# JOB MARKET ANALYSIS BASED ON HANDSHAKE

Jieer Chen

Syracuse University

IST 652 M002 Spring 2020

Instructor: Benjamin Nichols

## Table of Contents

## Project Description

This project aims to help students major in Applied Data Science and Business Analytics in making an accurate career plan and goal and writing informative resumes by providing them answers to what skills they should develop throughout their academic careers, what experiences they should highlight on their resumes, and which areas will offer them better chances to land a job.

The collection of datasets I used in this analysis contain targeted job search information. Theis collection of data has job information posted on the Handshake website between mid-February and early-April. The below tables summaries the questions they are going to answer in this report.

| **Handshake Dataset** | |
|---|---|
| **Job Location** | Q1: Where are the jobs come from?<br>Q2: What is the number of jobs in a city or cities in different countries? |
| **Job Type** | Q1: What is the number of jobs for each job type? What are the percentages? |
| **Employment Type** | Q1: What is the number of jobs for each employment type? What are the percentages? |
| **Salary Type** | Q1: What are the salary types and their percentage?<br>Q2: What is the number of jobs for each level of salary type within employment type for each job type? |
| **Accept OPT/CPT Candidates** | Q1: How likely it is to find an employer that will accept OPT/CPT candidates?<br>Q2: Where are the employers that accept OPT/CPT candidates in terms of location and industry? How many employers are there? |

| | |
|---|---|
| **Willing to sponsor candidate** | Q1: How likely it is to find an employer that is willing to sponsor a candidate?<br>Q2: Which employers are willing to accept OPT/CPT candidates and provide sponsorship? |
| **Expires** | Q1: What is the number of jobs expires in each weekday in different months? |

# Dataset and Source

- **Handshake Dataset**

The handshake dataset consolidates Handshake job data from five excel files. The five Handshake job excel files were originally created by the Career Center at Whitman School of Management and sent out to all Whitman students via email. Upon the completion of data consolidation, we have a Handshake dataset ( Handshake_Internship_Job.xlsx) that contains job information on various types of jobs including Business Analysis and Data Scientists. The dataset with a total of 11 columns and 6359 columns.

| | |
|---|---|
| **Employer** | Company Name |
| **Title** | The job title |
| **Job Location** | The job location |
| **Job Type** | Type of Job (ex. Job or Internship) |
| **Employment Type** | Type of Employment (ex. Full-time or Part-time) |
| **Employer Industry** | The industry |
| **Salary Type** | Paid / Unpaid |
| **Expires** | Expiration date |
| **Located in the US** | Yes or NaN |
| **Accepts OPT/CPT Candidates** | Yes or NaN |

| | |
|---|---|
| **Willing To Sponsor Candidate** | Yes or NaN |

# Collation of Data

- **Handshake Dataset**

  ***Data Cleaning***

  - **Step1: Find records related to Business Analytics and Data Scientist positions**
    First used the map() function and the lower() function to lowercase the “Title” column. Then used the str.contains() function to find the jobs that contain any of the selected keywords, “big data”, “data”, “analyst”, “analytic”,“ intelligence”, “data scientist”, “computer science”, “data engineer”, and “machine learning”.
  - **Step 2: Drop Duplicates**
    Four rows were removed at this stage using drop_duplicates().
  - **Step 3: Format Column labels**
    Spaces within the column names were replaced by “_” using str.repalce().
  - **Step 4: Missing values**
    Five columns have missing values. The eight missing values in “Located_In_US” were placed by “No” because those 8 jobs locate in foreign countries. The Employer_Indsutry, Accepts_OPT/CPT_Candidates, and Willing_To_Sponsor_Candidate columns have more than 100 missing values and hose missing values were replaced by "Unknown". The only row with missing values in Salary_Type column was removed.
  - **Step 5: Type of Value**
    The type of “Expires” from object to pandas DateTime object.
  - **Step 6: Further data cleaning and processing on column values**
    ***Salary_Type Column***: Based on the context, two irrelevant records were found and deleted.
    ***Job Type Columns***: An irrelevant record was deleted and the values of two records were changed to “Job” from “Cooperative Education” based on the context.
    ***Employer Industry Column***: Both “NGO” and “Non-profit - Other” values were found. They were changed into “Non-Profit” for consistency purposes.

***Job Location Column***: The job column was split into multiple columns using separator ‘,’ and then stored in a new data frame, in which we had “United States of America”, “United States”, “USA” and “NY” values in the country column. We also found that the names of some states were written differently in the “State_or_Territory” column. Therefore, we change some of the values to have consistent naming.

