BAX 422 – Final Project Report

Data Scientist Job Popularity Analysis

Section 1 - Group 5

Alexa Aguirre, Bernardo Arambula, Yumi Jin

Master of Science in Business Analytics

University of California, Davis

1. Executive Summary

This project provides valuable insights for recruiters and job seekers in the data scientist job market, helping them understand the market dynamics and factors influencing job popularity.

Based on Python language, we used Beautiful Soup and Requests libraries to scrape job listings for data scientists from LinkedIn and stored the acquired data in an SQL database for further analysis. During this process, we encountered challenges in avoiding LinkedIn's blocking mechanisms and accessing the HTML files, but we managed to overcome these obstacles in the end.

Central to our investigation is the Data Scientist Job Popularity Analysis. By building a model that considers the number of applicants as the dependent variable against various independent variables, we can identify the elements most impactful to a job's appeal. This analytical framework enables a quantifiable assessment of job attractiveness, offering actionable insights for companies in their recruitment strategies and, thus, is very meaningful.

2. Project Background

In the rapidly evolving data job market, both employers and job seekers strive to understand the dynamics that govern job popularity and competitiveness. This project is set against the backdrop of the technology and analytics industry, which is characterized by its high demand for skilled data scientists.

Companies within this sector are continually seeking to optimize their recruitment strategies to attract top talent, while job seekers aim to enhance their visibility and desirability to potential employers. Through analyzing job listings on LinkedIn, this study provides critical insights into the factors that make certain data science positions more attractive than others, thus informing recruitment and job search strategies in this competitive field.

3. Database Building Process

Data Sources Choosing

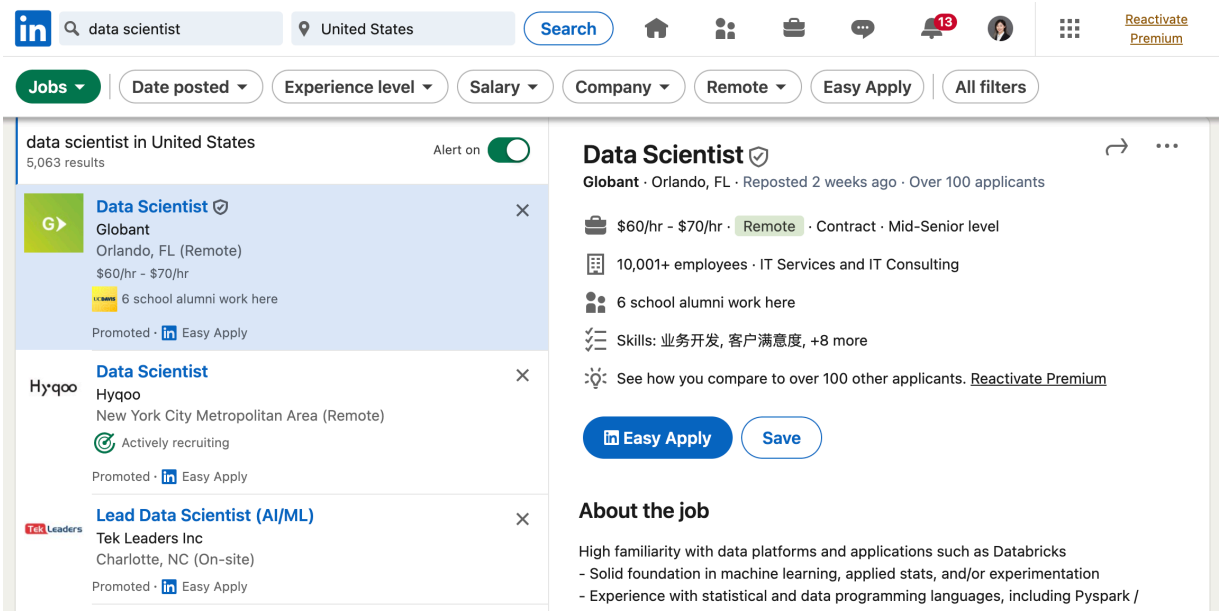
Our dataset is derived from job listings on LinkedIn, a premier platform for professional networking and job searches. We chose LinkedIn for its wide range of job postings and its user-friendly interface, making it an excellent resource for job market analysis. A critical factor in selecting LinkedIn was the availability of the **'number of applicants'** for each job listing. This information is pivotal for our analysis, as not all recruitment platforms provide such data, rendering LinkedIn a unique and indispensable source for our study.

