

BAX 422 Final Project

Web Scrape Analyst Job Posting

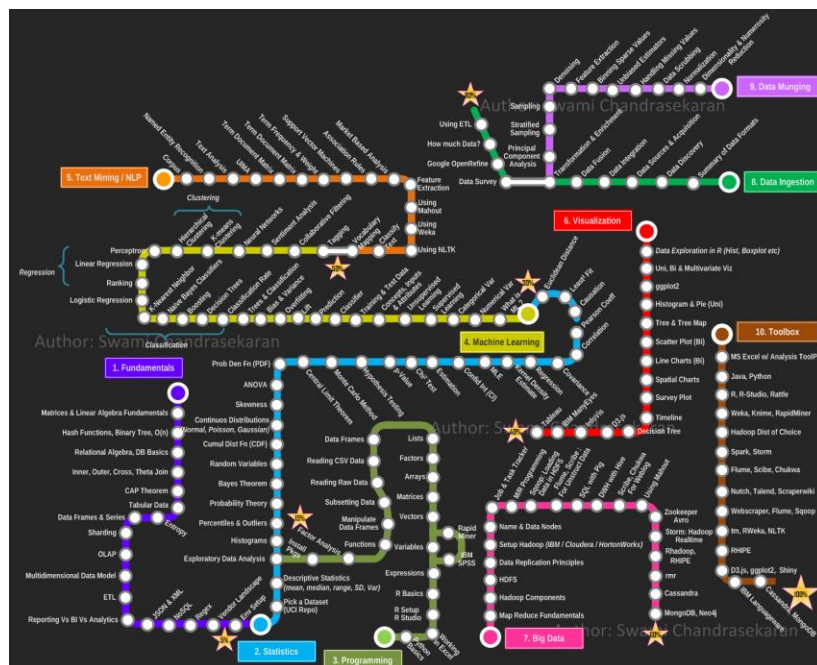
Executive Summary

Employment websites provide aggregated information on the job market, providing information on job listings. Scraping data from employment websites can provide insights into the market trends. Specifically, the data gathered from employment websites could be useful for MSBA programs, and graduating students.

In this project, we carefully assessed all available employment websites and finalized our web-scraping on Monster.com due to its advantages. We scraped data for Analyst related positions, which are: *'data analyst'*, *'business analyst'*, *'data scientist'*, *'marketing analyst'*, *'product analyst'*, and *'strategy analyst'* in the Bay Area, which we designed to center around Palo Alto in a 50 miles radius. We used BeautifulSoup to parse the information on the search query page as well as each individual job listings page. The queries returned a total of 458 results, which are stored in a MySQL database. We chose to store it this way as it maximizes flexibility to cross-reference and join with other primary and secondary data, such as the data gathered from the tableau dashboard produced by UC Davis MSBA alumni last year.

We have showcased how MSBA programs and graduating students could make use of the data to find out the top skills required for each job designation. Program staff and graduating students can pinpoint the most wanted skills, the former can seek to have experts from the industry to teach those skills and incorporate those into the curriculum; the latter could prioritize polishing those skills to better prepare for interviews or further customize their resume. The interested parties could perform further data analysis to gather insights on the job salary changes, top hiring companies, and how the jobs have shifted over time.

Have you pondered what are the required skills for your career?



The above is a well-known picture by Swami Chandrasekaran highlighting the roadmap for a data scientist.

You can see that it is comprehensively delineated, but it will probably take a lifetime to master all the skills in the roadmap. We remember looking at this picture and were daunted by the number of things data scientists need to learn. This brings us to a question: do we need to understand everything in the roadmap, or was it just superfluous? Eventually, we decided that a more efficient way to finding the skills needed to kickstart our career is to discover what are the hottest skills that are in need by the market.

That is a pertinent question waiting to be answered for MSBA Programs, particularly our UC Davis MSBA Program. This is because the UC Davis MSBA program is heavily career-oriented, for one of its main competitive advantages lies in its program location in the Bay Area. To capitalize on its strength, the program needs to continuously cater and update our curriculum to the latest market trend. Hence, the admission team needs to understand which are the hiring companies, the top skills that are listed in the

job postings as well as salary range. So that they could answer the important questions, such as how the jobs have shifted, and where the shifts are happening, when comparing data from the past years. The findings could enhance the program offering and the student experiences. For instance, the career teams could better advise the students on which crucial skills to highlight, organize career events by inviting the top hiring companies to share their hiring tips, and fine-tune the curriculum over the following years to cater to the industry needs.

