FINAL REPORT

ON

DATA SCRAPING

BY

R.AKHILESH SAGAR

18STUCHH010266

AT

INTENSEL, HONG KONG

An Internship Program - I station of



**FACULTY OF SCIENCE AND TECHNOLOGY**

**ICFAI FOUNDATION FOR HIGHER EDUCATION**

**(Deemed to be University under Section 3 of UGC Act. 1956)**

**HYDERABAD**

**AUGUST 2020**

A REPORT

ON

DATA SCRAPING

BY

R.AKHLIESH SAGAR

18STUCHH010266

PREPARED IN PARTIAL FULFILLMENT OF

INTERNSHIP PROGRAM-I

AT

INTENSEL , HONG KONG

An IP-I station of



**FACULTY OF SCIENCE AND TECHNOLOGY**

**ICFAI FOUNDATION FOR HIGHER EDUCATION**

**(Deemed to be University under Section 3 of UGC Act. 1956)**

**HYDERABAD**

**AUGUST 2020**

**Faculty of Science & Technology, IFHE Hyderabad**

**Station:** Intensel

**Duration:** 8 Weeks **Date of start:** 1st July 2020

**Date of Submission**: 26th August 2020

**Title of Project:** Data Scraping

**ID NO./Name/Discipline**:18STUCHH010266 R.Akhilesh Sagar CSE

**Name and designation of the expert:** Dr.Entela Benz (Co-Founder & CEO) and Dr.Sourabh Singh (Co-founder & CTO)

**Name of the IP faculty:** Sirisha Potluri

TABLE OF CONTENTS

1. **Acknowledgment5**
2. **Abstract6**
3. **Introduction6**

3.1 How it works6

* 1. Why Web Scraping?7

3.3 Advantages & Disadvantadges7

3.4 GIS and its uses8

1. **Components9**

4.1 Geo Pandas9

* 1. Beautiful Soup10

4.3 Matplotlib10

4.4 Selenium10

4.5 What are CSV files?11

4.6 What are Shapefiles?11

1. **Screenshots12**
2. **Working of Code14**
3. **Final Work15**
4. **About Intensel17**
5. **Conclusion18**
6. **References19**

**1.ACKNOWLEDGMENT**

In this acknowledgment,I would like to thank all the people who helped me, whose guidance and encouragement made it possible.The internship opportunity I got in INTENSEL was a great chance in learning so much and a professional development of my skills.Therefore, I’m very lucky to be a part of it. I had a chance to meet wonderful people and professionals who led me through this internship period.

I would like to express our thankfulness to our director **Mr. M. Srinivasa Reddy** of Faculty of Science and Technology (ICFAI Foundation for Higher Education, Hyderabad) for giving us such kind of opportunity.

I am grateful to our Placement and Training Manager **P.Lakshmi Narayana** for giving me an opportunity for participation in this Internship Program.

I am also very much thankful to the IP coordinator **Sirisha Potluri** for her constant supervision, guidance and co-operation throughout the project.

We have taken efforts in this project. However, it would not have been possible without the kind support and help of **INTENSEL**, Hong Kong.

I would like to express our special gratitude and thanks to **Dr.Entela benz & Dr.Sourabh Singh**,**Co-Founders, Intensel,Hong Kong** for giving us such attention and time.

And a huge thanks to our Mentor **Aditya Allamraju, GIS Analyst,** without which we would not have completed our intership successfully.

At last I would like to extend my sincere thanks to all of them who helped us to complete this project.

**2.ABSTRACT**

Data scraping, also known as web scraping, is the process of importing information from a website into a spreadsheet or local file saved on your computer. It's one of the most efficient ways to get data from the web, and in some cases to channel that data to another website.The software for web scraping may access the planet’s Wide web directly Hypertext Transfer Protocol and retrieve the data.Usually data scraping is done manually, but including automation into it makes it more efficient and saves lot of time.

Since modern business runs on high volume of data, these data is unstructured, and extracting them can be an intense labour work.So using these traditional methods of web extraction with an automation tool bridges the gap between unstructured data and business applications.And when these two things are combined they allow us to read, write, and update a wide variety of data sources automatically.

**3.INTRODUCTION**

Data scraping refers to a technique of extracting data from output generated from another program. Data scraping is commonly manifest in web scraping, the process of using an application to extract valuable information from a website. Webscraping abilities helps us to create and power a number of the world’s most revolutionary business applications.

### **How It Works ?**

1. First the piece of code used to pull the information, which we call a scraper bot, sends an HTTP GET request to a specific website.
2. When the website responds, the scraper parses the HTML document for a specific pattern of data.
3. Once the data is extracted, it is converted into whatever specific format the scraper bot’s author designed.

**Why Web Scraping ?**

Typically companies do not want their unique content to be downloaded and reused for unauthorized purposes. As a result, they don’t expose all data via a consumable API or other easily accessible resource. Scraper bots, on the other hand, are interested in getting website data regardless of any attempt at limiting access. As a result, a cat-and-mouse game exists between web scraping bots and various content protection strategies, with each trying to outmaneuver the other.

### **Several different uses of web scraping are:**

Scraping stock prices into an app API

Scraping data from a store locator to create a list of business locations

Scraping product data from sites like Amazon or eBay for competitor analysis

Scraping sports stats for betting or fantasy leagues

Scraping site data before a website migration

Scraping financial data for market research and insights

**Advantages:**

* Inexpensive
* Easy to Implement
* Online Presence can be tracked
* To gather public opinion
* To bulid vertical-specific search engines

**Disadvantages:**

* Difficult to analyse
* Time taking process
* Speed and protection policies
* Constrained by your limits in bandwidth,processing and storage