Web-Scraping Method Choosing

Given LinkedIn's robust anti-scraping measures and consistent page layout, we opted for the Beautiful Soup and Requests libraries for data extraction, bypassing Selenium to navigate and parse content more efficiently. Another reason for selecting Beautiful Soup and Requests over Selenium was the scope of our data collection, aiming for hundreds to thousands of job listings. Selenium, though effective for browser automation, demands more time and system resources. For our extensive dataset, these requirements would have been impractical, making Beautiful Soup and Requests the more efficient choices for our project's needs.

Web-Scraping Process

Our process began by initializing a session and configuring the headers, essential steps for simulating a browser session that adheres to LinkedIn's web standards. We then proceeded to analyze LinkedIn's URL structure to collect job listings. Since each page displays only 25 job listings, identifying the pattern in URL changes between pages was crucial. Our strategy involved breaking down the URL into a base component and query parameters, with the 'start' parameter key to navigating through pages. By incrementally adjusting the 'start' value, we could systematically access and extract the URLs for individual job listings, accumulating them in a list for subsequent analysis. This methodical approach enabled us to efficiently gather URLs for 1,560 data scientist job listings.



P1 - LinkedIn Job Searching Interface

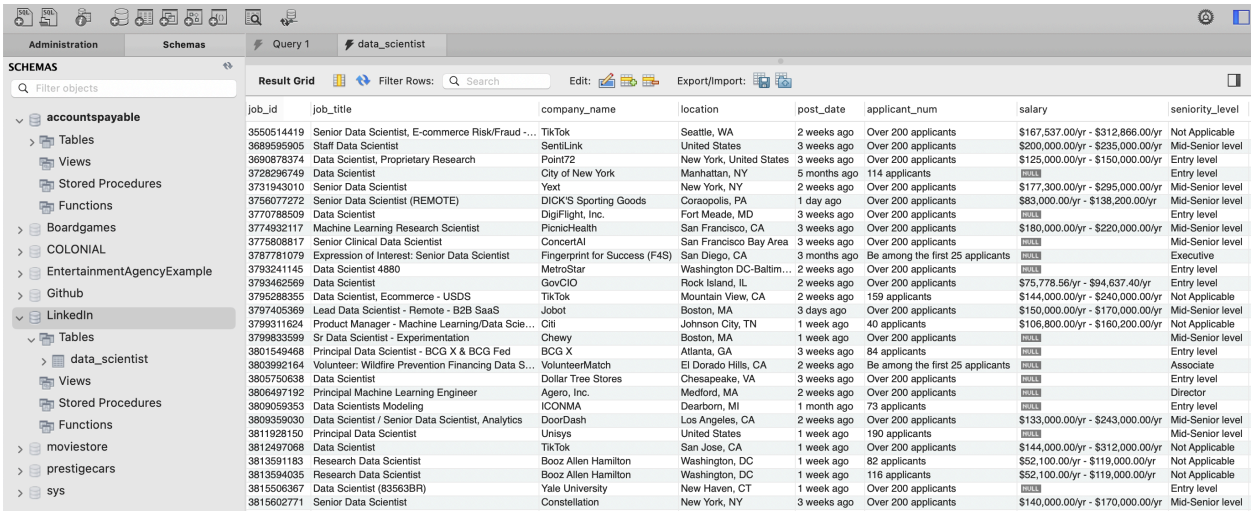
After compiling the list of URLs, we divided it into smaller segments to conduct web scraping incrementally, minimizing the risk of being blocked. We utilized the Requests library to retrieve each job listing page, saving the content as HTML files locally.

Next, we iterated through these saved HTML files, extracting essential details such as job titles, company names, salaries, and the number of applicants, and organizing this information into dictionaries. Due to repeated listings and LinkedIn's anti-scraping measures, the final tally of unique datasets was reduced to 265 by the time we processed and stored the information in HTML format.