### *Data aggregation and summarization*

Upon the completion of data processing, the clean dataset named df_relevant has 599 rows and 13 columns. Among 13 columns, we have one pandas data time object and the rest of 12 columns are string objects.

### *Data aggregation and summarization*

The type of all the columns in Data.csv and Business.csv is object. The data of salary and job type in Data.csv only has 341 and 76. Compared to the total data of this dataset, which is 1310, is very small. The same problem also exists in the Business.csv. So, we will not do analysis on salary and job types.

```
Data.describe()
```

| | Title | Company | Location | Salary | Summary | Link | Description | Job Type | State |
|---|---|---|---|---|---|---|---|---|---|
| count | 1310 | 1310 | 1310 | 341 | 1310 | 1310 | 1310 | 76 | 1310 |
| unique | 257 | 484 | 327 | 57 | 561 | 1177 | 564 | 19 | 40 |
| top | Data Scientist | Apple | Seattle, WA | $101,231 a year | With one application you can be considered for... | https://www.indeed.com/rc/clk?jk=cdf043cdf0349... | how the position worksthe data scientist repor... | Contract | CA |
| freq | 521 | 84 | 156 | 69 | 69 | 2 | 69 | 25 | 231 |

```
Business.describe()
```

| | Title | Company | Location | Salary | Summary | Link | Description | Job Type | State |
|---|---|---|---|---|---|---|---|---|---|
| count | 1289 | 1289 | 1289 | 355 | 1289 | 1289 | 1288 | 237 | 1289 |
| unique | 307 | 506 | 376 | 101 | 580 | 1160 | 567 | 23 | 44 |
| top | Business Analyst | Shelter Insurance | Columbia, MO | 43,735–63,677 a year | Collecting new business requests from stakehol... | https://www.indeed.com/company/Amick-Brown,-LL... | a company built to serve you. it’s your career... | Contract | NY |
| freq | 267 | 80 | 81 | 80 | 39 | 3 | 40 | 77 | 167 |