**GIS (Geographic Information System)**

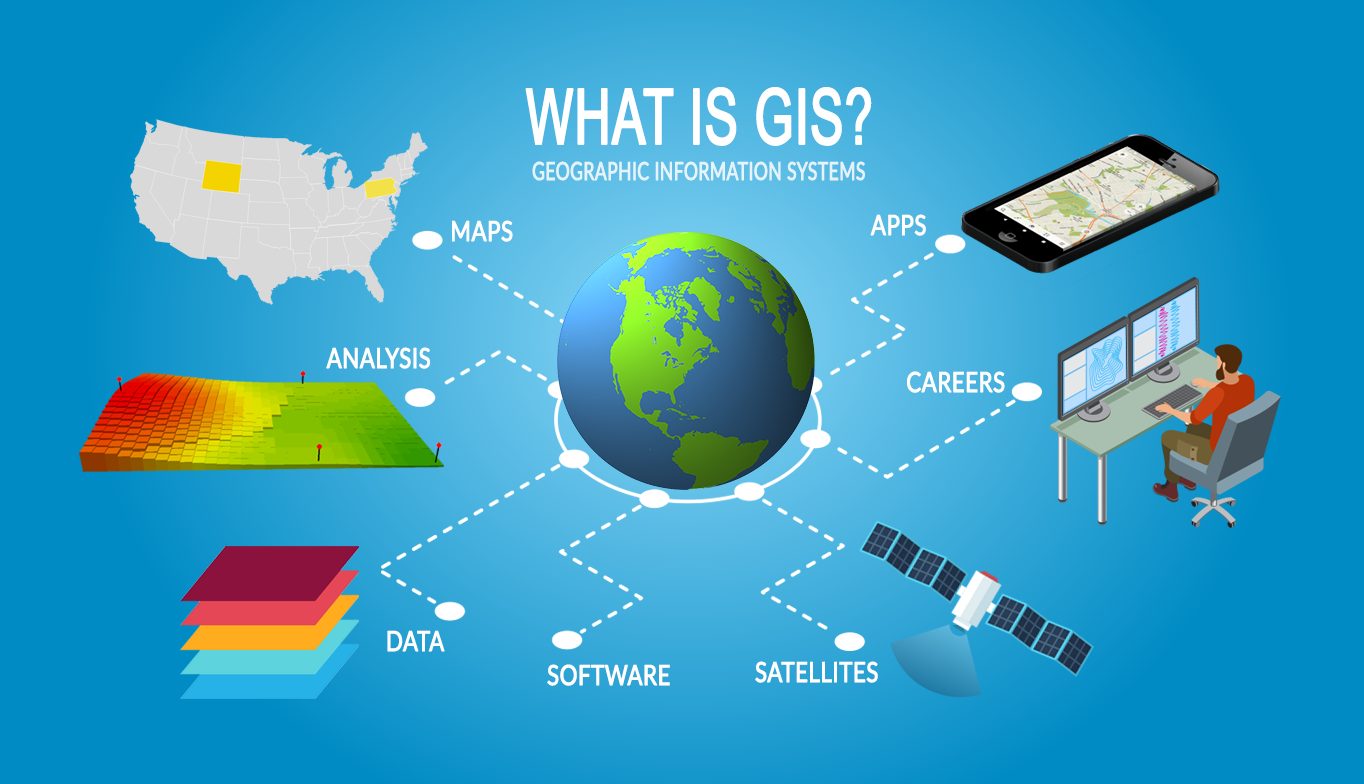
A geographic information system (GIS) is a computer system for capturing, storing, checking, and displaying data related to positions on Earth’s surface. By relating seemingly unrelated data, GIS can help individuals and organizations better understand spatial patterns and relationships.

GIS technology is a crucial part of spatial data infrastructure, which the White House defines as “the technology, policies, standards, human resources, and related activities necessary to acquire, process, distribute, use, maintain, and preserve spatial data.”

GIS can use any information that includes location. The location can be expressed in many different ways, such as latitude and longitude, address, or ZIP code.

**Common uses of GIS include:**

Inventory and management of resources, crime mapping, establishing and monitoring routes, managing networks, monitoring and managing vehicles, managing properties, locating and targeting customers, locating properties that match specific criteria and managing agricultural crop data, addressing public health concerns, mapping wildfire risk and preparedness, modeling hazmat risk, first response, and mapping/monitoring/mitigating invasive species.



**4.COMPONENTS**

**Python**

Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. An important goal of Python's developers is keeping it fun to use.

Python is meant to be an easily readable language. Its formatting is visually uncluttered, and it often uses English keywords where other languages use punctuation. The latest version of python is 3.8. Python even has the fastest debugger which inspects local and global variable, evaluation of arbitrary expressions as the debugger itself is built in the python language.It’s used by every programmer because it can handle big data and complex mathematics.

The **Python Package Index** (PyPI), the official repository for third-party Python software, contains over 200,000 packages with a wide range of functionality, including:

Automation, Data analytics,Databases, Documentation, Graphical user interfaces, Image processing, Machine learning, Mobile App, Multimedia, Networking, Scientific computing, System administration, Test frameworks, Text processing, Web frameworks, Web scraping.

**The Packages and Libraries Used:**

**1) GeoPandas:** GeoPandas is an open source project to make working with geospatial data in python easier. GeoPandas extends the datatypes used by pandas to allow spatial operations on geometric types. Geometric operations are performed by shapely. Geopandas further depends on fiona for file access and descartes and matplotlib for plotting.

**2) BeautifulSoup:** Beautiful Soup is a Python library for pulling data out of HTML and XML files. It works with your favorite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work. It is a Python library designed for quick turnaround projects like screen-scraping. Three features make it powerful:

* Beautiful Soup provides a few simple methods and Pythonic idioms for navigating, searching, and modifying a parse tree: a toolkit for dissecting a document and extracting what you need. It doesn't take much code to write an application
* Beautiful Soup automatically converts incoming documents to Unicode and outgoing documents to UTF-8.
* Beautiful Soup sits on top of popular Python parsers like lxml and html5lib, allowing you to try out different parsing strategies or trade speed for flexibility.

**3) Matplotlib:** Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. It is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+.

**4) Selenium:** Selenium is an open-source tool that automates web browsers. It provides a single interface that lets you write test scripts in programming languages like Ruby, Java, NodeJS, PHP, Perl, Python, and C#, among others. It is most famous for enabling rapid, repeatable web-app testing, which allows developers to ship new releases faster and with confidence.

