Syracuse University

Developer Profiles and Job Market Analysis

Sandra Tang and Emma Woods

2020-0108:37884 IST 652 Scripting for Data Analysis

Dr. Debbie Landowski

March 15th, 2019

TABLE OF CONTENTS

[**INTRODUCTION**](#_23aafwx84ckc) **3**

[Research Questions & Datasets](#_g89v9urmhbgt) 3

[2019 Stack Overflow Developer Survey Results](#_3datbpj514mo) 3

[Developer Job Markets (Stack Overflow, GitHub, Working Nomads)](#_tcribpovgi98) 4

[**2019 STACK OVERFLOW DEVELOPER SURVEY**](#_6rl4nf8plyfy) **5**

[About the Data](#_3tswgyvpu2hh) 5

[Data Import, Exploration, and Pre-Processing](#_z52z5sd0mu9) 5

[Data Analysis & Results](#_iafxgbby6vqz) 7

[Developer Demographic Profiles](#_8dpswprfp37d) 7

[Compensation by Developer Demographics](#_480wpg4ic07t) 8

[Age at which Developers First Start Learning to Code](#_etidaucr0089) 11

[Compensation Comparison According to Professional Coding Experience](#_3928leoork4) 11

[Analysis of Job Factors Ranked Most Desirable by Developers](#_i15ec8px3cki) 14

[Developers Technical Working Experience and Desires](#_rw04ytjo9ivw) 15

[**STACK OVERFLOW JOBS**](#_nmftxj1dc7ta) **16**

[About the Data](#_rwi1u21t2i1v) 16

[Data Analysis & Results](#_t5ivqqs2k226) 18

[Python Job Listings by Day Published or Location](#_psmdaddsxah1) 18

[Python Job Roles Comparison in the Top 2 States](#_kumloqaf7e49) 19

[Analysis of R Job Postings](#_rms9mifjo8f0) 20

[Python Job Descriptions Analysis](#_7gfp7pr9krld) 22

[**GITHUB JOBS**](#_xutx3ivlq1jf) **25**

[About the Data](#_bt6gnq43r7vd) 25

[Data Import, Exploration, and Pre-Processing](#_4woh48rjdia7) 25

[Data Analysis & Results](#_k14tsnwpt5o9) 27

[Github Job Listings by Region](#_91bx9reqenuy) 27

[Analysis of Job Titles for U.S. Based Positions](#_jgybtccmjaut) 27

[Github Job Listings by Day of the Week](#_kwmuximp8wqj) 28

[Exploration of Github Job Descriptions](#_zhplztm4ywwu) 29

[Unstructured Text Processing](#_4pv21znnxpl5) 29

[Analysis and Comparison of Job Factors listed within Github Jobs Postings](#_limhnq5g6ny0) 31

[**WORKING NOMADS**](#_8cjmm4jj89xh) **33**

[Data Import, Exploration, and Pre-Processing](#_kyok52roycmo) 33

[Data Analysis & Results](#_3we5xh1c33kc) 34

[Working Nomads Job Listings by Region](#_k7lxflky1gqp) 34

[Analysis of Job Titles for U.S. Based Positions](#_gndhum64i39u) 35

[Comparison of Job Categories in USA versus Remote](#_7ahg5vf9tezs) 36

[Working Nomads Job Listings by Day of the Week](#_bluwxhxp0fzn) 37

[Exploration of Working Nomads Job Descriptions](#_mazl93m91hdy) 37

[Unstructured Text Processing](#_s3wuxi25imir) 38

[Analysis and Comparison of Job Factors within Working Nomads Postings](#_xx1ndlpw0p80) 39

[**FINAL CONCLUSION**](#_ajegjipyi9vh) **41**

[**PYTHON PROGRAM DESCRIPTION**](#_88iq34t5rku) **43**

[**TEAM MEMBERS, ROLES & TASKS**](#_36xr12ilh3y6) **45**

# INTRODUCTION

U.S. News reports that of the “25 Best Jobs of 2020”, the number one spot goes to software developers. After gathering data on nearly 200 jobs from the federal Bureau of Labor Statistics, this research concluded that currently software developers enjoy higher median salaries, more job growth, and a greater ability to maintain a good work-life balance than any other job analyzed (<https://money.usnews.com/money/careers/slideshows/the-25-best-jobs>).

This project analyzes the demographics, traits, preferences and other factors present within the developer community as identified by the Stack Overflow Developer Survey results published in 2019. The current job market for developers is also explored in order to ascertain whether the demographics, preferences and other factors identified within this developer community are reflected in the job opportunities currently on offer by the companies seeking to hire.

## 

## Research Questions & Datasets

In order to profile software developers and the current job market environment, the research for this project was conducted using a combination of sources:

1. Structured dataset containing all published results of the 2019 Stack Overflow Developer Survey
2. Unstructured job market data sources retrieved via API

Following is a high-level summary of each dataset, including methods of analysis and research questions explored.

### 2019 Stack Overflow Developer Survey Results

Stack Overflow publishes their annual developer survey results as CSV files; therefore, descriptive analysis is the main method used to analyze this structured dataset. By exploring averages and summaries of different data dimensions and aggregation levels, it is possible to identify different demographics, salary, and coding background of the survey participants and see how different groups compare and contrast with each other. The key questions explored with this data are:

* Who are the participants of the 2019 Developer Survey? What are their demographics and background? Are they a good representation of developers and people who code?
* For respondents in software development and data science-related professions, how does salary correspond with geographical location, gender, age, organization size, and professional coding experience?
* Does gender have any impact on the age respondents first start learning how to code, whether they are in a coding profession, and in the salary compensation they receive?
* What additional job factors do survey respondents find most important?

### 

### Developer Job Markets (Stack Overflow, GitHub, Working Nomads)

To provide “real-world” and up-to-date context around the career market for developers, this research also explores current job postings. There are many job boards dedicated to attracting developer talent; a small number enable download access to job search results as semi-structured datasets (XML and JSON) which contain job descriptions stored in unstructured fields (HTML).

Diving into job postings requires the use of analysis methods for semi-structured dataset to traverse and capture the tree and element tags. Once the data is stored in a database or a dataframe, patterns can be identified by querying the organized schema. In addition, text analysis methods are also helpful to analyze any unstructured data by looking at word frequency distributions like raw word counts or percentage of words within a job posting or groupings. Applying natural language processing methods can help to uncover insights that are not otherwise retrievable, such as:

* What is the developer job market landscape?
* How do different job sites compare for Python and R job seekers?
* What key development requirements and skills are companies hiring for?
* Are there correlations between the skill sets reflected in the developer survey and what hiring companies are looking for?

# 

# 2019 STACK OVERFLOW DEVELOPER SURVEY

## About the Data

Stack Overflow is a large, trusted online community for developers to learn, share programming knowledge, and build careers. With nearly 90,000 responses fielded from over 170 countries and dependent territories, their 2019 Annual Developer Survey examines all aspects of the developer experience from career satisfaction and job search to education and opinions on technology. Official published survey results can be found in this URL on their website:<https://insights.stackoverflow.com/survey>**.** There are 2 main data files associated with the survey results:

1. survey\_results\_public.csv - CSV file with main survey results, one respondent per row and one column per answer
2. survey\_results\_schema.csv - CSV file with survey schema, i.e., the questions that correspond to each column name

## Data Import, Exploration, and Pre-Processing

The 2019 survey results file is loaded into a pandas dataframe for exploration. Here is some high-level information obtained by examining the dataframe using different pandas functions and methods.

* There are 88,883 rows. Each row represents one survey submission by respondent.
* There are 85 columns. The first column is an ID field that uniquely identifies each respondent. Each of the remaining 84 columns represents a different survey question.
* The ID field is called “Respondent” column does not contain nulls or duplicates and therefore, can be leveraged as index. The remaining fields may contain null data where the answer was left blank.
* Many questions are multiple choice where the answers are stored in the same field as a list of items separated by semicolons (;)

For the purpose of this research, the entire response population is used but only a smaller scope of survey questions is selected. The following table represents the subset of 19 survey questions, the definition of each column (survey question) along with the completeness based on the number of null values from the 88,833 responses.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Id** | **Column** | **NaN** | **Completeness** | **Definition** |
| 0 | Respondent | 0 | 100.00% | Randomized respondent ID number (not in order of survey response time) |
| 1 | MainBranch | 552 | 99.38% | Which of the following options best describes you today? Here, by "developer" we mean "someone who writes code." |
| 2 | Employment | 1702 | 98.08% | Which of the following best describes your current employment status? |
| 3 | Country | 132 | 99.85% | In which country do you currently reside? |
| 4 | Student | 1869 | 97.90% | Are you currently enrolled in a formal, degree-granting college or university program? |
| 5 | EdLevel | 2493 | 97.19% | Which of the following best describes the highest level of formal education that you completed? |
| 6 | UndergradMajor | 13269 | 85.06% | What was your main or most important field of study? |
| 7 | YearsCode | 945 | 98.94% | Including any education, how many years have you been coding? |
| 8 | YearsCodePro | 14552 | 83.62% | How many years have you coded professionally (as a part of your work)? |
| 9 | CareerSat | 16036 | 81.95% | Overall, how satisfied are you with your career thus far? |
| 10 | JobSat | 17895 | 79.86% | How satisfied are you with your current job? (If you work multiple jobs, answer for the one you spend the most hours on.) |
| 11 | JobSeek | 8328 | 90.63% | Which of the following best describes your current job-seeking status? |
| 12 | LastHireDate | 9029 | 89.84% | When was the last time that you took a job with a new employer? |
| 13 | JobFactors | 9512 | 89.29% | Imagine that you are deciding between two job offers with the same compensation, benefits, and location. Of the following factors, which 3 are MOST important to you? |
| 14 | ConvertedComp | 33060 | 62.78% | Salary converted to annual USD salaries using the exchange rate on 2019-02-01, assuming 12 working months and 50 working weeks. |
| 16 | LanguageWorkedWith | 1314 | 98.52% | Which of the following programming, scripting, and markup languages have you done extensive development work in over the past year? |
| 17 | LanguageDesireNextYear | 4795 | 94.60% | Which of the following programming, scripting, and markup languages do you want to work in over the next year? |
| 18 | Age | 9673 | 89.11% | What is your age (in years)? |
| 19 | Gender | 3477 | 96.09% | Which of the following do you currently identify as? Please select all that apply. |

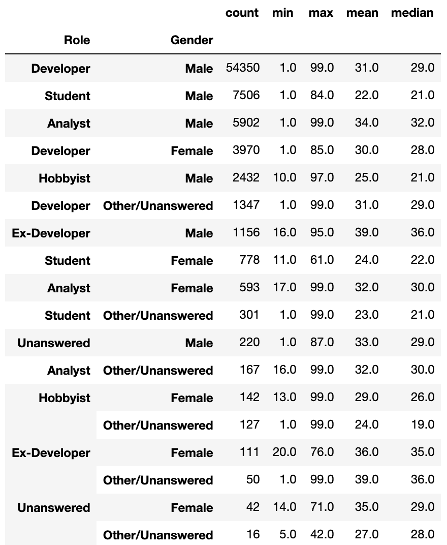
## Data Analysis & Results

### Developer Demographic Profiles

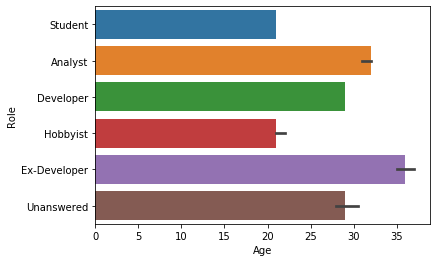
Even though Stack Overflow claims that their 2019 Developer Survey is the largest and most comprehensive survey of people who code around the world, they acknowledge that the survey results may not represent everyone in the developer community evenly.

