**GEO-441 PROJECT SUBMISSION**

Application of Web Scraping & GIS to

Map Open Data Science Position Around

US Sates

Based on Indeed and Career Builder Websites

(Technical Report)

by

**Pradeep Sathyamurthy[GitHub]**

Under the guidance of: Prof. Nandhini Gulasingam

DePaul University

15th August 2017

Contents

[1. Project Summary: 3](#_Toc490725568)

[Abstract 3](#_Toc490725569)

[Introduction 3](#_Toc490725570)

[Project Scope 3](#_Toc490725571)

[Geographic Extents 3](#_Toc490725572)

[2. Project Methodology Layout 4](#_Toc490725573)

[Web Data Scraping 4](#_Toc490725574)

[Project Excel Workbooks and Attribute Data Explanation 5](#_Toc490725575)

[Geographic’s Information System and Shape file data 5](#_Toc490725576)

[3. Map1 – Data Science Position in US States based on Indeed website 6](#_Toc490725577)

[4. Map1 Findings 7](#_Toc490725578)

[5. Map2 – Insight on US States with High number of Data Science Position 8](#_Toc490725579)

[6. Map2 Findings 9](#_Toc490725580)

[Map 2.1 Data Science Positions with respective to Counties in California State 9](#_Toc490725581)

[Map 2.2 Data Science Positions with respective to Congressional Districts in New York State 9](#_Toc490725582)

[Map 2.3 Data Science Positions with respective to Congressional District in Washington State 10](#_Toc490725583)

[Map 2.4 Data Science Positions with respective to Congressional District in Illinois State 10](#_Toc490725584)

[7. Map3 - Data Science Position in US States based on Career Builder website 11](#_Toc490725585)

[8. Map3 Findings 12](#_Toc490725586)

[9. Future Works 13](#_Toc490725587)

[Web data mining with Attribute data and mapping its detail 13](#_Toc490725588)

[Unsupervised Learning with Review Data and map user based result 13](#_Toc490725589)

[Generalizing the Map with user Search 13](#_Toc490725590)

[10. Timetable 14](#_Toc490725591)

[11. Key Personnel 14](#_Toc490725592)

[12. Deliverables 14](#_Toc490725593)

[13. Reference 14](#_Toc490725594)

[14. Zip File with Map Packages 15](#_Toc490725595)

[15. Appendix 15](#_Toc490725596)

[Python-Code for Indeed Website Data Scraping 15](#_Toc490725597)

[Python-Code for Career Builder Website Data Scraping 18](#_Toc490725598)

# 1. Project Summary:

## Abstract

In this project, I have mapped the number of open Data Science jobs available across United States of America (USA) based on one day (09 Aug 2017) data scraped[[1]](#_Reference) from two major job portals, Indeed and Career Builder. Indeed data is rich when compare to Career builder based on the count of open positions available on the day mentioned. Hence, I have used data obtained from ***Indeed website to visualize US states with maximum job opportunities for Data Science position*** which would be our first map for Analysis. Based on [*Map-1*](#_Map1_–_Data), I have looked for States with high job postings available for Data Science and looked at each of these states individually at a granular level to see which congressional district is rich with Data Science job position, this would be our second map ([*Map-2*](#_Map2_–_Insight)) for analysis. Our third map ([*Map-3*](#_Map3_-_Data)) is based on the data scraped from Career builder website which is used for cross validation and as well, this career builder website has an additional information about position type, whether a job position is a Full-Time or Contract position, with this ***we can hypothesize whether industries are more inclined towards full time employee for Data Science position or do they trust to share their data with employee who is available part-time***. We will conclude our analysis by recommending future scope with which the data can be further analysed deep down.

## Introduction

“*Data Science*” is a word which makes a big buzz during recent years across all industries. It is an interdisciplinary field about scientific methods, processes and systems to extract insight from data in various form. Data can be a structured or Unstructured data and learning from these data can be a Supervised or Un-supervised learning. People who does this kind of data insight are generally called as Data Scientist. People who look to apply for this kind of positions highly depend on job search portals like Indeed, LinkedIn, Career Builder, Dice, etc., which provides necessary information about open Data Science position available in a company under various names like Machine Learning Engineers, Data Insight engineers, etc., along with the location where the position is available. As part of this project we scrape these data from web and produce a map which would answer “***Which state in US has more Data Science position?***”

## Project Scope

In this project data is scraped only from two major job portals:

***1. Indeed***[[w]](https://www.indeed.com/)

***2. Career Builder***[[w]](http://www.careerbuilder.com/)

There are other job portals in web like LinkedIn, Dice, etc., whose data are out of scope for this project and the maximum geographical extent is restricted to Unites States of America.

We use vector spatial data and map types like Choropleth[[2]](#_Reference_1) and Charts Map are in scope for this project. Raster data and map thematic types like Single Symbol, Unique Values, graduated symbol and Dot Density are out of scope for this project to keep our maps simple and effective.

## Geographic Extents

In this project, we will construct totally 3 maps, in this map-2 is multi-framed:

|  |  |  |  |
| --- | --- | --- | --- |
| SI.NO | Map Name | Geographic Extent | Unit of Analysis |
| 1 | Map-1 | United States of America | State |
| 2.1 | Map-2.1 | California State | Counties |
| 2.2 | Map-2.2 | New York State | Congressional District |
| 2.3 | Map-2.3 | Washington State | Congressional District |
| 2.4 | Map-2.4 | Illinois State | Congressional District |
| 3 | Map-3 | United States of America | State |

# 2. Project Methodology Layout

This project make use of two methodologies to obtain attribute data and mapping it with respective shape file:

1. ***Web Scraping***[[1]](#_Reference_1) using Python 3.5

2. ***Geographic’s Information Systems (GIS)*** using ArcMap from ArcGIS 10.5 a licensed product of Esri Inc.

## Web Data Scraping

Web data scraping[[1]](#_Reference_1) is a technique to extract data from a website. It is an activity which needs to be done with responsibility and it is subjected to the terms and condition of each website. Few websites like Facebook or twitter enforce its users to login before a data is scraped from their website while there are few websites like job portal, flight booking websites do not ask user to login while displaying any information. However, if any action needs to be performed on that information displayed, then website will request user to register & login. Web scraping, we do as part of this project is from Indeed and Career Builder which is a Job search engine and do not ask for user credentials to display any information from their data base and extracting their information for non-profit activity is legal.

While we extract data from a website, the format of data obtained plays a major role in our analysis. Data obtained can be in one of below formats mostly:

***1. Raw HTML page***

***2. XML Page***

***3. JSON Object***

Indeed and Career Builder website that we scrape provides us a raw html page for general user. Only those websites like google maps, Twitter, Facebook, etc., expose their respective API’s like Google Geolocation API, Twitter API, Facebook API respectively to share their data in the form of Extensible Mark-up Language ‘XML’ or Java Script Object Notation ‘JSON’. (*With latest update even Indeed as exposed its API, focusing publishers to promote Indeed job content, which we haven’t taken into scope for this project)*

I used python to scrape data from Indeed and Career Builder website. Below python packages where mainly used as part of this data scraping activity:

***1. URLLIB Package***

***2. Beautiful Soup Package***

***3. Pandas Package***

Python’s *URLLIB package* is used to access the files from Indeed and Career Builder web server and get them to local system.

