Trends in Data Science Job Postings on Stack Overflow

Benjamin Ackerman

October 9, 2017

Introduction

- Increased demand for data scientists, hot new field, Stack Overflow's popularity

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 upon request from David Robinson, 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. 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), and the full text of job descriptions and requirements.

While a fair amount of variables were already provided in a dataframe, additional information was extracted from the data and cleaned. The `geocode` function in the `ggmap` package was used to gather latitude and longitude coordinates for each job location. Preferences of academic backgrounds were extracted from the job requirements section. 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**. 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 years of postings and ranked to determine the most common skills sought by employers, and where the most employment opportunities were geographically located.

- Exploratory data analysis:
  - Counted number of times each degree type was listed
  - Selected "highest degree" as "degree most preferred" in job listings
  - Tabulated number of job postings by month and year
- Generated hex maps of US and Europe to view geographic distribution of jobs

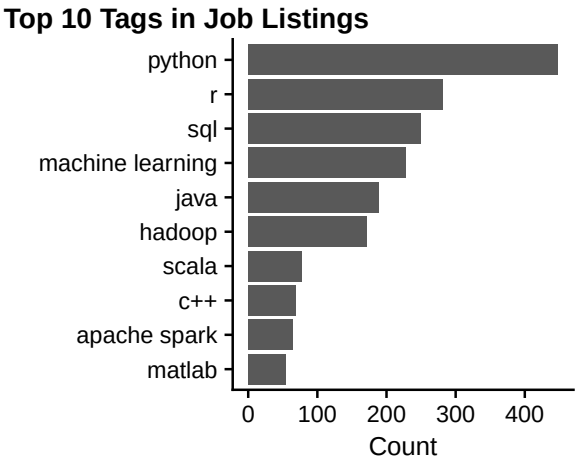
Statistical Analysis

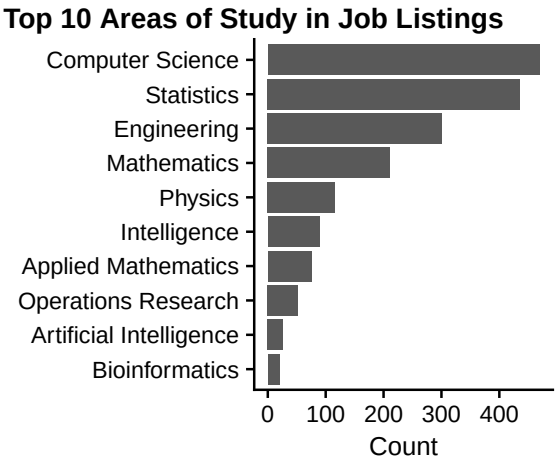
In order to assess any differences in jobs by location, proportions of jobs that offer visa sponsorship, allow remote work, and assist with relocation were compared between jobs listed in the US and Europe, the two geographic regions with the highest numbers of job listings, using two-sample t-tests.

