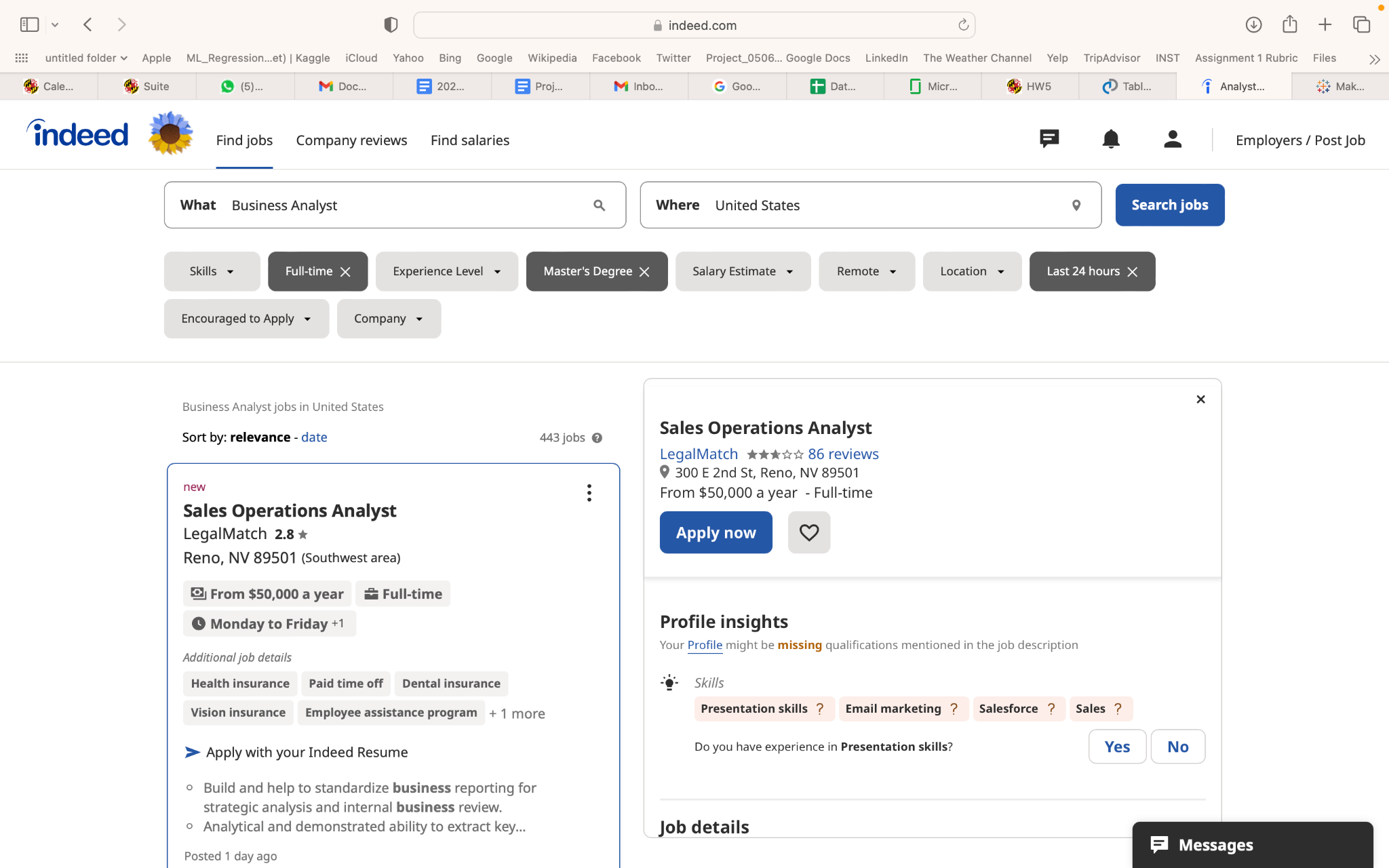
**Readme File**

1. Data Collection:

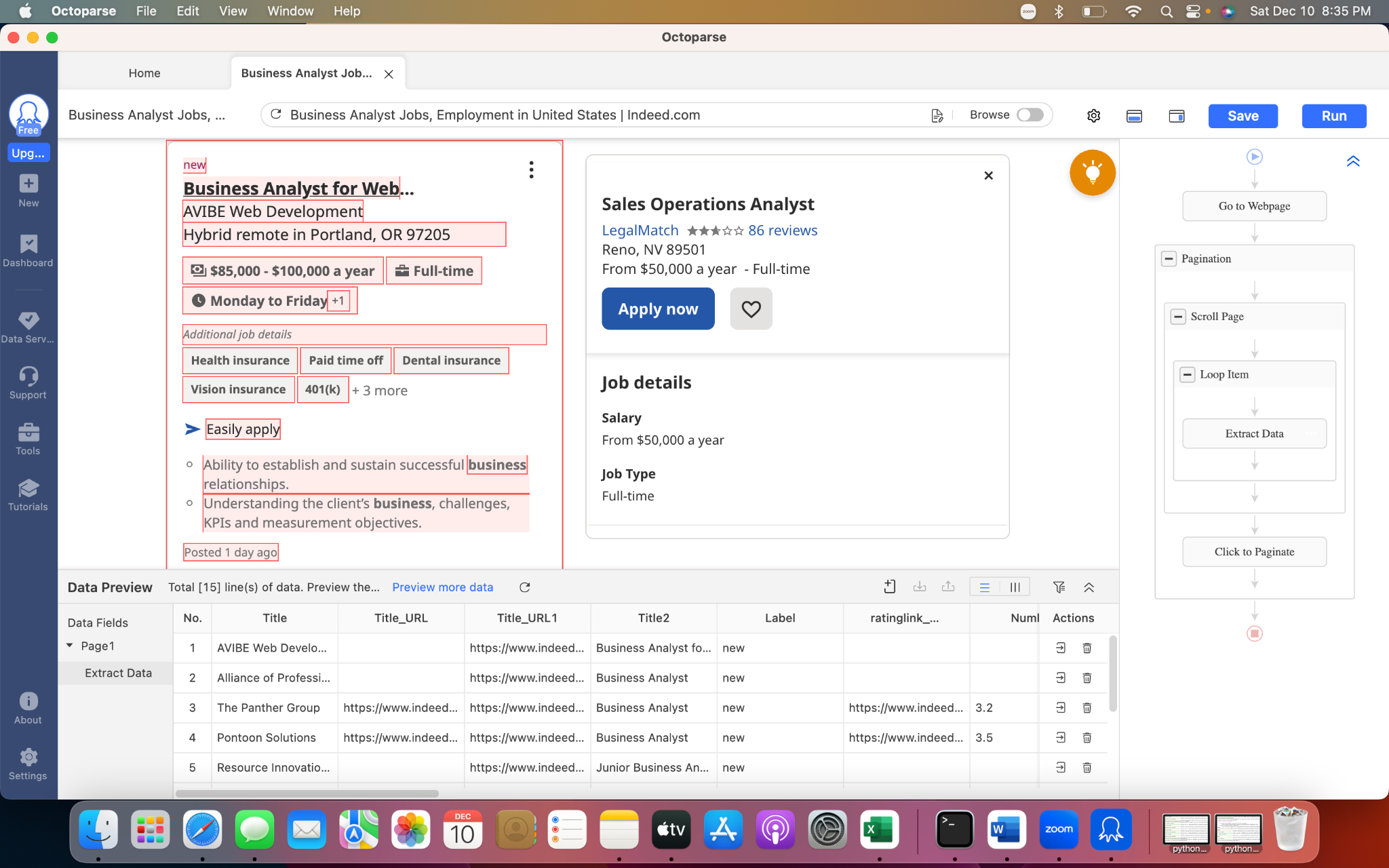
Collected most of the accessible data required for our project, from the following two websites - Indeed & Glassdoor

