Introduction:

Midst the pandemic, entering the work force as graduates can be difficult. The difficulty level increases further if one does not understand the demographics of the job market. Landing a job at a company that complies well with your work ethics and houses likeminded people can be achieved only via a good market study. This project aids you in getting a concise summary of the data science job market demographics, actively hiring companies and evolving needs of data scientists in different cities across Australia using data from SEEK.

All major job seeking/hunting platforms offer services that keep you updated with new jobs being added to their portal, but they often do not give you a summary of the current state of the market. At times getting a feeler of where the job market is heading could be beneficial in setting goals.

Background:

Being in the final semester of my data science degree brings me job anxiety. Like me, most of the students in their final year feel the same. Everyone is sceptical about landing a job at their desired company or even getting a decent job. I figured most of the times this was because I did not understand what the job trend looked like and lacked market understanding.

Henceforth I decided scrape data off SEEK to overcome these bottlenecks so that me or someone else will never have to struggle to get a feeler of the job market.

Problem Statement:

Before applying for jobs, it is always beneficial to do a good market research, but there are not many tools or services available to help you do so. For example, we see LinkedIn posts saying, “Sydney is aggressively hiring for data scientists”, but this does not give us proper information. Even if the stated information in posts is correct, we are unaware of considered timeline adding up to the information. The objective of this project is to avoid these problems and get a short and sweet timely update.

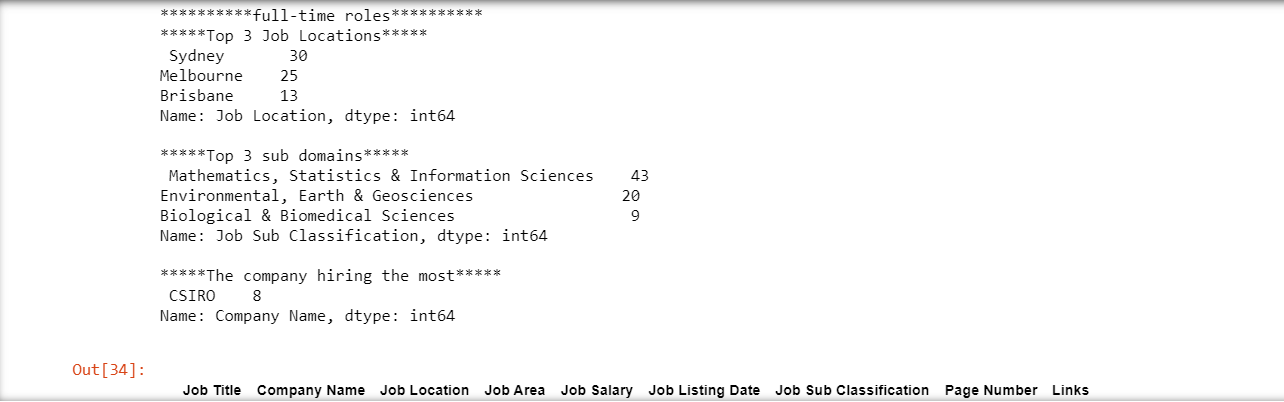
The Project:

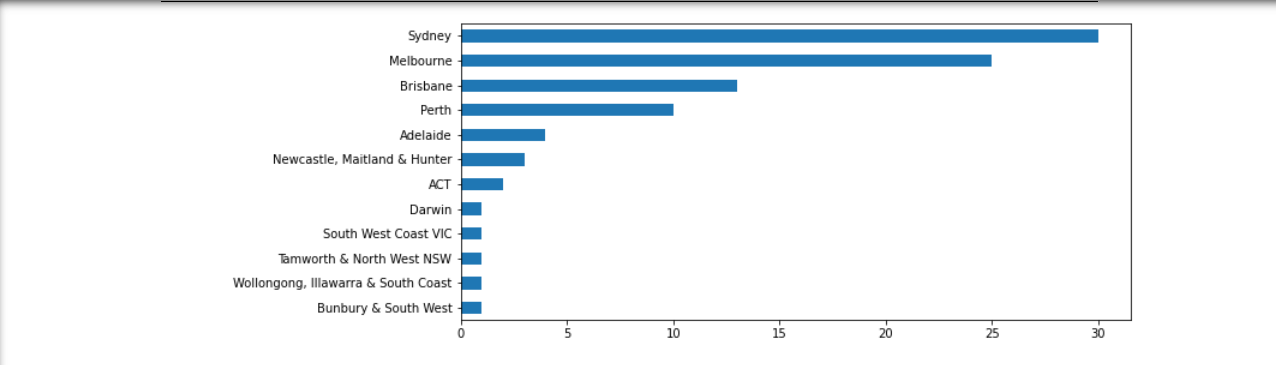
At the core, the project is based on BeautifulSoup for scrapping the data off SEEK for data scientist roles. The program iterates through all jobs across different pages and stores their details in an csv file for future reference. Since SEEK has unrestricted data access measures in place, the code is not subject to change and can be reused whenever needed.

