Final Project report – Job Descriptions

Ruxin Liu

12/13/2020

Introduction

As we are now in a big data world, more and more data related jobs have been created recently among different fields. When we are searching these job positions online, it is very common to see diverse descriptions, titles and requirements among different companies and job boards. The main goal of this project is to explore some of these varieties through the use of the application programming interface (API).

Data Procesing

The data sources of this project are from Adzuna and The Muse, which are both online career platforms. After registering for the API ID and key, we could write queries to obtain the data and process the data in R and all the detailed codes for data cleaning and processing are in the file Data Processing.Rmd.

Adzuna is an employment website where its headquarters located in the United Kingdom (Wikipedia 2020). From this site, we obtain 50-page of job information in Britain that has keywords of data, statisticians, statistician, analyst or analysts. Also, we obtain 50-page of job information in America with the same list of key words. Only the job information from Britain has salary listed, which is reasonable, since the majority of the job boards do not show the salary on their pages.

The Muse is another employment website founded in 2011, where its headquarters located in the United States (Wikipedia 2020). From this site, we obtain 60-page of job information from the data science category, which is very close to the keywords being searched from Adzuna.

EDA (Exploratory Data Analysis)

Salary Level & Job Title

Since the UK job information on Adzuna website has predicted maximum salary in pounds, we could explore whether the salary level has any relationship with the job title. In order to do this, an indicator variable is created to determine whether or not the position name includes “Statistician”. From Fig.1 below, we can see that overall companies predict a relatively higher salary for job titles containing “Statistician” (indicator = 1). However, it is also very interesting to see that there are much more outliers with very large values for the job titles not containing “Statistician” (indicator = 0). This suggests that on average the jobs that does not contain “Statistician” in the title may have a lower salary, but at the same time, there are large opportunities to reach a very high and less limited salary level.

Table 1: Number of job posting in each level

level	n
entry	1
internship	178
management	153
mid	485
senior	422

Required Skills & Level

On The Muse website, we could obtain the level of the positions for each job posting, which includes internship, entry, mid, senior, management. From Table 1 , we can see the number of job posting in each level among the sampled pages, and it seems that under the data science category there are very few positions available for the entry level.

Since programming skills are very crucial in data related positions, we could explore the occurrence frequency of specific skills in the job description among different position levels. More specifically, 4 commonly-used programing languages are tested, which are SQL, Python, SAS and R.

From Fig.2 below, it is clear that instead of internships, about half of the job descriptions among other levels include SQL as a skill.