I used *Beautiful Soup package* to parse the HTML page that was downloaded above.

Pandas package was used to create data frames and write the data frame to an excel sheet as CSV or XLS format which are one of the major file type supported by ArcMap for mapping.

|  |  |  |
| --- | --- | --- |
| SI.NO | Scraped File | Treated Excel Work Book |
| 1 | Indeed\_Raw\_Data\_Scientist\_09Aug2017.csv | Indeed\_Treated\_Data\_Scientist\_09Aug2017.xls |
| 2 | CB\_Raw\_Data\_Scientist\_8Aug.csv | CB\_Treated\_Data\_Scientist\_\_09Aug2017.xls |

To construct maps, I will be using files under ‘*Treated Excel Work Book’*. These are files which are cleaned based on the shape files selected for respective Geographical extent and Unit of Analysis. File format is maintained as XLS because it is best supported by ArcMap and as well easy to manipulate data with rich functionalities provided by Microsoft Excel.

## Project Excel Workbooks and Attribute Data Explanation

As mentioned above, for this project we will use 2 excel work book which has multiple sheets that holds data with respective to shape file that we will discuss in next section. (*DS = Data Science*)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SI.NO | Excel Workbook Name | Sheet Name | Used for Mapping | Data Description |
| 1 | ***Indeed\_Treated\_Data\_Scientist\_09Aug2017.xls*** | Scrapped\_Data | No | Raw Scraped Data with Title, Company Name, User Review Count about Company, City, State and Country where job is posted. |
| Data\_To\_Map | No | Frequency Statistics – Job posting counts based on US States and top 4 states with high DS position. |
| StateCnt | Yes | US States & Its Total DS Job Count |
| Cali\_City | Yes | California Counties & DS Job Count |
| NY\_City | Yes | New York Congressional District & DS Job Count |
| WA\_City | Yes | Washington Congressional District & DS Job Count |
| IL\_City | Yes | Illinois Congressional District & DS Job Count |
|  |  |  |  |  |
| 2 | ***CB\_Treated\_Data\_Scientist\_\_09Aug2017.xls*** | Scrapped\_Data | No | Raw Scraped Data with Title, Company Name, Position Type (Full Time or Contract), Salary, City, State and Country where job is posted. |
| Data\_To\_Map | Yes | US States and respective total DS job posting, Number of Fulltime position and Number of Contract Position |

## Geographic’s Information System and Shape file data

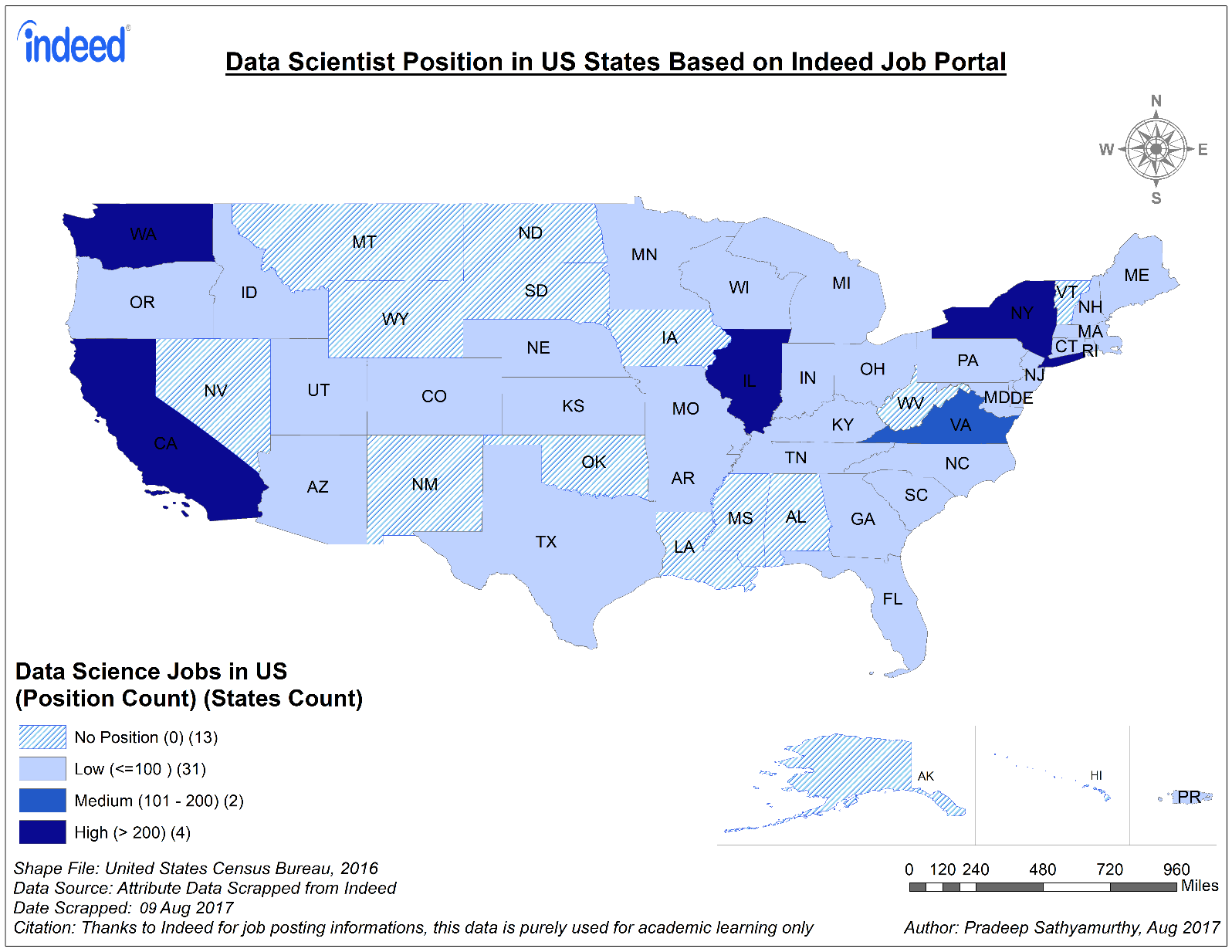
GIS is a set of computer tool used to collect, store, analyse and represent/map spatial data and its corresponding attribute data. We use vector data to represent our spatial features and in specific all our shape files are of Polygon type spatial model that represent states, counties and congressional districts. Below are some information’s about the shape files used in this project.