Aside from its business value, graduating students from the various MSBA programs are stepping into our graduating term. They would be keen to find out the insights from this research to kickstart and optimize their job hunt process. For example, the student may wish to rewrite its resume to feature more keywords that lead to the greatest probability of ending up on the requirements list of a job posting and be picked up by the AI resume screen.

Likewise, it is also valuable to us personally as we approach our final quarter with the MSBA program. We want to familiarize ourselves with the market outlook and increase our probability of finding a good job.

Hence, we turn our eyes to employment websites. We would like to extract the hiring company name, position name, tally key skills, and salary from the job postings. To extract the aforementioned information efficiently, we need to web scrape the information from the website.

Data Sources, Web-Scraping Routine, and Database Design

Introduction of the Data Sources:

For the employment website, we chose Monster.com operated by Monster Worldwide, Inc.. Before choosing Monster.com, we explored other more well-known options such as LinkedIn, Indeed.com, and Glassdoor. However, we finalized on Monster.com due to three main reasons to ensure that this is a reliable tool that the admission teams could use.

Firstly, some websites such as Glassdoor do not allow the scraping of jobs or reviews, but only company information. Additionally, if we went aggressive about getting requests, Glassdoor would start rejecting or throttling our connections, or may even block our IP. The admission teams would definitely want to avoid this from happening.

Secondly, Monster.com uses a relatively consistent HTML tagging syntax, without frequent changes to the HTML structure. This enabled a consistent scraping process without the use of web browser automation, such as Selenium WebDriver. As an illustration, LinkedIn uses dynamic JavaScript content and actively redirects users, which makes it difficult for web scrapers to navigate and obtain information. It also uses Infinite scrolling to load content continuously as the user scrolls down the page, which makes it difficult to crawl.

Lastly, we avoided using selenium as we have run into Captcha issues on the other major employment websites such as indeed.com. Captcha challenge-response test used in computing to determine whether the user is human¹, we cannot bypass them using selenium easily.

Monster.com does not have those issues, therefore it became our platform of choice.

Additionally, we also searched online for the past job listing data for comparison. We found a [tableau dashboard](#) produced by our UC Davis MSBA alumni, Vardhini. She showcased the job trends, top skills overview, and job posting trends. We used it to compare the shift in job skills, and could be used more extensively for further studies.

Description of Web-Scraping Routines

Tools used and set up:

¹ <https://www.pluralsight.com/guides/advanced-web-scraping-tactics-python-playbook>

We used Python to scrape the website, as it is a good general-purpose programming language that has good web scraping packages such as BeautifulSoup. We used BeautifulSoup as the library for parsing HTML and XML contents, and the *'requests'* package to handle and make HTTP requests.

We then fetch the web page and store it in a BeautifulSoup object. To prevent websites from blocking the requests, we added a header when requests are sent to appear as humanly as possible. We also used the default HTML parser to parse the HTML on the web page.

We have generated a list of relevant job titles that MSBA students could apply for upon graduation. The job titles contain *'data analyst'*, *'business analyst'*, *'data scientist'*, *'marketing analyst'*, *'product analyst'*, and *'strategy analyst'*. We break our search into different job titles and enclosed the search terms with quotation marks to force the search to look for the exact order of search phrases. This is to ensure the accuracy of job comparison.

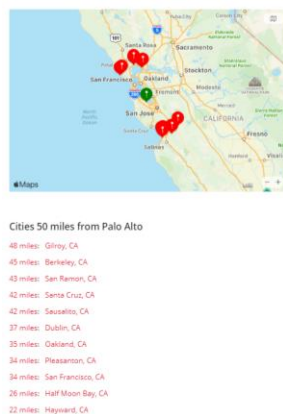


Figure 1 Cities 50 miles from Palo Alto²

We then established the criteria that are used in our search queries. We finalized on looking only at full-time positions with a search radius of 50 miles from Palo Alto. We chose Palo Alto because it is approximately the center of the bay area.