Fig. 2: Occurrence Frequency of SQL in Job Description

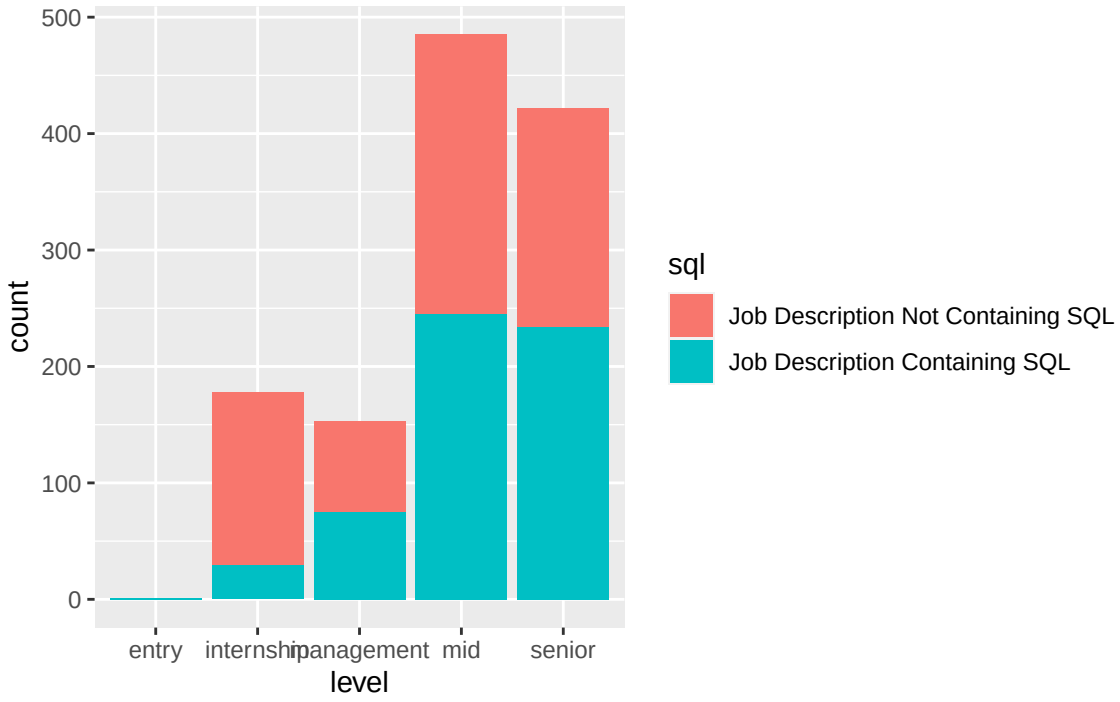


Fig.3 clearly shows that instead of the management level, more than half of the job descriptions among other levels include Python as a skill.

Fig. 3: Occurrence Frequency of Python in Job Description

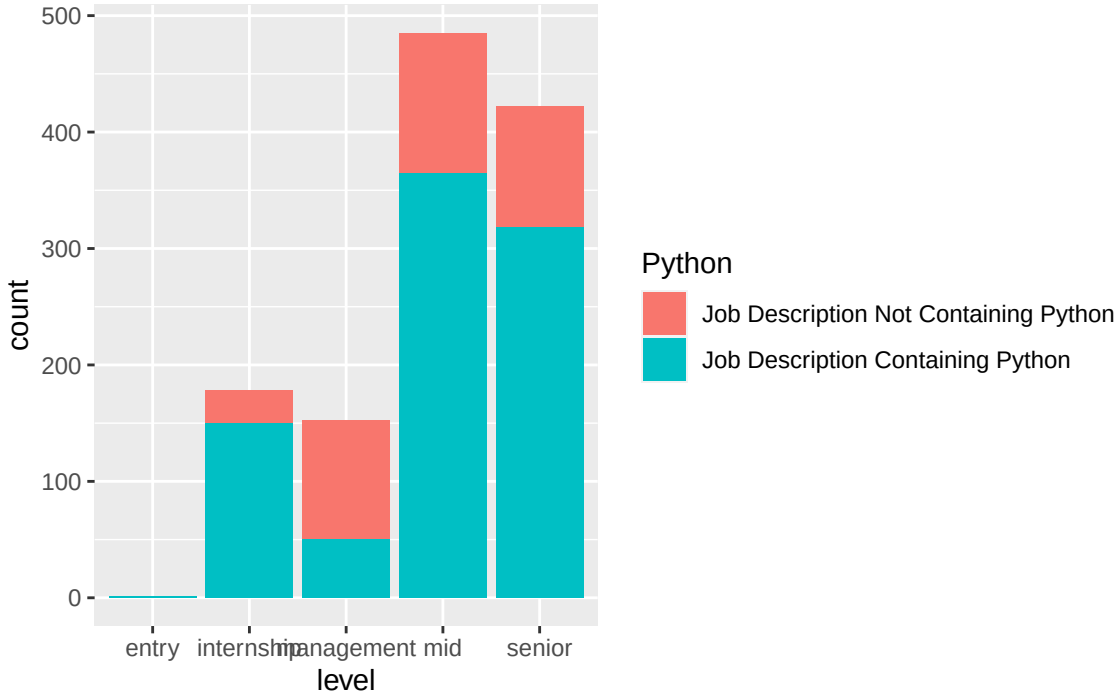
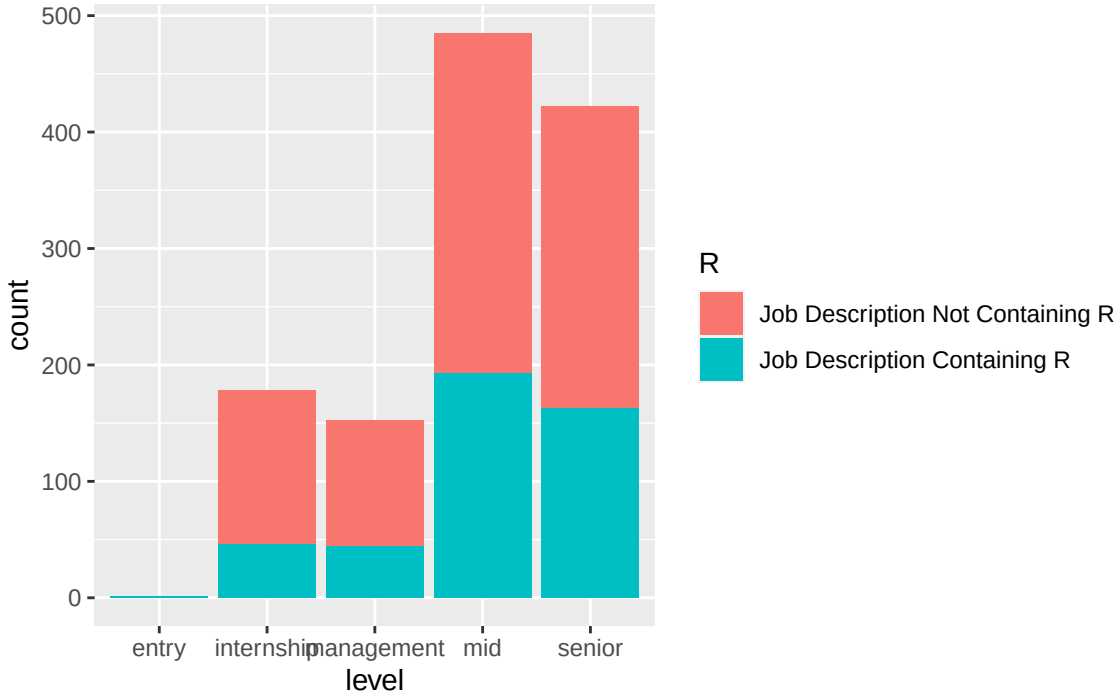