|  |  |  |  |
| --- | --- | --- | --- |
| SI.NO | Shape File Name | Source Link | Shape File Description |
| 1 | tl\_2016\_us\_state.shp | [Link-1](https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2016&layergroup=States+%28and+equivalent%29) | US States shape file information |
| 2 | CaliforniaCounty.shp | [Link-2](http://52.26.186.219/internships/useit/content/california-counties-shapefiles) | California counties shape file information |
| 3 | nycgwi.shp | [Link-3](http://www1.nyc.gov/site/planning/data-maps/open-data/districts-download-metadata.page) | New York Congressional District shape file information with area covering water bodies |
| 4 | nycg.shp | [Link-4](http://www1.nyc.gov/site/planning/data-maps/open-data/districts-download-metadata.page) | New York Congressional District shape file information without area covering water bodies |
| 5 | cd2012.shp | [Link-5](http://www.ofm.wa.gov/pop/geographic/tiger.asp) | Washington Congressional District shape file information |
| 6 | PA 97-14 Congressional Districts.shp | [Link-6](http://www.ilhousedems.com/redistricting/?page_id=554) | Illinois Congressional District shape file information |

With above attribute data (count of Data Science positions by state and cities) and spatial data (Polygon type) obtained we can well represent our maps with Thematic maps types. In order to keep it simple and as well effective[[5]](#_13._Reference) I chose Graduated color/Choropleth to represent my data uniform across my project. To represent the Position type (Full-Time/Contract) with respective to Career builder data I have used Pie Chart Map along with choropleth.

Let us now visualize and analyse each Maps constructed with the data explained above as part of this project and within the scope defined.

# 3. Map1 – Data Science Position in US States based on Indeed website



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SI.NO | Map Name | Geographic Extent | Unit of Analysis | Thematic Type |
| 1 | Map1 | United States of America | State | Graduated Colours |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SI.NO | Map Name | Shape File Name | Attribute Excel Work Book Name | Sheet Name |
| 1 | Map1 | tl\_2016\_us\_state.shp | ***Indeed\_Treated\_Data\_Scientist\_09Aug2017.xls*** | StateCnt |

# 4. Map1 Findings

Indeed is an American worldwide employment based search engine. As per Indeed[[W]](https://www.indeed.com/about), they claim to be the #1 job site as on date with over 200 million unique visitors every month from over 60 different countries and 28 languages. Indeed is intensely passionate about delivering the right fit for every hire. Indeed helps companies of all sizes hire the best talent and offers the best opportunity for job seekers to get hired. Thus, becoming the highest traffic job website in the United States.

Above map is plotted based on the data extracted from Indeed on ***9th August 2017*** with:

* **Job Title:** Data Scientist[[W]](https://www.indeed.com/jobs?q=Data+Scientist&l=)
* **Location:** NULL (search jobs in all US states). Domain (*.COM*)restrict its maximum search boundary to USA only by default.

Data from *Indeed\_Treated\_Data\_Scientist\_09Aug2017.xls work book and* StateCnt is used as attribute data and tl\_2016\_us\_state.shp file is used as spatial data to render this Map. With this map we can visualize the US states with High, Medium, Low and No Data Scientist Job position. Any states having more than 200 open Data Science position are considered to be the states with high demand for Data Scientist. States with open Data Science position between 101 and 200 are states with medium demand for Data scientist. Any states with open Data Science position less than or equal to 100 are considered to be states with low demand for Data Scientist while states with zero open positions are states with No demand for Data Scientist role.

Based on the Map and details explained in legend we can figure out that there are totally 4 states[[3]](#_Reference_1) with high demands for Data Scientist. They are:

1. ***California (CA)***
2. ***New York (NY)***
3. ***Washington (WA)***
4. ***Illinois (IL)***

There are the states having more than 200 open Data Science position. It is so evident that ***California***, Silicon Valley of United States with companies like Google, Apple, Intel, etc., generate more Data Scientist position due to presence of more ***IT industries*** around this area. ***New York*** being the financial capital for United States with major ***financial sectors*** like JP Morgan Chase, Citi, Gold Man Sacs also generate high Data science position for their financial analysis needs. With ***E-Commerce and software product*** giants like Amazon, Microsoft, HP, etc., being situated in ***Washington*** also becomes a high generator for Data Science position in US. ***Illinois***, windy City of US which is also know for the presence of major ***Consulting firms*** like KPMG, McKinsey & Company, etc., generates more Data Science position. Thus, based on Indeed these are the 4 major states with high Data Science Job openings.

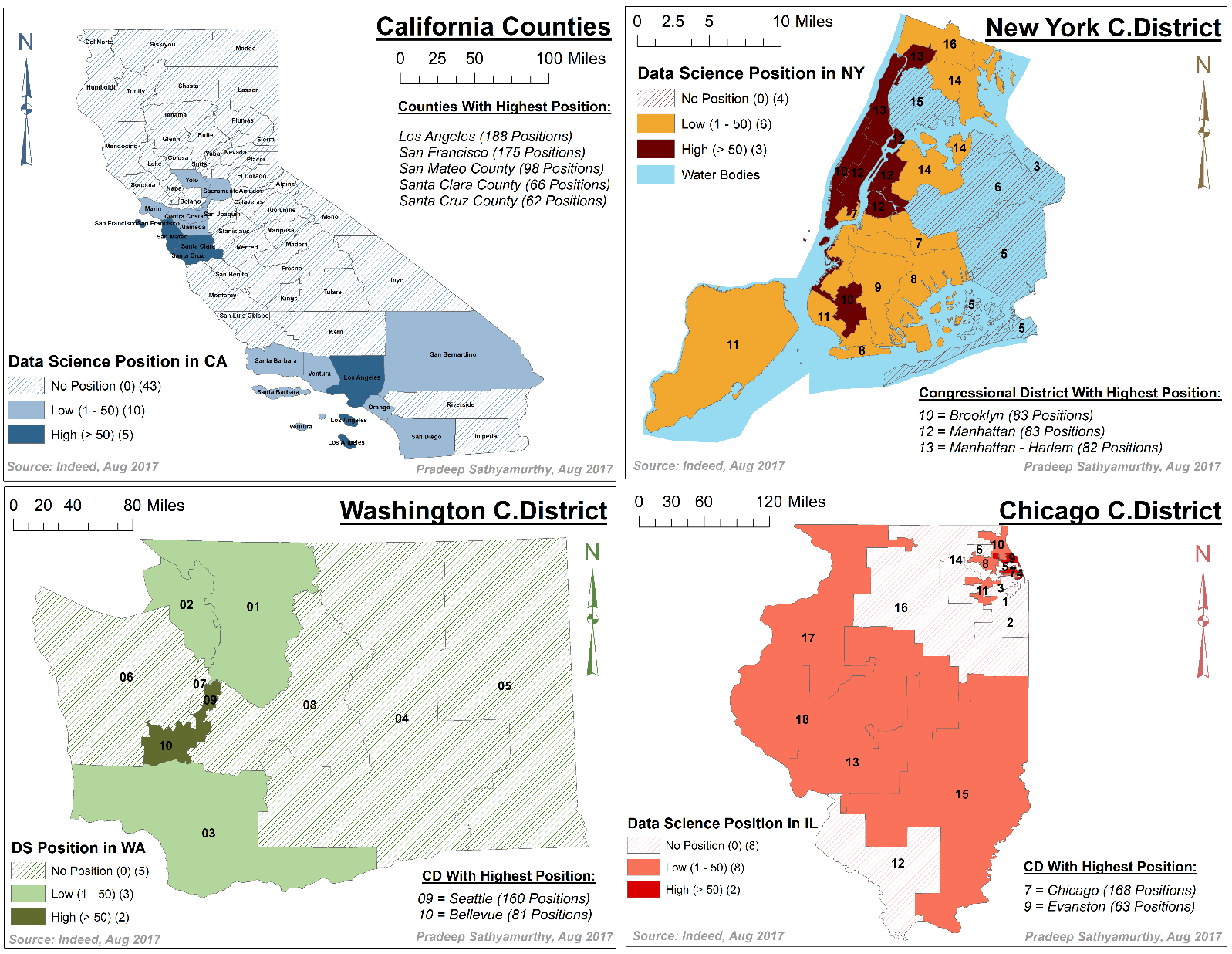
We see there are 2 states ***Virginia (VA)*** and ***Washington***, ***D.C. (DC)*** which generates medium Data Science position ranging between 101 to 200. Companies like Verizon, Capital One, etc., generate good demand for Data Scientist in Virginia while companies like IBM, CGI, etc., generates good demands for Data Scientist in DC. While there are 31 states like Florida (FL), Texas (TX), Utah (UT), etc., generates low Data Science jobs less than 100, we can hypothesise that these states have more tourism, manufacturing and Energy sectors who outsource their Data Science work to other states like CA, NY, IL, WA, DC, VA, etc., as discussed above which have more IT consulting firms. We also see 13 US states like North and South Dakota (ND & SD), New Mexico (NM), Oklahoma (OK), Vermont (VT), etc., shown in shaded lines generate zero Data Science positions based on the data extracted from Indeed.

