Advandced Data Science Project RMD

Nikita Stempniewicz

September 8, 2017

Note: I am not done yet but wanted to take the opportunity to have you look over the general idea and contents. Before the final draft I will work on cleaning up the text and figures, and finishing the last section and conclusion.

Introduction

In general, a data scientist primary function is extracting insights from data, and communicating those insights in a clear and efficient manner. This requires a diverse skill set based in math, statistics, and computer, and information sciences. There is no typical technical or educational skill set required of data scientist, and different employers will require varrying skills. Some companies only require a bachelors degree while others a doctoral for data scientist. Most require programming and statistics skills, but often vary in the programming language or statistical software, e.g., SQL, R, STATA, Python, etc. Some data scientist position are more focused on data vizualization and communication, with an emphasis on skills with data vizualization software such as tableau, and others on exploratory analysis whith more of an emphasis on statistical methods such as machine learning.

Specific Aims

- Describe educational, technical, and professional requirements for data scientist positions
- Describe estimated salaries for data scientist
- Investigate the relationship between the different requirements and estimated salary

Data Source

All the data for this analysis was scraped from Glassdoor.com. To minimize the variation in salaries due to geographic differences, the location when searching for data scientist positions was restricted to New York City, which includes the city itself and surrounding areas.

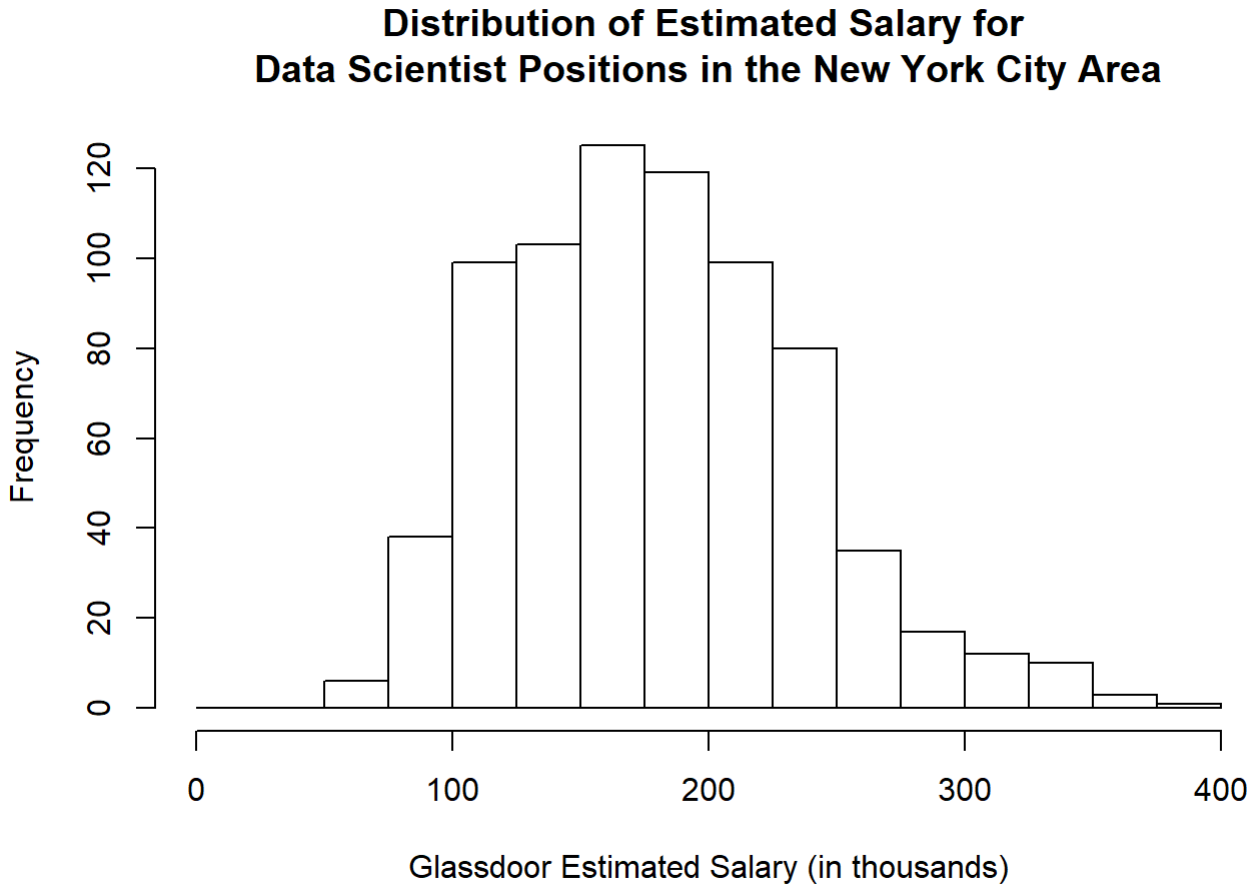
The methods used to scrape data is described below:

- First, I get data, i.e., job id, employer id, job title, salaries (if available), locations, employer from the search results for data scientist.
- Next, using the job id from the search results in the previous step, I build the URL for the individual job posts on glassdoor and get the raw job descriptions which includes text information on qualifications which I will later query to build structured fields.
- Finally, using the employer id from the search results, I build the URL for the employer page on glassdoor, and get the raw text from the employer description field, which includes things like, size, industry, year founded.

