a powerful toolchain for transforming LinkedIn web data into meaningful, actionable insights. This workflow not only enhances data accessibility and usability but also empowers stakeholders to leverage data for strategic advantages. By continually refining these techniques and tools, the potential for data-driven decision-making across various domains will continue to grow, providing a significant edge in today's competitive landscape.