Web Scraping Project Report

Shubham Joshi 437962

Rohit Mundhra 444464

1. Description of the Project:

Every Parents want their children to study in a school/university which has a good reputation and excellent education standard for their bright future. **Career360** is a data-enabled and technology-driven Educational Products and Services Company. It seamlessly integrates millions of student and institutional data points with the user-generated preferences of its more than **15 million+ monthly visitors,** to build sophisticated **Prediction and Recommendation products** for the students to explore and achieve career plans, based on their interests and abilities. Hence , we decided to scrape this website to pull the data and do some analysis.

2. Different approach used to scrape the website:

We used 3 methods to scrape the website as required i.e. Beautiful Soup, Scrapy & Selenium . These are common methods that are often used to scrap the data from website. After scraping data from these approaches , we realized that all the different approaches has its own advantages and disadvantages . In this case, we scrape 100 hundred pages using all the 3 different methods to get the same result .At the end of this project we learned to scrap data in 3 different ways. Let us describe each of them one by one.

2.1 Beautiful Soup

Beautiful Soup is a Python package for parsing HTML and XML documents. It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping. To access the website, it uses request library. In addition to clean and match the pattern in some string, it requires regular expression (RE) library to fetch raw HTML . We used Boolean parameter to set the page limit, Used lists to store the different values and dictionary to store the scraped data and Finally stored the data into csv.

2.2 Scrapy

Scrapy is a free and open-source web-crawling framework written in Python. Originally designed for web scraping, it can also be used to extract data using APIs or as a general-purpose web crawler. We extracted the same data that we extracted using beautiful soup and stored it in a dictionary and then sent it to a csv file. We defined a function that phrase the HTML content. It uses Spiders which are classes that you define and that Scrapy uses to scrape information from a website .

2.3 Selenium

Selenium is an open source umbrella project for a range of tools and libraries aimed at supporting browser automation. It provides a playback tool for authoring functional tests without the need to learn a test scripting language. First, We import the necessary packages, set the driver path, and define the starting page URL. Then we create an empty list for every information we are going to scrap. We create a loop that goes from the starting page to the 100th page. In the loop, we searched for the elements by XPATH and appended the results into our empty list. We scroll down to get where the next page link to-cates, and then click on the link. These processes will be repeated until the first 100 pages are scraped.

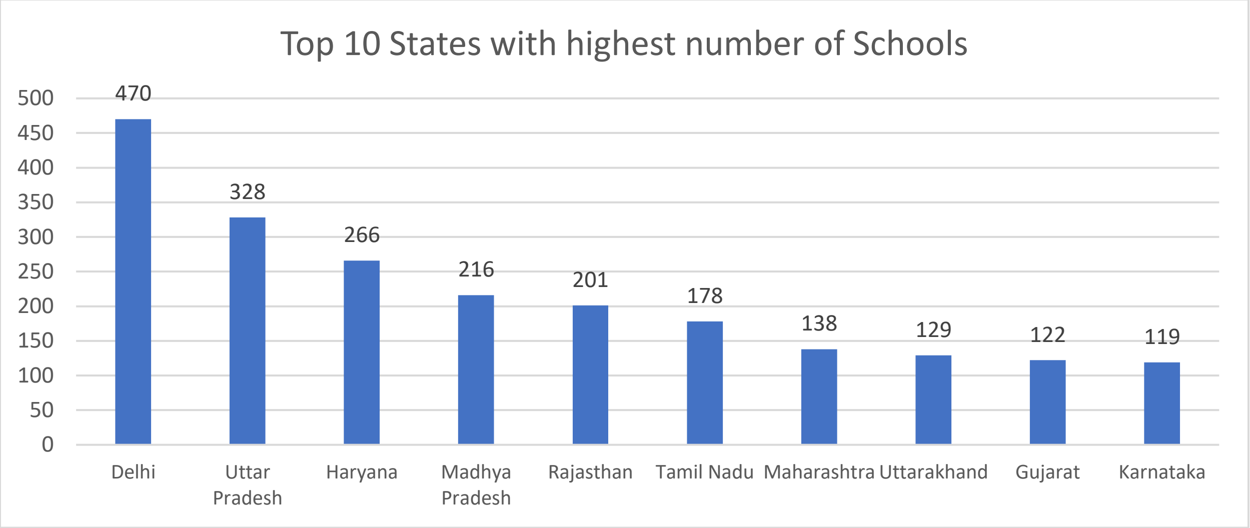
3. Analysis of the output

We append the data to the list every time it runs for the next link. Then save the data in a CSV file. In the table below, you can see the description of the output.