² <https://withinhours.com/50-miles-of-palo-alto-ca>

The URL format to query is quite simple to interpret as it uses a GET Method. To obtain the first page of the result with our criteria for data analyst:

'https://www.monster.com/jobs/search/Full-Time_8?q=__22Data-Analyst__22&where=Palo-Alto__2c-CA&rad=50.

However, to load more results onto the page, we need to click on the 'Load more jobs' button as shown.

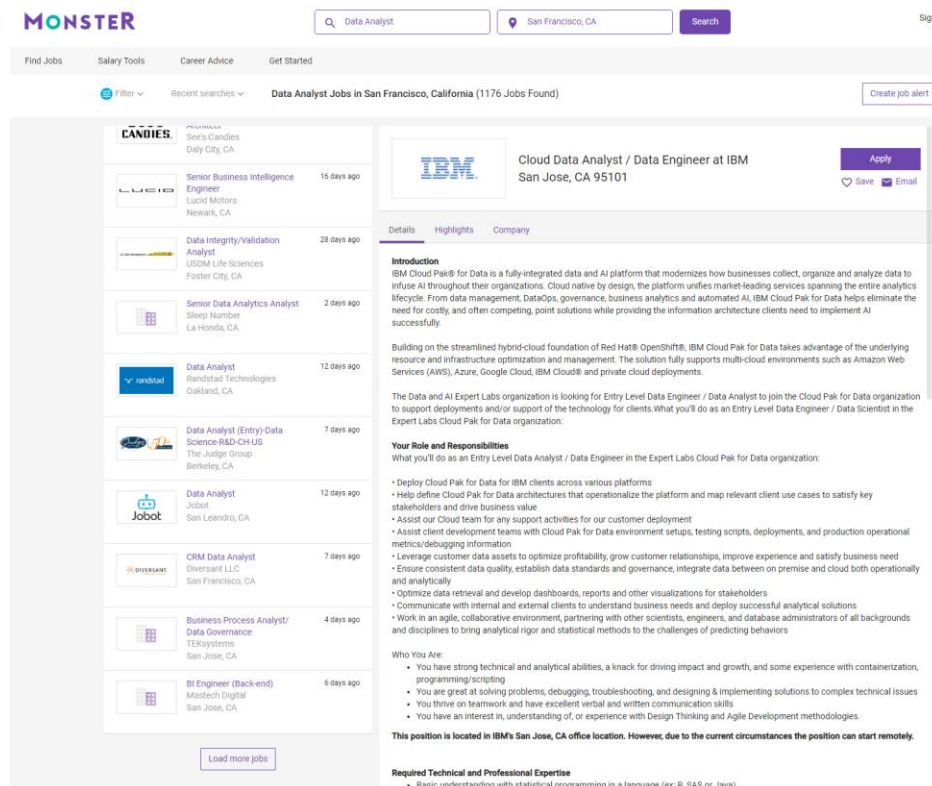


Figure 2 Monster.com

The second page's URL looks as such: 'https://www.monster.com/jobs/search/Full-Time_8?q=__22Data-Analyst__22&where=Palo-Alto__2c-CA&rad=50&stpage=1&page=2', where we need to add the start page and end page to the URL. Considering some job titles may only return a few results, and each page returns around 25 job postings, we choose to scrape 4 pages for each job title to keep the number

consistent for each position type in this project. As such, we can obtain around 100 job postings for each job title.

Page Inspection:

Upon inspecting the source page, we found that the *'data-job-id'* attribute in the *'section'* tag to be very useful, as it is a unique identifier that corresponds to the job listing IDs. By locating this attribute, we can identify each job listing and obtain a snippet of it. We can obtain the job title, company, and location of every job posting from the snippet of the job listing. We store the data in a data frame for further use. From the identifier, we can also obtain the URL to access the static listing page. The ability to access the static listing page is very important for two reasons. Firstly, using static webpages allow us to avoid the use of the selenium web driver. Secondly, it allows us to obtain more in-depth information on the job listing.