Modern developers and product teams don’t have that kind of time to allot for testing, but they can’t set aside exhaustive testing in a hurry to release either. This is why they super-charge their testing with automation, powered by Selenium.

**5) What are CSV files?**

These files serve a number of different business purposes. They help companies export a high volume of data to a more concentrated database, for instance.

They also serve two other primary business functions:

* CSV files are plain-text files, making them easier for the website developer to create
* Since they're plain text, they're easier to import into a spreadsheet or another storage database, regardless of the specific software you're using
* To better organize large amounts of data.

**6) What are Shapefiles?**

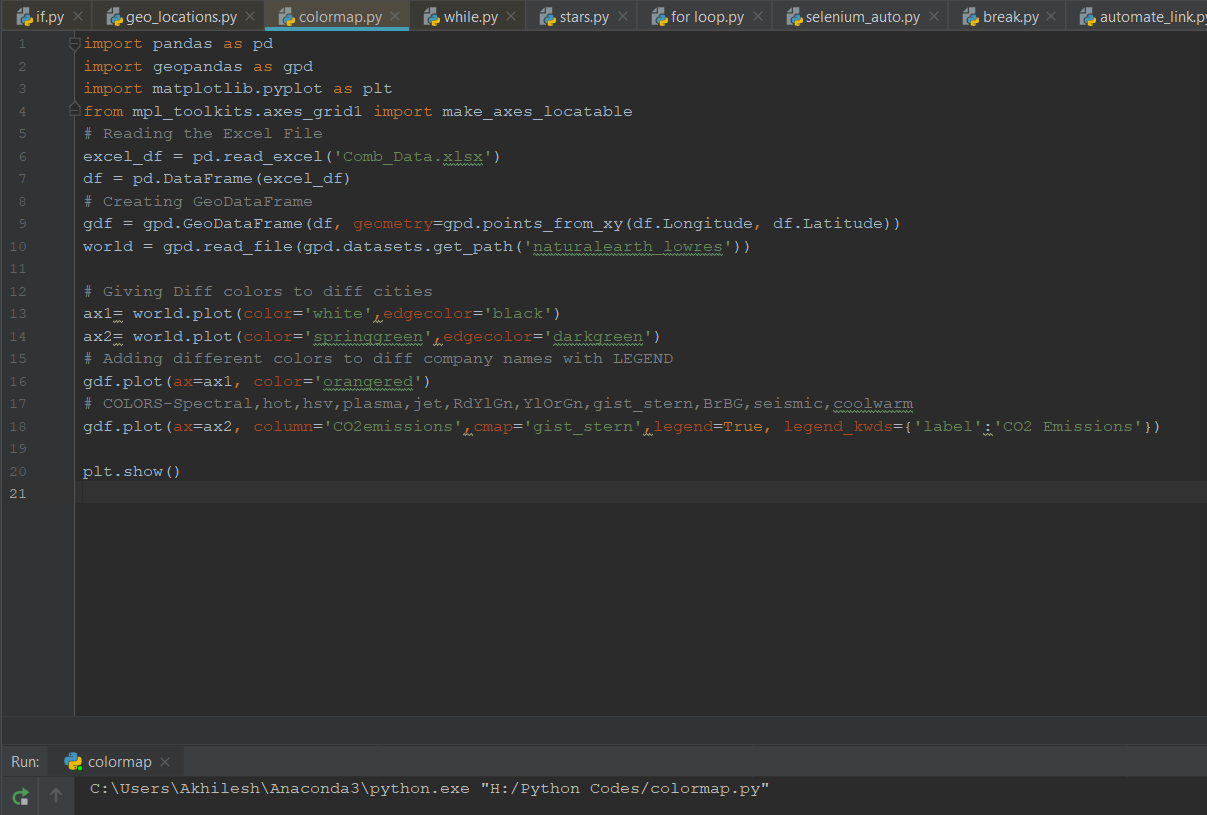
A shapefile is a simple, nontopological format for storing the geometric location and attribute information of geographic features. Geographic features in a shapefile can be represented by points, lines, or polygons (areas). The workspace containing shapefiles may also contain dBASE tables, which can store additional attributes that can be joined to a shapefile's features.

The following list summarizes how you can use shapefiles in ArcGIS Online and provides links to instructions:

