Trends in Data Science Job Postings on Stack Overflow

Benjamin Ackerman

October 24, 2017

Introduction

Coined in 2008 by data analytics leads at Facebook and LinkedIn, the term "data science" refers to a profession for people with quantitative training, computer programming experience and intellectual curiosity to make discoveries about the world through big data [7]. As computers and technology have evolved, so too have methods of defining, collecting and processing data. The recent emergence of the field of data science has accompanied researchers' abilities to access and store unprecedented amounts of information. There is a massive (and growing) demand for data scientists to make discoveries with the estimated 2.5 quintillion bytes of data that are created *daily* [16]; however, the growth in available data has far exceeded the rate at which people have entered data science positions. On average, data science jobs currently remain open for 45 days, five days longer than other job types, and it is predicted that there will be a 39% increase in demand for data scientists and data engineers by 2020, corresponding to around 364,000 more job openings [5].

With such a high demand for data scientists, it is crucial for employers to identify the best platforms for advertising their job listings. Traditional online job boards, like Craigslist, Monster and Indeed, contain ample data science job postings, but sometimes lack the search capabilities for candidates to easily identify and apply for relevant positions. Therefore, individuals interested in data science opportunities often utilize smaller, more specialized job boards [2]. Stack Overflow is a popular online community for developers and computer programmers to learn, share knowledge and discover career opportunities, with over 40 million visits to the site every month. A recent survey showed that about 15% of the site's users are actively searching for a job on Stack Overflow (26% of whom are students), and 78% are interested in hearing about job opportunities. Additionally, 57% of visitors check Stack Overflow multiple times a day, making it a desireable site for employers seeking data scientists to advertise job openings [6].

Research Aim

The purpose of this paper is to examine trends in job postings for "data scientists" on the Stack Overflow job board. This involves determining the most common computing skills that employers look for, along with their preferences of degree types and areas of study. This paper will also locate geographic regions where data science jobs are in highest demand, and if there are substantial differences in job

characteristics by location. Finally, trends of these characteristics of job listings will be explored over time.

Methods

Data Collection

Data were made privately available from a data scientist at Stack Overflow. The provided data consist of information from jobs posted on the Stack Overflow job board that either have "data scientist" or "data analyst" in their title between August 25, 2010 and September 25, 2017 (n = 995). While company names were censored from the data, the following attributes of each posting were provided in a data frame: job title, original posting date (YEAR-MM-DD), associated tags indicating relevant skills, job location (City, State, Country), salary (when included), whether a company would sponsor a visa, allow remote work, or offer assistance with relocation, and the full text of job descriptions and requirements.

Additional information was manually extracted from the provided data. Latitude and longitude coordinates for each listing were determined by the specified location [3]. Preferences of academic backgrounds were extracted from the job requirements section [14]. This included any mentions of type of degree (Bachelors, Masters, PhD) along with mentions of favorable majors and departments. To detect relevant majors, a dictionary was compiled using a comprehensive list of STEM fields provided by Stemdegreelist.com. Additionally, for jobs that mentioned multiple degrees (i.e, "Bachelor's degree required, Master's degree preferred"), the "highest degree preferred" for a job listing was determined. For listings that did not provide job requirement sections, the job descriptions section was used to check for these attributes.

Exploratory Data Analysis

Exploratory data analysis was conducted to summarize the most commonly listed attributes in the job postings. Skill tags, areas of study, and job locations were tabulated across all postings and ranked to determine the most common skills sought by employers, and where the most employment opportunities were geographically located [15]. Hex maps were generated to view the distribution of the number of jobs posted by geographic location [13]. Change-in-ranking plots were created to visualize the changes in the top ten tags, areas of study, and job locations over the last five years [12]. Number of job postings were also tabulated by year and geographic region to determine if there were any changes in frequency of postings by region over time.

Statistical Analysis

In order to assess any differences in jobs by location, proportions of jobs that offer visa sponsorship, allow remote work, and assist with relocation were compared between jobs listed in the United States and Europe, the two geographic regions with the highest numbers of job listings, using two-sample t-tests. The distributions of highest degree preferred were compared across regions with a Pearson's Chi-squared test [8].