* Data was filtered for all the full time, Masters level jobs posted 1 day ago across Job Functions of Business Analyst, Data Analyst, Data Engineer & Data Scientist from Indeed

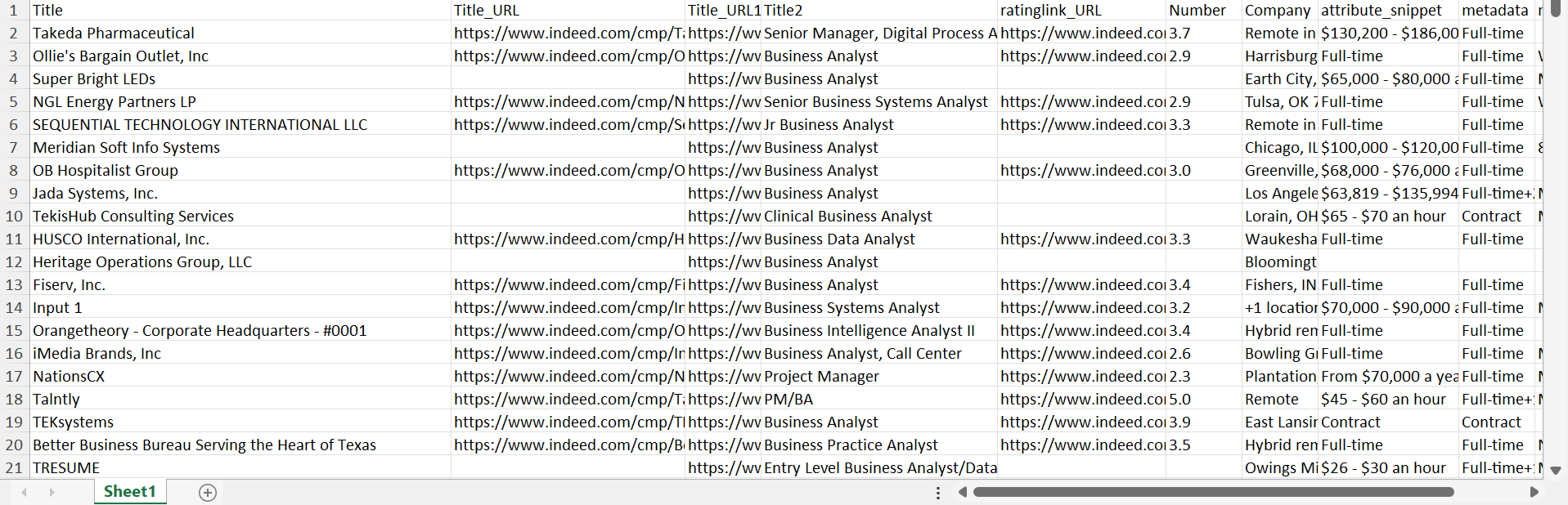


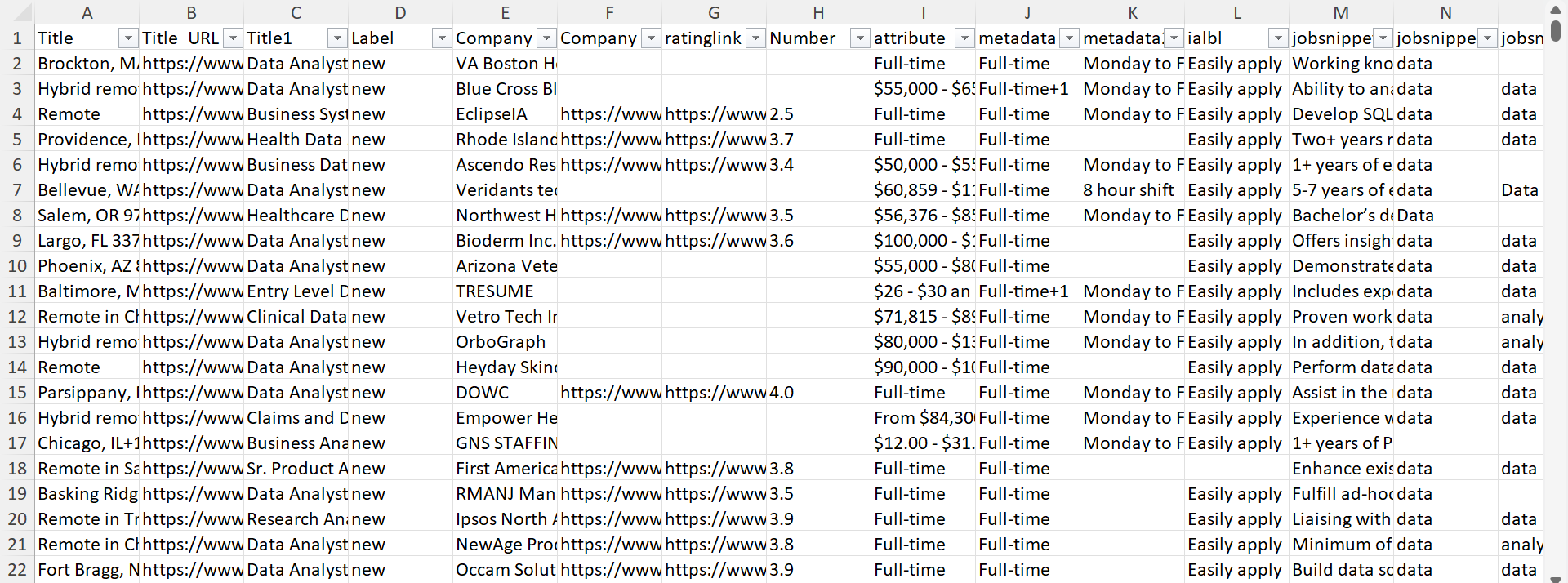
* Inserted the website in Octoparse,APIfy and created a workflow to extract information required to build our database such that all the required data (highlighted in red in below screenshot) is collected in the form of columns and automatically clicks load more jobs button, when it comes to the page end and re-collects the information again till it exhausts all the job listings on the website.