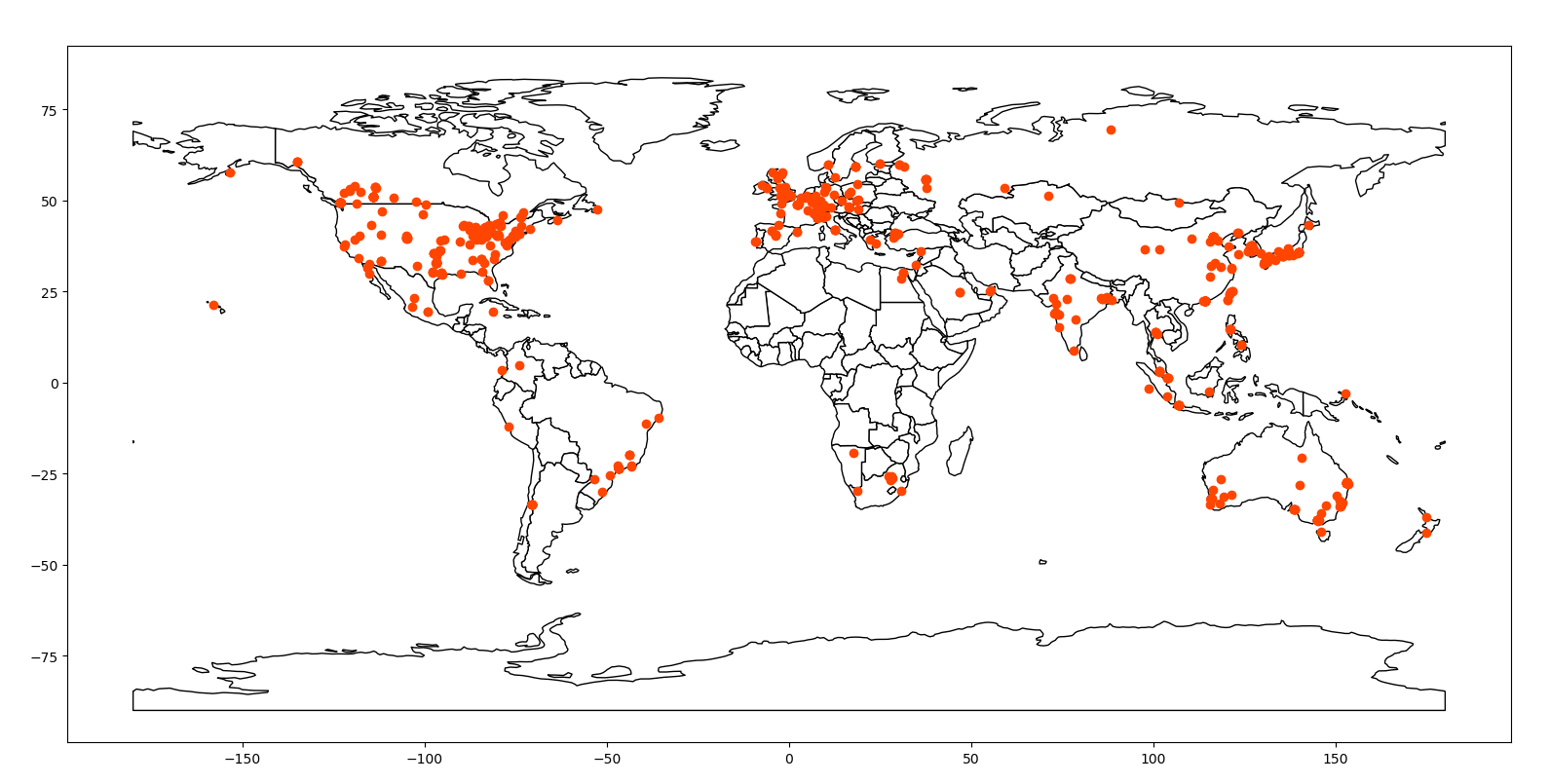
* Publish a hosted feature layer from a shapefile (.zip).
* For hosted feature layers that have export enabled, the people who have access to the hosted feature layer can export the layer data to a shapefile.
* Add shapefiles (in a .zip file) to My Content, and share the file with others so they can download it.
* Add a shapefile (.zip) as a layer to Map Viewer.

**5.SCREENSHOTS**

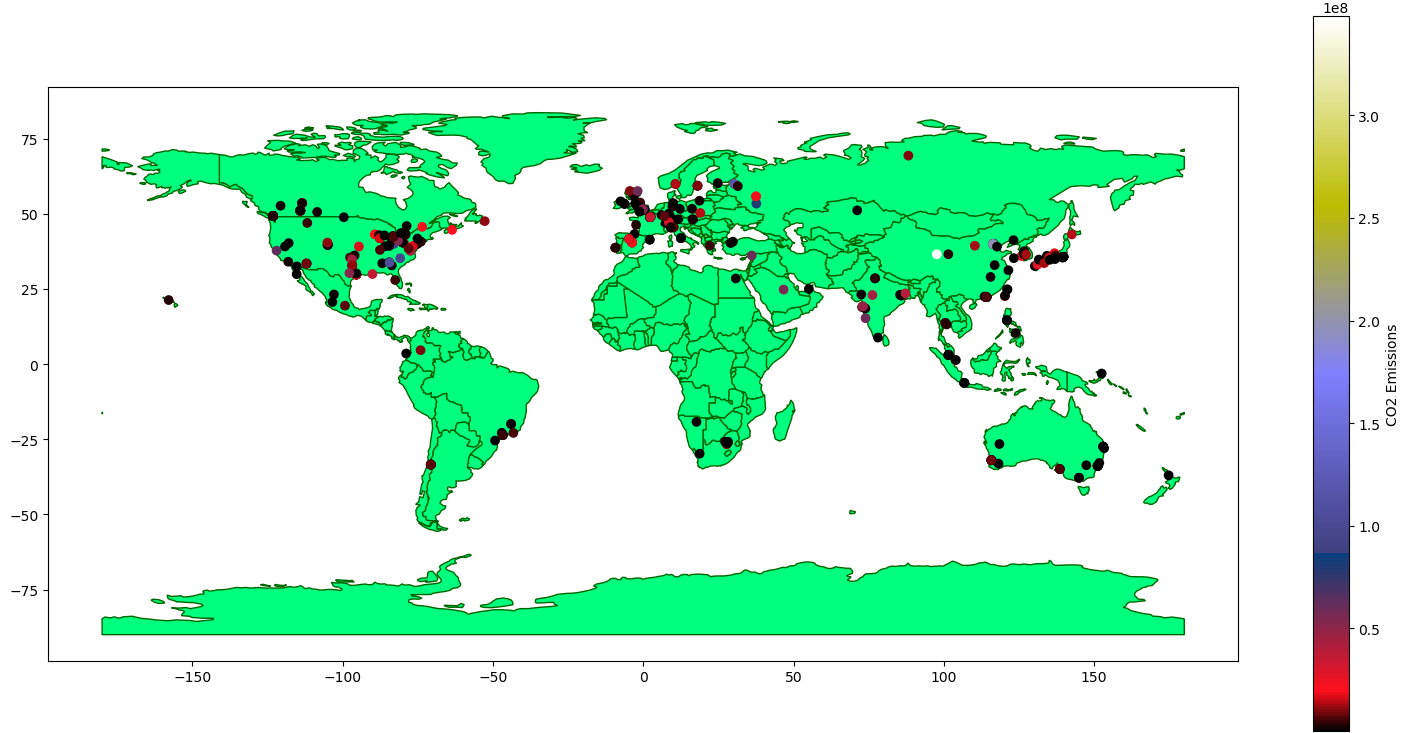
**Code for ploting the geo locations using Geo-Pandas.**