Demographics of the 2019 survey respondents can be obtained by summarizing respondent counts by role, gender, and age. The majority of survey respondents are male developers. The remaining respondents identify themselves as students, analysts, ex-developers, and hobbyists.

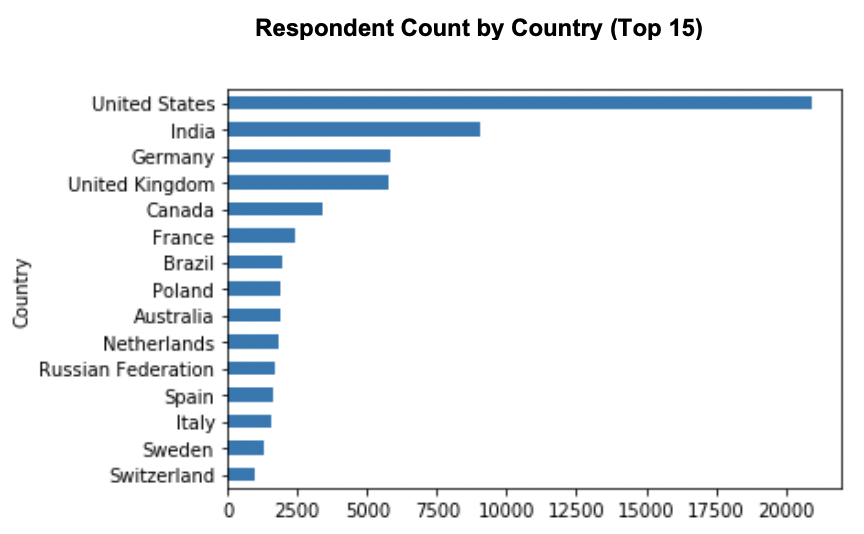
Below is a breakdown of respondent counts by role and gender, along with the minimum, maximum, mean, and median age for each group. Due to some irregular responses (e.g. minimum age of 1 and maximum age of 99), the median age is displayed in the chart on the left. Regardless of using median or average age, the comparison remains similar. Students and hobbyists are the youngest age groups; developers are in the middle; analysts are slightly older than developers, and ex-developers represent the oldest group.



**Median Respondent Age by Role**

****

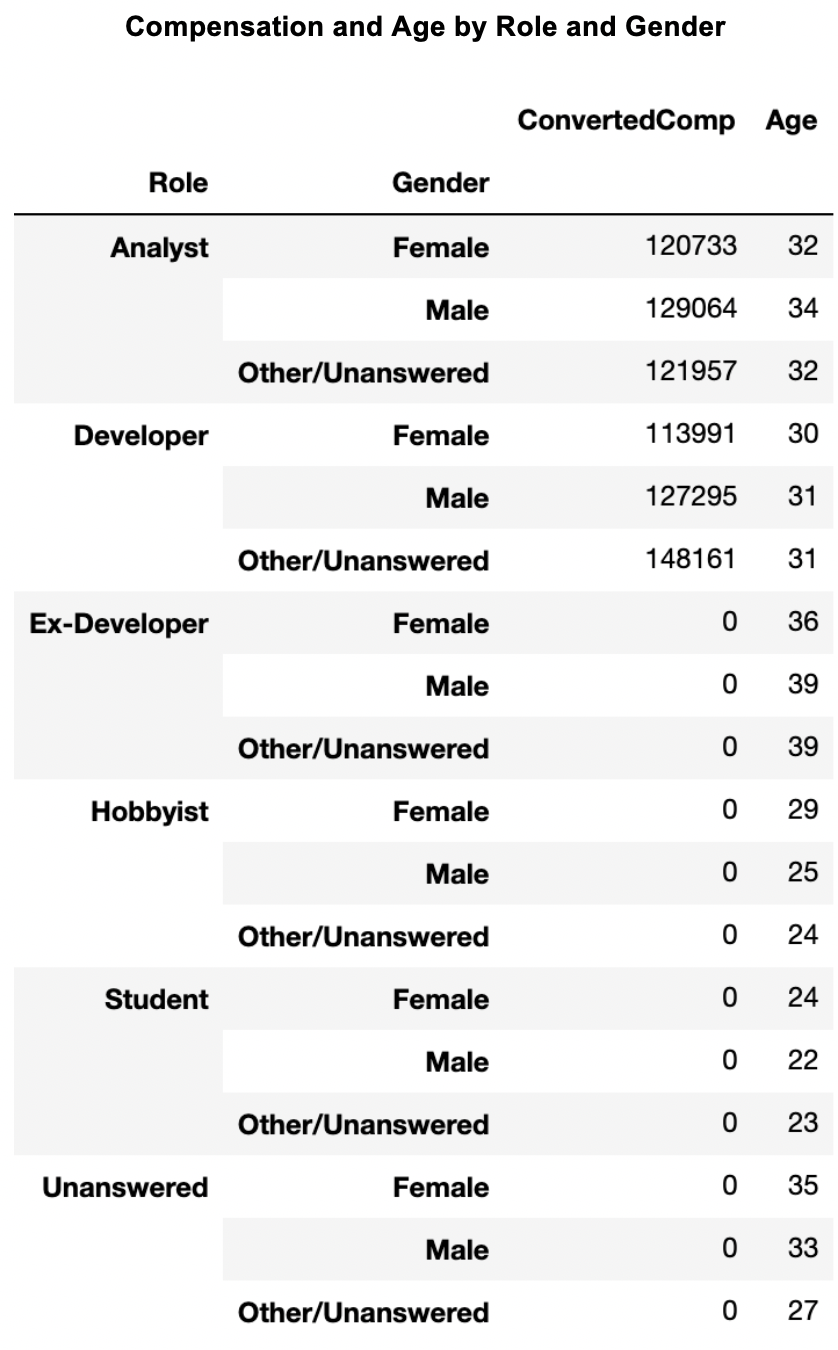
As for country of residence, more than 20% of survey respondents reside in the United States. Other popular countries are in India, Germany, United Kingdom, Canada, and France. The chart below shows the top 15 countries where survey respondents live and how many of them live there.



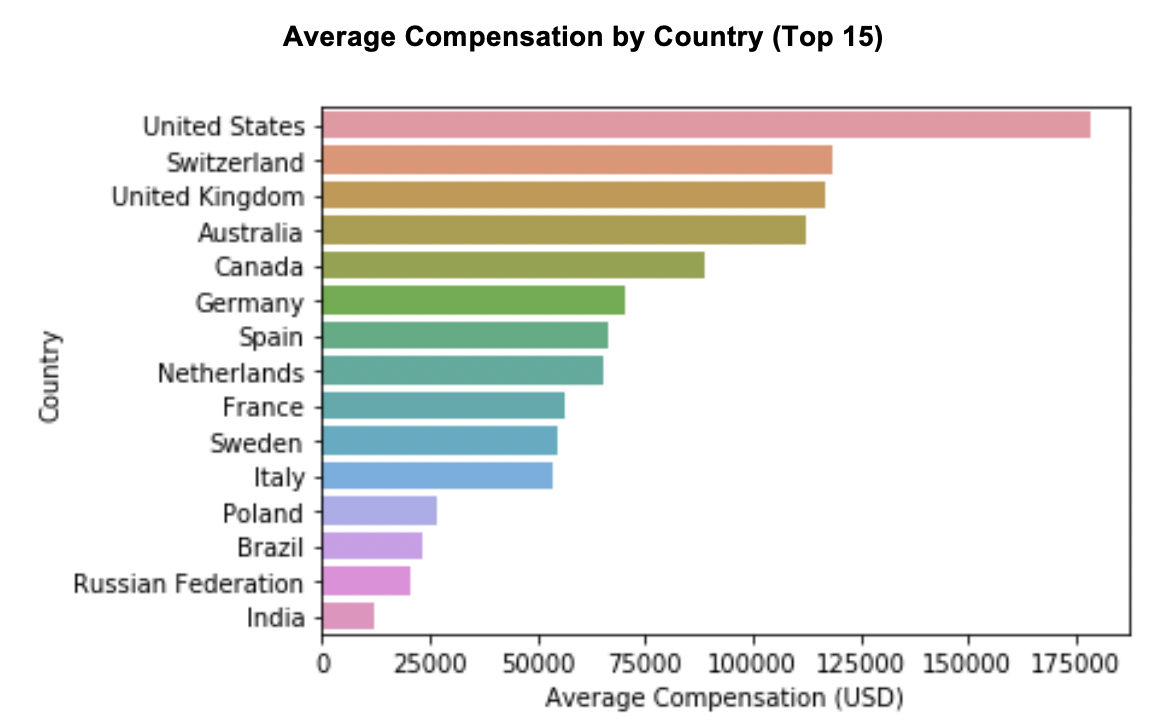
### 

### Compensation by Developer Demographics

The following table illustrates the average compensation and age by role and gender. From a comparison standpoint, it makes sense that compensation data is only available for the developer and analyst groups. After reviewing the compensation comparison, women seem to get paid less than men regardless of their job role. It is also noted that close to 24% of developers and analysts (82% of survey respondents) did not disclose their compensation information.



Besides reviewing compensation by role and gender, it is also interesting to see how compensation differs by country of residence. The following graph shows the average compensation for the top 15 countries ranking from highest to lowest. From the list of top 15 countries, respondents from the United States have the highest average compensation, followed by Switzerland and the United Kingdom.



A more detailed view of compensation by country can be obtained by incorporating a breakdown of gender. Reviewing the minimum and maximum values reveals certain underlying data inconsistencies. For example, the max compensation for several countries are listed as 1 or 2 million USD. This warrants further understanding as to why such even numbers appear as responses.

For the majority of countries in this list, women’s compensation still appears lower than men’s. This is consistent with the prior findings of compensation by gender and role type.

### 

The following table shows the average compensation broken down by gender for the top 15 countries. The median compensation per group is also included as reference since the mean may be sensitive to outliers.