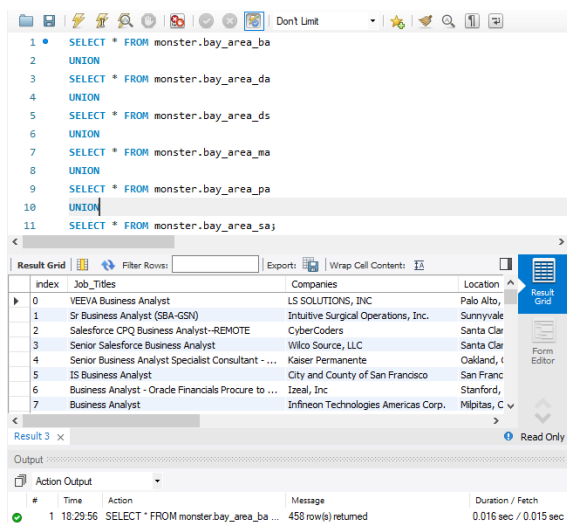
Using the URLs stored in the *new_urls*, we can open the individual static job listing description pages. We identified the attribute *'name'* with a value of *'value description'* to retrieve each job description. In querying for the job description, we have taken into consideration of potential problems, such as job listings used a preformatted format or job listings that are no longer available but still exist on the server. We remedied the problem using if conditions.

[Explanation of the Dataset/Database Design Choices](#)

We repeated the process for each job title and appended all data to separate datasets. The variables we included in each dataset includes the 'Job Title', 'Company Name', 'Location', 'Job Role', 'Job Description', 'Salary', 'URL'. We believe these are fitting information to extract from the data.

We have chosen to store the data in MySQL data after processing each dataset for each job title, and we chose to save the data in a tabular format as the columns in each dataset are consistent with proper structure.

We use an RDBMS database that is MySQL instead of NoSQL, as we want to keep the ability to use joins to cross-reference the different tables. An equally important consideration is to retain the robustness to cross-reference other locations or datasets for further studies, which is more essential than the flexibility offered by the NoSQL database. This shows data returning a total of 458 rows.



The screenshot shows a SQL query editor with a query window and a results grid. The query is a UNION of SELECT statements from various tables in the 'monster' database. The results grid displays a table with 4 columns: index, Job_Titles, Companies, and Location. The first 7 rows are visible, showing job titles like 'VEEVA Business Analyst' and 'Sr Business Analyst (SBA-GSN)' along with their respective companies and locations.

```

1 SELECT * FROM monster.bay_area_ba
2 UNION
3 SELECT * FROM monster.bay_area_da
4 UNION
5 SELECT * FROM monster.bay_area_ds
6 UNION
7 SELECT * FROM monster.bay_area_ma
8 UNION
9 SELECT * FROM monster.bay_area_pa
10 UNION
11 SELECT * FROM monster.bay_area_sa;

```

index	Job_Titles	Companies	Location
0	VEEVA Business Analyst	LS SOLUTIONS, INC	Palo Alto,
1	Sr Business Analyst (SBA-GSN)	Intuitive Surgical Operations, Inc.	Sunnyvale
2	Salesforce CPQ Business Analyst-REMOTE	CyberCoders	Santa Clara
3	Senior Salesforce Business Analyst	Wilco Source, LLC	Santa Clara
4	Senior Business Analyst Specialist Consultant - ...	Kaiser Permanente	Oakland, CA
5	IS Business Analyst	City and County of San Francisco	San Francisco
6	Business Analyst - Oracle Financials Procure to ...	Izeal, Inc	Stanford,
7	Business Analyst	Infineon Technologies Americas Corp.	Milpitas, CA

Figure 3 Full dataset

Discussions

Using the data, MSBA program staff and graduating students can get a better grasp of the job market. To illustrate what the data can do, we conducted a comparison of the top skills required for the most popular 3 job titles - Business Analyst vs. Data Analyst vs. Data Scientists, each with approximately 100 job listings. We cleaned the texts from the job descriptions and tokenized them. Then we removed the typical stopwords, Unicode, lowered all cases, and formatted the results. Next, we came up with a list of skills, that we think is important to any of the three jobs, and scan each word to see if it tallies. Using a counter, we count the number of times that these words appear in our data, and tabulated it as shown:

