US Job Market for Data Science

Quan Gan

1. Introduction

Data Science is the job section I've chosen. There are numerous careers in the field of data science as a whole. I review the 4 websites [1-4] to conclude the top 9 data science occupations, Data Scientist, Data Analyst, Data Architect, Data Engineer, Database Administrator, Statistician, Machine Learning Scientist, Machine Learning Engineer, and Business Analyst. Indeed.com is a job search website based in the United States that lists jobs all over the world. This website offers the research data. I'll focus more on Data scientists, which I'm interested in, after briefly introducing the 9 data science occupations.

2. Categories

What are the differences among these 9 occupations? Data Scientist basically knows most knowledge of data science from data collecting and analyzing to finally visualizing and presenting including machine learning skills and software engineering. Data analyst almost does the same job as Data Scientist but not involves complex models and too many programming skills. Data Architect focuses on the companies' databases to guarantee well-formatted and accessible data. This occupation always required advanced data science knowledge and experience. Data Engineer concentrates on data preprocessing to prepare tidy and cleaned data for further analysis. Database Administrator focuses more on managing the database. Statistician just like the name needs to have strong statistics knowledge to offer valuable insights and statistical models. Machine Learning Engineer needs to be familiar with all kinds of machine learning algorithms and techniques to employ them in the job. Machine Learning Scientist not only requires the techniques for operating the machine learning algorithms but also has the ability to design new algorithms and approaches based on different requirements. Business Analyst's key difference from other data science occupations is this job requires business knowledge to analyze the market and business trends.

3. Data

3.1 Occupations Distribution Data

To analyze the distribution of the nine vocations, I used the Indeed website to search for the nine job titles using only one parameter: location as “United States”. Because recruitment information changes over time, the scope of this distribution is limited to currently available positions until September 14. Because the amount of job information on Indeed represents the pages, the actual job amount should be page number multiple job numbers on one page (15 jobs per page).