### 

### 

### Age at which Developers First Start Learning to Code

The age at which developers report writing their first line of code is analyzed to see whether trends exist either by age or gender. The following chart shows that in all age brackets, females report starting to code at a slightly older age than males. This is true for all age groups, with the **exception** of the youngest age bracket (18-30):

# 

The indication is that in 2019, girls are becoming interested in coding at a younger age, with the hopeful result that we will see a wave of young female coders entering the profession over the next few years.

### Compensation Comparison According to Professional Coding Experience

Salary averages according to professional coding experience are compared, grouping by organization size and country the developer resides in.

|  |  |  |
| --- | --- | --- |
|  | **Country** | **Median Salary** |
| 79 | Liechtenstein | 811188.0 |
| 118 | San Marino | 301788.0 |
| 140 | Timor-Leste | 229500.0 |
| 3 | Andorra | 171862.0 |
| 37 | Democratic Republic of the Congo | 110484.0 |
|  | --- | --- |
| 32 | Cuba | 3666.0 |
| 21 | Burundi | 3606.0 |
| 135 | Syrian Arab Republic | 2568.0 |
| 131 | Sudan | 2016.0 |
| 54 | Guinea | 1956.0 |
|  | | |

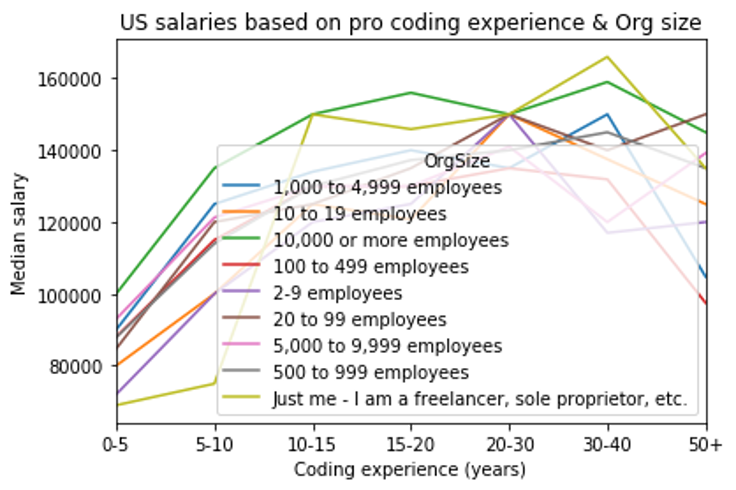
Liechtenstein reports the highest average salary of all countries. The United States does not appear in the top 5 (or bottom 5!) of the list.

Developer value counts are then assessed by Country, in order to provide a comparison of overall counts with the salary median values reported. The following table confirms that the United States has by far the highest number of respondents in the survey. Liechtenstein, which placed first according to salary average, does not make the top 10 list of respondents.

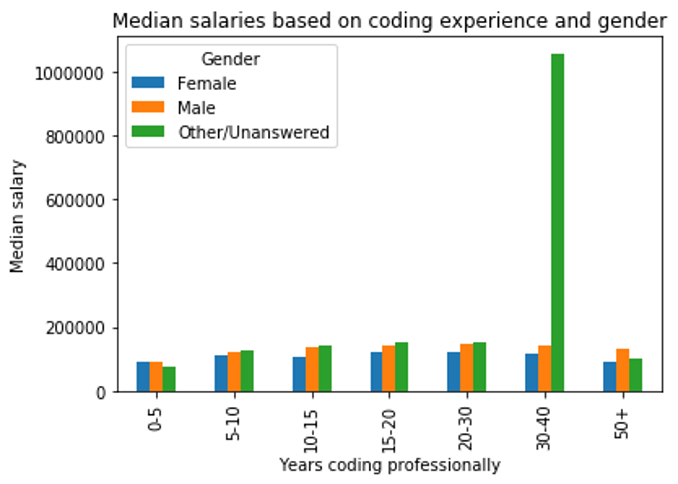
|  |  |
| --- | --- |
| **Country** | **Value Counts** |
| United States | 13748 |
| United Kingdom | 3772 |
| India | 3577 |
| Germany | 3529 |
| Canada | 2067 |
| France | 1563 |
| Brazil | 1327 |
| Poland | 1214 |
| Australia | 1150 |
| Netherlands | 1132 |

This indicates that while Liechtenstein reports a high median salary, there are a low number of developers that participated in the survey.

Diving deeper into the environment for U.S. based developers, salary results based on professional coding experience and gender are isolated to this region only. The chart generated from the results indicate that respondents with more than 10+ years of professional coding experience show the highest peaks for median salary values when they are freelancers or work at a company that has 10k+ employees.



Within the U.S. it appears that these career options provide the best route to improving salary earnings as developers continue to increase in their professional coding experience.



Comparing compensation by professional experience and gender reveals similar trends seen earlier; females receive lower compensation than males across all categories of professional coding experience.

# 

# 

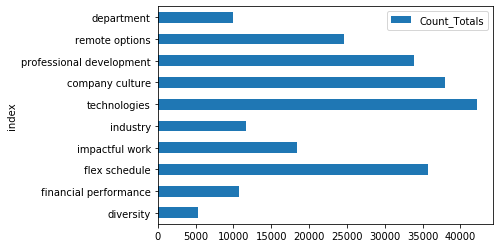
### 

### Analysis of Job Factors Ranked Most Desirable by Developers

Following analysis of the demographics, background, compensation, and coding experience of developers represented in the survey, their opinions regarding the most desirable job benefits in the workplace are analyzed. The ‘JobFactors’ variable within the dataset contains up to three entries from respondents, each selected from a predefined list in answer to the question:

*“Imagine that you are deciding between two job offers with the same compensation, benefits, and location. Of the following factors, which 3 are MOST important to you?”*

Entries within this variable are separated, categorized, sorted and counted in order to rank the job factors that respondents most frequently state are highly valuable to them.



The job factor listed with the highest frequency amongst respondents is the “languages, frameworks, and other technologies” that they will work with on the job. This indicates that the majority of developers are highly focused on applying the skills they have learned and that they rank the opportunity to develop new skills as highly desirable.

The next three most important factors within the rankings are:

* Office environment or company culture
* Flex time or a flexible schedule
* Opportunities for professional development

When the basics between jobs are comparable, the ability to develop technically and professionally, the office and company culture, and an ability for companies to provide flexibility rank most highly with respondents. These factors should therefore be taken into consideration by hiring companies when soliciting jobs within the highly competitive software development field.

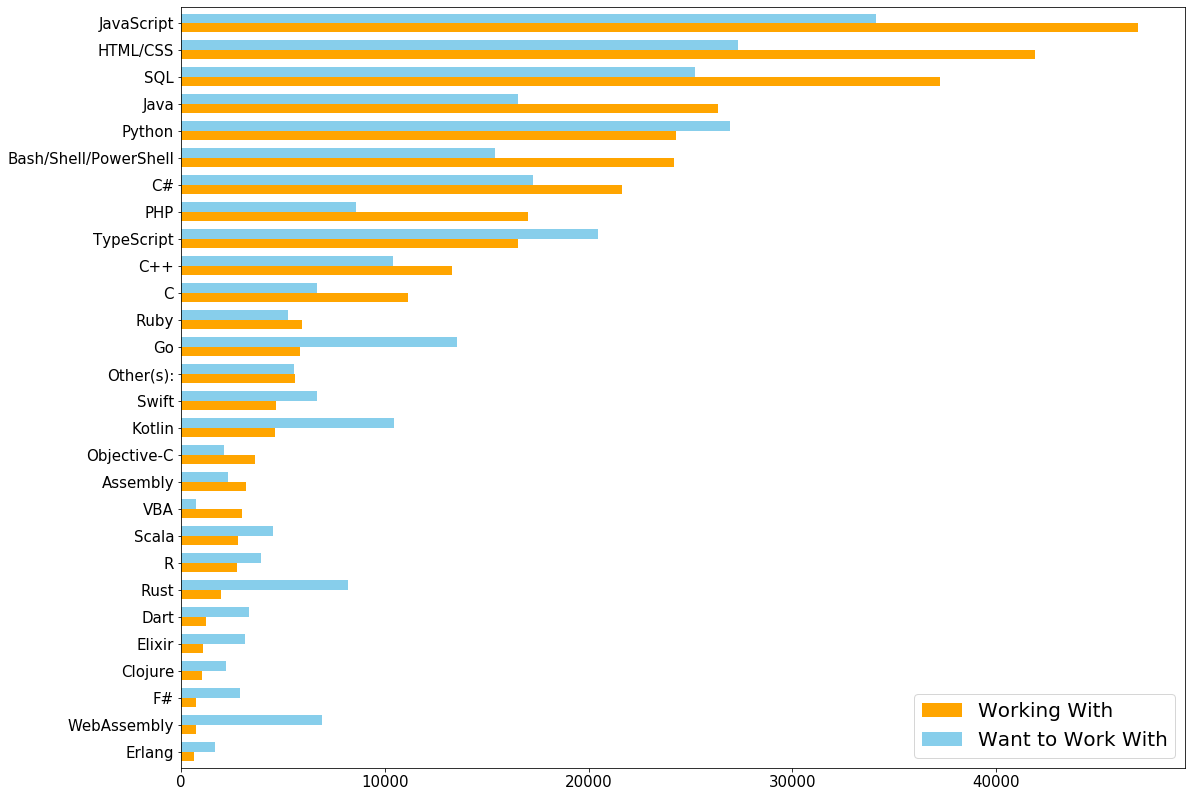
### Developers Technical Working Experience and Desires

As companies evolve to embed tech into the heart of their products, the need for skilled developers is growing exponentially. The Stack Overflow survey contains a set of questions that focuses on technology and tech culture and can be used to gauge the types of technical skills possessed by developers and coders. For this research question, answers to these 2 survey questions are analyzed.

*Which of the following programming, scripting, and markup languages have you done extensive development work in over the past year, and which do you want to work in over the next year?*

Since these are questions that allow multiple choice answers, the responses must first be flattened into lists before applying further grouping and frequency.

Below is a chart of the different programming, scripting, and markup languages that the developer and analyst groups say they work with compared to what they would like to learn over time. JavaScript is the most commonly used programming language, but Python, the fastest-growing major programming language today is surpassing Java in overall ranking. It is notable to see a surge of learning interest in TypeScript, which is an open-source programming language by Microsoft that extends JavaScript for application-scale development and Google Go, an open-source language with an “easy to learn” appeal for software development.



# STACK OVERFLOW JOBS

Online communities such as Stack Overflow not only help developers around the world connect with each other, they also serve as a marketplace for hiring companies and job seekers. Exploring information about career opportunities reveals insights ranging from which skills hiring companies find the most valuable to what they look for in an applicant and during the interview process.

## About the Data

Stack Overflow hosts a career site (<https://stackoverflow.com/jobs>) designed to help developers find the jobs they love. Even though the primary responsibility of Stack Overflow jobs is “writing software”, there are also openings for people who work closely with developers or have strong programming knowledge. Stack Overflow claims that their job site is designed to put the candidates first and that they will never present recruiter spam or fake job listings.