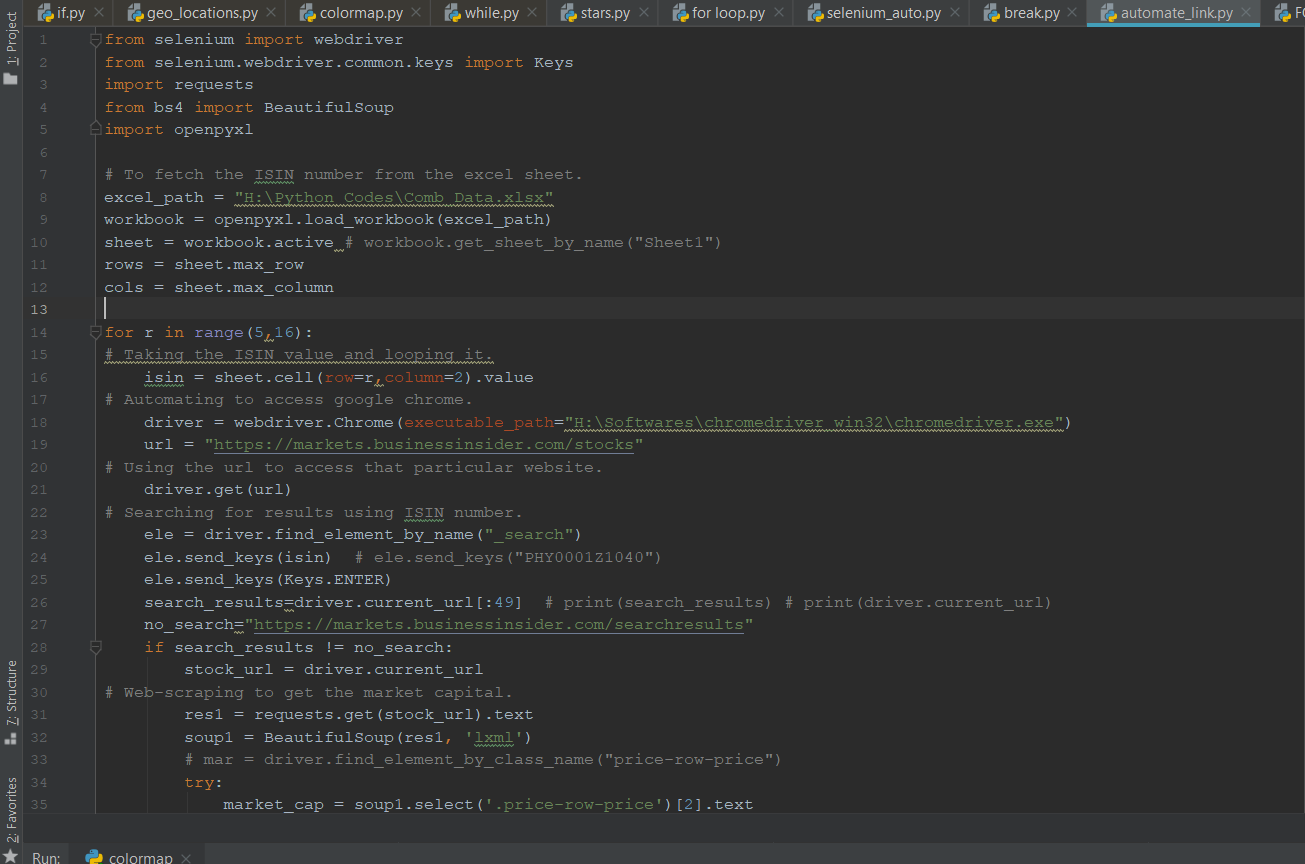
**A) Output 1: Locations of companies on map using Matplotlib.**



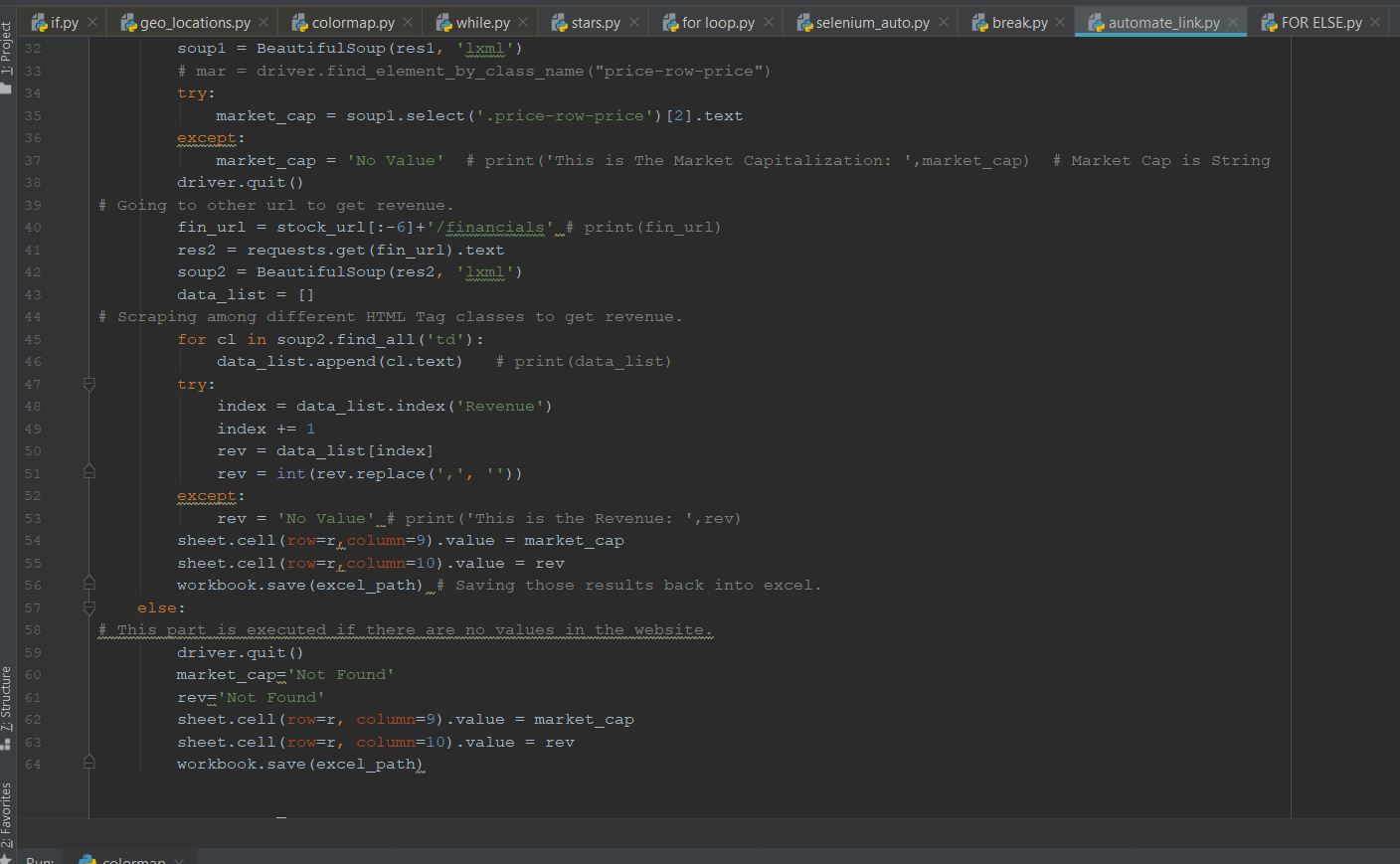
**B) Output 2: The CO2 Emissions of the Companies**