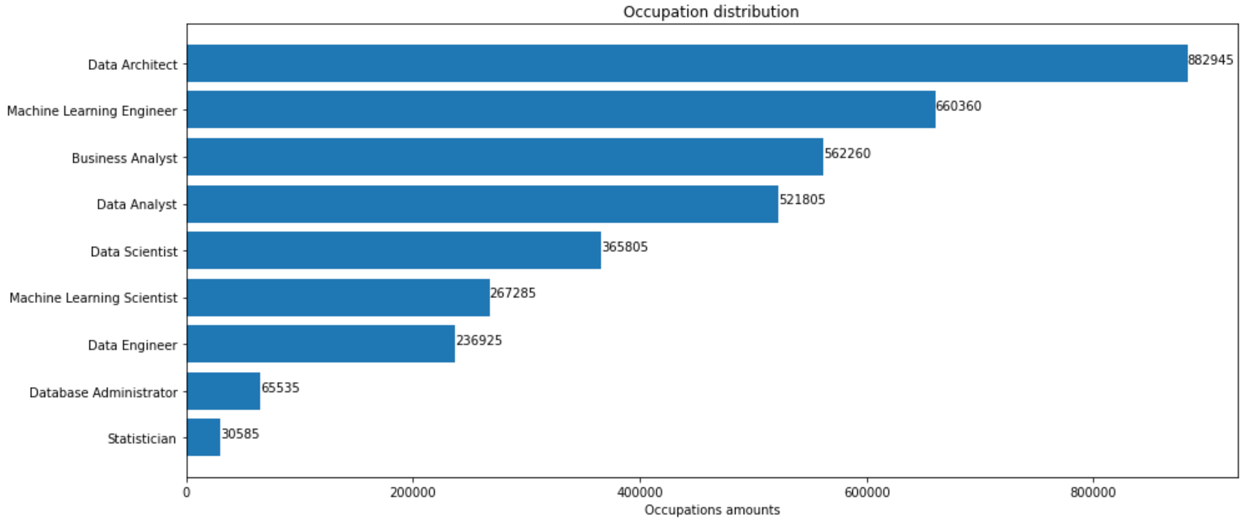
3.2 Data Scientist Data

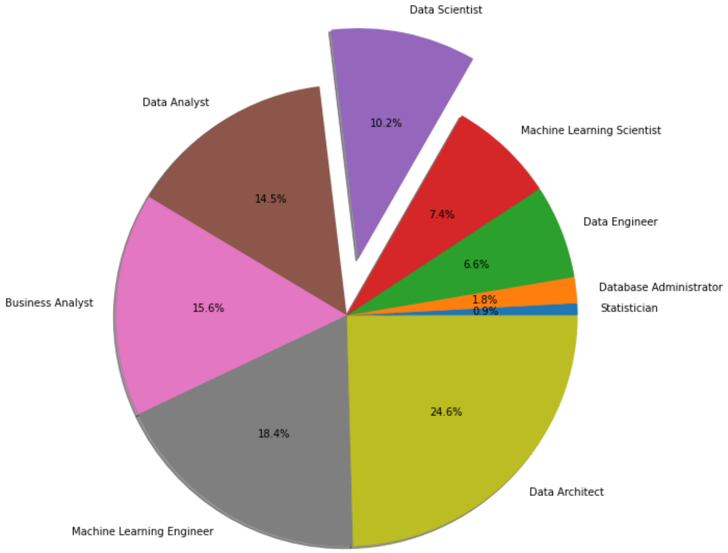
By web crawlers, I collected 4391 recruitment information from indeed.com. The data constraints are 1) posted from 2021/09/11 to 2021/09/18, 2) United States positions, 3) job titles containing “Data sci”. The last constraint is to guarantee the jobs related to data scientists. However, the fewer job postings have the salary and qualification sections. From 4391 recruitment information, I only found 83 postings have qualification sections and 108 postings have salary sections. In addition, only full-time salary will be analyzed, which means the salary unit is a year.

4. Method and Result

4.1 Occupations Distribution Analysis

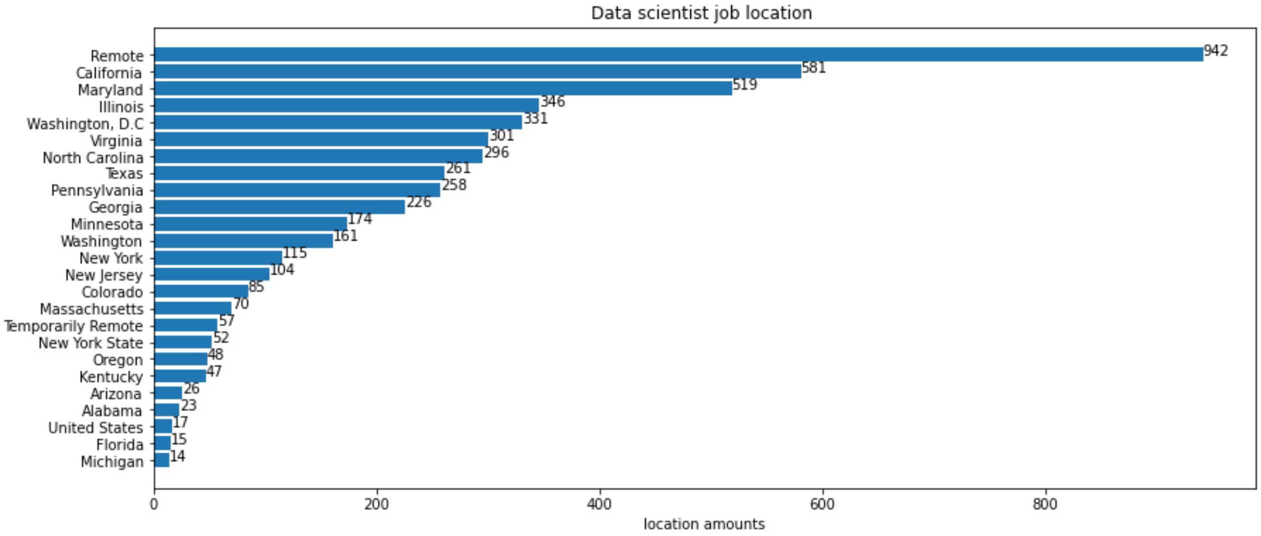
Because I only collect the number of distinct occupations, the data for 9 occupations are collected manually. The illustration is then rendered using the Python Jupyter Notebook. Data Architect is currently the most demanding occupation, which is not what I expected to see. Because of the increasing growth of machine learning technology, the second-highest need is for a Machina Learning Engineer. In all categories, barely 10% of people work as data scientists. The overlapped workflows of the Data Scientist and Data Analyst could be the reason for some companies' inability to distinguish them effectively.