The datasets for this research section come from these 2 job searches performed on Feb 23rd, 2020.

1. Search for all Python jobs in the United States (within 100 miles – maximum selectable distance for a location)<https://stackoverflow.com/jobs/feed?l=United+States&d=100&u=Miles&q=python>
2. Search for all R jobs in the United States (within 100 miles – maximum selectable distance for a location)

<https://stackoverflow.com/jobs/feed?l=United+States&d=100&u=Miles&q=R>

Stack Overflow provides job search results in an XML format accessible via the above RSS/Atom feed URLs. Both XML outputs are first converted to JSON feeds using a free utility from<https://feed2json.org/>. This is the only manual step in the data acquisition process.

Data Import, Exploration, and Pre-Processing

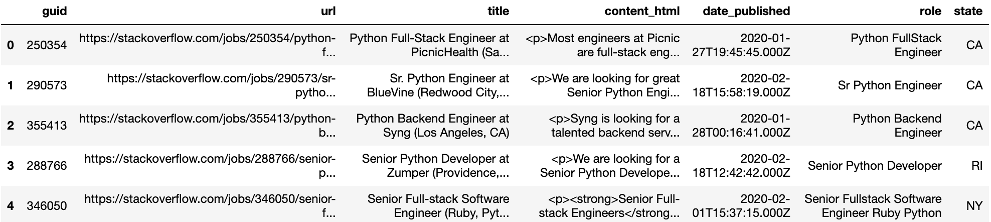
Once each RSS/Atom XML feed has been converted to JSON, they are loaded into JSON dictionaries and then transformed into pandas dataframes for exploration. Within each JSON feed structure, the job postings are located in a key called “items”, which requires further navigation and unpacking to expose more data elements about each job such as title, content, and posting date. Initial review of the data reveals the following information about both job feeds.

* There are 825 entries in the job collection for “Python”. Each line is a unit representing one job posting.
* There are 88 entries in the job collection for “R”. Each line is a unit representing one job posting.
* The following fields are present in each feed as JSON keys:
  + guid – unique identifier of a job posting
  + url – link to the job posting
  + title – contains the job title, location, and hiring company, ex: “Python Engineer at PicnicHealth (San Francisco, CA)”
  + content\_html – description of the job posting in HTML
  + summary – same as content\_html
  + date\_published – date and timestamp of when the job is posted

After gaining a basic understanding of the initial data elements, several steps are performed to pre-process the data for analysis.

* Role – The job role is extracted from the “title” field to capture only the role, ex: “Python Engineer” is extracted from “Python Engineer at PicnicHealth (San Francisco, CA)”
* Punctuation and the word “and” are excluded from the newly created “role” field.
* State – The state abbreviation of the job location is extracted from the “title” field, ex. “CA” is extracted from “Python Engineer at PicnicHealth (San Francisco, CA)”.
* The “summary” field is removed from the dataframe since it contains duplicate information as content\_html.

Below is a sample output of the dataframe after pre-processing.



## 

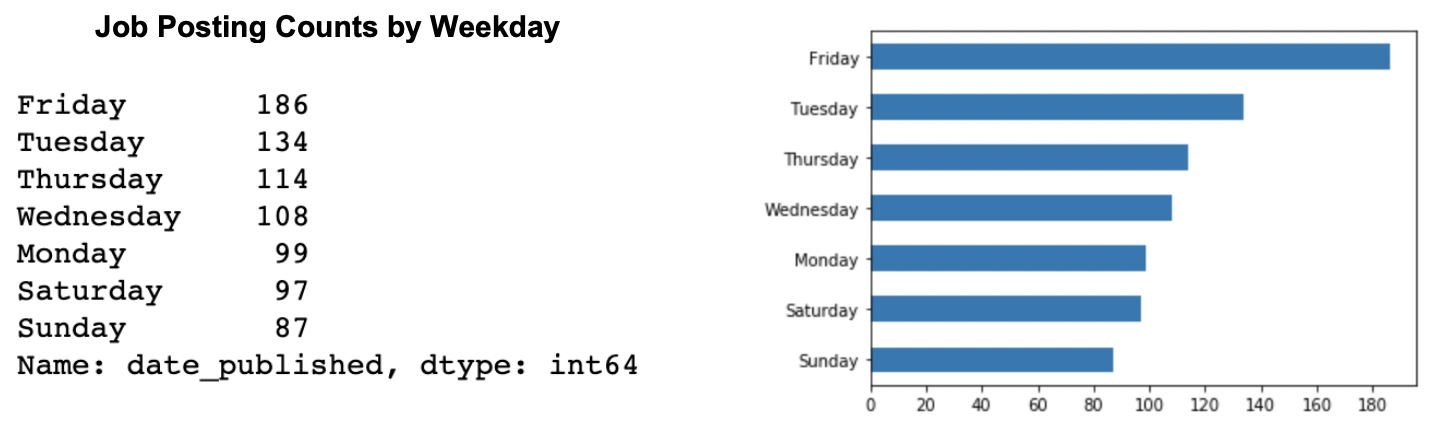
## Data Analysis & Results

### Python Job Listings by Day Published or Location

The first analysis aims to better understand the timing of Python job postings on Stack Overflow and where they are located, i.e. How many Python jobs are returned in the search feed and how far back do the job postings go? Where are Python skillset in demand?

Using counts, min, and max functions, the program is able to determine that the Python job search result contains 825 postings published between January 24th, 2020 and February 23rd, 2020.

After summarizing the publish date of each listing by weekday, it is possible to see that more jobs are posted on Fridays and Tuesdays. Below are visual output comparing the number of Python jobs by the day of the week that they are published.



Since the Python job search does not specify a particular city or state location, it is expected to return job postings from anywhere within the United States. During the data pre-processing phase, a field called “state” is extracted from the job title. With this field, the program can summarize the count of listings by state and report a breakdown of jobs by state. For visualization purposes, only the top 15 states having the highest number of jobs are displayed.

From the bar chart below, about 30% of Python jobs posted on Stack Overflow within the last month are located in California as it has the greatest number of jobs, followed by New York and Massachusetts. Virginia, Washington, and Texas have similar demand. Further research indicates that state “te” is specified for remote jobs. This means the Python job search also returned 35 jobs that can be done away from the hiring company location.

### 

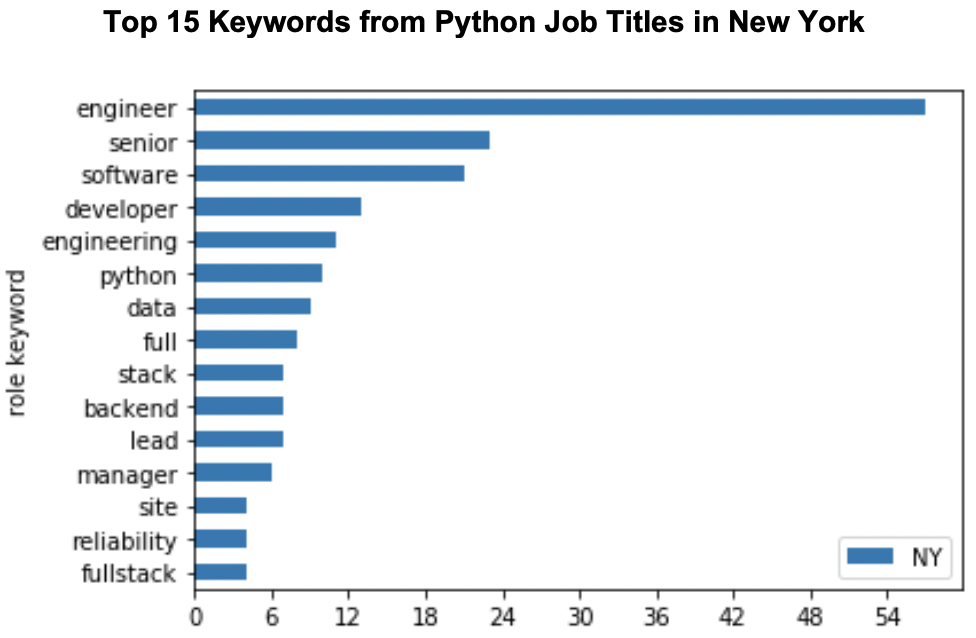
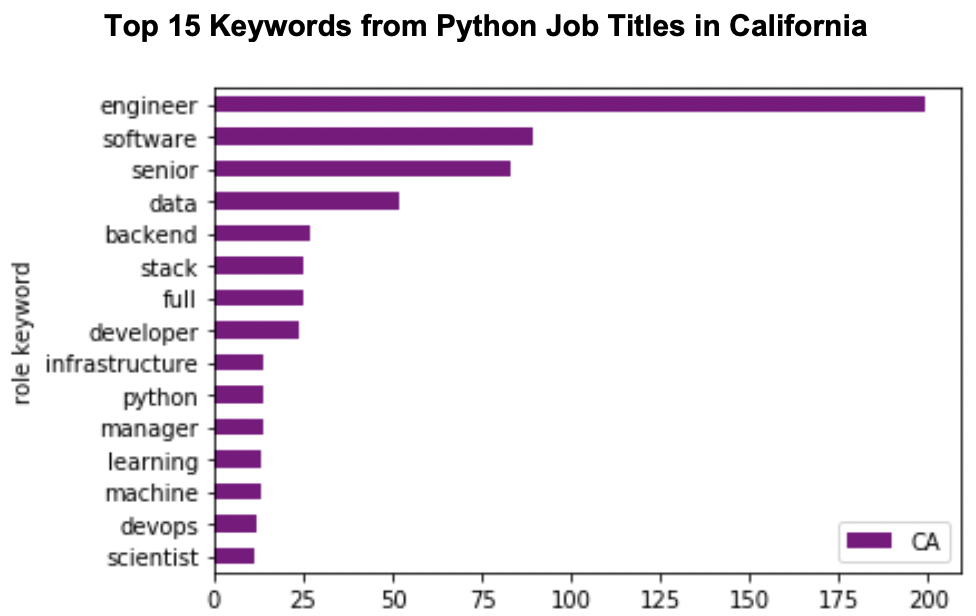
### 