With this map, we can infer that anyone who would like ***to pursue a career in Data Science, can plan to apply for companies situated in CA, NY, WA and IL as their first choice*** because we see there are high count of open data science position that increase the probability in getting into one in the course of time. Also, ***getting connected with networks in these states through LinkedIn*** can make the referral & relocation easy once the job gets confirmed. With this we can also ***recommend universities*** *in these major states like CA, NY, WA and IL to* ***open-up more Data Science program*** *and* ***get more industry collaboration*** *which will make fresh graduates to land in their dream job* with no delay.

Now that we know the states with high open position for Data Science, let us try to reduce our geographic extent to each of these (CA, NY, WA & IL) states and map the data science availability to see the dispersion across these states.

# 5. Map2 – Insight on US States with High number of Data Science Position

*Map2.1 Map2.2*



*Map2.3 Map2.4*

|  |  |  |  |
| --- | --- | --- | --- |
| SI.NO | Map Name | Geographic Extent | Unit of Analysis |
| 2.1 | Map-2.1 | California State | Counties |
| 2.2 | Map-2.2 | New York State | Congressional District |
| 2.3 | Map-2.3 | Washington State | Congressional District |
| 2.4 | Map-2.4 | Illinois State | Congressional District |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SI.NO | Map Name | Shape File Name | Attribute Excel Work Book Name | Sheet Name |
| 1 | Map2.1 | CaliforniaCounty.shp | Indeed\_Treated\_Data\_Scientist\_09Aug2017.xls | Cali\_City |
| 2 | Map2.2 | nycgwi.shp | Indeed\_Treated\_Data\_Scientist\_09Aug2017.xls | N/A |
| 3 | Map2.2 | nycg.shp | Indeed\_Treated\_Data\_Scientist\_09Aug2017.xls | NY\_City |
| 4 | Map2.3 | cd2012.shp | Indeed\_Treated\_Data\_Scientist\_09Aug2017.xls | WA\_City |
| 5 | Map2.4 | PA 97-14 Congressional Districts.shp | Indeed\_Treated\_Data\_Scientist\_09Aug2017.xls | IL\_City |

# 6. Map2 Findings

## Map 2.1 Data Science Positions with respective to Counties in California State

Based on Map-1 discussed above we could infer that California state tops the table with 633 open Data Science positions spread across the state. As part of this section I thought to give more insight for this particular state and hence ***shrink the Geographical Extent to California State and Unit of Analysis being Counties in California***.

There are totally 58 counties represented as part of Map-2.1 in which there are 5 counties with high number of Data Science position open. To compare maps with in the data frame of Map-2, I have maintained same legends across all the maps. However, I would suggest to compare Map2.2, 2.3 and 2.4 with each other while Map 2.1 cannot be directly compared with other maps as its Unit of Analysis is Counties which is different from the Unit of Analysis for Map-2.2 to 2.4 whose unit of analysis is Congressional District.

From Map-2.1 we can infer that there are 5 counties[[4]](#_Reference_1) with open Data Science position greater than 50, they are:

1. ***Los Angeles with 188 positions***
2. ***San Francisco with 175 positions***
3. ***San Mateo with 98 positions***
4. ***Santa Clara with 66 positions***
5. ***Santa Cruz with 62 positions***

Thus, with 188 and 175 open Data Science job openings, Los Angeles and San Francisco are two places in California that requires more Data Scientist. So, ***students or working professionals who would like to pursue Data Science career irrespective to any Domain specifications can have Los Angeles and San Francisco as their main destination to search for Data Science jobs*** followed by other counties like San Mateo, Santa Clara and Santa Cruz in California.

Other than this we see 10 counties like San Diego, Santa Barbara, etc., have low job positions ranging between 1 and 50. While there are 43 counties, more towards the north and central California like Trinity, Kings, etc., have no Data Science job positions open.

## Map 2.2 Data Science Positions with respective to Congressional Districts in New York State

After California it is New York State which has high number of Data Science job openings in US. In order to represent this map one step above the granular level shown for California, I selected unit of analysis as Congressional District while the Geographic extent remains the same ‘State’.

From Map-2.2 we could infer that there are 3 congressional districts[[4]](#_Reference_1) which have high number (more than 50) of Data Science job positions open, they are:

1. ***Brooklyn (10) with 83 Positions***
2. ***Manhattan (12) with 83 Positions***
3. ***Harlem (13) in Manhattan wit 13 Positions***

With 83 open Data science position Brooklyn and Manhattan in New York can also be considered as a next destination for people who want to be a Data Scientist. These are the two main places where headquarters for many financial sectors like Citi Bank Global Headquarters, JP Morgan Chase Headquarters, etc., are present. Thus, ***for people who would like to pursue their Data Science career more specific in financial domain can concentrate to apply for jobs in Brooklyn and Manhattan*** to grab one.

There are 6 congressional Districts like Richmond (11) and Kings (9), etc., with low open data science position, while there are 4 congressional districts away from Manhattan and Brooklyn like Nassau (3), Queens (6), etc., have no open Data Science position.

## Map 2.3 Data Science Positions with respective to Congressional District in Washington State

Based on Indeed data extracted on 09 Aug 2017, Washington holds the third rank with respective to open Data Science position with a total of 249. To observe the spread of these position across Washington we constructed Map-2.3 based on Washington’s Congressional District.