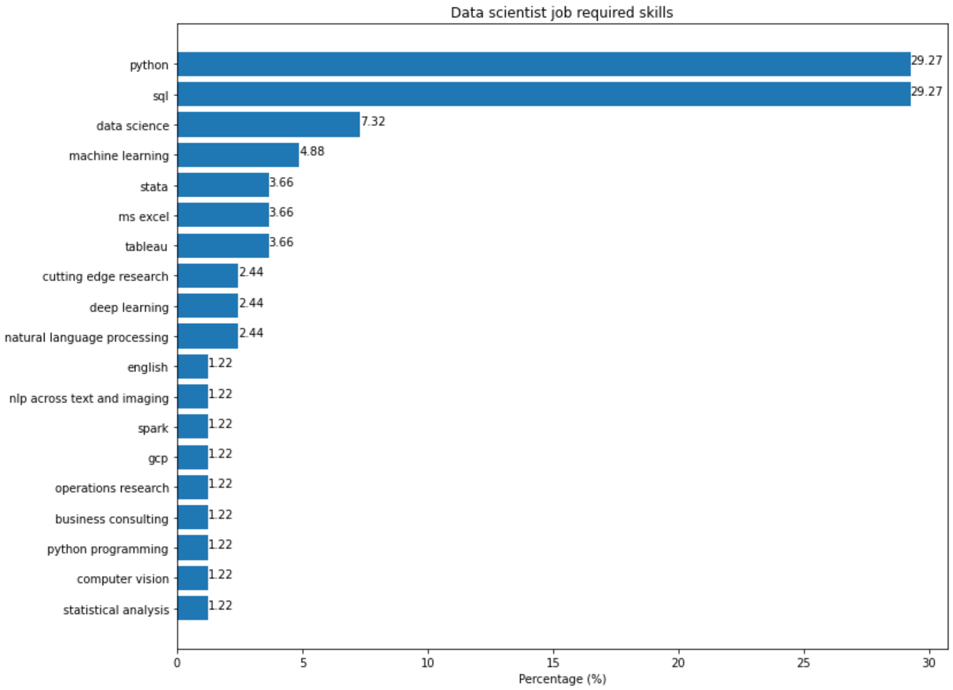


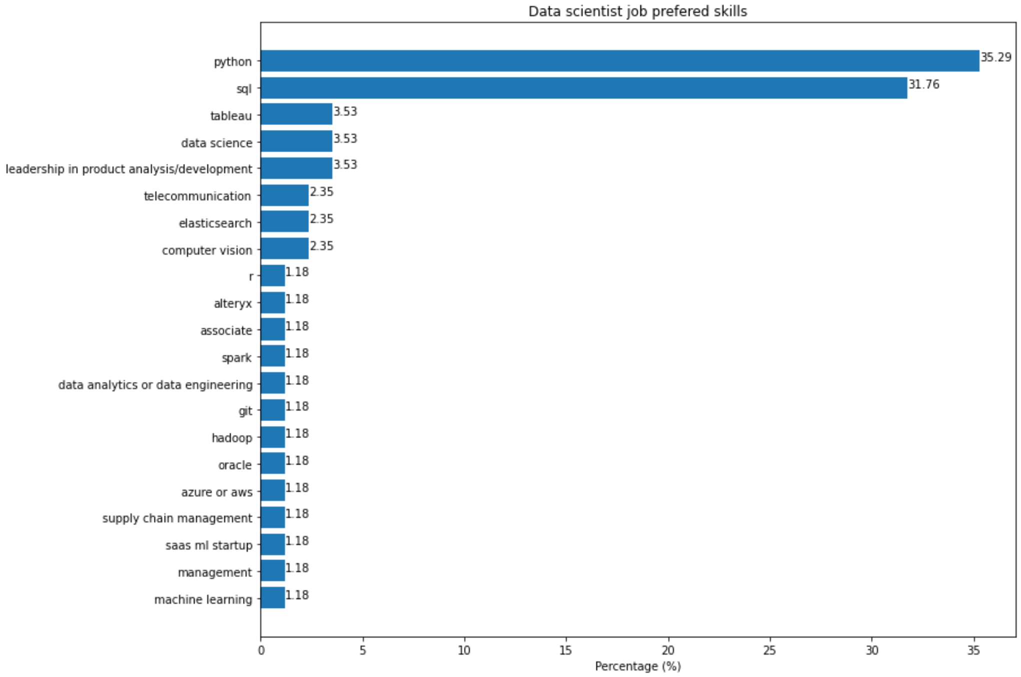


4.2 Data scientist Analysis

The main technique in this research is python and the code is written in the Jupyter Notebook. Data scientist data is collected by web scraping, which requires HTML/CSS knowledge and python library, requests. Next, I use the python libraries, pandas and Matplotlib, to store, analyze, and visualize the data. The following graphs illustrate the job market. And the mean salary is $139610.02 based on limited data.







5. Limitation

5.1 Occupations distribution

Because the last page may not include 15 tasks, the accurate amount may differ from the experiment amount up to 15 records, the outcome cannot be guaranteed. Furthermore, after reviewing the recruitment information for various vocations, I discovered that not every query result will be the job information for the specified occupation since the indeed website queries job lists based on keywords. For example, if I search for the job title "Data scientist," I'll get "Data analyst" and "Data Engineer" among the results. As a result, the outcome suffers from duplicate recruitment data.

5.2 Data Scientist

This study is additionally limited by the restriction on web scraping. I was only able to scrape information about small-scale recruitment from the indeed website. The website will then prohibit my program from continuing. Another stumbling block is the inconsistency of recruitment data. Because not all job postings include a section for job description and a section for qualifications, the amount of available data is reduced. Different companies have different ways of describing work details. For example, Hours or years may be used as salary units.