**C) Automation** & **WebScraping** Code for getting Market Capital and Revenue



Continuation…



**6.Working of this Code:**

**A)** First we need to fetch the data from excel in-order to read the input (**ISIN** number) using **Openpyxl** libraries. Then we use ‘for loop’ to repeat automation process until all the input data is is covered.

**B)** In this loop,using **Selenium** libraries we open chrome and access the ‘url link’ we give it.Then we call the **ISIN** number from excel and use it in the search bar of the website to get the company details.(This process is automated and repeated)

**C)** Then using **Requests** librarywe send the request to the website to parse its webpage content. Later we use “try and exception” and we do **Web-Srcaping** (search for the required data on the site we want) with the help of **BeautifulSoup** library.

**D)** Atlast we save the data we have retrieved from website and store it in the same excel sheet, and this processed is repeated until all the ISIN numbers are searched..

**NOTE:** **There were some cases where we didn’t get proper results, like:**

* Search results of a company, so assigned its value as “Not Found”.
* Some Market Capital & Revenue values were empty, so assigned them “No Value”, and we are further working to find other possible solutions.

**7. FINAL WORK**

Further more companies details were given to find their multiple locations along with their addresses and company website link.

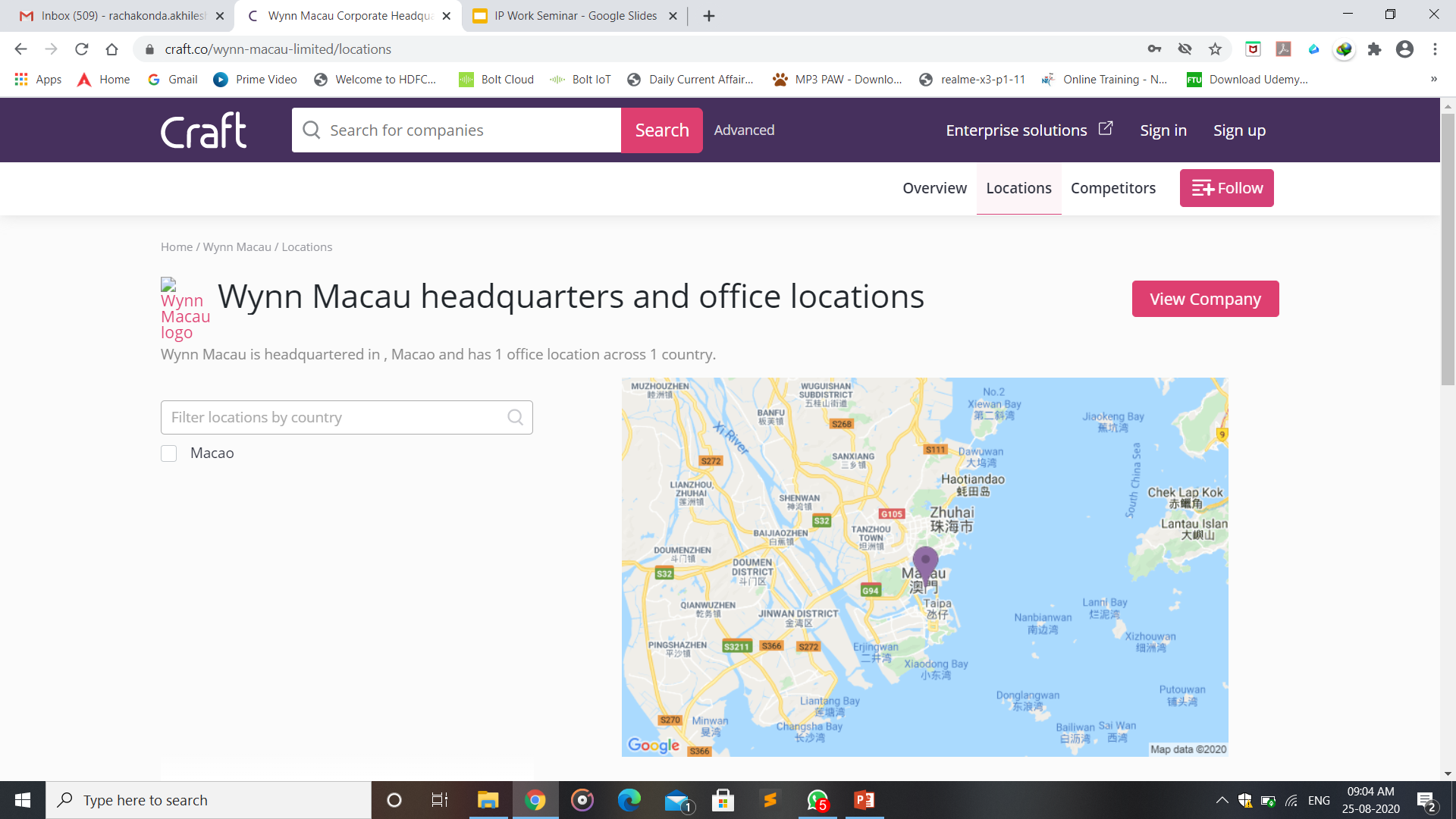
Here we have used various tools and referred some websites to find the company information and these are:

* Aeroleads
* Hunter
* Skrapp
* Snov
* Lusha
* Linkedin Sales Navigator

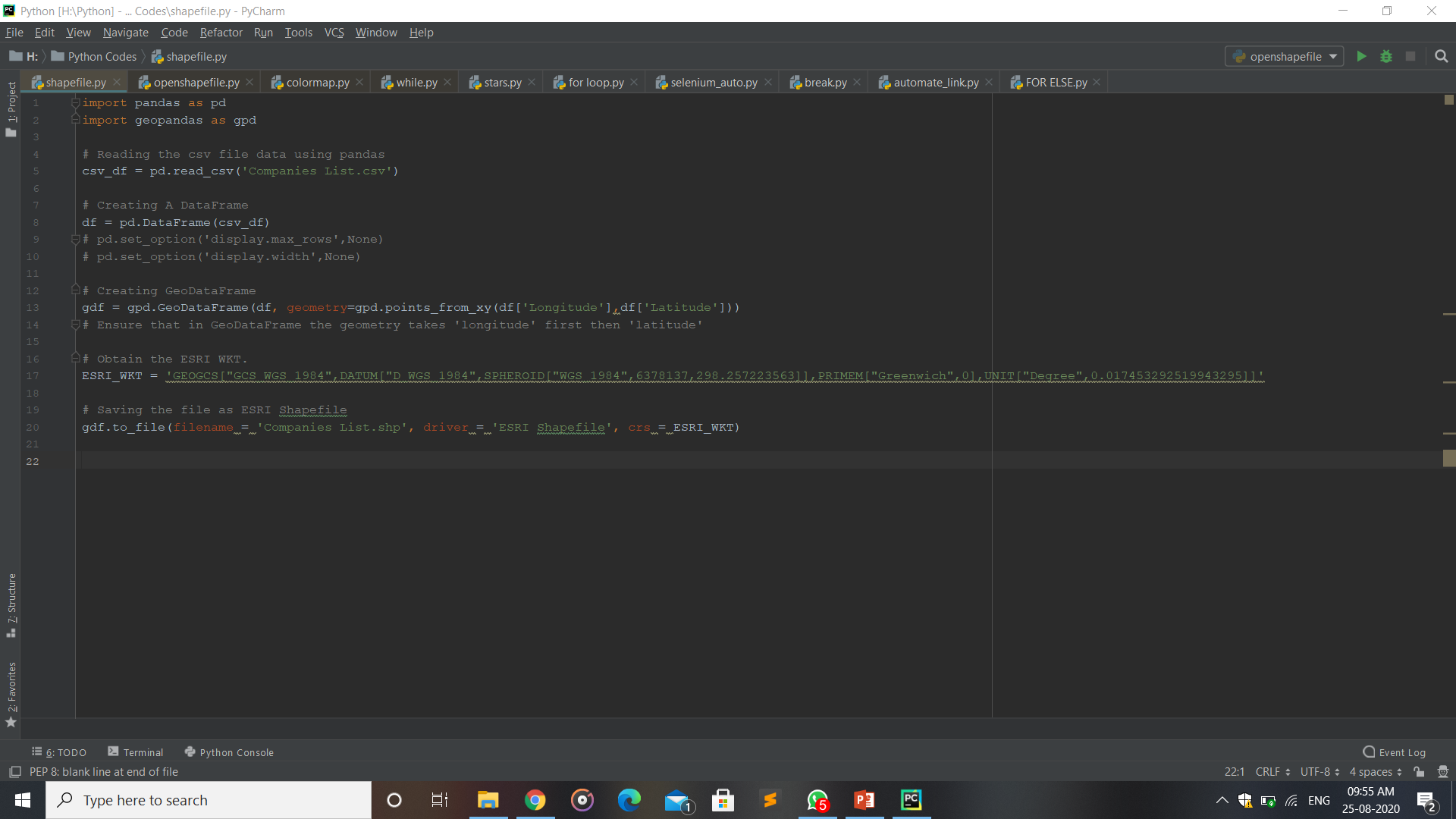
[**https://www.corporateinformation.com/**](https://www.corporateinformation.com/)

[**https://www.getcompanydetails.com/**](https://www.getcompanydetails.com/)