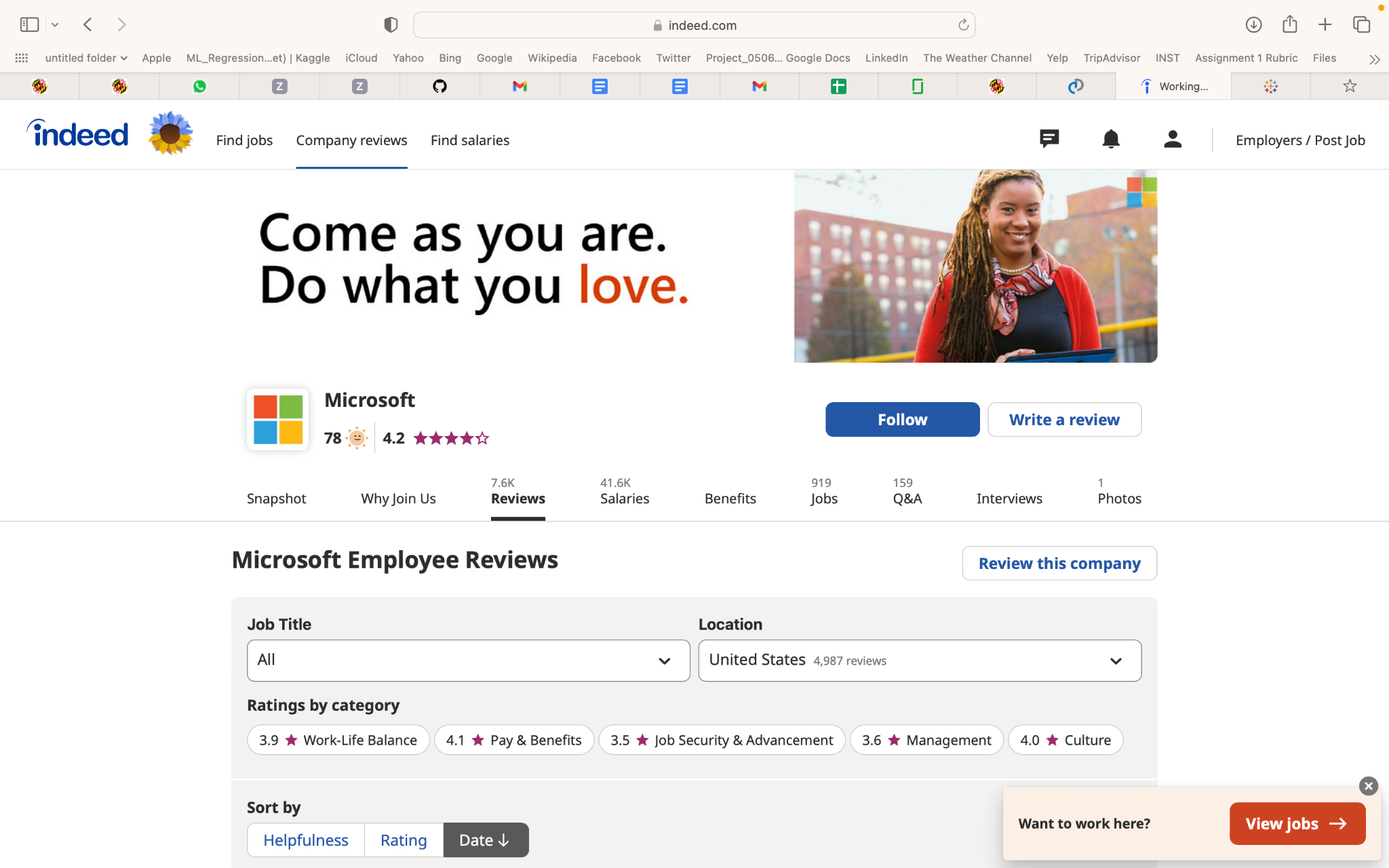


* The web scraped data for each of the job functions - business analyst, data analyst, data scientist and data engineer was extracted to csv files