6. Idea job

The ideal job title is Junior Data Scientist. This job needs to 1) participate in the design and build of modern data analytics solutions using state-of-the-art technologies; 2) design, build, and test modern data pipelines using cloud technologies; 3) build lakehouse architectures using data warehouses, and data lakes, delta lakes.

The requirements of this job include 1) 1-3 years of data science or technical consulting experience (INFO 523, RA, GA, Internship) 2) experience with python (CSC 110, CSC 120, ISTA130). 3) Experience developing the data pipeline (INFO 531). 4) experience building Machine Learning Models (INFO 510, INFO 521, INFO 557, INFO 523). 5) experience with SQL, NoSQL databases, and cloud data platforms (INFO 531, INFO 570, CSC 560). 6) experience with visualization tools (MIS 561, CSC 544, INFO 526). 7) Strong analytical, verbal, and written communication skills (INFO 505, INFO 507, RA, GA, Internship)

Reference

1. *Types Of Jobs In Data Science in 2021 [Updated] | Data Science with Henry Harvin Blogs*. (n.d.). Retrieved September 14, 2021, from **https://www.henryharvin.com/blog/types-of-jobs-in-data-science/**
2. Chatterjee, M. (2020, April 29). *Top 9 Job Roles in the World of Data Science for 2021*. GreatLearning Blog: Free Resources What Matters to Shape Your Career! https://www.mygreatlearning.com/blog/different-data-science-jobs-roles-industry/
3. *13 Data Science Careers That Are Exploding Now*. (2020, November 12). Mallory. https://mallory.com.au/data-science-careers-guide/
4. Metwalli, S. A. (2020, November 9). *10 Different Data Science Job Titles and What They Mean*. Medium. https://towardsdatascience.com/10-different-data-science-job-titles-and-what-they-mean-d385fc3c58ae