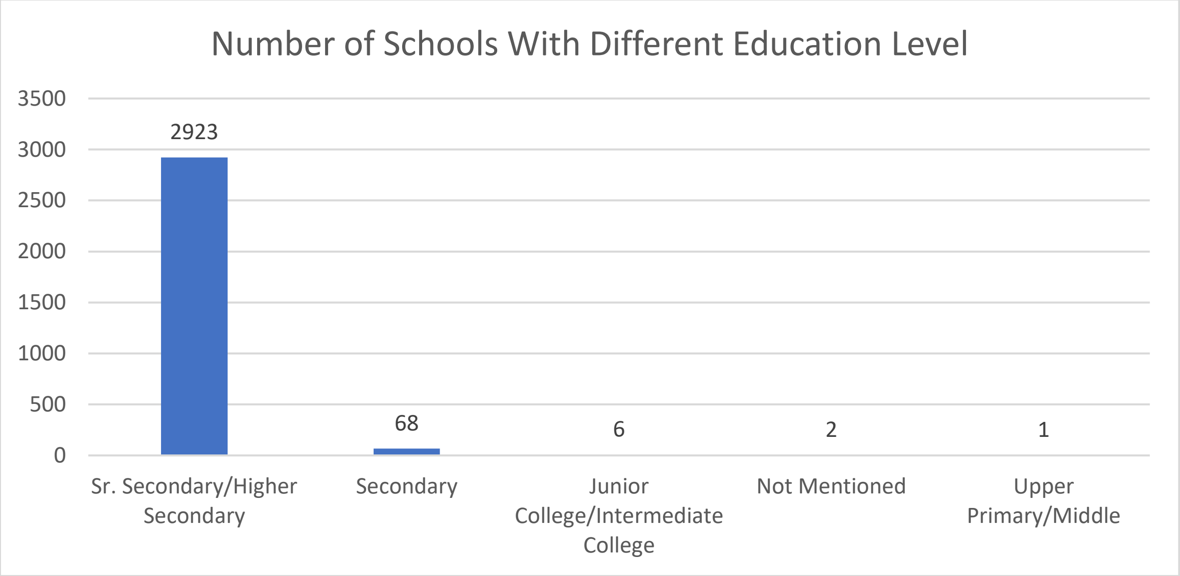
Output	Description
Name	It contains the name of the school
City	It contains the city that the school is located.
State	It contains the State where the city belongs.
Affiliation	An affiliated school is an educational institution that operates independently, but also has a formal collaborative agreement with another, usually larger institution that may have some level of control or influence over its academic policies, standards or programs.
Education Level	It contains the education level of the school.

3.1 Statistical analysis

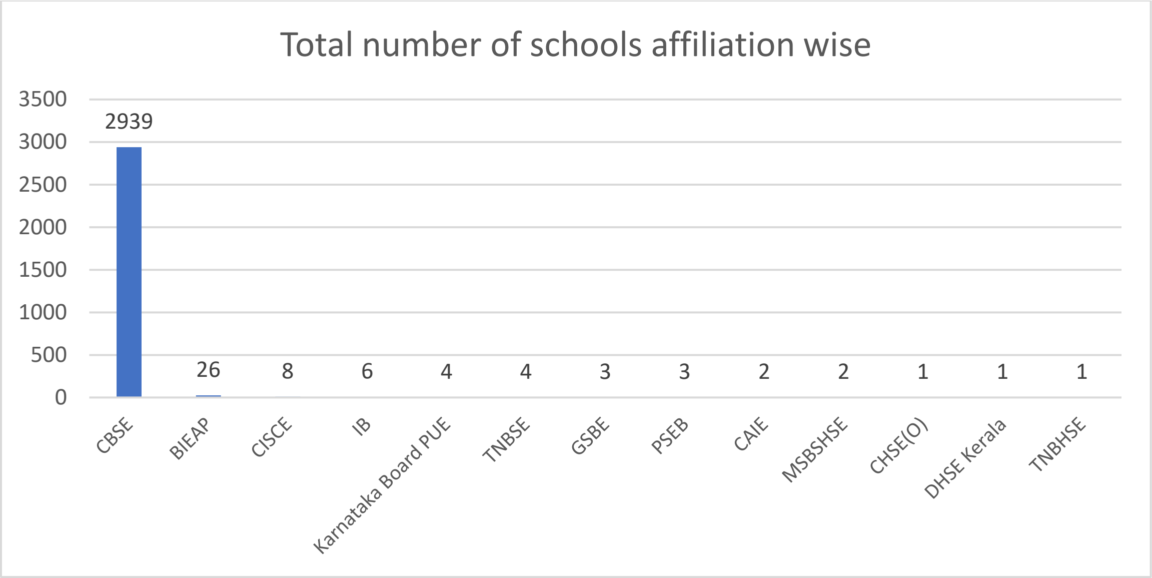
This website provides us information about the schools and their location and hence after pulling the data from the website we did data visualization using graphs and pivot tables which are displayed below.



This graph shows us the top 10 states out of 38 states with highest number of schools and it says that that Delhi has the most number of schools.



The above graph shows number of schools for different levels of education type and clearly higher secondary school are the most .



The above graph shows us different type of affiliation and it shows that CBSE board is most adopted in every state of the country.

4. How was the project divided between project partners

We initially did one method each and the third one which was little complicated i.e. scrapy , we decided to do it together so that the fundamentals could be clear for both of us.

5. Conclusion

We found this website very interesting because it gives us information about the different types of schools and their affiliation board in carious cities across different states. It is a very important decision making website for parents who wants to move to different cities so that their child can get better quality education.

For us , it helped us understand the different concepts of web scrapping and 3 main methods to scrape the website. It will help us in our future projects and our analytical career since it is a very important tool to extract data from any website and so the analysis.