Fig.4 shows very different patterns compared to the plot for Python, where very few of the job descriptions among all the levels include SAS as a skill.

Fig. 4: Occurrence Frequency of SAS in Job Description

Lastly, Fig.5 displays similar patterns as what is seen from the plot for SQL, where more than half of the job descriptions among all the levels do not include R as a skill.

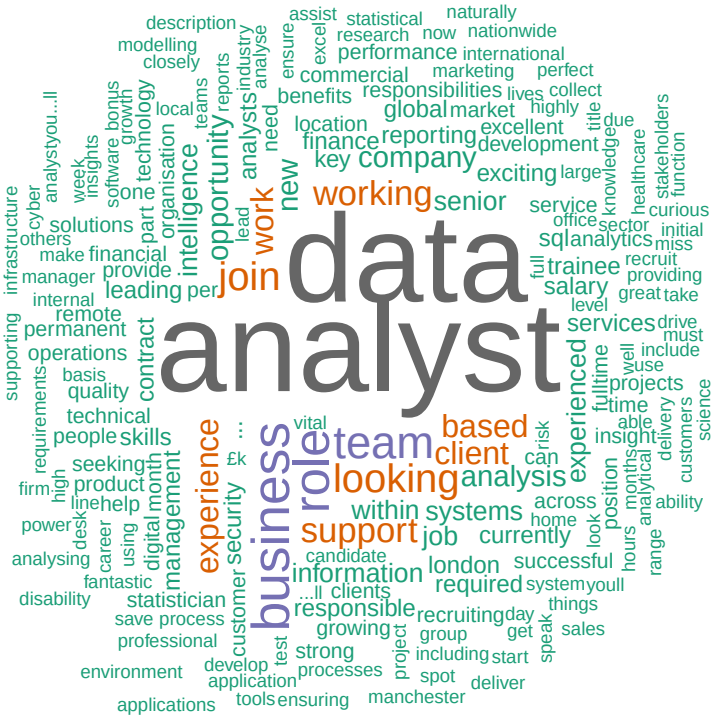
Fig. 5: Occurrence Frequency of R in Job Description



Word Cloud

We could analyze the text in job descriptions by creating the word cloud. Below are the word clouds for the 3 data sources, which show the frequency of words by visualization.