### Python Job Roles Comparison in the Top 2 States

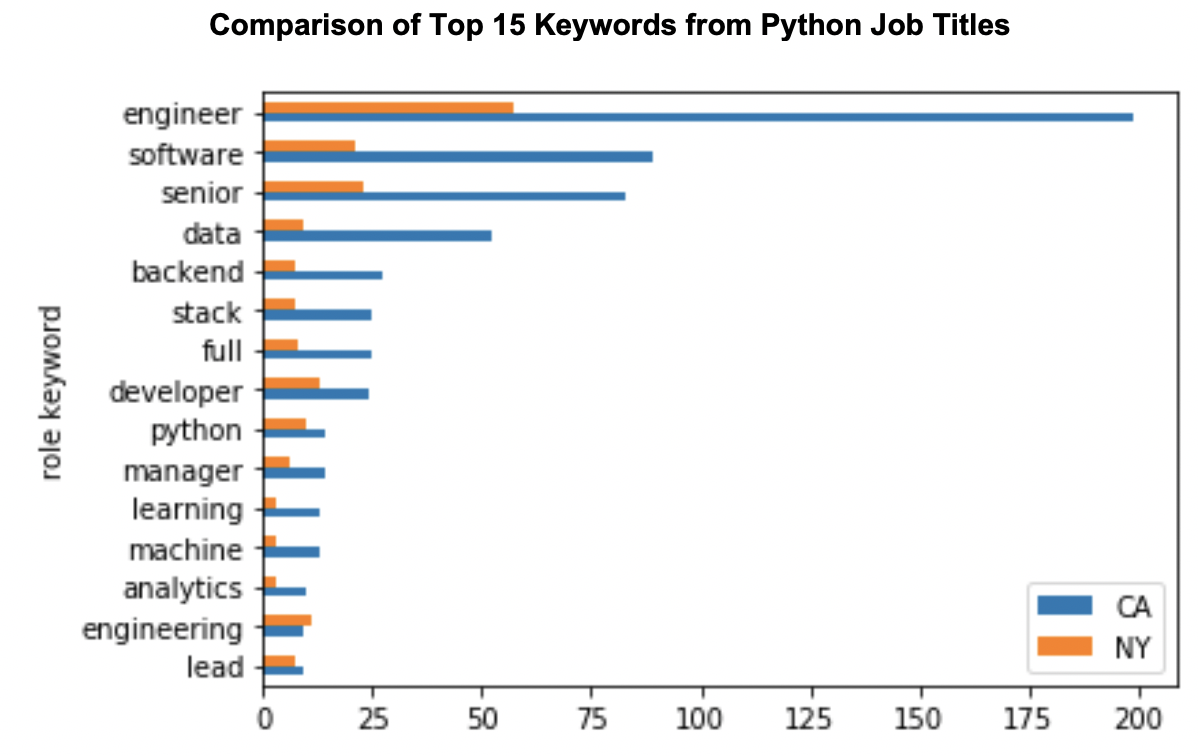
After learning that most Python job openings are located in California and New York, it would be insightful to compare the data from these 2 states, seeking answer to the question: “Is the market for Python skill similar or dissimilar in California and New York?”

To accomplish such a comparison, the program examines keywords from the job titles by state. First, each job role is parsed into a list of words, which is then aggregated to create a dictionary of words and their frequency counts. This process is performed twice, once for each state, to obtain 2 separate dictionaries of keyword frequency.

Below are top 15 keywords and the number of times they appear in Python job titles for the states of California and New York.

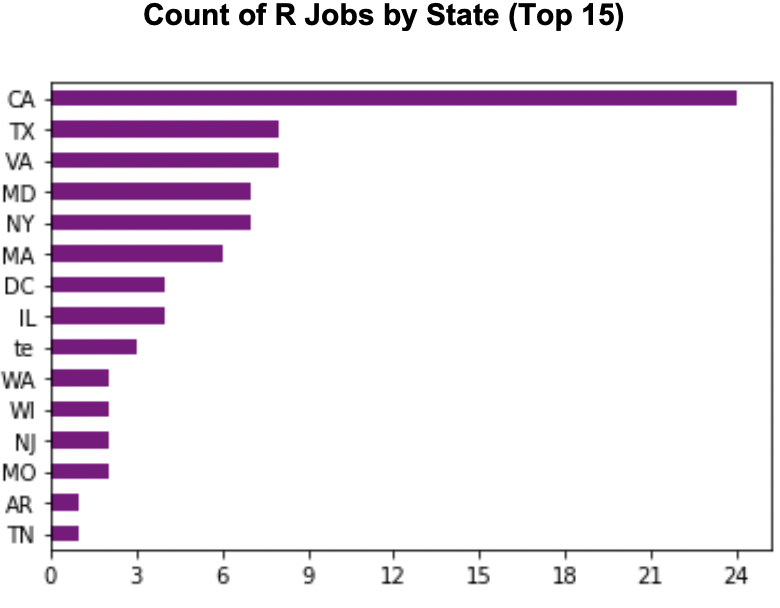


When comparing the keyword results for each state side by side, it appears that both states are mostly hiring senior software engineers and developers. Backend and full-stack developers are highly advertised in both states, while machine learning, data, and analytics jobs seem more popular in California since only 1 out of those 4 keywords is in the top 15 for New York.



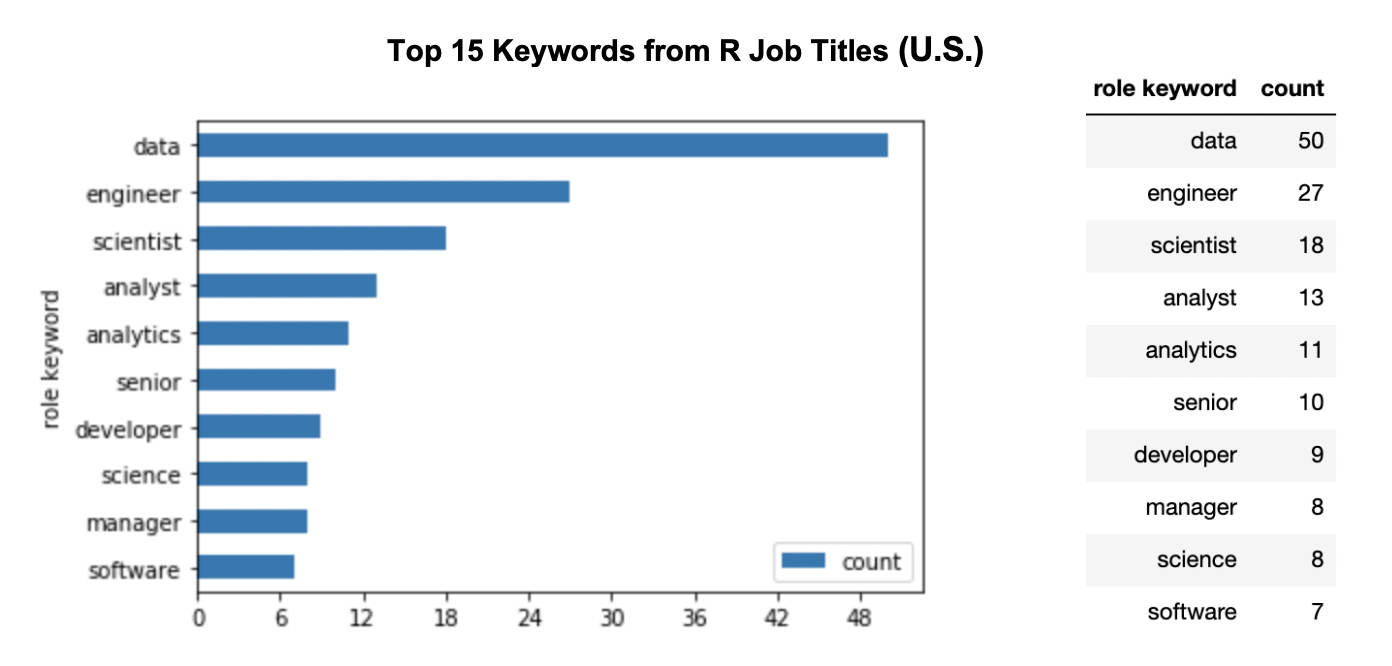
### Analysis of R Job Postings

Similar evaluation can be done on the R job search results to enable comparison of Python and R job markets. The R job feed contains only 88 job listings published from January 25h, 2020 to February 23rd, 2020. Once again, California has the highest number of job openings for R, but followed by Texas and Virginia.



In addition to reviewing where most of the R jobs are located, a keyword analysis on R job titles reveals that R job roles are indeed different from Python ones.

Below are charts illustrating top keywords found in R job titles within the United States. R job listings are much more oriented towards candidates with data and analytics skill set as most companies advertise to hire data engineers, scientists, and analysts.



### 

### Python Job Descriptions Analysis

In order to better understand the different types of skill and knowledge hiring companies are looking for in their Python job openings, it is important to look beyond job titles and analyze the data within the job descriptions. With over 800 Python job postings, each with an extensive write-up, it is more efficient to analyze only list items within the descriptions, as companies usually highlight key job requirements as bullet points. After identifying and extracting all list items using the HTML <li> tag from all Python job postings, n-grams are generated and analyzed along with their frequency distributions.

The first analysis focuses on unigrams that comprise of lemmatized words with common stop words removed (english stop words and certain words that are too common to provide context, for example: experience, knowledge, ability, work, use, etc.). Using the histogram below, it is possible to see that “data” is the most frequently mentioned word. Overall, the top words are pretty aligned with job responsibilities expected of software developers and engineers. However, it is interesting to note that “cloud” and “aws” appear more than 300 times. It can be deduced that cloud technology such as AWS is a sought-after skill in Python job positions. In addition, Java is a also complementary skill in Python candidates.

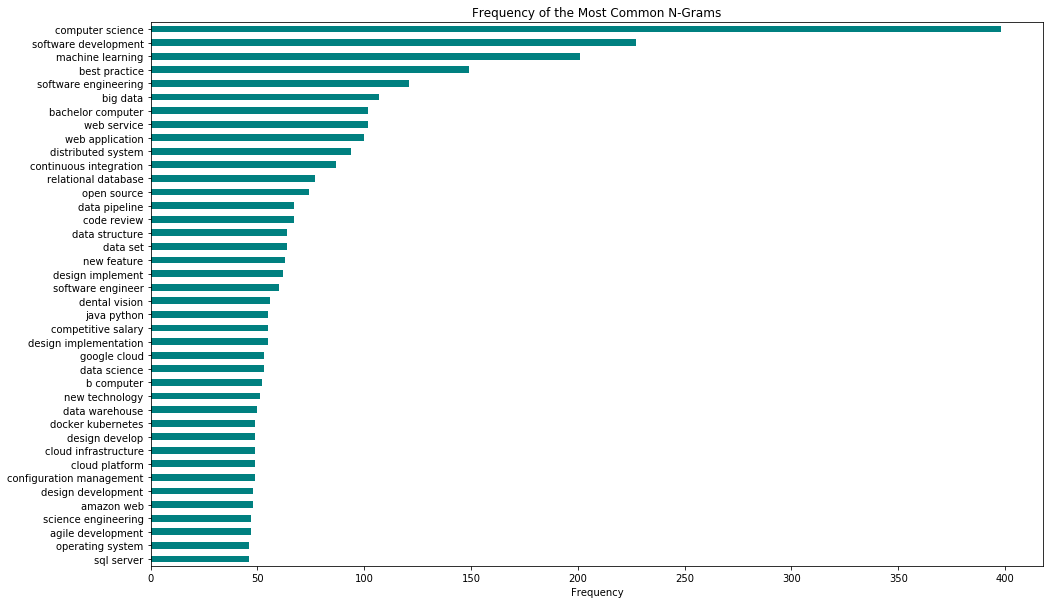
Following are the top 30 most frequently used words found in bullet lists from the Python job collection.