Ultimately, we transferred the gathered data into a MySQL database. During this process, we carefully defined the database schema, ensured alignment between the data and its corresponding columns, and designated 'job_id' as the primary key to

uniquely identify each job listing, acknowledging its distinctiveness as an identifier.

Dataset Explanation



job_id	job_title	company_name	location	post_date	applicant_num	salary	seniority_level
3550514419	Senior Data Scientist, E-commerce Risk/Fraud -…	TikTok	Seattle, WA	2 weeks ago	Over 200 applicants	$167,537.00/yr - $312,866.00/yr	Not Applicable
3689595905	Staff Data Scientist	SentiLink	United States	3 weeks ago	Over 200 applicants	$200,000.00/yr - $235,000.00/yr	Mid-Senior level
3690878374	Data Scientist, Proprietary Research	Point72	New York, United States	3 weeks ago	Over 200 applicants	$125,000.00/yr - $150,000.00/yr	Entry level
3728296749	Data Scientist	City of New York	Manhattan, NY	5 months ago	114 applicants	NULL	Entry level
3731943010	Senior Data Scientist	Yext	New York, NY	2 weeks ago	Over 200 applicants	$177,300.00/yr - $295,000.00/yr	Mid-Senior level
3756077272	Senior Data Scientist (REMOTE)	DICK'S Sporting Goods	Coraopolis, PA	1 day ago	Over 200 applicants	$83,000.00/yr - $138,200.00/yr	Mid-Senior level
3770788509	Data Scientist	DigiFlight, Inc.	Fort Meade, MD	3 weeks ago	Over 200 applicants	NULL	Entry level
3774932117	Machine Learning Research Scientist	PicnicHealth	San Francisco, CA	3 weeks ago	Over 200 applicants	$180,000.00/yr - $220,000.00/yr	Mid-Senior level
3775808817	Senior Clinical Data Scientist	ConcertAI	San Francisco Bay Area	3 weeks ago	Over 200 applicants	NULL	Mid-Senior level
3787781079	Expression of Interest: Senior Data Scientist	Fingerprint for Success (F4S)	San Diego, CA	3 months ago	Be among the first 25 applicants	NULL	Executive
3793241145	Data Scientist 4880	MetroStar	Washington DC-Baltim…	2 weeks ago	Over 200 applicants	NULL	Entry level
3793462569	Data Scientist	GovCIO	Rock Island, IL	2 weeks ago	Over 200 applicants	$75,778.56/yr - $94,637.40/yr	Entry level
3795288355	Data Scientist, Ecommerce - USDS	TikTok	Mountain View, CA	2 weeks ago	159 applicants	$144,000.00/yr - $240,000.00/yr	Not Applicable
3797405369	Lead Data Scientist - Remote - B2B SaaS	Jobot	Boston, MA	3 days ago	Over 200 applicants	$150,000.00/yr - $170,000.00/yr	Mid-Senior level
3799311624	Product Manager - Machine Learning/Data Scie…	Citi	Johnson City, TN	1 week ago	40 applicants	$106,800.00/yr - $160,200.00/yr	Not Applicable
3799833599	Sr Data Scientist - Experimentation	Chewy	Boston, MA	1 week ago	Over 200 applicants	NULL	Mid-Senior level
3801549468	Principal Data Scientist - BCG X & BCG Fed	BCG X	Atlanta, GA	3 weeks ago	84 applicants	NULL	Entry level
3803992164	Volunteer: Wildfire Prevention Financing Data S…	VolunteerMatch	El Dorado Hills, CA	2 weeks ago	Be among the first 25 applicants	NULL	Associate
3805750638	Data Scientist	Dollar Tree Stores	Chesapeake, VA	3 weeks ago	Over 200 applicants	NULL	Entry level
3806497192	Principal Machine Learning Engineer	Agero, Inc.	Medford, MA	2 weeks ago	Over 200 applicants	NULL	Director
3809059353	Data Scientists Modeling	ICONMA	Dearborn, MI	1 month ago	73 applicants	NULL	Entry level
3809359030	Data Scientist / Senior Data Scientist, Analytics	DoorDash	Los Angeles, CA	2 weeks ago	Over 200 applicants	$133,000.00/yr - $243,000.00/yr	Mid-Senior level
3811928150	Principal Data Scientist	Unisys	United States	1 week ago	190 applicants	NULL	Mid-Senior level
3812497068	Data Scientist	TikTok	San Jose, CA	1 week ago	Over 200 applicants	$144,000.00/yr - $312,000.00/yr	Not Applicable
3813591183	Research Data Scientist	Booz Allen Hamilton	Washington, DC	1 week ago	82 applicants	$52,100.00/yr - $119,000.00/yr	Not Applicable
3813594035	Research Data Scientist	Booz Allen Hamilton	Washington, DC	1 week ago	116 applicants	$52,100.00/yr - $119,000.00/yr	Not Applicable
3815506367	Data Scientist (83563BR)	Yale University	New Haven, CT	1 week ago	Over 200 applicants	NULL	Entry level
3815602771	Senior Data Scientist	Constellation	New York, NY	3 weeks ago	Over 200 applicants	$140,000.00/yr - $170,000.00/yr	Mid-Senior level

P2 – Data Scientist Database Overview

The dataset encompasses a wide range of information and has 11 columns. Including job ID, job title, company name, location, posting date, applicant number, salary, seniority level (Entry level / Senior level), employment types (Full-time / Volunteer / Contract), job function, and industry. It has 265 rows in total.

4. Business Question & Value

Business Question:

How can employers optimize their recruitment strategies to attract top data science talent, and how can job seekers enhance their profiles to increase their visibility and competitiveness for these positions?

This inquiry delves into understanding the dynamics that influence the attractiveness of data scientist positions from both the recruiters' and job seekers'

perspectives. By analyzing various factors such as job title, company name, location, salary, seniority level, employment type, job function, and industry, this project aims to uncover actionable insights that can inform effective recruitment and job search strategies in the highly competitive technology and analytics job market.

Business Value

a. Recruitment Strategy Optimization for Employers:

By identifying the elements that make data scientist roles more appealing to potential candidates, companies can tailor their job listings and recruitment strategies to attract top talent. Insights from this analysis can guide employers on aspects such as competitive salary ranges, preferred job functions, and desirable employment types, enabling them to stand out in a crowded market.

b. Enhanced Job Search Effectiveness for Job Seekers:

Job seekers can leverage the findings to understand what makes a data scientist role attractive and tailor their applications accordingly. This includes focusing on roles that match their level of expertise, desired job functions, and preferred industries, thereby increasing their chances of securing positions that align with their career aspirations.

c. Strategic Planning and Market Analysis:

Beyond immediate recruitment and job search strategies, the insights gained from this project can serve as valuable input for strategic planning and market

analysis within the technology and analytics sector. Understanding the dynamics of job popularity and competitiveness can help companies anticipate future trends in talent acquisition and retention, shaping the development of long-term human resource strategies.

d. Database Implementation Advantages:

The decision to use a MySQL database for storing and analyzing the scraped data ensures efficiency, scalability, and ease of access. Compared to alternative data storage solutions, MySQL offers robust data management capabilities, facilitating complex queries and analysis. This choice supports a streamlined analysis process, enabling the extraction of meaningful insights from the dataset. Additionally, the structured approach to data collection and storage enhances the reliability and validity of the analysis, contributing to the project's overall value proposition.

5. Conclusions

In conclusion, our study on the data scientist job market could offer valuable insights for both recruiters and job seekers. By collecting and analyzing job listings from LinkedIn, we could find out what makes data scientist jobs appealing. We could suppose that our model shows that things like salary, job title, and the type of work are important for attracting the best candidates, then…

For employers, this means creating job ads that highlight these attractive features can help them find top talent more easily. Job seekers, on the other hand, can use

this information to focus on applying for jobs that fit their skills and career goals best.

Our decision to use a MySQL database for organizing our data helped us analyze the information efficiently and get meaningful insights. This makes our findings not just useful for immediate job hunting or hiring but also for long-term planning in the tech and analytics industries.

In short, our project could help employers attract the right candidates and help job seekers find the best opportunities in the data scientist job market.