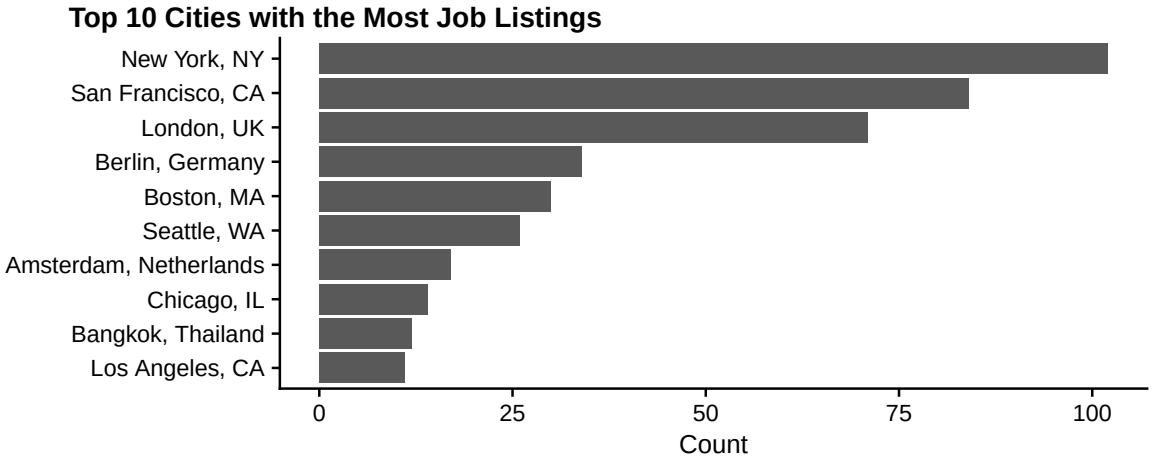
- Chi-Squared test for highest degree

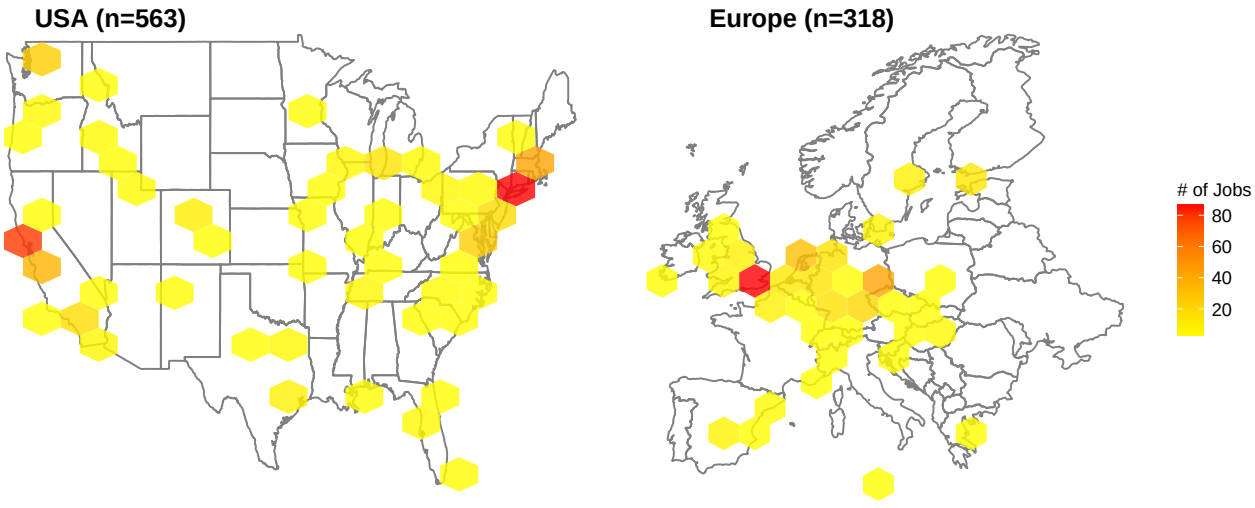
Results

Most Popular Attributes of Job Listings from 2010-2017:









Changes in Attributes of Job Listings over the Last 5 Years (2013-2017):

Table 1: Differences in Job Listings in the US vs. 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 Required:[1]			3.32e-03
Bachelors	122 (35%)	26 (19.4%)	
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 2: Top 10 Tags in Job Listings by Year

2013	2014	2015	2016	2017
python	python	python	python	python
hadoop	sql	r	r	r
java	hadoop	machine learning	machine learning	sql
sql	machine learning	sql	sql	machine learning
r	r	java	java	java
machine learning	java	hadoop	apache spark	hadoop
analytics	data	c++	hadoop	scala
hive	analytics	matlab	scala	apache spark
mapreduce	data mining	data	c++	c++
mysql	mysql	mysql	cassandra	matlab

Table 3: Top 10 Cities with the Most Job Listings by Year

2013	2014	2015	2016	2017
San Francisco	New York	New York	San Francisco	Berlin
New York	London	San Francisco	New York	London
London	San Francisco	London	London	Amsterdam
Palo Alto	Seattle	Seattle	Bangkok	New York
Seattle	Boston	Boston	Berlin	San Francisco
Boston	Toronto	Berlin	Bengaluru	Bangkok
Exeter	Chicago	Chicago	Boston	Boston
Toronto	Cupertino	Amsterdam	Hamburg	Nürnberg
Washington	Amsterdam	Cambridge	Portland	Philadelphia
Amsterdam	Los Angeles	Baltimore	Amsterdam	Hamburg

Discussion/Limitations

Discussion:

- More quantitative fields dominate top areas of study requested
- While Python has consistently been most tagged skill, R and machine learning have become increasingly more important to data science jobs in the last three years (as they move up the rankings)
- Most US jobs are in big cities like New York, San Francisco, Boston, and Chicago. While most jobs overall are located in the US, in the last two years, European cities have been becoming increasingly more popular.
- There are significant differences between US and European jobs: US allows more remore work, and more jobs for people with Bachelors degrees, while European jobs sponsor more visas and offer more compensation for relocation.

Limitations:

- results do not generalize to all data science job boards - Stack Overflow could attract jobs from particular industries, could be biased in this way
- skills are detected through tags and not through what is mentioned in the text

Conclusion