Results

The number of yearly data science job listings posted on Stack Overflow has increased over time, most notably more than doubling between 2013 (n=75) and 2014 (n=174) (Figure 2). Figure 1 displays the overall top ten skill tags, areas of study, and cities, while Figure 4 contains the "top ten" lists by year, and depicts how the rankings have changed over time. Differences between jobs listed in the United States and jobs listed in Europe are described in Table 1, and the geographic distributions of jobs in both regions are portrayed in Figure 3.

Skill Tags

The top three computing skills listed as tags on job listings are Python, R and SQL (Figure 1a). Of the 995 jobs listed, 448 jobs (45%) use the `Python` tag, 281 jobs (28.2%) use the `R` tag and 249 jobs (25%) use the `SQL` tag. While Python has consistently been the most tagged skill in data science job postings, R has become increasingly more important to employers hiring data scientists over the last three years, as it has jumped from the 5th-most frequent tag to the 2nd-most frequent tag from 2013 to 2017 (Figure 4a). Similarly, knowledge of machine learning algorithms has gained more popularity among employers hiring data scientists in a similar timeframe.

Areas of Study

Employers' focus on coding abilities are also highlighted in the top three preferred areas of study for data science job candidates (Figure 1b): Computer Science (n = 469, 47.1%), Statistics (n = 436, 43.8%) and Engineering (n = 301, 30.3%). Few to no changes in preferred areas of study of job candidates have occurred in the past five years, indicating that employers hiring data scientists have consistently sought candidates with quantitative and programming-based backgrounds (Figure 4b).

Location

In both the United States and Europe, the highest number of job listings appear to cluster regions where there are major cities. As seen in Figure 3, darker reds appear over big cities like New York and San Francisco (Figure 3a), and London and Berlin (Figure 3b), indicating higher numbers of job listings, while less dense and less industrial areas either have lighter shades of yellow or are white, indicating few to zero data science job opportunities. Interestingly, while New York and San Francisco are the two cities with the most overall job listings (Figure 1c), and have consistently ranked in the top two cities between 2013-2016, they have dropped to the fourth and fifth spots in 2017, as European cities like Berlin and London have risen to the top (Figure 4c). This geographic trend is also noticable in figure 2, where it is apparent that the proportion of jobs posted in 2017 that are located in Europe is much larger than that of earlier years.

While there are similarities in the types of cities with the most jobs in the US and Europe, there are several differences between the job benefits and characteristics by region. European employers are more likely to offer visa sponsorship (EU: 20.8%, USA: 5.9%, $p < .01$) and to offer assistance with relocation (EU: 35.5%, USA: 26.6%, $p < .01$) than US employers. US employers are more likely to allow employees to work remotely (USA: 9.2%, EU: 2.8%, $p < .01$) and offer more jobs for candidates with Bachelor's Degrees only ($p < .01$) than European employers (Table 1).