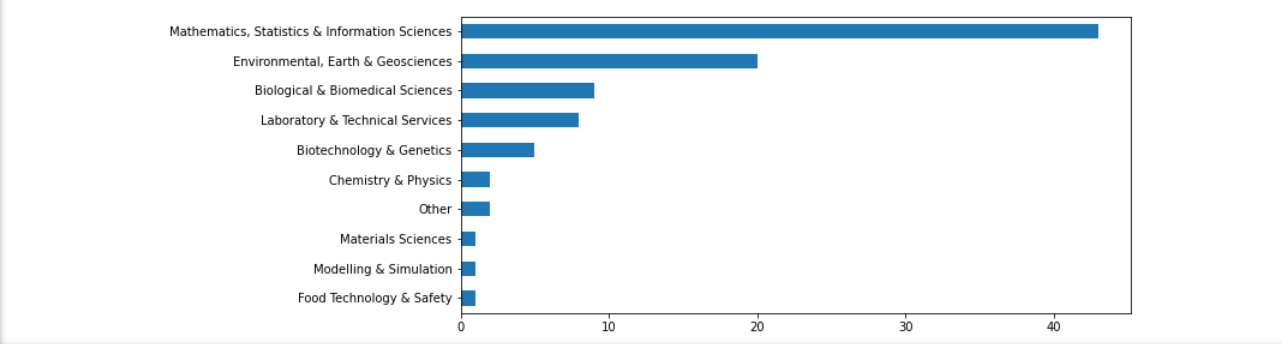
After the data collection, python libraries like pandas and matplotlib are used to clean, format, and analyse the data.

The current version of the program scrapes full-time, part-time and contract/temporary jobs for data scientist roles in the science and technology domain in Australia and is triggered manually when necessary.

Below given is the sample output for full-time data science job roles in the science and technology centre.







Graphical user interface

Description automatically generated with low confidence

From the above output we have found the following:

* Sydney needs the most data scientist in comparison to other cities of Australia.
* In the science and technology field of data science, the most popular subfield is Mathematics, Statistics, and Information Sciences.
* Australia wide, CISRO is hiring maximum number of data scientists.
* The output for recent jobs is empty, as on the current time, the most recent job was posted 2 days ago.

One could drill down further into the data to answer questions like, “Which sub domain is popular in Sydney?”

The Users:

This project can be used by students willing to understand the demand of their profession across different cities of Australia and get a list of recently added jobs. Indirectly it could be used by analysts as an additional tool in their tool kit to get a brief summary of the ever-changing job market.

Pre-Condition:

For best results,

* run this program in Anaconda environment on Jupyter Notebook with python version 3+
* make a new virtual environment and install requirements.txt.

Post-Conditions:

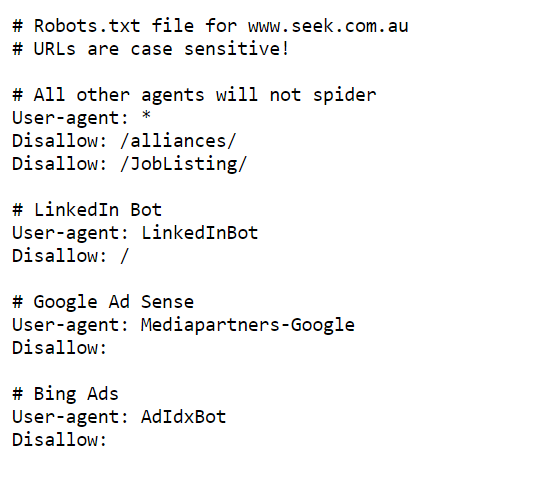
Had there been no problems in the execution, an output on the Jupyter notebook should be visible and an csv file saved in your current working directory.

Main Flow:

A picture containing diagram

Description automatically generated

Ethical Issues:



The above image represents robots.txt file for SEEK. From the image it can be seen that, user (scrappers like me) are allowed to scrape anything from SEEK except for the following URLS [www.seek.com.au/JobListing/](http://www.seek.com.au/JobListing/) and [www.seek.com.au/alliances](http://www.seek.com.au/alliances) and whereas LinkedIn Bot, Google Ad Sense and Bing Ads are allowed to scrape anything on the website.

However other websites have additional attributes present in the robots.txt file to limit a crawler’s behaviour like,

* crawl late limiting (e.g. Crawl-delay: 11) - This is used to limit crawlers from hitting the site too frequently. As frequent hits by crawlers could place unwanted stress on the server and make the site slow for human visitors, many sites add this line in their robots file. In this case, the site can be crawled with a delay of 11 seconds.
* Visit time (e.g. *Visit-time: 0400-0845*) - This tells the crawlers about hours when crawling is allowed. In this example, the site can be crawled between 04:00 and 08:45 UTC. Sites do this to avoid load from bots during their peak hours.
* Request rate (e.g. *Request-rate: 1/10) -* Some websites do not entertain bots trying to fetch multiple pages simultaneously. The request rate is used to limit this behaviour. 1/10 as the value means the site allows crawlers to request one page every 10 seconds.

This project abides by the **Python Web Scrapping Code of Conduct** and DOES NOT,

* Download copies of document that are clearly not public.
* Share downloaded content illegally.
* Share what you cannot.
* Overload the server.
* Scrape the same data repeatedly.
* Violate rules mentioned in the robots.txt file.