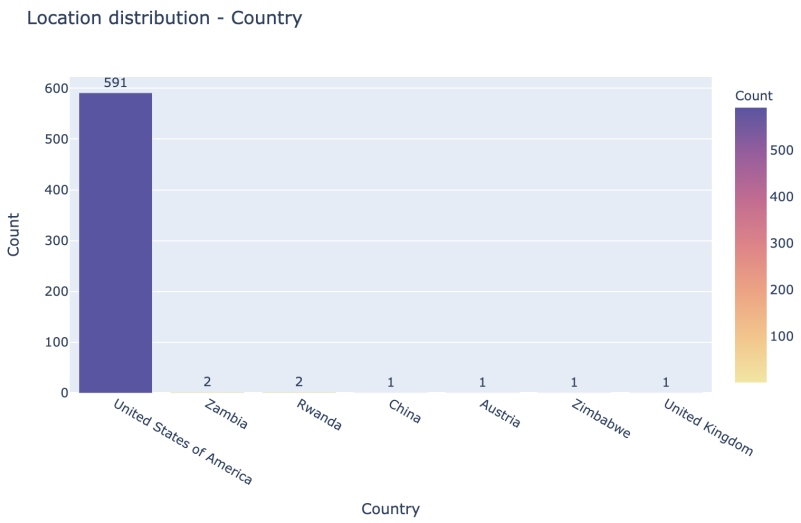
# Data Analysis

- **Handshake Dataset**

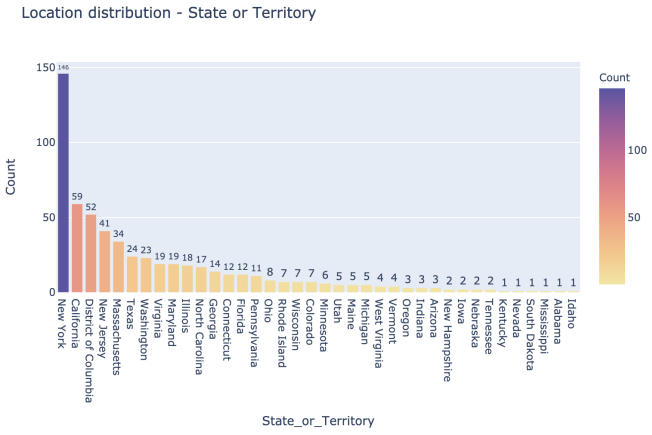
  - ***Job location***

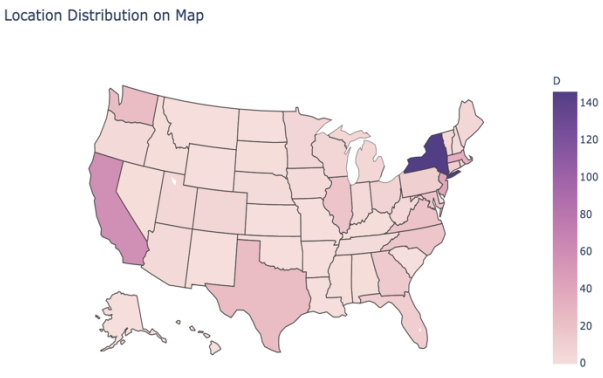
**Q1: Where are the jobs come from?**

According to our analysis, the working location of the majority of Business Analytics and Data Scientist related jobs are in the United States. Only 8 of them located in foreign countries.




For jobs located in the U.S, around 25% of them located in New York State while around 10% of them located in California. District of Columbia, New Jersey, and Massachusetts ranked 3rd, 4th, and 5th respectively.



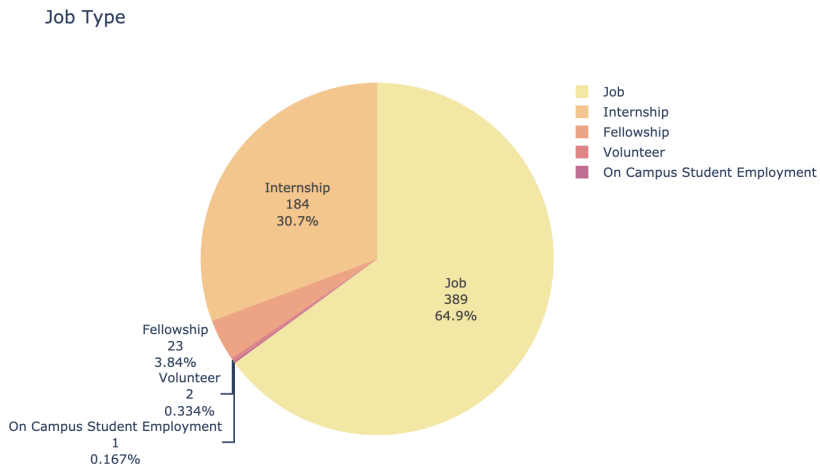
**Q2: What is the number of jobs in a city or cities in different countries?**

- **Field of analysis:** Country & State_or_Territory
- **Process:** Group the data by “Country” and “State_or_Territory” and then count the number of job titles (equivalent to the number of jobs).
- **Output file:** **handshake_job_location.csv**
- **Description:** The table has three columns and 44 rows. Each row has a country name, a state or territory name, and the number of jobs in that place.

### ○ *Job Type*

**Q1: What is the number of jobs for each job type? What are the percentages?**

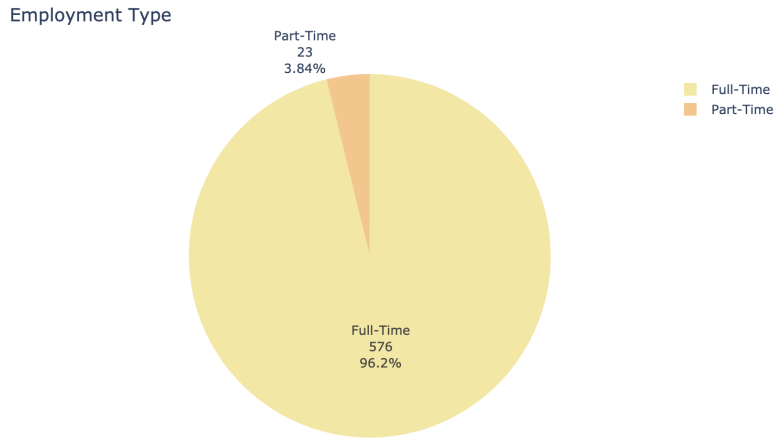
The five types of jobs in the handshake dataset are “Job”, “Internship”, “Fellowship”, “Volunteer”, and “On Campus Student Employment”. Regular jobs account for around 64.9 % of the total while internships account for around 30.7% of the total. The rest three categories combined takes up below 5% of the total.