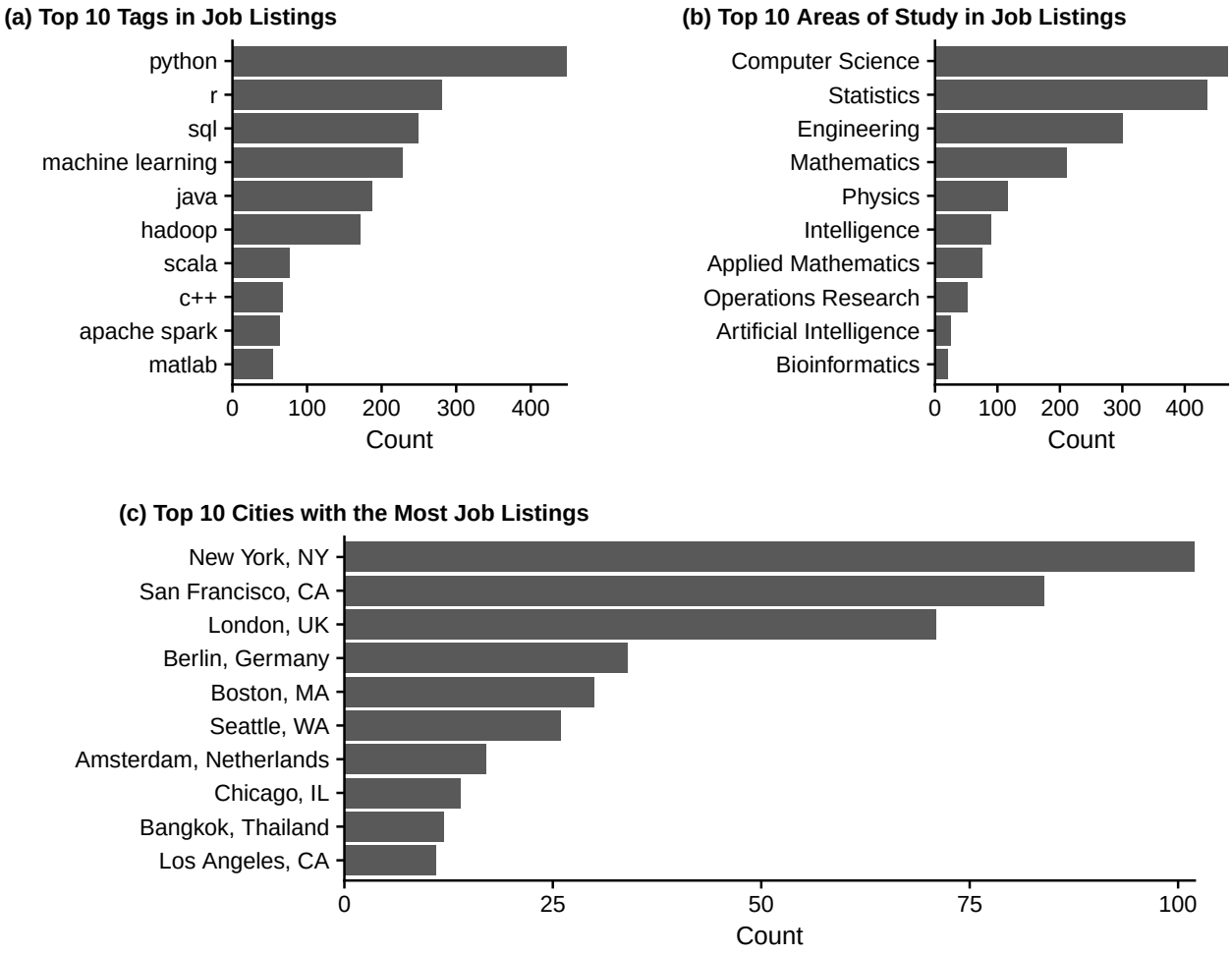

Figure 1. Most Popular Attributes of Job Listings. (a): The most commonly tagged computing skills are Python, R, SQL and Machine Learning. (b): The most commonly sought academic backgrounds are Computer Science, Statistics, Engineering and Mathematics. (c): The cities with the most overall job listings are New York, San Francisco, London and Berlin.

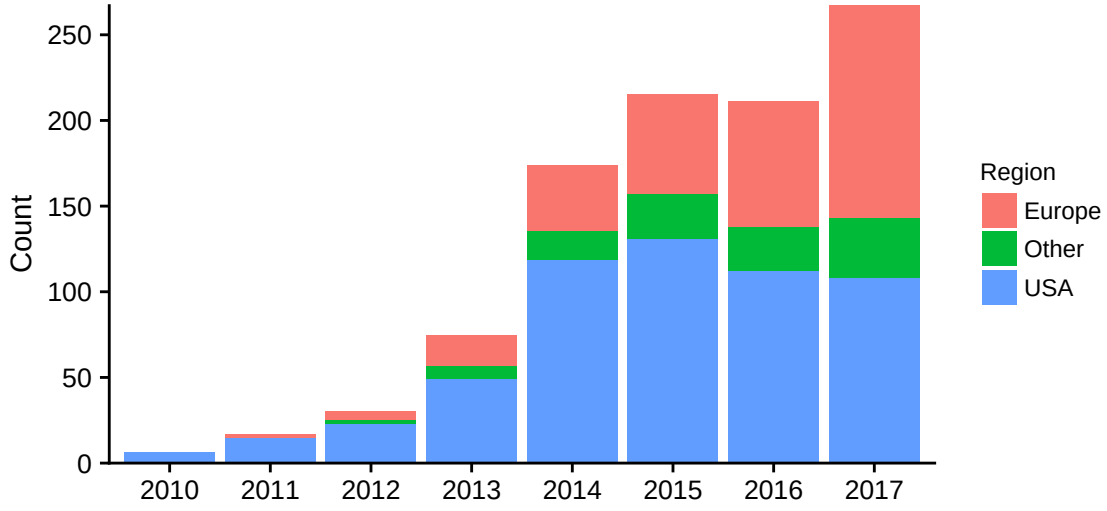

Figure 2. Number of Job Listings by Year and Geographic Region. The number of jobs posted on Stack Overflow increased yearly (except in 2016), with the greatest increase occurring between 2013 and 2014. The proportion of jobs located in Europe has increased, particularly over the last four years, while the proportion of jobs located in the United States has remained steady or declined in the same timeframe.

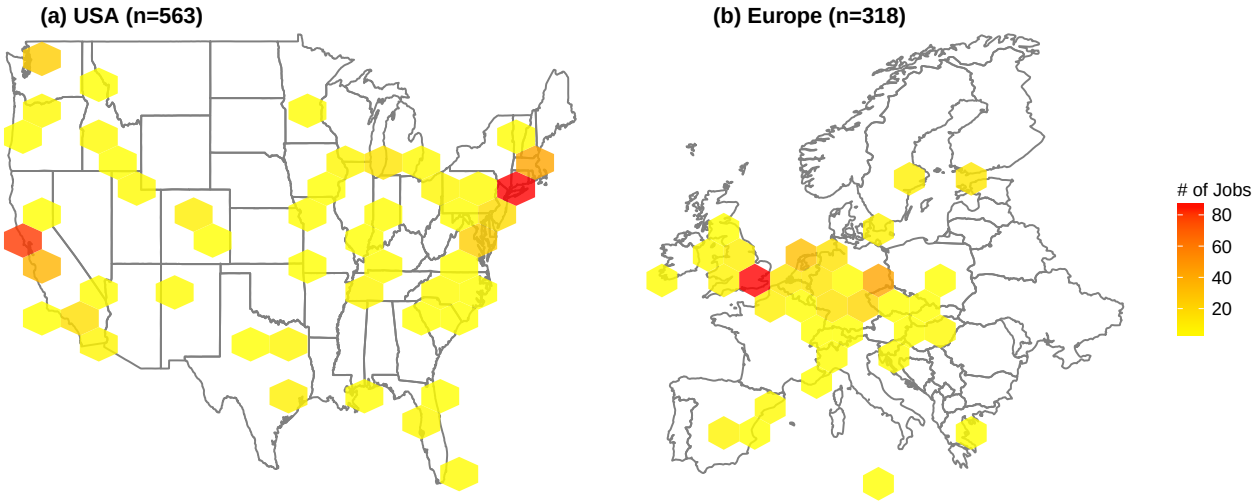

Figure 3. Geographic Distribution of Jobs in the United States vs. Europe. (a): Of the jobs located in the United States, the most jobs appear to be in major cities on the east and west coasts. (b): Of the jobs located in Europe, the most jobs appear to be in Western and Central Europe.

	USA	Europe	P-value
Visa Sponsorship	33 (5.9%)	66 (20.8%)	3.81e-11
Allows Remote Work	52 (9.2%)	9 (2.8%)	5.42e-04
Offers Relocation	150 (26.6%)	113 (35.5%)	7.08e-03
Highest Degree Preferred:[1]			
Bachelors	122 (35%)	26 (19.4%)	3.32e-03
Masters	79 (22.6%)	34 (25.4%)	
PhD	148 (42.4%)	74 (55.2%)	

[1] Due to missingness, percents are calculated from totals of 349 for USA and 134 for Europe.

Table 1. Differences between Job Listings in the United States and Europe. European job opportunities offer more visa sponsorships than those in the United States (20.8% vs. 5.9%), as well more assistance with relocation (35.5% vs. 26.6%). More job listings in the United States allow candidates to work remotely than those in Europe (9.2% vs. 2.8%), and there are more job listings for candidates with only Bachelor's degrees in the United States than in Europe (35% vs. 19.4%).

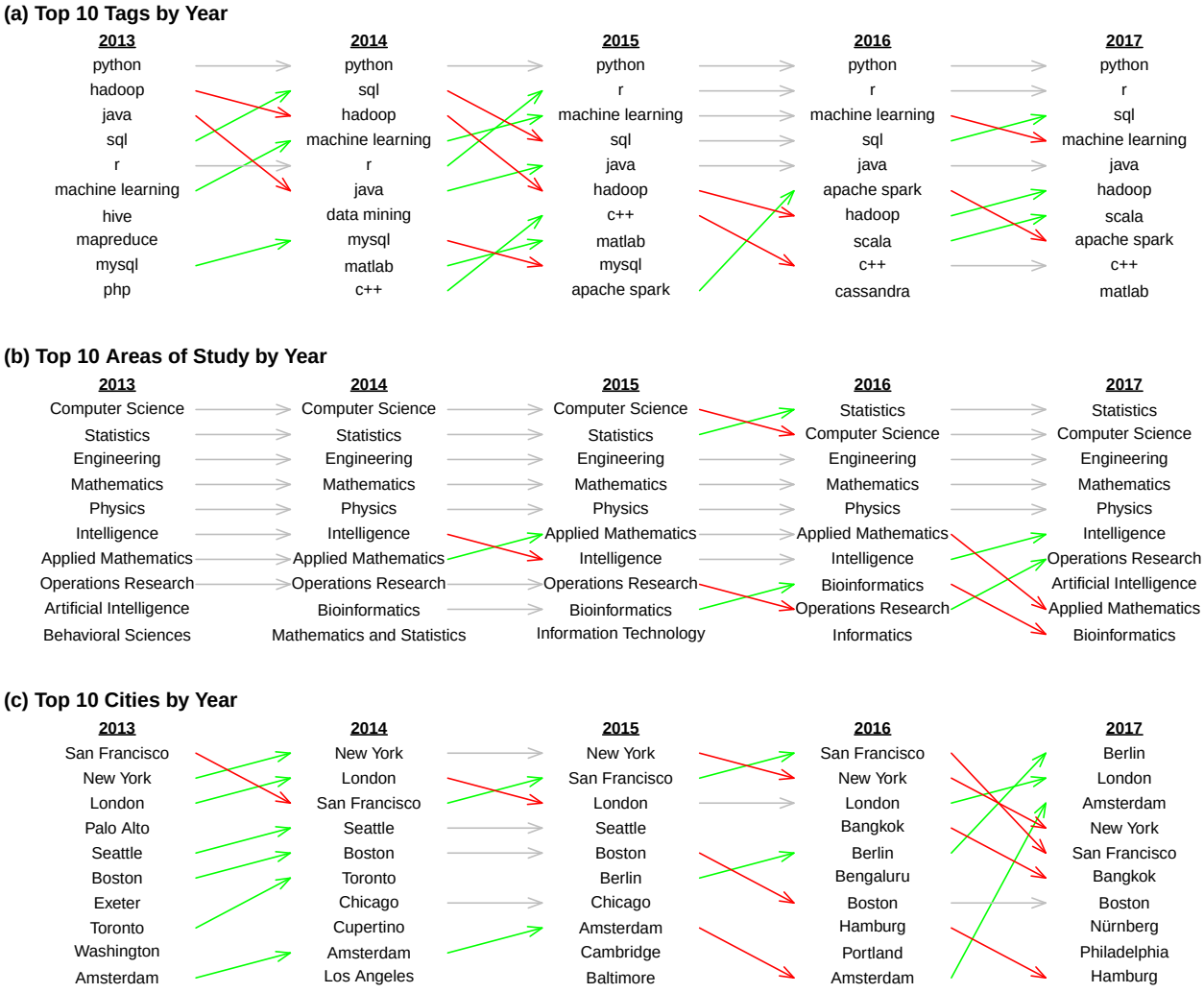

Figure 4. Changes in Job Listing Attributes over the Last Five Years. (a): While Python has consistently been the top skill tag on job listings, programming languages like R and SQL, along with Machine Learning algorithms, have risen to the top over the last three years. (b): There has been little to no change in the top ten areas of study preferred by employers over the last five years. (c): While New York and San Francisco have had the most job listings between 2013-2016, 2017 is dominated by European cities like Berlin, London and Amsterdam.

Limitations/Discussion

There are several limitations to this work. First, these results do not generalize to all data science job listings on job boards and company pages. The trends seen here are specific to the job board on Stack Overflow, and may be biased towards particular industries, geographic regions, or any mechanism by which Stack Overflow obtains job listings to post on their site. Second, there is potential for under-estimation in the prevalence of computing skills in job listings. Skills were detected through tags on the job listings; it is possible that additional skills and computing languages were mentioned in the job

description or requirements sections. Similarly, there is potential for error in scraping job descriptions for areas of study and degree types, since there may be some variability in how job descriptions are written and structured. Finally, since the data for 2017 are incomplete, trends over time may change once full data for the year are available. It is possible that different regions have different seasonal trends in the frequency of posting job opportunities.

Given these limitations, this paper provides some interesting insight into the optimal characteristics of a qualified candidate for a data science position. Furthermore, the trends presented in this paper are consistent with broader changes in the popularity of statistical analysis technologies. Python has been claimed to be the fastest growing major programming language, with a year-over-year-growth in online traffic to Stack Overflow questions related to Python of 27% [10]. R programming has grown at a similar year-over-year rate, though its growth in tech and data science has been less than that of Python [9] [11]. Still, as employers list both Python and R programming more frequently as desired skills, individuals interested in data science jobs should be sure to supplement their quantitative backgrounds with coding expertise.

This paper also highlights the ongoing rise of startup culture in Europe, as both investment in tech and opportunities for data scientists increase in the region. London and Berlin, which lead 2017 with the most data science job listings, have promising growth in their tech sectors [1], while many believe that Eastern Europe may be the next location for talent and investment in tech as well [4]. With higher rates of visa sponsorship and relocation assistance than job listings in the United States, European opportunities should not be overlooked by those searching for data science jobs.

Acknowledgements

A big thank you to David Robinson for providing job listings data from Stack Overflow.

References

[1] Caroline Baldwin. *Berlin plays growing role in Europe's startup race for innovation*. URL: http://www.computerweekly.com/feature/Berlin-and-the-European-startup-race-for-innovation.

[2] Manu Jeevan. *20 Websites To Find Data Science Jobs*. May 2017. URL: https://www.springboard.com/blog/data-science-jobs/.

[3] David Kahle and Hadley Wickham. "ggmap: Spatial Visualization with ggplot2". In: *The R Journal* 5.1 (2013), pp. 144–161. URL: http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf.

[4] Michael Kahn. *Eastern European tech entrepreneurs are reviving the region's startup scene*. May 2017. URL: http://www.independent.co.uk/Business/indyventure/eastern-europe-tech-secotr-entrepreneurs-startup-global-investors-funds-growth-a7752546.html.

[5] Steven Miller and Debbie Hughes. *The Quant Crunch: How the Demand for Data Science Skills Is Disrupting the Job Market*. 2017.

[6] Stack Overflow. *Developer Survey Results 2016*. 2016. URL: https://insights.stackoverflow.com/survey/2016.

[7] D.J. Patil and Thomas H. Davenport. "Data scientist: the sexiest job of the 21st century". In: *Harvard Business Review* (2012).

[8] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2017. URL: https://www.R-project.org/.

[9] David Robinson et al. *The Impressive Growth of R*. Oct. 2017. URL: https://stackoverflow.blog/2017/10/10/impressive-growth-r/.

[10] David Robinson et al. *The Incredible Growth of Python*. Sept. 2017. URL: https://stackoverflow.blog/2017/09/06/incredible-growth-python/.

[11] David Robinson et al. *Why is Python Growing So Quickly?* Sept. 2017. URL: https://stackoverflow.blog/2017/09/14/python-growing-quickly/?utm_source=so-owned&utm_medium=blog&utm_campaign=gen-blog&utm_content=blog-link&utm_term=growth-r.

[12] *Simplest way to plot changes in ranking between two ordered lists in R?* URL: https://stackoverflow.com/questions/25781284/simplest-way-to-plot-changes-in-ranking-between-two-ordered-lists-in-r.

[13] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. ISBN: 978-0-387-98140-6. URL: http://ggplot2.org.

[14] Hadley Wickham. *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.2.0. 2017. URL: https://CRAN.R-project.org/package=stringr.

[15] Hadley Wickham et al. *dplyr: A Grammar of Data Manipulation*. R package version 0.7.4. 2017. URL: https://CRAN.R-project.org/package=dplyr.

[16] Paul Zikopoulos et al. *Harness the power of big data The IBM big data platform*. McGraw Hill Professional, 2012.