**Project 2**

***Ideas for the project***

**Job Aggregator**

Job aggregation can be your next big scraping project. There are people actively looking for jobs and there are companies looking to hire suitable manpower. The problem is there are a ton of job boards with a lot of listings. What if you can scrape the job links and title, put it in a single place from where job seeker can get the details.

A job aggregator will help all 3 parties: Job board, Company and Job seeker. However when crawling, don’t crawl all the data and bypass the job board. I would suggest scraping just the title and link would be enough. Job seeker can be redirected to the original site where the job was posted.

There are many job portals online, and you can pick anyone for this project. Here are some places to get you started:

* *Naukri.com*
* *Indeed.co.in*
* *Timesjobs.com*

**How to work on this project**

In this project, you can build a tool that scrapes a job portal (or multiple job portals) and checks the requirements of a particular job. For example, you can look at all the ‘data analyst’ jobs present in a job portal and analyze its job requirements to see the most popular criteria for hiring one such professional.

You can add more jobs or portals in your search to add more difficulty to this project. It’s a fantastic project for anyone who wants to apply data science in management and relevant streams

**Get Financial Data**

The finance sector uses a lot of data. Financial data is useful in many ways as it helps investors analyze a company’s performance and reliability. Similarly, it helps a company in analyzing its position and where it stands in terms of finances. If you want to use your knowledge of data and web scraping in the finance sector, then you should work on this project.

**How to work on this project**

There are multiple ways to go about this project. You can start by scraping the web for the performance of a company’s stock in a set period and the news articles related to the company of that period. This data can help an investor figure out how different things affected that particular company’s stock price. Apart from that, this data will also help the investor understand what factors affect the company’s stock price, which factors don’t.

Financial statistics are crucial for any company’s health. They help the stakeholders of a company understand how well (or how badly) their business is performing. Financial data is always helpful, and this project will allow you to use your skills in this regard.

You can start with a single company initially and make the project more challenging by adding the data from more companies. However, if you want to focus on one particular company, you can increase the timeline and look at the data of a year or more.

**College data**

<https://www.kaggle.com/wsj/college-salaries?select=salaries-by-region.csv>

*“Where it Pays to Attend College”* This data was collected by the well-known Wall Street Journal as well as data collected through PayScale Inc.

Contains 3 CSV type files named and their analysis:

1. **degrees-that-pay-back.csv**

* The Wall Street Journal stated “A year-long survey of 1.2 million people with only a bachelor's degree conducted by PayScale Inc. shows that graduates in these subjects earned 103.5% and 97.8% more, respectively, about 10 years post-commencement. Majors that didn't show as much salary growth include Nursing and Information Technology.”

1. **salaries-by-college-type.csv**

* In this persons research they took schools based on ivy, party, and engineering status’s to see which type of schools saw the highest earning increase since receiving job positions newly graduated from college. In their description they state “You already know your starting salary will be different based on the type of school you attend. But increased earning power shows less disparity”. In which they mean the salary will more than likely be based on the type of school you go to, in contrast to believing the higher the degree the higher the salary.

1. **salaries-by-region.csv**

* The research conducted by PayScale Inc. found that attending college in the Midwest leads to a lower salary both at graduation and at mid-career compared to those graduated in the Northeast and California.

**Project 2**

***Project Overview and Objective***

**The idea we chose**

* We chose to work with “College Data” data collected by The Wall Street Journal and PayScale Inc.
* Use our second data source containing college rankings with tuition.
* Salary based school ranking and tuition

**Extracting our data**

* We will load our CSV files and read them using Pandas

**Transforming Our Data**

* We will transform our data into 2 new data frames that are cleaned by filtering out the columns we don’t need and use the columns we do using Pandas.

**Loading Our Data**

To load in our data we cleaned using pandas we will use PgAdmin and then we created tables using pgadmin and then merged the tables as well.

* + Postgres (relational)
  + School table:
    - school\_name, (string)
    - state (string(2))
    - avg\_salary (numeric)
    - ranking (integer)

**Requirements**

* Show only the ranked universities
* Show only the starting median salary
* Show only the school name

**Data Sources**

* <https://www.kaggle.com/wsj/college-salaries?select=salaries-by-region.csv>
* https://data.world/ian/united-states-university-rankings