### ○ *Employment Type*

**Q: Where is the number of jobs for each employment type? What are the percentages?**

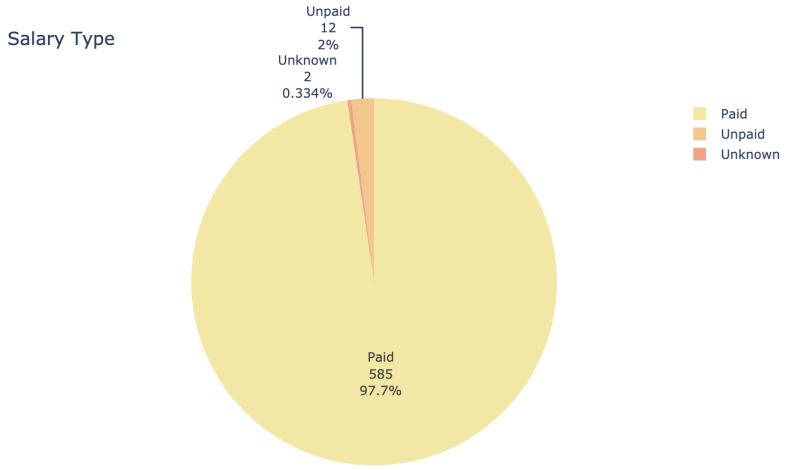
Part-time and Full-Time are the two employment types in this dataset. As shown by the graph, the number of full-time employment accounts for as high as 96.2% of total jobs while that of part-time employment is around 3.84%.




- ***Salary Type***

**Q1: What are the salary types and their percentage?**

Around 97.7% of the jobs are paid while 2% of the jobs are unpaid.




**Q2: What is the number of jobs for each level of salary type within employment type for each job type?**

- **Field of analysis:** Salary_Type & Job_Type & Employment_Type
- **Process:** Group the data by “Job_Type” and “Employment_Type” and then count the number of job titles that are equivalent to the number of jobs. After creating the dataset, add a percentage column.
- **Output file:** **Jobtype_Employtype_Saltype.csv**

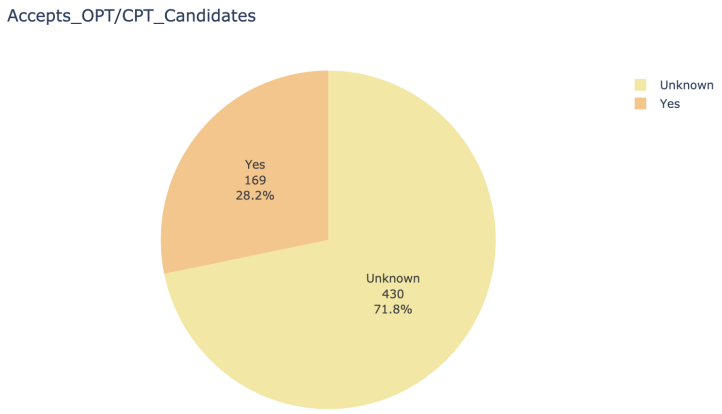
- **Description:** The dataset has 9 rows and four columns. Each row has a job type, an employment type, a salary type, and the number of jobs available for that type and its percentage.

| | Job_Type | Employment_Type | Salary_Type | Title | percentage |
|---|---|---|---|---|---|
| **0** | Fellowship | Full-Time | Paid | 23 | 0.038 |
| **1** | Internship | Full-Time | Paid | 159 | 0.265 |
| **2** | Internship | Full-Time | Unknown | 2 | 0.003 |
| **3** | Internship | Full-Time | Unpaid | 3 | 0.005 |
| **4** | Internship | Part-Time | Paid | 13 | 0.022 |
| **5** | Internship | Part-Time | Unpaid | 7 | 0.012 |
| **6** | Job | Full-Time | Paid | 389 | 0.649 |
| **7** | On Campus Student Employment | Part-Time | Paid | 1 | 0.002 |
| **8** | Volunteer | Part-Time | Unpaid | 2 | 0.003 |

- ***Accept OPT/CPT Candidates***

**Q1: How likely it is to find an employer that will accept OPT/CPT candidates?**

Around 28.2% of employers indicate that they are willing to accept OPT or CPT candidates when they post jobs on Handshake. The rest of the employers didn't reveal their willingness to take OPT/CPT candidates.