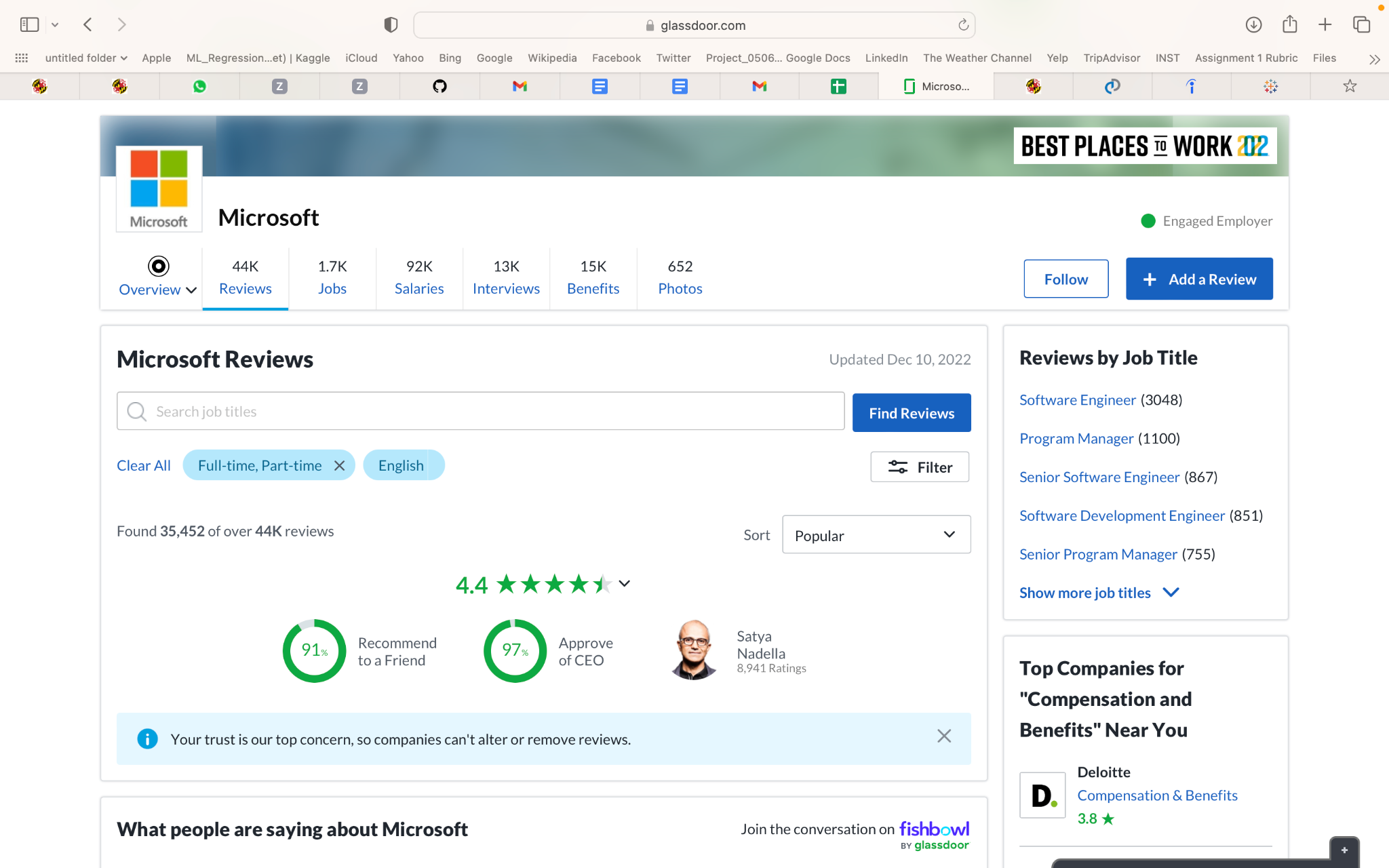




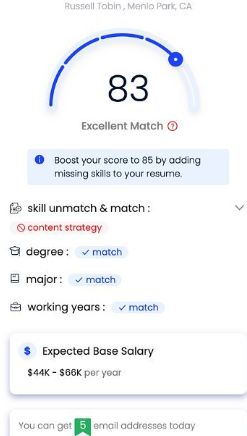
* Collected Firm Ratings for companies from the web scraped data for Rating table in our database from both Indeed and Glassdoor
* Indeed



* Glassdoor



* Collected Resumes and Information from few University of Maryland, CP students and matched their profiles with JobSeer tool to the extracted Job data.