Data set

- 990 data scientist positions in the New York City area
    - No job descriptions (n=1)
    - No estimated salary (n=242)
- 747 data scientist positions in the New York City area with a job description and estimated salary

Estimated Salaries



The average estimated salary for data scientist in the new york city area was 179.9 and the standard deviation 57.3

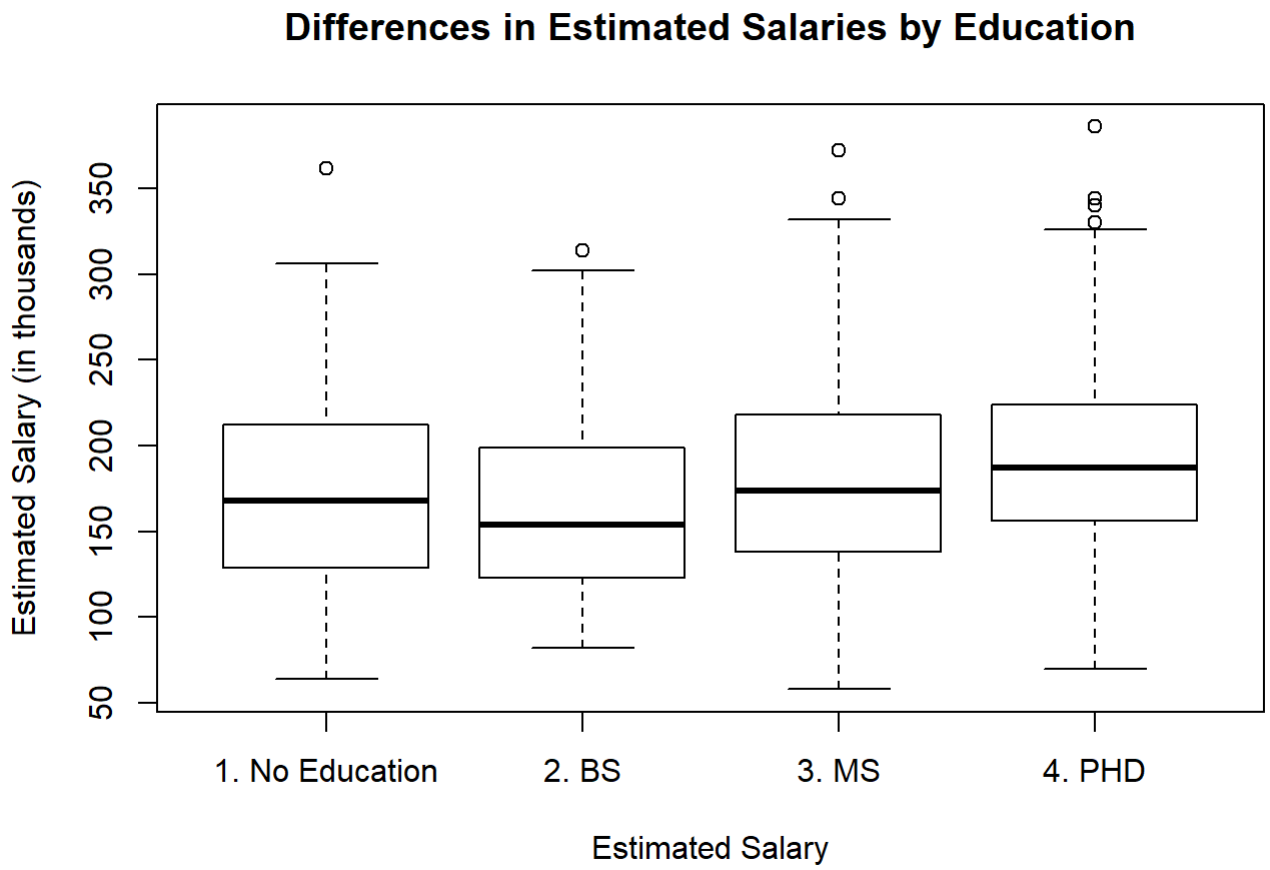
Education

It is not uncommon for employers to have required and preffered qualifications which often include multiple possibilities for educational requirements, e.g., requiring a bachelors at a minumum, but specyfying a preference for candidates with a masters or PhD. When more than 1 educational requirements are mentioned we classify jobs to the highest level, e.g., posts that included BS and MS are considered MS.

Degree	Jobs	Salaries
No Education	99 (13.3%)	177.7 (59.7)
Bachelors (BS)	175 (23.4%)	165.2 (50.8)
Masters (MS)	255 (34.1%)	179.9 (56.9)
PhD	218 (29.1%)	192.8 (59)

Overall, we were not able to get educational requrements for 13.3% of job posts. Among job descriptions where education was specified, 23.4%, 34.1%, 29.2%, mentioned a bachelors, masters, and PhD degrees respectively.