# 

The next analysis focuses on bi-grams, which are composed of a sequence of two adjacent words found in the Python job description lists. Bigrams are helpful when analyzing text data as they can provide additional context that may be missing from unigrams.

The below histogram displaying the top 40 bigrams depicts a better picture of the key skills requirements from Python jobs.

For Python jobs, companies are mostly looking for candidates with bachelor degrees. In addition to web application or web service development, it is not surprising that data skills like machine learning, big data, and data pipeline, database, data science are in high demand. Google, Amazon, along with Docker, Kubernetes are popular cloud platforms and devOps technologies that Python developers would need to work with.



An analysis of trigrams reveals even more interesting insights about Python job descriptions. The top qualification for Python jobs is ‘bachelors degree in computer science’ or ‘engineering’. Beside the degree requirement, the next most frequently mentioned skill is around communication. Hiring companies are constantly citing excellent written, verbal communication skills in their job descriptions. Aside from all the technical acumen, communication is the number one sought after skill for Python developers.

As far as other programming languages go, Java and C++ both seem complementary to Python. Once again, cloud technologies and data skills (big data, machine learning, data structures) are in demand. In addition, the top trigrams also confirm the need for developers to know software development practices such as devOps, configuration management, and SDLC.

Following is a chart of the top 40 trigrams from the Python jobs collection. For the purpose of trigram analysis, lemmatization is not applied in order to preserve the original context found in the job descriptions.

# 

# GITHUB JOBS

Github is an online community where more than 40 million people learn, share, and work together to build software. Github’s main platform provides private repositories online for developers to share code; according to their company timeline by April 2019, 50% of 2018 Global Fortune 100 companies were using the Github Enterprise platform for software development. In addition to code collaboration, Github also provides a fee-based job marketplace enabling companies to create and preview their own job listings and publish them live on the site.

## About the Data

Github Jobs features an API enabling job seekers to perform a query which then returns relevant job listings via URL. Accessing the JSON formatted version of the search results simply requires making a few adjustments to the URL returned from the query performed on the jobs.github.com website. Search queries returning more than 50 results are paginated into separate pages, therefore requiring a separate URL to be accessed for each set of results.

The data for the following section comes from a job search on “All” job categories and for “All” locations generated on 3/11/2020.

## Data Import, Exploration, and Pre-Processing

In order to preserve the search results, once the JSON-formatted search data is returned via URL it is then saved to MongoDB for storage and easy retrieval. For Github jobs data this is particularly important, as the API supports pagination then multiple URLs are retrieved for the entirety of the search result data. Storing within MongoDB enables this data to be compounded together in order to form one complete document collection within the database.

A database is created in MongoDB named ‘jobsdb,’ containing a collection named ‘jobs’ for storing the Github subset of data.

Initial exploration of the ‘jobs’ database collection within MongoDB shows:

* There are **249** documents. Each document represents one job posting.
* There are 12 fields of data present in each document as JSON key-value pairs.
* The key-value pairs are as follows:

o Key: \_id, Value: Object ID unique identifier created by MongoDB

o Key: id, Value: Unique ID field assigned by Github jobs

o Key: type, Value: Job type. Options include “Full Time”

o Key: url, Value: Direct web link to the job posting on Github jobs

o Key: created\_at, Date Time stamp representing when job was posted

o Key: company, Value: Name of the company providing the job opportunity

o Key: company\_url, Value: Website address of the company providing the job

o Key: location, Value: Name, and/or state, and/or country where job is located

o Key: title, Value: Title of the job

o Key: description, Value: Free-form text field in html describing the job

o Key: how\_to\_apply, Value: Specific directions for how to apply for the job

o Key: company\_logo, Value: Website address returning logo for company

The ‘jobs’ collection is then exported to a Pandas dataframe for further exploration and analysis. During this process unnecessary fields are pruned in order to retain only those required for the research questions. For the Github job dataframe, the following fields are retained:

*‘type’,* ‘*created\_at’, ‘company’, ‘location’, ‘title’, and ‘description’*

Reviewing the first few rows of the dataframe shows that the variables most accessible for summarization and other tasks are ‘location’, ‘title’, and ‘created\_at’. Several steps are now performed to pre-process the data for analysis.

Data Cleaning:

* **Location** – This field is unstructured and entries show great variation; cleaning and transformation steps are applied to separate entries into city and state values (where possible).
* **Created\_at** – This field is updated to datetime format

Data Transformation (new variables):

* **Region** – Transformation steps performed on the location field generate a Region variable containing the region (US, Europe, etc) according to the city and state extrapolated from Location (in entries where multiple are provided, the first value is taken).
* **Days** – Transformation steps performed on the converted ‘created\_at’ field generate a new day column containing the day of the week each job was posted on.

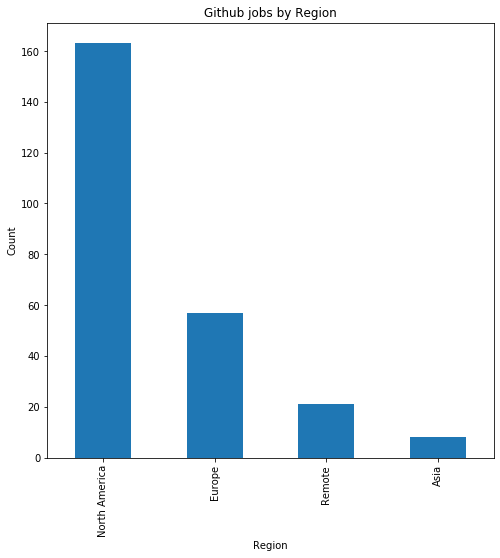
## 

## Data Analysis & Results

### 

### Github Job Listings by Region

The first analysis seeks to summarize where the majority of opportunities are located within the current job environment, and to see whether career opportunities are regional or global in nature. This analysis also seeks to compare what opportunities are available for job candidates to work remotely, as opposed to requiring residence in any particular fixed geographic location. Does this trend appear prevalent within tech-heavy fields?



Results show that **66%** of the job opportunities posted on Github are located within North America.

**10%** of jobs advertised are labelled “Remote” and do not require a fixed geographic abode.

# 

# 

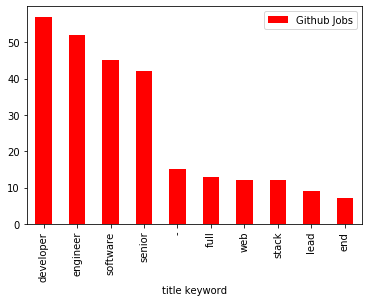
# 

### 

### Analysis of Job Titles for U.S. Based Positions

The frequency of job titles posted for North American-based applicants was analyzed in order to uncover whether keywords relevant to developers (such as “developer”, “programmer”, “software engineer”, etc.) appear more or less frequently than other jobs posted for locations within the United States.

Results based on generating a counter of the Top 10 most frequent keywords in the Title field:



developer: 57

engineer: 52

software: 45

senior: 42

-: 15

full: 13

web: 12

stack: 12

lead: 9

front: 7

# 

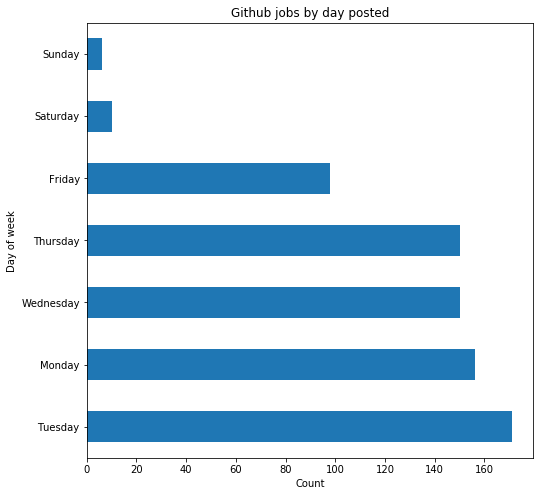
As may be expected, the top 10 title keywords from Github job postings contain keywords relevant to the tech industry. For our survey respondents it is heartening to note that ‘developer’ appears at the top of the list. This title is closely followed by ‘engineer’, and then ‘software’.

Interestingly the words ‘senior’ and ‘lead’ also feature within the Top 10, suggesting Github jobs may be a good source for experienced developers and software engineers when looking for more highly skilled and/or management positions.

### Github Job Listings by Day of the Week

Job postings are summarized by the day on which they are created, in order to investigate whether there are any trends regarding when jobs are most regularly posted.

As RSS feeds are updated frequently, understanding any significant patterns regarding when jobs are regularly published enables job seekers to be prepared and have a head start when actively looking for new opportunities.



Github posts jobs most frequently during weekdays, with **Tuesday** having the highest count of postings at 23%. This is closely followed by Thursday. Far fewer jobs are posted on Fridays.

As may be expected the lowest job posting counts are to be found on the weekend, with Sunday accounting for less than 1% of all job postings

### 

### Exploration of Github Job Descriptions

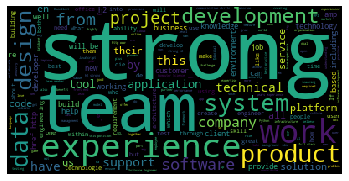
Following exploration of the job search data regarding timing frequency, types of jobs, and locations, the description for each job is unpacked in order to analyze the information used by employers to describe each job opportunity. The information found can then be correlated back to the factors cited as most important to Stack Overflow survey respondents, in order to see if there are any gaps.

The job ‘description’ field within the Github postings consists of unstructured text, providing descriptions for each job post that can contain a wide variety of elements. The field is free-form on the site, containing no apparent structure or requirements for data entry. Applying Natural Language Processing techniques to parse out the unstructured text provides an opportunity to identify the information each employer chooses to use in promoting their company and job opportunities.

#### Unstructured Text Processing

After tokenizing the text, filters and stopwords are applied to remove any unnecessary characters and words that are found frequently within the descriptions yet which do not prove meaningful to analysis.

Prior to analyzing the text for the job factor keywords, frequency distribution lists of filtered bigrams and trigrams are generated. An overview of all top Github job keywords are also visualized using word cloud:



The most emphasized word in the cloud is “strong”, indicating this is the most frequently cited word according to frequency within the job descriptions. ‘team’, ‘experience’, ‘product’, and ‘development’ also rank highly on the list.