	Keyword	Frequency		Keyword	Frequency		Keyword	Frequency
0	Analysis	115	0	Analysis	203	0	Analysis	151
1	Communication	87	1	SQL	96	1	Python	118
2	Excel	43	2	Communication	68	2	SQL	78
3	SQL	42	3	Tableau	65	3	Statistics	72
4	Tableau	27	4	Visualization	57	4	R	71
5	Cloud	25	5	Excel	51	5	Cloud	69
6	Oracle	24	6	Python	48	6	ML	41
7	Python	10	7	Statistics	48	7	Communication	41
8	Statistics	8	8	R	38	8	Spark	36
9	R	8	9	Cloud	35	9	Hadoop	23
10	Visualization	6	10	SAS	24	10	Scripting	21
11	Java	4	11	Hadoop	12	11	Java	19
12	Scripting	4	12	Scripting	11	12	AWS	19
13	MySQL	3	13	Storytelling	9	13	Scala	18
14	JavaScript	2	14	Oracle	9	14	Tableau	15
15	ML	1	15	MySQL	6	15	Visualization	14
16	SAS	1	16	AWS	6	16	Wrangling	12
			17	NoSQL	5	17	NoSQL	11
			18	Looker	5	18	Prediction	10
			19	PostgreSQL	4	19	Looker	5
			20	Spark	4	20	SAS	5
			21	Java	2	21	Excel	3
			22	Scala	2	22	PostgreSQL	3
			23	Prediction	2	23	JavaScript	2
			24	JavaScript	2	24	MySQL	2
			25	MongoDB	1	25	Oracle	2
						26	MongoDB	1

Figure 4 Top Skills for Business Analyst in Bay Area (Left), Figure 5 Top Skills for Data Analyst in Bay Area (Middle), Figure 6 Top Skills for Data Scientists in Bay Area (Right)

We can see that all three job listings have ‘Analysis’ as the most frequently appeared word. This commonality is expected as we expect all jobs to perform analysis, so it also serves as a sanity check for us. Interestingly, while they are all the top word, their frequency largely varies. ‘Analysis’ appeared almost twice as frequent for Data Analysts than Business Analysts, and quite a bit more than Data Scientist. This unravels the differences in job nature.

Indeed, Business Analysts have much lesser skills mentioned in the job listings compared to the other two job titles. This suggests that it requires a different set of skills compared to the other two, with a potential focus on business domains. While Data Analysts and Data Scientists have almost equally long lists, the

skills utilized in the respective jobs are different. Data Analysts require more 'softer skills' such as '*communication*' and '*visualization*' than Data Scientists, while the latter focuses on tools such as '*Python*', '*R*', and '*Cloud*', as well as 'harder skills' like '*Statistics*' and '*ML*'. Machine Learning ('*ML*') is not mentioned for Business Analysts and Data Analysts.

Likewise, the job differences naturally unveil in the main tools required for each job. We can see that the main tool for Business Analysts is a combination of '*Excel*' and '*SQL*', '*SQL*' for Data Analysts, and '*Python*' for the Data Scientists. Additionally, we can observe that the usage of '*Python*' and '*R*' (to a lesser extent) increases exponentially from Business Analysts to Data Scientists. Excel is almost not mentioned for Data Scientists. Communication is least mentioned for Data Scientists, less mentioned for Data Analysts, and is the 2nd most frequently appearing keyword for Business Analysts.

The past year's data revealed a similar storyline to our above findings. However, R tends to play a lesser role this year at all levels. This is an interesting finding, which could be attributed to geographical needs.

Summary

Based on the above results, program staff and graduating students can pinpoint the most wanted skills, the former can seek to have experts from the industry to teach those skills and incorporate those into the curriculum; the latter could prioritize polishing those skills to better prepare for interviews.

Further extensions of this project could include exploring other employment websites to get more aggregated scraping results across multiple platforms. This could be desirable as different companies may prefer one employment over another, we could more accurate judgment of the metrics. We could also feature more locations to find out how job skills and salaries could differ across the country. These could help to create a more comprehensive database.