From map we see that there are 2 congressional district[[4]](#_Reference_1) (09 & 10) having high job openings for Data Science:

1. ***Seattle (09) with 160 positions***
2. ***Bellevue (10) with 81 positions***

With 160 open positions Seattle seems to be the best place to pursue Data Science career in Washington. There are major giants like Amazon in Seattle and Getty Image which generates more Data Science jobs in Seattle. There are other giants like Microsoft and Concur which generates more Data Science job in Bellevue.

There are 3 congressional district like Lewis (03) and Whatoom (01) and 02 have low Data Science position with less than 50 open Data Science position. While there are other congressional districts like Kings (08) and Clallam (06), etc., which are shaded in Green have no open Data Science position in these congressional districts. Thus, ***to pursue a career in Washington it is Seattle which would be the best place to apply and plan for***.

## Map 2.4 Data Science Positions with respective to Congressional District in Illinois State

With 248 open Data Science position Illinois is the 4th state with more job openings for Data Scientist. Map-2.4 is constructed to see the distribution of these 248 jobs across Illinois Congressional Districts.

Based on the map we see that there are 2 congressional district[[4]](#_Reference_1) (7 & 9) have more job openings for Data Scientist:

1. ***Chicago (7) with 168 Positions***
2. ***Evanston (9) with 63 Positions***

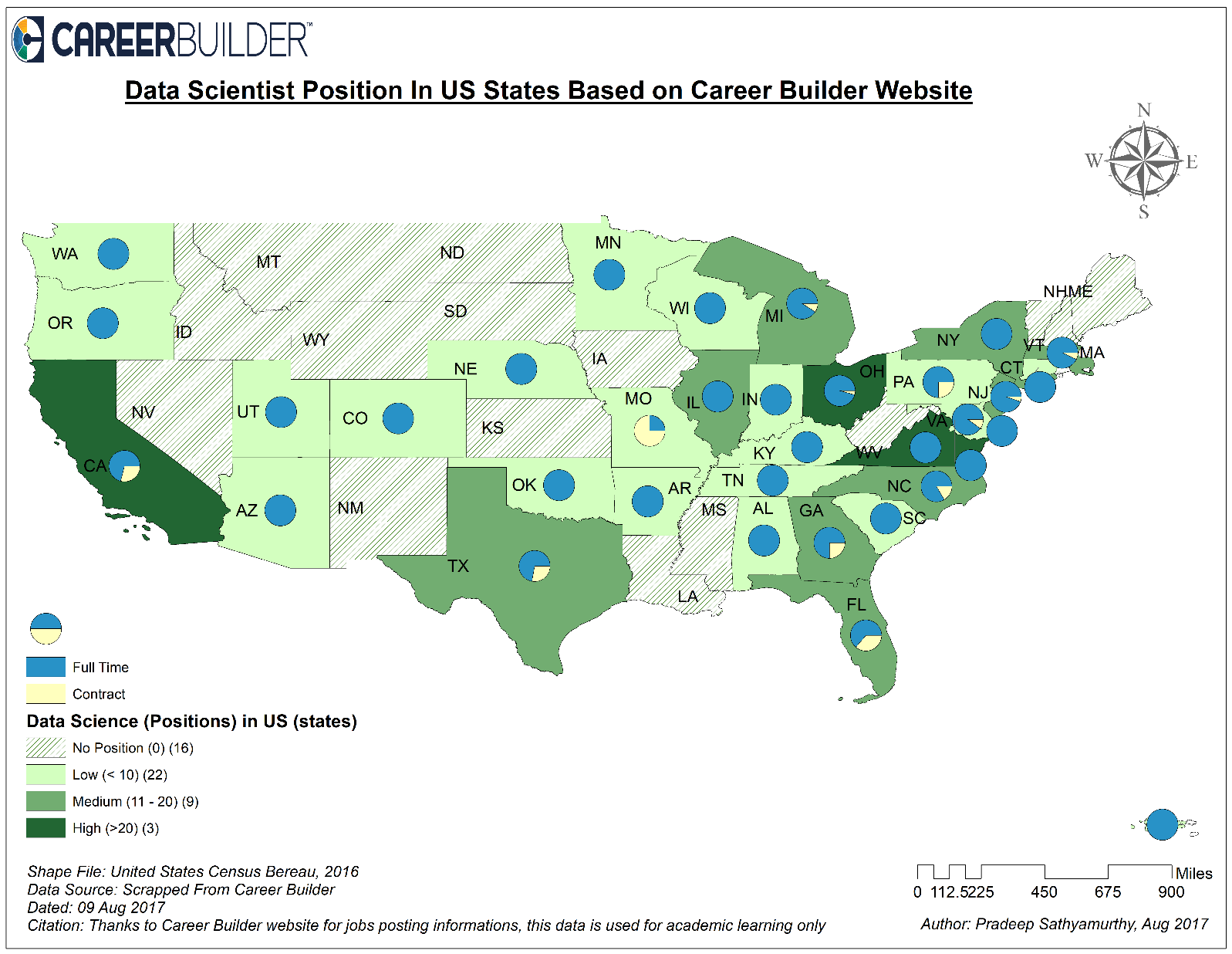
As expected, Chicago, the most populous city of Illinois and the third populous city in US holds the highest number of job openings when compare to other congressional districts. There are major consulting firms like KPMG, McKinsey & Company that open-up more Data Science position for people who would like to pursue Data Science career in Chicago. Irrespective to these, there are other firms like Google, General Electric (GE), Nielson, Uptake which generates more Data Science positions in Chicago.

North Western University in Evanston generate more frequent Data Science position for their research needs, though we couldn’t observe any other major industry around Evanston district, from map we can observe even this being a major contributor for Data Science jobs in Illinois.

We do have 8 congressional districts like 18, 13, 15, 17, etc., which comprises areas like Springfield, Schaumburg, Riverwoods, Bloomington, Des Plaines, etc., with industries like Discover Financial Services, State Farm Mutual Automobile Insurance Company, ABBOTT LABORATORIES, etc., which open-up more Data Science position for people who want to pursue Data Science career in the state of Illinois.

From map, we observe that there are 8 congressional districts represented in shade like 1, 2, 12, etc., which comprises areas like Herrin, Marion, Kankakee, etc., have zero data science positions. Thus, ***for students and professionals who search for Data Science jobs in Illinois, Chicago would be the ideal place to apply to excel in Data Science Career***.

# 7. Map3 - Data Science Position in US States based on Career Builder website



|  |  |  |  |
| --- | --- | --- | --- |
| SI.NO | Map Name | Geographic Extent | Unit of Analysis |
| 1 | Map-3 | United States of America | State |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SI.NO | Map Name | Shape File Name | Attribute Excel Work Book Name | Sheet Name |
| 1 | Map3 | tl\_2016\_us\_state.shp | CB\_Treated\_Data\_Scientist\_\_09Aug2017.xls | Data\_To\_Map |

# 8. Map3 Findings

Based on [Map-1](#_3._Map1_–) and [Map-2](#_5._Map2_–) we can conclude that ***California (CA) generates more Data Science position*** when compared to other states in US. After California, we also have states like ***New York (NY), Washington (WA) and Illinois (IL) which have more number of open positions for Data Scientist***. Besides, from these 4 big states we can infer that places like ***Los Angeles, San Francisco, Brooklyn, Manhattan, Seattle and Chicago are the few of the major cities where there are high number of jobs for Data Scientist***.