Adzuna Data – UK



Adzuna Data – US

The Muse

There is no surprise to see that word "data" is always the one with the highest word frequency since we are focusing on data-related jobs. Except for this, the 3 different word clouds also show very interesting patterns. Some of the most commonly used descriptors from Adzuna UK are: "experience", "analyst", "team" and "business". And some of the most commonly used descriptors from Adzuna US are: "analyst", "business" and "support". Although these are job information for different countries, it seems that the words used in descriptions are relatively consistent.

For The Muse, some of the most commonly used descriptors are: "experience", "team", "learning", "research", "computer", "models", "quantitative", "science" and "statistics". The words "experience" and "team" have very high frequency for both job websites, which suggests that practical working experience in the fields and teamwork skills are always crucial things to have. Except for this, there are more academic/statistics related descriptors appears on The Muse, which suggests that different employment websites may have a slightly different focus when describing the position.

Discussion

- On average, jobs that contain "Statistician" in the titles have a higher salary, however, some jobs that do not contain "Statistician" in the titles have the potential to reach a much higher maximum salary.
- Based on the EDA, we can see some patterns of the occurrence frequency of programming languages in the job description. Overall, among the 4 tested programming skills, SAS is the least popular one. R and SQL both are mentioned in job descriptions for about half of the time. And Python is now shown to be the most required skill
  for data-related jobs. This could be very helpful when we are applying for jobs in the future.
- The job information for different countries from the same job board is more likely to have similar descriptions, while similar job information from different job boards may have a different focus when describing the position.

Limitations & Future Direction

- Since for some job sites, there are daily limitations in utilization when using API to query the data, it takes long time to get the data. Also, since the job sites have new updates every day, it is almost impossible to have the complete/full data. Therefore, it could been developed into a more long-term project in the future, which can have some updates on a regular base.
- Instead of increasing the sample size, more different online job boards across different countries could also be considered.

Shiny APP Description

The shiny app for this project is published on the bu-rstudio-connect.bu server: https://bu-rstudio-connect.bu.edu/connect/#/content/listing?filter=min_role:viewer&filter=content_type:all

The shiny app has 3 tabs. The 1st tab provides reactive word clouds. The 2nd tab provides reactive histograms to show the occurrence distribution of different programming skills. The 3rd tab can be selected to show the link and the API for several online job boards.

Bibliography

1. Jeroen Ooms (2014). The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. arXiv:1403.2805 [stat.CO] URL https://arxiv.org/abs/1403.2805.
2. Hadley Wickham (2019). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. https://CRAN.R-project.org/package=stringr
3. Wikipedia (2020). *Adzuna* [online]. Available from: https://en.wikipedia.org/wiki/Adzuna [accessed 13 December 2020].
4. Wikipedia (2020). *The Muse* [online]. Available from: https://en.wikipedia.org/wiki/The_Muse_(website) [accessed 13 December 2020].
5. Hao Zhu (2019). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.1.0. https://CRAN.R-project.org/package=kableExtra
6. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
7. Ian Fellows (2018). wordcloud: Word Clouds. R package version 2.6. https://CRAN.R-project.org/package=wordcloud
8. Erich Neuwirth (2014). RColorBrewer: ColorBrewer Palettes. R package version 1.1-2. https://CRAN.R-project.org/package=RColorBrewer
9. Ingo Feinerer and Kurt Hornik (2020). tm: Text Mining Package. R package version 0.7-8. https://CRAN.R-project.org/package=tm
10. Towards Data Science (2019). *How to Generate Word Clouds in R* [online]. Available from: https://towardsdatascience.com/create-a-word-cloud-with-r-bde3e7422e8a [accessed 13 December 2020].
11. Stats And R (2020). *Draw a word cloud with a R Shiny app* [online]. Available from: https://www.statsandr.com/blog/draw-a-word-cloud-with-a-shiny-app/ [accessed 13 December 2020].
12. Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra