Analyzing Data Science Job Market in U.S Metropolitan Areas

– Saani Rawat, TDI Fellow

**Overview**

Data Science, according to Harvard Business Review, is the sexiest job of the 21st century**[[1]](#footnote-1).** Despite its popularity, according to a LinkedIn article[[2]](#footnote-2), there is a clear shortage of data science talent available in the market. When there are only a limited number of people to select from, companies need to make sure that they attract the best data science talent. To ensure this, companies need to understand the current market status and have insights on job market trends. Through this project, we will look at real data scraped from a famous web search engine to analyze the gain insights on prevailing data science market conditions.

**Data**

*Data Source*

For this analysis, data were scraped from <https://www.indeed.com>. Indeed is an [employment](https://en.wikipedia.org/wiki/Employment)-related search engine comprising mostly of job listings within United States. The site aggregates job listings from thousands of websites, including job boards, staffing firms, associations, and company career pages. They generate revenue by selling premium job posting and resume features to employers and companies hiring[[3]](#footnote-3).

*Data scraping*

Indeed’s home page allows users to search for any job title within a location. Once a job title and a location are entered, user can view 15 job snapshots[[4]](#footnote-4) at a single page. For each job snapshot, user can click on the job and expand for more details. During the data scraping process, the following fields were extracted for each job posting –

1. Job Title: position or title of the job being offered
2. Company Name: name of the company posting the job opening
3. Company Address: physical address of the job opening  
   The following cities were considered as U.S. metropolitan areas during scraping –
   1. San Francisco
   2. Boston
   3. Houston
   4. Washington DC
   5. Dallas
   6. Seattle
   7. Charlotte
   8. New York Metro area
   9. Chicago
   10. Philadelphia
   11. San Diego
   12. Atlanta
4. Salary: range of salary offered by the company
5. Job Summary: a short description of the responsibilities or the most important skills required

These fields were utilized during the analysis phase after data cleaning.

*Data Cleaning*

Job posting data, after web scraping, was not in the most useful form. Data cleaning steps were needed to make the data ready for analysis. The following data cleaning steps were executed –

1. Observations lacking salary information were removed
2. Observations lacking company location were removed
3. All duplicate rows were removed. Hence, one job posting can exist only once in the dataset
4. Original salary variable was a range. Ranges also differed between hourly, daily, weekly, monthly and yearly. All these ranges were converted into yearly ranges. For each observation with salary range information, the maximum of the range (after time conversion) was used as the highest possible salary.

**Analysis**

After cleaning, data was analyzed based on job title and location. For a respective job title and a particular location, the following aspects were analyzed –

1. Number of related job postings in a location – the number helps in identifying the level of competition in the market. Companies can be aware, by looking at this information of the number of competing job openings that might exist in a particular location.
2. City’s Market share – For a particular city, the proportion of related jobs in that city vs total related jobs. This helps in gauging how attractive the city is to prospective data scientists.
3. Top 10 Salaries offered at a location – highest maximum salaries offered in particular city or region. By having this information, the companies will be able to see the salaries offered by the peers and use this information to decide their own compensation packages.

**Deployment**

The results of the analysis were presented and deployed using Heroku. Heroku is a cloud platform as a service (PaaS) supporting several programming languages. Python was used for deployment.

App link – <https://dashboard.heroku.com/apps/saani-rawat-tdi-capstone>

**Future Scope**

1. Perform in-depth inferential analysis on data science job market statistics and make recommendations based on job location, type, salary and company.
2. Use Natural Language Processing (NLP) to identify the most common words (or buzz words) used by employers

**Conclusion**

Through this project, the prevailing conditions of the data science market can be analyzed. The data scraped from indeed.com, after processing and cleaning, can be used by employers to understand the latest job market situation and also the strategies that other businesses might be undertaking; thereby allowing the employers to make more informed decisions.

**Appendix**

*GitHub*

The project can be found here on GitHub: <https://github.com/saannidhya/TDI-Capstone-Project>

1. HBR article – <https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century> [↑](#footnote-ref-1)
2. LinkedIn article – <https://news.linkedin.com/2018/8/linkedin-workforce-report-august-2018> [↑](#footnote-ref-2)
3. For more information, see <https://en.wikipedia.org/wiki/Indeed> [↑](#footnote-ref-3)
4. Job snapshot – a high-level summary of most important job-posting related attributes [↑](#footnote-ref-4)