**Top Bigrams by Frequency**

'computer', 'science', 0.0003600267448439027

'software', 'development', 0.0003485973243726677

'software', 'engineer', 0.00030287964248772766

'communication', 'skills', 0.00029145022201649266

**Top Trigrams by Frequency**

'senior', 'software', 'engineer', 9.715007400549755e-05

'excellent', 'communication', 'skills', 7.429123306302753e-05

'flexible', 'working', 'hours', 7.429123306302753e-05

'personal', 'development', 'budget', 6.857652282741003e-05

'proven', 'track', 'record', 5.1432392120557526e-05

'full', 'stack', 'web', 4.571768188494002e-05

'amazon', 'web', 'services', 4.000297164932252e-05

'equal', 'employment', 'opportunity', 4.000297164932252e-05

Note: several of the top trigrams within the Github job descriptions are in German. Evidently there are many job postings that are based in Germany!

The Trigram list proves interesting, as terms such as “excellent communication skills” and “flexible working hours” point to ‘soft skills’ and job factors that would not otherwise be identifiable within the other fields of the job posting. These allow for full characterization of the nature of the postings and what employers look for in addition to development-specific skills.

### Analysis and Comparison of Job Factors listed within Github Jobs Postings

With these filters applied, the raw text is investigated by applying similarity and concordance measures to select words. The job factors ranked as the Top 3 by survey respondents are analyzed using these measures in order to see what results can be found:

**1st job factor**: “languages, frameworks and other technologies I will be working with”

*The word “frameworks” is used as most representative within the Github postings text as relevant to this job factor.*

66 concordance matches are found.

Select examples of the types of phrases this word is contained within are:

**the most common javascript frameworks such react or angular**

**in particular angular v5+ frameworks html5 css3 and javascript**

**projects using react and other js frameworks depending on your interests**

**scripting and javascript frameworks jquery basic understanding**

**2nd job factor**: “office environment or company culture” :

71 concordance matches.

Select examples of the types of phrases this word is contained within are:

**establishing a culture of openness positive communication**

**having a great company culture is really important to us**

**strong remote culture work wherever you are most producti**

**3rd job factor**: “flex time or a flexible schedule”

9 concordance matches.

Select examples of the types of phrases this word is contained within are:

**we offer schedule flexibility**

**ability to have a flexible schedule**

**in control of your own schedule**

The top 2 job factors listed by Stack Overflow respondents have a fairly good representation within the text, which is encouraging for those using the Github job site as a career resource. “Flex time or a flexible schedule” however is only represented 9 times as a cited phrase, suggesting that this factor is not offered as frequently or considered as highly as company culture and technology frameworks.

# 

# WORKING NOMADS

This website aims to connect professionals wanting to work remotely with innovative companies offering “independent positions”. The site curates lists of the most interesting remote job offers in a whole range of professional careers. Job listings are provided by category, with a diversity of areas listed such as Development, Marketing, System Administration, Finance, and more. The site features an ‘API’ page which, when accessed provides a raw RSS feed of JSON-formatted data representing every job posting currently listed on their website.

The data for the following section comes from a job search on “All” job categories and for “All” locations generated on 3/11/2020.

## Data Import, Exploration, and Pre-Processing

The JSON-formatted search data is first saved to MongoDB for storage and easy retrieval. Within the ‘jobsdb’ database already established in MongoDB, a collection named ‘nomads’ is created for storing the Github subset of data.

Initial exploration of the ‘nomads’ database collection within MongoDB shows:

* There are **73** documents. Each document represents one job posting.
* There are 9 fields of data present in each document as JSON key-value pairs.
* The key-value pairs are as follows:

o Key: \_id, Value: Object ID unique identifier created by MongoDB

o Key: url, Value: Direct web link to the job posting on Working Nomads jobs

o Key: title, Value: Title of the job

o Key: description, Value: Provides summary of the job opportunity, including title

o Key: company\_name, Value: Name of the company offering the job

o Key: category\_name, Value: Name of the job specialty (e.g. “Finance”)

o Key: tags, Value: Skills provided as ‘tags’ on the job posting (e.g. “accounting”)

o Key: location, Value: Geographical location for the job

o Key: pub\_date, Value: Date Time stamp representing when job was published

The ‘nomads’ collection is exported to a Pandas dataframe for further exploration and analysis. During this process unnecessary fields are pruned in order to retain only those required for the research questions. For the Nomads job dataframe, the following fields are retained:

‘*title’, ‘description’, ‘company name’, ‘category name’, ‘tags’, ‘location’ and ‘pub date’.*

Reviewing the first few rows of the dataframe shows that the variables most accessible for summarization and other tasks are ‘location’, ‘title’, and ‘created\_at’. Several steps are now performed to pre-process the data for analysis.

Data Cleaning:

* **Location** – This field contains several overlapping entries (for example, “ANYWHERE (100% REMOTE)” and “Anywhere”). A function is created to standardize the values.
* **Pub date** – This field is updated to datetime format.

Data Transformation (new variable):

* **Days** – Transformation steps performed on the converted ‘pub date’ field generate a new day column containing the day of the week each job was posted on.

## Data Analysis & Results

### Working Nomads Job Listings by Region

The first analysis seeks to summarize where the majority of opportunities are located. Working Nomads jobs are distinguished by those located in the USA versus jobs that are available remotely. This analysis seeks therefore to compare the proportion of remote jobs to those with a fixed geographical location.



**72%** of the jobs listed are located within the USA.

**28%** of jobs are considered ‘remote’ and do not require residence in a specific physical location.

### 

### 

### 

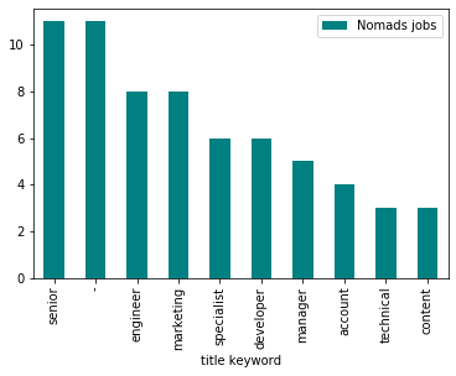
### 

### 

### Analysis of Job Titles for U.S. Based Positions

The frequency of job titles posted for “USA” only is generated by first filtering the dataframe on the relevant records.

Results based on a counter of the Top 10 most frequent keywords in the Title field yields the following:



senior: 11

-: 11

engineer: 8

marketing: 8

specialist: 6

developer: 6

manager: 5

account: 4

technical: 3

content: 3

The title ‘Senior’ appears with the highest frequency within the Working Nomads listings. Engineer features second, however ‘developer’ is listed at 6th place. The higher diversity of keywords demonstrates that the Working Nomads job site is less tech-heavy than Github jobs, although it may still provide a good source of opportunity for those looking for senior-level positions.

### 

### Comparison of Job Categories in USA versus Remote

As a diversity of categories exist within the Working Nomads data, a cross-tabulation was computed in order to compare job categories across opportunities located in the USA versus those advertised as ‘remote’.

|  |  |  |
| --- | --- | --- |
| **category\_name** | **Remote** | **USA** |
| Administration | 0 | 2 |
| Consulting | 0 | 2 |
| Customer Success | 1 | 1 |
| Design | 2 | 0 |
| **Development** | **8** | **15** |
| Education | 0 | 3 |
| Finance | 0 | 4 |
| Human Resources | 1 | 2 |
| Management | 0 | 3 |
| Marketing | 4 | 8 |
| Sales | 2 | 4 |
| Systems Administration | 2 | 3 |
| Writing | 1 | 6 |

While ‘developer’ appears 6th in the keyword list for Working Nomads, the largest category of jobs posted on this site are categorized as in the “Development” field. This is true for jobs located in the USA and those listed as ‘Remote’. Within the listings, there are almost twice as many development jobs located in the USA than there are Remote postings.

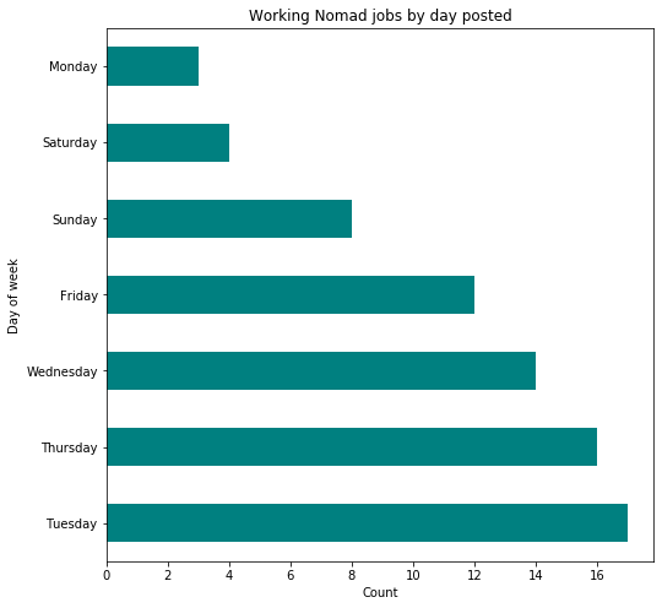
Of 21 Remote jobs, 8 are in the ‘Development’ category (38%).

Of 53 jobs located in the USA, 15 are listed as ‘Development’ (28.3%).

These results appear to suggest that a higher tendency towards remote work may exist for those in the software development field.

### Working Nomads Job Listings by Day of the Week

Job postings are summarized by the day on which they are created, in order to investigate whether there are any trends regarding when jobs are most regularly posted.



Results are consistent with Github:

Working Nomads posts jobs most frequently on **Tuesday** (23%). Thursday is next.

However, unlike Github the fewest jobs are posted on Mondays.

# 

### 

### 

### Exploration of Working Nomads Job Descriptions

Following exploration of the job search data regarding timing frequency, types of jobs, and locations, the description for each job is unpacked in order to analyze the information used by employers to describe each job opportunity. The information found can then be correlated back to the factors cited as most important to Stack Overflow survey respondents, in order to see if there are any gaps.

The job ‘description’ field within the Working Nomad postings consists of unstructured text similar to Github jobs. Each description provides a variety of information for each job post, depending on what employers choose to include. Applying Natural Language Processing techniques to parse out the unstructured text provides an opportunity to identify the information each employer has chosen to promote their company and the specific job opportunity.

#### 

#### Unstructured Text Processing

After tokenizing the text, filters and stopwords are applied to remove any unnecessary characters and words that are found frequently within the descriptions yet which do not prove meaningful to analysis.

Prior to analyzing the text for the job factor keywords, frequency distribution lists of filtered bigrams and trigrams are generated. An overview of all top Working Nomads job keywords are also visualized using word cloud:



The most emphasized word in the cloud is “strong”, indicating this is the most frequently cited word according to frequency within the job descriptions. This was also the most frequent unigram in the Github job description field. ‘Experience’, ‘work’, ‘team’, and ‘experience’ feature next on the list.