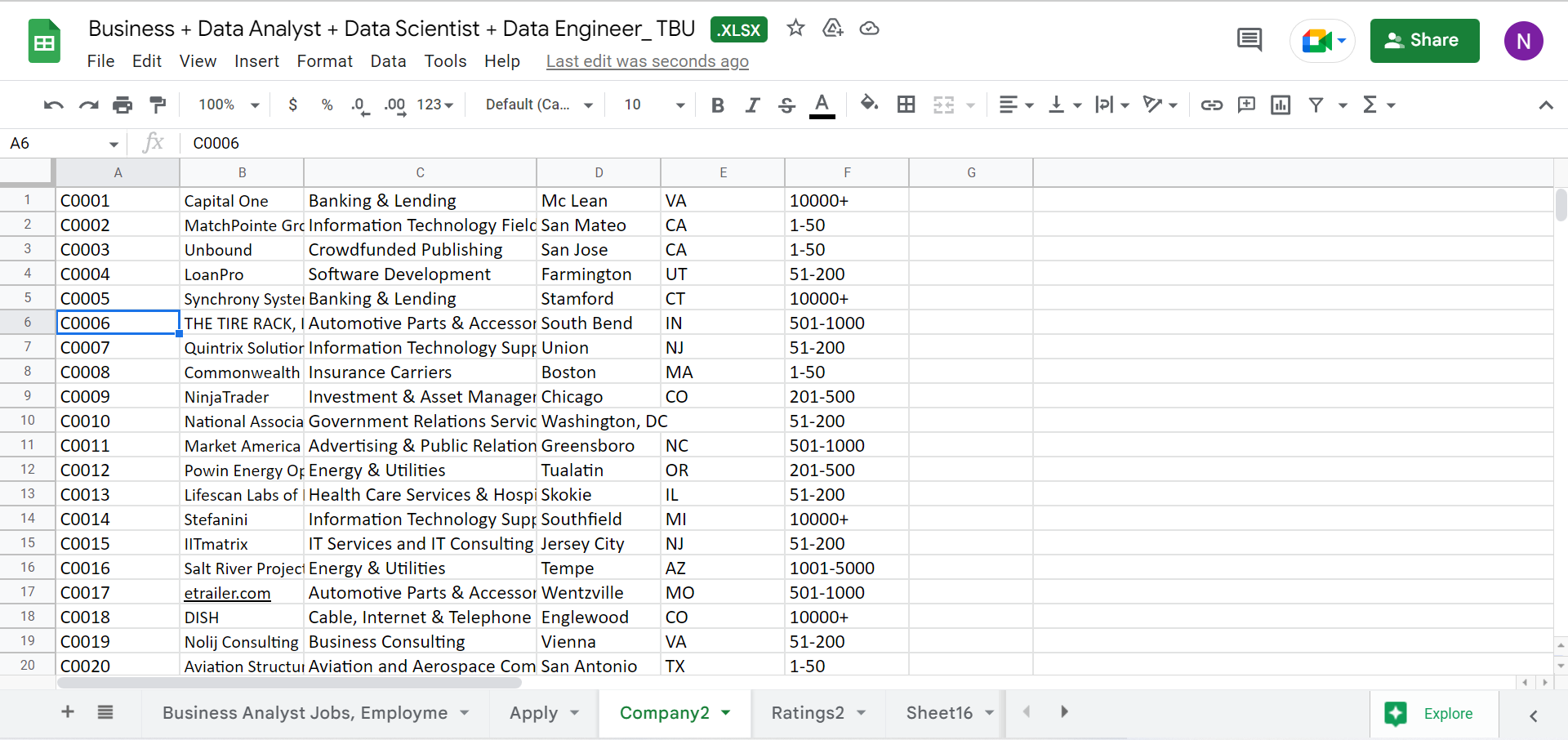
1. Data Pre-processing:

* The raw data extracted in csv format was further processed to remove duplicate values, null values and redundant information using **Remove Duplicates** in Microsoft Excel.
* The job Location was mapped to either Remote or 2 letter State Code.
* The job Salary was converted to per annum from monthly or hourly formats and averages were taken in cases of salary range in order to make the data uniform for analysis purposes.
* Assigned unique Identifiers for each of Job, company, JobSeeker, Rating, ProfileMatch tables etc.
* The data, in its raw form was converted to a more readable format, that could be then entered into database tables. The new finalized data has headings, no null values and no duplicate values.

Post cleaning of the data, the below is the finalized set of data stored in excel.

**Example:**

Company:



1. REFERENCES

* <https://www.indeed.com/>
* [https://www.glassdoor.com/](https://www.glassdoor.com/member/home/index.htm)

1. TEST METHODS

To test our database, and to be able to view and extract required information, we have used sql queries and tableau visualizations. Below are the associated results.

Using the dashboard created , users can see the data associated with each job function for different business transactions. One such example is shown below