Though, we know the states and cities that create high demand for Data Scientist, we are not sure if all these positions are Full-time or part-time because the job type information was not available as part of attribute data extracted from Indeed. In order to visualize the Data Science job nature with respective to job type, I scraped data from Career Builder website.

Career Builder is another employment based search engine famous in market. In January 2013, based on web traffic ranking, Career Builder second behind Indeed.com. However, we are not aware of the current ranking of Career Builder. Though the number of job posting obtained from this website is comparatively less in count when compare to Indeed, they do frequent update on job listing. ***One main reason for me to choose this website is, this is one of the few job search portals that has information about Job type***.

Job Type is classified into two main categories by career builder:

1. ***Full-Time***
2. ***Contract***

Based on this extra information obtained, we mapped Map-3 shown above to observe the states with high Data Science job openings and the pattern that exist in this role when it comes to Job type. ***We would like to infer whether industries trust contract worker to share their data and if so what is the existing pattern of such case across US***.

Since the range of open job position count differs between Map-1 and Map-3 direct comparison of these Map is not advisable. However, we can do a cross validation between these two map. Thus***, looking at Map-3 it is still evident that California holds highest job positions*** with 38 vacancies based on the data extracted on 09 Aug 2017. However, we could see states like NY, WA and IL fall behind OH, VA, NJ, TX and GA based on Map-3.

There are totally 3 states California (CA), Ohio (OH) and Virginia (VA) have high job positions that is greater than 20 openings on 09 Aug 2017 in Career Builder. We see there are 9 states like Texas (TX), New York (NY), Illinois (IL), etc., shown in light green color are states with medium data science job openings ranging between 11 and 20. However, it was quite surprising to see states like Washington (WA) falls behind these states with respective to Career Builder data. Thus, there are 22 states like Michigan (MI), Utah (UT), Arizona (AZ), etc., have low Data Science positions and there are 16 states like North Dakota (ND) and South Dakota (SD), etc., have no Data Science job opening which are almost similar to those with no data science position in Map-1.

Also in Map-3 we have used Pie Chart to represent the proportional difference between Full-time and Part time job openings with respective to Data Science. In most of the States with high Data Science position like California, Ohio and Virginia we see there are more full time positions when compare to part-time positions***. Except Missouri (MO) all other states with open Data Science position have more Full-time jobs when compare to Contract jobs.*** Thus, from map we see very few states like California, Texas, Florida, Missouri have contract positions greater than 3 on 09 Aug 2017 based on Career Builder website, while the other ***states are more dominated with Full-time Data Science positions only***.

Thus, from this [Map-3](#_7._Map3_-) it is quite evident that California is the state with more Data Science job openings and ***most of the companies across United States look for Full-Time data scientist for their company when compare to contract workers***. With this we can ***hypothesise that companies do not like to share their data to outsiders and would like to have their own Data Scientist team in every organization*** to find insights from data shared.

# 9. Future Works

As part of this project I have kept the graph simple[[5]](#_Reference_1) without making it complex for user to read and understand. This goes as per my course was thought ‘Keep it Simple Silly’. However, with the data extracted from Indeed, we can make advance to further level in plotting map with user review and salary information’s.

## Web data mining with Attribute data and mapping its detail

As part of the data scraped from Indeed website, User review is one of the attribute obtained along with the position count for a Data Science job. These are reviews given by the users who have applied for a company and given a review about that company based on their overall experience. Over all experience includes interview process, communication, salary parameter, etc., Thus, better the company is higher the rating & review it is obtained. This information is one of the main attribute for web data mining.

Web data mining is the application of data mining and machine learning technique to extract useful knowledge from the web content, structure and usage of web resources. Importance of web mining in this space using a ***predictive user modelling dynamically serve customized content (ads, products, deals, recommendation, etc.,) to users based on their profile, preferences or expected interest.*** Need for this web data mining for such websites arises to grow customer loyalty.

Web data mining has 3 different flavours, Content mining, usage mining and structure mining. ***As part of this project we have explored one of the flavour of content mining at a very basic level***. User review data is one such data which helps to do more advance web content mining to:

1. Document clustering or categorization
2. Topic Identification and tracking
3. Concept discovery
4. Focused Web Crawling
5. Content based personalization
6. Intelligent search tool

## Unsupervised Learning with Review Data and map user based result

Most Web content mining and information retrieval applications involve measuring similarity among two or more documents. Vector representation facilitates similarity computations using vector-space operations (such as Cosine of the angle between two vectors) with the data obtained from indeed and career builder which are employment based search engine we can ***measure the similarity between a query and index document vector to return a ranked list of relevant document***.

Also, with the review data we can apply some unsupervised learning like clustering techniques to group the companies based on similarity and dissimilarity distance among them. ***Result obtained from clustering result can be used in plotting map in the US states shape file to show the domain richness and domain based Data Scientist need from each state***.

This way, a data scientist who is inclined toward any particular domain like Banking, Consulting, Retail, manufacturing, health care, etc., can prioritize state based on high needs and start with his application process and plan his future career accordingly. This ***way we can provide a focused and personalized search for Data Science jobs to user***.

## Generalizing the Map with user Search

I would like to submit this idea to Indeed and career builder by creating a web application and providing a search text box which is independent of job title and not restricting to only Data Science jobs. This way user who search for a particular job title can visualize a map to see the maximum number of job position in a particular state and in particular city to be more specific and accordingly make their job search effort more efficient and productive.

# 10. Timetable

|  |  |  |
| --- | --- | --- |
| **Phases** | **Description of Work** | **Start and End Dates** |
| **Phase One** | Idea Proposal | 17-Jul to 18-Jul 2017 |
| **Phase Two** | Validating the best Job search portals | 20-Jul to 25-Jul 2017 |
| **Phase Three** | Studied Career Builder Job Portal | 08-Aug to 10-Aug 2017 |
| **Phase Four** | Studied Indeed Job Portal | 08-Aug to 10-Aug 2017 |
| **Phase Five** | Scraped Data from Career Builder | 09-Aug to 11-Aug 2017 |
| **Phase Six** | Scraped Data from Indeed Job Portal | 09-Aug to 11-Aug 2017 |
| **Phase Seven** | Shape File Identification for US states | 11-Aug to 12-Aug 2017 |
| **Phase Eight** | Shape File Identification for Top 4 cities with high Data Science Position (CA, NY, WA, IL) | 11-Aug to 12-Aug 2017 |
| **Phase Nine** | Data Cleanup and Attribute Data Creation | 12-Aug to 13-Aug 2017 |
| **Phase Ten** | Built Join on US States Shape file with Indeed Data to produce map | 13-Aug to 14-Aug 2017 |
| **Phase Eleven** | Built Join on US States Shape file with Career Builder Data to produce map | 13-Aug to 14-Aug 2017 |
| **Phase Twelve** | Built Join on US States Shape file with Career Builder Data to produce map | 13-Aug to 14-Aug 2017 |
| **Phase Thirteen** | Analyze Map and Reporting | 14-Aug to 15-Aug 2017 |