**Top Bigrams by Frequency**

‘experience', 'working', 0.00036850116157974847

'equal', 'opportunity', 0.0003204357926780421

'communication', 'skills', 0.00028839221341023794

'opportunity', 'employer', 0.00028839221341023794

**Top Trigrams by Frequency**

'equal', 'opportunity', 'employer', 0.0002723704237763358

'equal', 'employment', 'opportunity', 0.00011215252743731475

'time', 'business', 'hours', 9.613073780341265e-05

'video', 'monetization', 'platform', 9.613073780341265e-05

'chief', 'operating', 'officer', 6.408715853560843e-05

'excellent', 'problem', 'solving', 6.408715853560843e-05

'opportunity', 'employer', 'committed', 6.408715853560843e-05

'verbal', 'communication', 'skills', 6.408715853560843e-05

In both Working Nomads lists, terms relevant to job factors are listed with more frequency than was seen in the Github job descriptions. While the Github terms are more tech-focused in nature, the bigrams and trigrams here feature “equal opportunity” in two different formations (bigrams) and three different formations (trigrams). This particular factor is obviously a key issue with the employers that use Working Nomads as a job resource.

### 

### Analysis and Comparison of Job Factors within Working Nomads Postings

With these filters applied the raw text is investigated by applying similarity and concordance measures to select words. The job factors ranked as the Top 3 by survey respondents are analyzed using these measures in order to see what results can be found:

**1st job factor**: “languages, frameworks and other technologies I will be working with”

**11** matches found.

*The word “frameworks” is used as most representative within the text as relevant to this job factor.*

Select examples of the types of phrases this word is contained within are:

**application frameworks libraries architectures and tools**

**frontend javascript frameworks and libraries**

**exposed to new languages frameworks and ideas from your peers**

**2nd job factor**: “office environment or company culture”

**26** matches found.

Select examples of the types of phrases this word is contained within are:

**we have a strong culture of mentorship**

**collaborative teach and learn culture**

**a positive company culture that understands work/life balance**

**3rd job factor**: “flex time or a flexible schedule”

**29** matches found.

*Note: for this phrase, “flexible” rather than “schedule” is found to yield the best results.*.

Select examples of the types of phrases this word is contained within are:

**strong benefits flexible hours competitive pay**

**fully remote with flexible working hours**

**group gym discounts flexible work hours**

The Working Nomads job factor keyword results appear to show a higher proportion of phrases centered around the job factors that developers are interested in. In general, flexibility and positive company culture feature highly. Flexibility in particular appears to make a lot of sense, as a higher proportion of Working Nomads jobs are listed as remote opportunities. The analysis makes it clear that these factors appear to be reinforced within the job descriptions.

To conclude: some interesting bigrams were found within job descriptions when generated by pointwise mutual information score that are fun to highlight (as these are generated by PMI, the score merely indicates the probability of the second word in the bigram following the first within the text):

**'morning'**, **'coffee**.', 15.09496012599459

**'ping'**, **'pong'**, 14.831925720160793

**'transcendental'**, **'meditation'**, 13.34464266793212

**'cloud'**, **'guru'**, 11.285748978878555

# 

# FINAL CONCLUSION

Analyzing the results of a survey such as the one conducted by Stack Overflow provides a unique peek into the demographics, backgrounds, and opinions of those who consider themselves part of the software development community. While the information retrieved proves insightful, it is representative only of those who took the time to respond. While the 2019 survey gathered opinions of over 88,000 respondents, this analysis shows that the majority are male developers living in the United States. While this may not seem particularly “global” or “comprehensive”, given that the survey takes at least 20 minutes to complete it may be assumed this demographic is at least representative of those spending the most time within the Stack Overflow community. Presumably those considering Stack Overflow most beneficial to their learning development and careers are most likely to “give back” in terms of time spent answering the survey questions.

Given these assumptions, this research analysis shows that overall women get paid less for development jobs than men. However, results also reveal that in the lowest professional coding experience category (0-5 years), women actually rate slightly higher in terms of compensation than men (not taking into account those in the ‘other’ or non-binary, gender fluid categories, or those who simply unanswered the gender question). This finding may correspond with the general push in tech towards hiring female programmers, and also corresponds to findings within this report demonstrating that in the youngest age group (0-18), girls report starting to code at an earlier age than males. These results may point to a cultural shift towards more encouragement in the coding profession for young women, with the result that more females are being employed in the technology industry at the lowest career levels. For women entering the tech field, the results of the analysis around compensation, professional coding experience and organizational size suggests that women may wish to consider seeking larger corporations or freelancing options in order to fulfill their maximum salary potential as they grow in their careers and experience.

Within the current job market for developers and technology-related careers, online job search sites appear to provide a distinct advantage to those seeking employment within today’s environment. As technology continues to evolve, the task of matching available jobs to aspiring applicants continues to evolve along with it. Job searches now begin most frequently through online screened sites, and as this analysis shows, more and more jobs are being conducted ‘remotely’ without requiring the need for workers and companies to be located within physical proximity. Given these factors, the task of matching qualified applicants to relevant job opportunities becomes most critical; in particular, when attempting to find the best fit in terms of inscrutable elements such as the atmosphere and culture of the hiring company and the expectations and desires of potential employees.

Based on our analysis around software developer profiles and the job market, full-stack developers are most highly sought after. This trend is likely to continue, although back-end developers and data scientists are not far behind. Although in high demand, the job of a developer is increasingly more dynamic. While new open-source programming languages appear yearly, developers are required to adopt new frameworks and technologies (such as cloud platforms and devOps practices), all while continuing to build data knowledge sufficient to handle big data and machine learning tasks. Even with the popularity of online learning tools such as MOOC and coding bootcamps, job descriptions provide evidence that there is still a preference towards candidates with formal computer science and engineering education backgrounds. In addition to technical acumen, today’s developers are expected to also be versed in software development processes, and perhaps the most referenced requirement of all, is the ability to have excellent verbal and written communication skills.

For those searching for development work located within the United States, a significant proportion of job opportunities exist, although many of these may not require a specific location in order to be eligible. This may be due to the fact that companies are using these sites more for senior-level and management-type positions, where the level of skill and experience necessary are most fundamental. For those seeking this type of work, online marketplaces appear to be a great resource for those with solid coding skills, in the most relevant programming languages, and with a higher level of experience. Job sites such as Stack Overflow and Github may not however work as effectively for those just starting out in their professional development career.

# 

# PYTHON PROGRAM DESCRIPTION

The python program authored in Jupyter notebook consists of 4 major sections:

1. Stack Overflow Survey Results
   1. Data Import, Cleaning, and Exploration – Using mainly pandas, seaborn, and matplotlib libraries, the CSV file is loaded into a dataframe for processing. Additional data cleaning steps leverage python lists and functions to transform and standardize response values. Counting is also used extensively to ensure a good understanding of data frequency, especially for those columns used as dimensions in the analysis, i.e. role, gender, age, and country.
   2. Analysis Question Subsections – Pandas crosstabs and group by methods are used to aggregate and summarize responses into counts. Bar or line charts are generated to provide visualizations of the results. Pandas group by methods are used to split data into groups for aggregations and other numerical functions such as averages, minimum, and maximum. In addition, sorting, slicing, and filtering functions are also used to display top or chosen results in the desired ranking order.
2. Stack Overflow Jobs
   1. Data Import, Cleaning, and Exploration – Using json and pandas libraries, JSON feeds are read and loaded into Python structures for data exploration and pre-processing. Additional data transformation steps leverage different list and string functions to parse and prepare the data for further analysis. The BeautifulSoup python library is used to extract data out of job descriptions that are formatted as HTML.
   2. Analysis Question Subsections – Pandas functions and methods such as counts, dates, filtering, and sorting are used to group and aggregate lower units of data into higher categories and groups. Bar charts are generated to provide visualization of the analysis results. In addition to all the aforementioned methods, the counter function is used to create a dictionary of word counts. Sorting, slicing, filtering, and other formatting functions are used to generate summarized outputs and display categorical comparisons in the desired ranking order. The NLTK library is used for text data analysis such as tokenization, stop words removal, lemmatization, and n-grams frequency distribution.
3. GitHub Jobs
   1. Data Import, Cleaning, and Exploration – Using RSS feeds, JSON-formatted data is saved into a MongoDB database collection for storage and easy retrieval. The database documents are filtered and loaded into a Pandas dataframe for further data exploration and pre-processing. Additional data cleaning functions are performed, and data transformation steps undertaken to create new features within the data. The NLTK python library is used for applying Natural Language processing techniques, including tokenizing, filtering, and applying stopwords to parse job description text for further analysis.
   2. Analysis Question Subsections – Pandas functions and methods such as counts, dates, filtering, and sorting are used to group and aggregate lower units of data into higher categories and groups. Bar charts are generated to provide visualization of the analysis results. In addition, the counter function is used to create a dictionary of word counts. Text comparison techniques such as text similarity, concordance, and frequency distribution counts of unigrams, bigrams and trigrams are applied to the unstructured text within the job description fields.
4. Working Nomad Jobs
   1. The same methodologies and techniques applied to the Github jobs data is also applied to Working Nomad Jobs.

Note: All analysis outputs are captured in this report so there are no external files.

# 

# TEAM MEMBERS, ROLES & TASKS

The project team is composed of Emma Woods and Sandra Tang, who collaborated extensively during the ideation and planning phase, including discussions regarding datasets to go after and interesting research questions to be tackled.

For data processing and analysis of the main survey dataset, research questions were divided up between team members. Each conducted independent work in order to maximize the amount of insights gained and to minimize duplication of effort. The job listings unstructured data sources were individually sourced; Sandra conducted all processing and analysis tasks for Stack Overflow jobs, and Emma carried out all research using Github Jobs and Working Nomads data.

Weekly discussions took place in order to divide the work effort appropriately and to discuss similarities and contrasts of each dataset. Particular consideration was given to how the skills and tools learned in class would be utilized and applied.

Following independent homework submissions, both team members worked closely together to formulate the project report. This included determining how to best incorporate the analysis into a cohesive unit in order to tell a compelling and insightful story. Both team members continue to work together to formulate a presentation flow that is intuitive, and that succinctly disseminates the highlights of the research that has been conducted.

|  |  |  |
| --- | --- | --- |
| **Project Tasks** | **Sandra** | **Emma** |
| **Data Acquisition and Analysis** | * Pre-processing and analysis of Stack Overview Survey file * Data acquisition, pre-processing, and analysis of Stack Overflow Python and R jobs | * Pre-processing and analysis of Stack Overview Survey file * Data acquisition, pre-processing, and analysis of GitHub jobs * Data acquisition, pre-processing, and analysis of Working Nomad jobs |
| **Project Report** | * Collaboration of introduction, final conclusion, program description * Write-up on Stack Overview Survey analysis subsections * Write-up on Stack Overflow jobs analysis subsections | * Collaboration of introduction, final conclusion, program description * Write-up on Stack Overview Survey analysis subsections * Write-up on Github jobs analysis * Write-up on Working Nomad analysis |