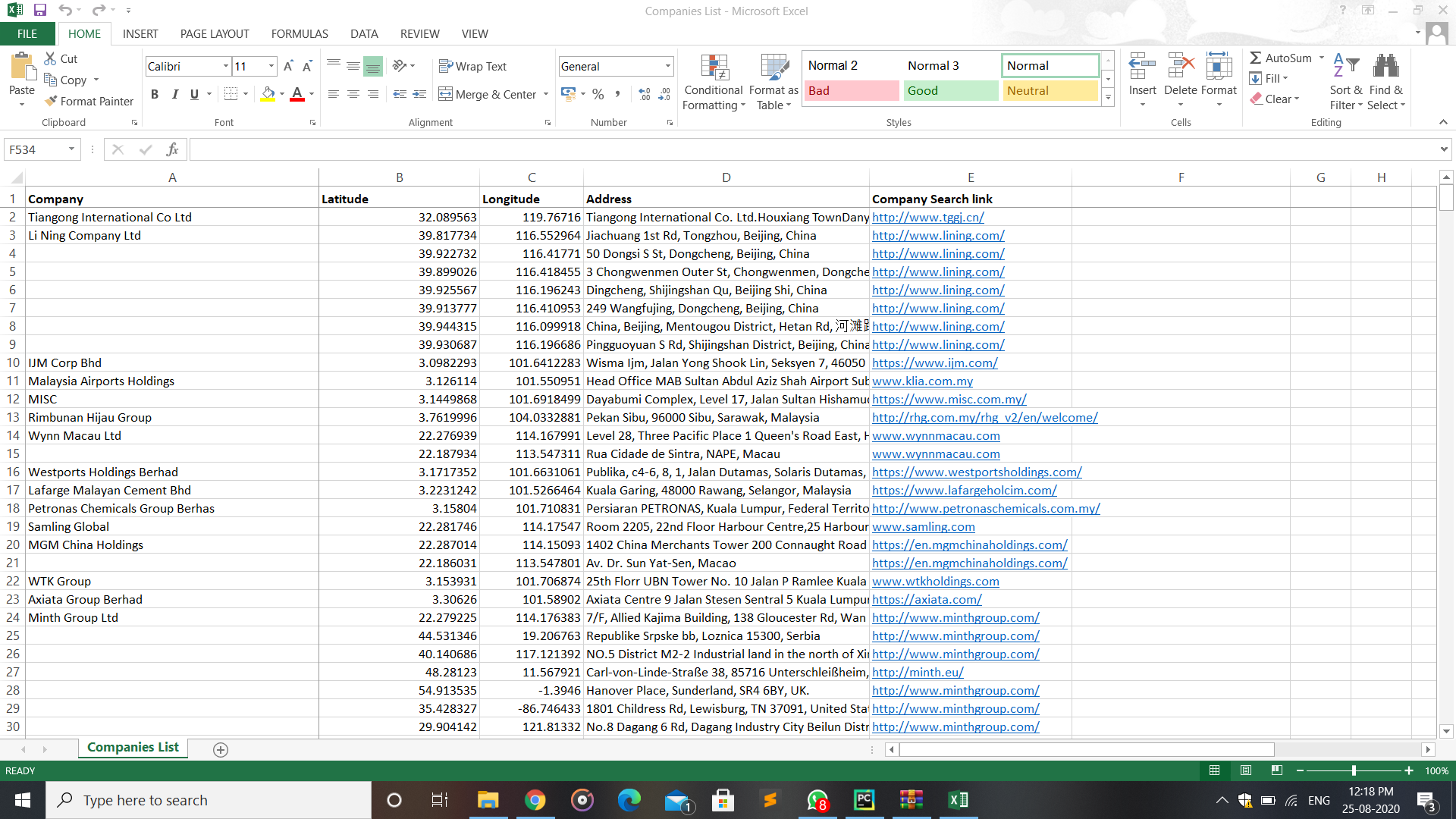
[**https://craft.co/**](https://craft.co/)



Later we converted the excel to a csv file and further into an ESRI Shapefile to work with it.



This is the code used for conversion of ‘CSV’ to ‘Shapefile’.



Here is the ‘CSV file’ (which contains coordinates, addresses, company website link)

**8.ABOUT INTENSEL**

As global weather makes the extreme risk more common, a new industry of climate risk prediction is born on the last 5 years. Using latest technology few startups are bringing unpriced risk into the investment world.They use cutting Edge AI power combined with numerical weather prediction for granular and precise weather forecast which helps their energy, reinsurance and banking clients assess risks and opportunities related to climate change.Their unique AI-NWP approach fills the missing historical, regional data gaps,thus enabling short, medium, and long-term accurate forecasts based on terabytes of processed data.

**WORKING PROCESS:**

**1.Data Modeling & Collection**

They run weather models plus collect datasets from satellite and observation stations thus a robust data assimilation process.

**2. Reinforcing with AI Models**

The datasets are then used to train different AI models to improve dynamicity, forecasts and automated models.

**3.SaaS**

They provide a ready-to-use SaaS that leverage the power of big data and fast AI simulations online.

**SERVICES THEY PROVIDE:**

* **Flooding:**Uses high-resolution data and AI-enhanced DEM to create flooding maps resulting from rainfall, taking into account both pluvial and fluvial aspects
* **Wind/Solar/Cloud Cover:**NWPs along with AI models are used to create short/medium/long term forecasts at the wind and solar power sites
* **Sea Level Rise:**Modelled and satellite-observed sea-level rise corresponding to different climate forcing and return periods
* **Hurricane:**AI trained and modelled forecasts thet are updated post every event, giving ensemble forecast based on several factors and climate forcing
* **Drought:**Intensel Drought Index (IDI) uses NWP output parameters such as ground-water depletion, snowmelt, rainfall index trained with AI and reinforced with satellite data
* **CO2 Emissions:**Uses alternative high-resolution satellite data coupled with NWP-Chem and AI to output revenue per tons of CO2 emissions, making companies greener

**9.CONCLUSION**

This project has given us real-world experience and hands on how data scraping takes place and how usefull it is in various applications and areas in real world with Web-scraping, coding and also automation.

Using our basic knowledge on Python, Selenium, Webscraping and Geopandas we were able to plot the geo-locations of companies and map them all at once and create an automation process which fetches the market capital and revenue of the company and stores their values in the Excel format.

From this project we have learned a lot of new things which we have never used previously, like mapping geolocations, webscraping, automation process.

And mostly, we spent a good time in this intership and we had a chance to work with a GIS Analyst who was mentoring us and learned nice stuff from her.It was very interesting to work with the data and manipulate it according to our need. Finally it was really a wonderful experience.

**10. REFERENCES**

<https://matplotlib.org/>

<https://geopandas.org/mapping.html>

<https://www.ftuforums.com/complete-python-3-masterclass-journey-udemy-course-free-download/>

Youtube Lectures:

* Selenium

<https://www.youtube.com/playlist?list=PLUDwpEzHYYLvx6SuogA7Zhb_hZl3sln66>

* Python

<https://www.youtube.com/playlist?list=PLsyeobzWxl7poL9JTVyndKe62ieoN-MZ3>

Websites To get Company Information:

<https://www.corporateinformation.com/>

<https://www.getcompanydetails.com/>

<https://craft.co/>