Data science positions that mentionded a PhD had an average salary of $192,800, compared to $179,900 for masters, and $165,200 when only a bachelors degree was mentioned. For jobs where no information on education was ascertained, the average salary was similar to the overall average.

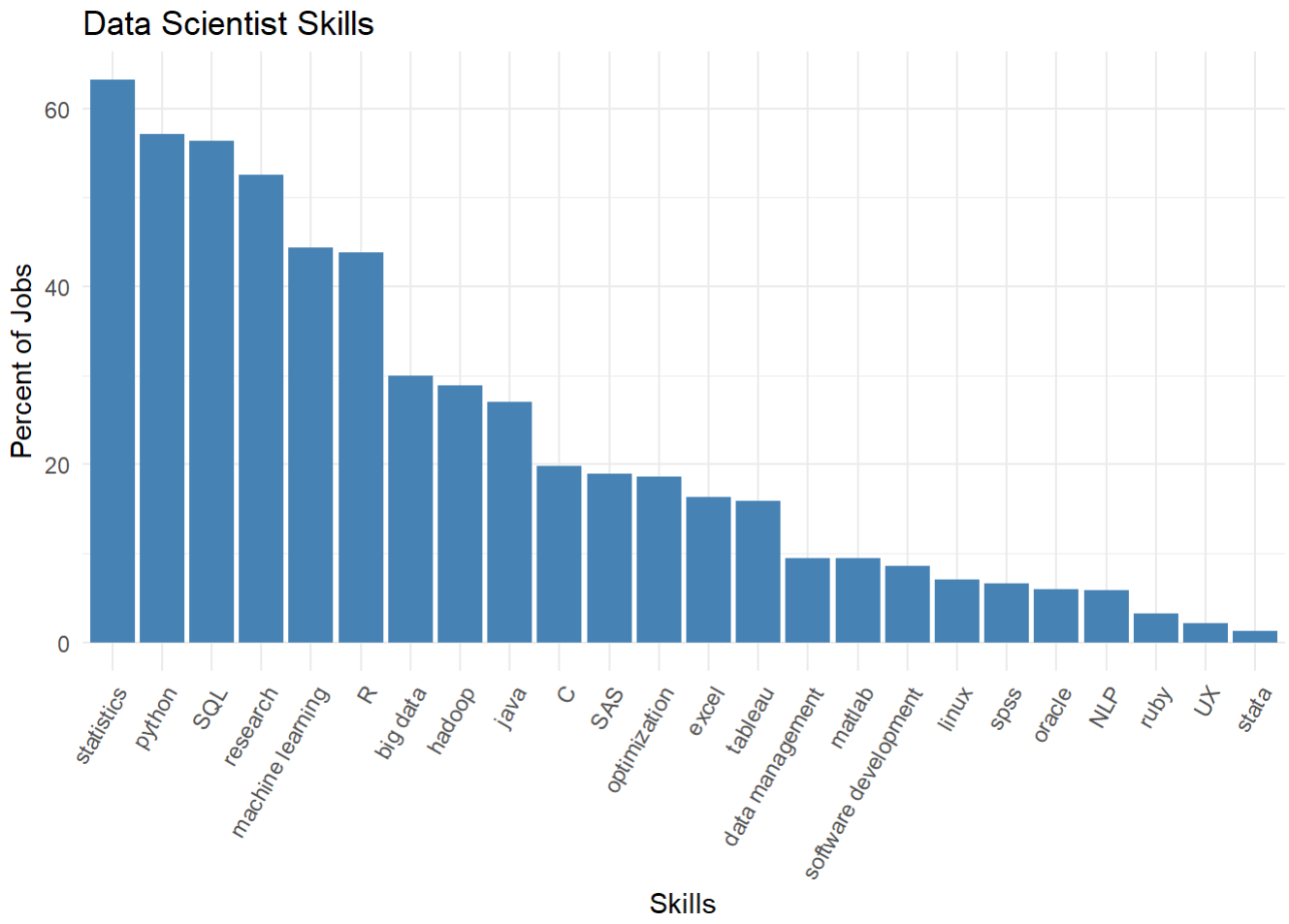


Using linear regression we confirmed differences in estimated salaries by educational requrements are statitically significant. Compared to jobs where bachelors degree is the only educational requirement mentioned, jobs that mentioned masters degrees had an estimated salary 14.7 (SE: 2.7) thousand higher, jobs that mentioned a PhD had an estimated salary 27.7 (SE: 4.9) thousand higher.

Job Skills

Using the raw text from the job and employer decsipription I created binary variables looking for text relevant skills specific to data scientists, e.g., R, SQL, and Python. Overall, 91.5% of the 990 job posts had at least 1 of the 15 skills that were investigated.

Overall, the most common skills are Python, SQL, and R. Some less common skills include Oracle, Ruby , and Stata.



Job Skills and Estimated Salary

Here I am going to include a figure looking at the associations of the individual skills and estimated salary (adjusting for differences in education). e.g., i will include the point estimates and 95% CI for the betas for the individual skills from each model.

The results make sense, Machine Learning had the largest estimate, and tableau and excel the lowest.

Conclusion

Here I will tie everything together