# 11. Key Personnel

|  |  |
| --- | --- |
| Team Member | Pradeep Sathyamurthy |
| Professor | Prof. Nandhini Gulasingam |
| Project for | GEO-441 |
| Target Team | DePaul CDM |

# 12. Deliverables

|  |  |  |
| --- | --- | --- |
| Final Report | **Prady\_GEO\_441\_Report.docx** | Contains final Technical Report |
| Map Package | **Prady\_Map\_Package.rar** | Contains important map packages |
| Deliverables | **Deliverables.rar** | Contains all Attribute and Shape file data |
| Map-1 | **Map1.png** | High Resolution Image File for [Map1](#_Exploratory_Data_Analysis) |
| Map-2 | **Map2.png** | High Resolution Image File for [Map2](#_Map2_–_Insight) |
| Map-3 | **Map3.png** | High Resolution Image File for [Map3](#_Map3_-_Data) |
| Python Script-1 | **Data\_Scraping\_Indeed.py** | Python Script to scrap Indeed Data |
| Python Script-2 | **Data\_Scraping\_Career\_Builder.py** | Python Script to scrap Career Builder Data |
|  |  |  |

# 13. Reference

*[1] Research paper on Social Media Analytics* <https://link.springer.com/article/10.1007/s00146-014-0549-4>

*[2] Occupational employment statistics* <https://www.bls.gov/oes/current/oes194092.htm>

*[3] Article from USA Today College on Top 10 US cities hiring new graduates –* [click here](http://college.usatoday.com/2017/05/11/apply-now-if-you-want-a-job-after-graduation-linkedin-data-shows/)

*[4] Indeed blog about top cities with STEM jobs in US –* [click here](http://blog.indeed.com/2015/03/09/find-your-next-stem-job-in-one-of-these-10-cities/)

*[5] Bilingual job opportunities in US –* [click here](http://blogs.sas.com/content/sastraining/2017/06/14/bilingual-job-opportunities-in-the-u-s/#prettyPhoto)

# 14. Zip File with Map Packages



# 15. Appendix

## Python-Code for Indeed Website Data Scraping

|  |
| --- |
| ***# -\*- coding: utf-8 -\*-***  ***"""***  ***Created on Tue Aug 8 20:14:25 2017***  ***@author: pradeep sathyamurthy***  ***Course: GEO-441 - Geographic Information Systems***  ***Data Scraing from indeed website***  ***"""***  import urllib.request as url  from bs4 import BeautifulSoup as bs  import pandas as pd  import os  os.chdir('D:\Courses\GEO441 - GIS For Community Development\Project')  # Functions for scraping data  def title\_data(obj):  bs\_obj0 = bs(obj,'html.parser')  job\_tags = bs\_obj0.find('a', {"data-tn-element": "jobTitle"})  if job\_tags:  job\_data1 = job\_tags.text  job\_data = job\_data1.replace(',','')  else:  job\_data = 'NoData'  return job\_data.strip()  def com\_data(obj):  bs\_obj1 = bs(obj,'html.parser')  com\_tags = bs\_obj1.find('span', {"class": "company"})  if com\_tags:  com\_data = com\_tags.text  company\_data = com\_data.replace(',','')  else:  company\_data = 'NoData'  return company\_data.strip()  def loc\_data(obj):  bs\_obj2 = bs(obj,'html.parser')  loc\_tags = bs\_obj2.find('span', {"class": "location"})  if loc\_tags:  loc\_data = loc\_tags.text  else:  loc\_data = 'NoData'  return loc\_data.strip()    def rev\_data(obj):  bs\_obj2 = bs(obj,'html.parser')  rev\_tags = bs\_obj2.find('span', {"class": "slNoUnderline"})  #print(rev\_tags)  if rev\_tags:  rev\_data = rev\_tags.text  review\_data = rev\_data.replace(',','')  else:  review\_data = 'NoData'  return review\_data  def whole\_data(obj):  whl\_lst = list()  bs\_obj = bs(obj,'html.parser')  whl\_tags = bs\_obj.findAll('div', {"data-tn-component": "organicJob"})  #print(whl\_tags)  for tag in whl\_tags:  tag = str(tag)  #print(tag)  title = title\_data(tag)  company = com\_data(tag)  review = rev\_data(tag)  location = loc\_data(tag)  single\_div\_tag\_data = title + ',' + company + ',' + review + ',' + location  whl\_lst.append(single\_div\_tag\_data)  return whl\_lst    # sending a get request to career builder site  # Site URL with job key is https://www.indeed.com/jobs?q=Data+Scientist&start=  # with pagination: https://www.indeed.com/jobs?q=Data+Scientist&start=10  # Pagination is a multiple of 10 in indeed, that is page 1 is referred as 10, which means the number of records retrived  # 10 records are retreived for a page  job\_key = 'Data Scientist'  job\_frame\_key = job\_key.strip().split()  job\_url\_key = 'jobs'+'?q='+job\_frame\_key[0]+'+'+job\_frame\_key[1]+'&'+'start='  #print(job\_url\_key)  final\_url= 'https://www.indeed.com/' + job\_url\_key  print(final\_url)  url\_read = url.urlopen(final\_url).read()  bs\_obj = bs(url\_read,'html.parser')  page\_tag = bs\_obj.find('div', {"id": "searchCount"})  page\_tag\_txt = page\_tag.text  page\_tag\_data = page\_tag\_txt.split()  total\_rec = int(page\_tag\_data[-1].replace(',',''))  total\_pages = round(total\_rec / 10)  print(total\_rec)  print(total\_pages)  # Data scraping from each page from 1 to 10 = range(1,11)  final\_data = list()  title\_lst = list()  com\_lst = list()  loc\_lst = list()  review\_lst = list()  whl\_lst = list()  for i in range(0,total\_pages,10):  final\_url\_paginated = final\_url + str(i)  data\_html\_cb = url.urlopen(final\_url\_paginated).read()  col0 = whole\_data(data\_html\_cb)  whl\_lst.append(col0)  final\_data.append(data\_html\_cb)  # Building data for each row listed as part of the website  final\_title\_lst = list()  final\_com\_lst = list()  final\_rev\_lst = list()  final\_city\_lst = list()  final\_state\_lst = list()  final\_country\_lst = list()  city = ''  state = ''  county = ''  for lst in whl\_lst:  for item in lst:  city = ''  state = ''  county = ''  items = item.split(',')  data0 = items[0].strip()  data1 = items[1].strip()  data2 = items[2].strip()  if (len(items)==4):  data3 = items[3].strip()  if (data3.upper() == 'UNITED STATES'):  country = 'United States'  else:  city = item[-1]  elif (len(items)==5):  country = 'United States'  #state = items[-1].strip()  state\_data = items[-1].split()  state = state\_data[0]  city = items[-2].strip()  elif (len(items)>5):  country = 'United States'  city = items[-2]  state\_data = items[-1].split()  state = state\_data[0]  final\_title\_lst.append(data0)  final\_com\_lst.append(data1)  final\_rev\_lst.append(data2)  final\_city\_lst.append(city)  final\_state\_lst.append(state)  final\_country\_lst.append(country)  # Data count from each extract  len(final\_title\_lst)  len(final\_com\_lst)  len(final\_rev\_lst)  len(final\_city\_lst)  len(final\_state\_lst)  len(final\_country\_lst)  # Creating a data frame  dtf = pd.DataFrame({'Title':final\_title\_lst})  dtf['Company'] = final\_com\_lst  dtf['Review'] = final\_rev\_lst  dtf['City'] = final\_city\_lst  dtf['State'] = final\_state\_lst  dtf['Country'] = final\_country\_lst  # Create a csv file  dtf.to\_csv('Indeed\_Raw\_Data\_Scientist\_09Aug2017.csv') |