**Q2: Where are the employers that accept OPT/CPT candidates in terms of location and industry? How many employers are there?**

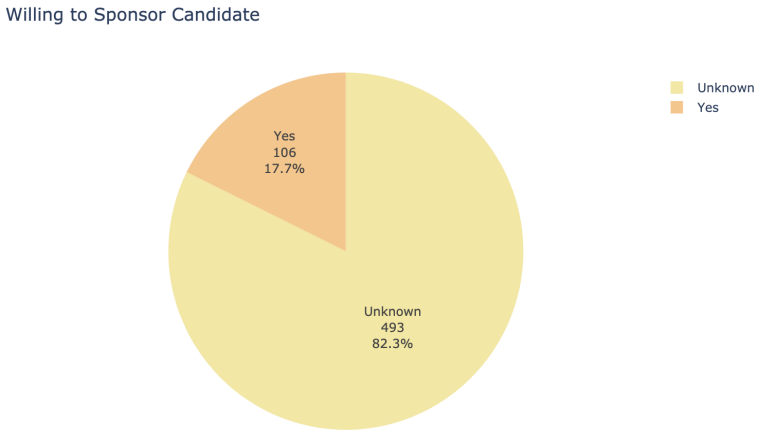
- **Field of analysis:** Willing_to_Sponsor_Candidate and Accepts_OPT/CPT_Candidates

- **Process:** First filtered the date to find employers that are willing to accept OPT/CPT candidates. Then grouped the data by “State_or_Territoryand” and “Employer_Indsutry”. Lastly, count the number of employers.
- **Output file: employer_opt_cpt.csv**
- **Description:** The dataset has 65 rows and 3 columns. Each row has a state name or a territory name, an industry name, and the number of employers in that given industry and location.

### ○ *Willing to sponsor candidate*

**Q1: How likely it is to find an employer that is willing to sponsor a candidate?**

Compared to 28.2% chances to find an employer willing to accept OPT/CPT candidate, only 17.7% of the employers indicate that they are willing to sponsor a candidate.




**Q2: Which employers are willing to accept OPT/CPT candidates and provide sponsorship?**

- **Field of analysis:** Accepts_OPT/CPT_Candidates, State_or_Territory, and Employer_Industry
- **Process:** Filtered data using two conditions: willing to accept OPT/CPT candidates and willing to sponsor a candidate.
- **Output file: Employer_sponsor_opt_cpt.csv**
- **Description:** The dataset has 101 rows and 13 columns. It has the employers that are willing to accept OPT/CPT candidates and provide sponsorship and their details.

**Q3: For each job type, how many jobs are there? how many of them will accept OPT/CPT candidate and how many of them will provide sponsorship?**

- **Field of analysis:** Job_Type
- **Process:** Count the number of jobs, the number of employers willing to accept CPT/OPT candidates, the number of employers willing to sponsor a candidate, and the number of employers that are willing to do both respectively. After that, using the job type and the values to create a data frame.
- **Output file:** **job_type_summary.csv**
- **Description:** A table with 5 rows and 5 columns as shown below. Each row has a job type and the number of jobs of that type, the number of jobs that accept OPT/CPT candidates, the number of jobs providing sponsorship, and the number of jobs that offers both options.

| | JobType | Total | OPT_CPT | Sponsor | Provide_Both |
|---|---|---|---|---|---|
| **0** | Job | 398 | 99 | 80 | 77 |
| **1** | Internship | 184 | 61 | 24 | 22 |
| **2** | Fellowhsip | 23 | 8 | 1 | 1 |
| **3** | Volunteer | 2 | 0 | 0 | 0 |
| **4** | On Campus Student Employent | 1 | 1 | 1 | 1 |

- ***Expires***

**Q1: What is the number of jobs expires in each weekday in different months?**

- **Field of analysis:** Expires
- **Process:** Extracted months and weekdays from the expires date column and added them as two new columns to the data frame. Grouped the data frame by “month” and “weekday” and then count the number of titles (equivalent to the number of jobs).
- **Output file:** **expires_time.csv**
- **Description:** A table with 23 rows and 3 columns. Each row has a month value, a weekday value, and the number of jobs available for that month and weekday combination.

  Through visualization, we found the Friday, Saturday, and Monday have the greatest number of jobs expires. However, when we looked into the weekdays in different

months, this conclusion isn't applicable. One of the possible reasons is that the amount and the length of data collected by the career center varies from month to month. Or, it could simply be that there is no pattern on job expiration dates.