## Python-Code for Career Builder Website Data Scraping

|  |
| --- |
| ***# -\*- coding: utf-8 -\*-***  ***"""***  ***Created on Sun Aug 6 00:57:02 2017***  ***@author: pradeep sathyamurthy***  ***Course: GEO-441 - Geographic Information Systems***  ***Data Scraping from Career Builder website***  ***"""***  import urllib.request as url  from bs4 import BeautifulSoup as bs  import pandas as pd  import os  os.chdir('D:\Courses\GEO441 - GIS For Community Development\Project')  # Functions for scraping data  def title\_data(obj):  job\_lst = list()  bs\_obj0 = bs(obj,'html.parser')  job\_tags = bs\_obj0.findAll('h2', {"class": "job-title hide-for-medium-up"})  for tag in job\_tags:  job\_data = tag.text  print(job\_data)  job\_lst.append(job\_data)  return job\_lst  def pos\_data(obj):  pos\_lst = list()  bs\_obj1 = bs(obj,'html.parser')  pos\_tags = bs\_obj1.findAll('h4', {"class": "job-text employment-info"})  for tag in pos\_tags:  pos\_data = tag.text  print(pos\_data)  pos\_lst.append(pos\_data)  return pos\_lst  def com\_data(obj):  com\_lst = list()  bs\_obj2 = bs(obj,'html.parser')  com\_tags = bs\_obj2.findAll('div', {"class": "columns large-2 medium-3 small-12"})  for tag in com\_tags:  com\_data = tag.text  print(com\_data)  com\_lst.append(com\_data)  return com\_lst    def loc\_data(obj):  loc\_lst = list()  bs\_obj = bs(obj,'html.parser')  loc\_tags = bs\_obj.findAll('div', {"class": "columns end large-2 medium-3 small-12"})  for tag in loc\_tags:  loc\_data = tag.text  print(loc\_data)  loc\_lst.append(loc\_data)  return loc\_lst  # sending a get request to career builder site  # Site URL with job key is http://www.careerbuilder.com/jobs-data-scientist?location=  # with pagination: http://www.careerbuilder.com/jobs-data-scientist?page\_number=1  job\_key = 'data scientist'  job\_frame\_key = job\_key.strip().split()  job\_url\_key = 'jobs'+'-'+job\_frame\_key[0]+'-'+job\_frame\_key[1]+'?'+'page\_number='  #print(job\_url\_key)  final\_url= 'http://www.careerbuilder.com/' + job\_url\_key  print(final\_url)  url\_read = url.urlopen(final\_url).read()  bs\_obj = bs(url\_read,'html.parser')  page\_tag = bs\_obj.find('span', {"class": "page-count"})  page\_tag\_txt = page\_tag.text  page\_tag\_data = page\_tag\_txt.split()  total\_pages = int(page\_tag\_data[-1])  print(total\_pages)  # Data scraping from each page from 1 to 10 = range(1,11)  final\_data = list()  title\_lst = list()  pos\_lst = list()  com\_lst = list()  loc\_lst = list()  for i in range(1,total\_pages):  final\_url\_paginated = final\_url + str(i)  #print(final\_url\_paginated)  data\_html\_cb = url.urlopen(final\_url\_paginated).read()  col0 = title\_data(data\_html\_cb)  col1 = pos\_data(data\_html\_cb)  col2 = com\_data(data\_html\_cb)  col3 = loc\_data(data\_html\_cb)  title\_lst.append(col0)  pos\_lst.append(col1)  com\_lst.append(col2)  loc\_lst.append(col3)  final\_data.append(data\_html\_cb)  final\_title\_lst = list()  for obj in title\_lst:  for item in obj:  data0 = item.strip()  #print(data1)  final\_title\_lst.append(data0)  final\_pos\_lst = list()  for obj in pos\_lst:  for item in obj:  data1 = item.strip()  #print(data1)  final\_pos\_lst.append(data1)  final\_com\_lst = list()  for obj in com\_lst:  for item in obj:  data2 = item.strip()  #print(data2)  final\_com\_lst.append(data2)    final\_loc\_lst = list()  for obj in loc\_lst:  for item in obj:  data3 = item.strip()  #print(data3)  final\_loc\_lst.append(data3)  # Data count from each extract  len(final\_title\_lst)  len(final\_pos\_lst)  len(final\_com\_lst)  len(final\_loc\_lst)  state\_lst = list()  address\_lst = list()  addr = ''  for items in final\_loc\_lst:  item = items.strip()  item1 = item.split()  if(len(item1)==2):  print(item1)  print(len(item1))  state = item1[1]  addr = item1[0]  print(state)  print(addr)  elif(len(item1)==3):  print(item1)  print(len(item1))  state = item1[-1]  addr = item1[0] + item1[1]  print(state)  print(addr)  elif(len(item1)==4):  print(item1)  print(len(item1))  state = item1[-1]  addr = item1[0] + item1[1] + item1[2]  print(state)  print(addr)  elif(len(item1)==5):  print(item1)  print(len(item1))  for item in item1:  if len(item) == 2:  state = item  else:  addr=addr+item  #addr = ''  print(state)  print(addr)  elif(len(item1)==6):  print(item1)  print(len(item1))  for item in item1:  if len(item) == 2:  state = item  else:  addr=addr+item  #addr = ''  print(state)  print(addr)  state\_lst.append(state)  len(state\_lst)  address\_lst.append(addr)  len(address\_lst)  # Creating a data frame  dtf = pd.DataFrame({'JOb\_Title':final\_title\_lst})  dtf['Position'] = final\_pos\_lst  dtf['Company'] = final\_com\_lst  dtf['Location'] = final\_loc\_lst  dtf['State'] = state\_lst  dtf['Address'] = address\_lst  # # Create a csv file  dtf.to\_csv('CB\_Raw\_Data\_Scientist\_8Aug.csv') |