Visualizing H1B petitions

CSCI55200 Spring 17 Project Report

Instructor: Dr. Shiaofen Fang

Adithya Morampudi and Yashwanth Reddy Kuruganti

May 8, 2017

Contents

1 Abstract

Today the world is driven by Data.There is an evidence to this statement. IBM results say, 90 percent of the data in the world is created in the last two years. Huge chunks of data are being generated every single day. This data is being stored for generating some useful insights and gain some valuable information from it. In this project, we are going to visualize a data set pertaining to H-1B dataset. We will make many helpful inferences from this dataset using different data visualization techniques. Some of the insights which we are going to find are which part of the United States has the largest number of Data Scientists, which part has the least number of Software engineers, which part of the United States has the highest number of high paid jobs and which part has the highest number of low paid jobs. We will also be finding which job has the highest number of H-1B petitions in a year.

2 Motivation

The main motivation behind considering this dataset is the fact that most of the strength are international students and we wanted to let them know what are the statistics of H-1B visas for a particular Industry and particular job type. Following the Recent Presidential Elections, there has been a lot of chaos in the United States about the H1-B visas and skilled immigrants, this was also a contributing factor in making us think about this dataset. Also, the resources which we have made us think about this huge dataset.

3 Introduction

H-1B Visa, what exactly is it? Well, let us have a closer look at what it is, H1-B, a non-immigrant visa which allows students and foreign professionals to work in the United States in specialized job positions, these job positions are given to professionals or students who have a high expertise in a related domain, these domains are mostly IT, Engineering, Mathematics etc. All the jobs require at least a minimum of US equivalent bachelors degree, people which meet these qualifications are eligible for applying H-1B visa. When we look at this the concept look good for non-immigrants but the catch here is that all the people who file a petition for H-1B may not be eligible or may not be granted a visa because of several visa caps. For a nonimmigrant to work in the united states H-1B is the most appropriate visa type when compared to others considering processing times and other overheads.

The present immigration law allows a total of 85,000 H-1B visas each year out of which 20,000 are reserved for students who are graduated from a US certified educational institution. There are also some eligibilities for qualifying for the H-1B visa, He/she must have a bachelors or equivalent degree, the nature of the job should be specialized. This visa type also has a limit on the length of stay, initially it is given for 3 years and later it can be extended to a maximum of 6 years. Our analysis would focus mainly on the number of petitions received each year, the weightage of each position towards the total number of petitions, the company which is filing highest number of petitions, and the present wage rate pertaining to the respected job position.

4 Goals of the system

The main objective of our project would be to find the positions which has a high success rate in the H-1B program, we need great insights to help us decide which company to choose from when we graduate, as it may affect our career heavily if not chosen properly. We also will get insights about the highest paying jobs, the regions which pay the highest salary for a job position, are the number of software developers growing year to year? Are the salaries for the data scientists growing yearly, which region has the highest number of software developers etc. this will be our main objective.

Apart from the result, we will be able to learn different visualization techniques from this project, we will be learning about geo visualization[1], stacked bar graphs[2], bubble graphs, playing with huge datasets, data preprocessing, donut charts, different transitions, functions used to create them, learnings about Geo Json, TopoJson etc.

5 Dataset Description

As mentioned earlier about the interest in H-1B petitions, we did a lot of extensive research to find the right dataset for our visualization. As we wanted the outcomes of this project to be precise, after some exhaustive search we found the dataset pertaining to H-1B petitions from the fiscal year $2011 - 2016$, this dataset was extracted from OFLC(office of foreign labor certification) and published in KAGGLE [3] for free. The dataset has 10 attributes with more than 30 lakh records, below is the description of the dataset.

5.1 Attributes and its description

i Case Status : This Describes the Details about status of an applicant after LCA processing, whether he/she is eligible for applying to H-1B. Typically there are four values for this attribute

- Certified: This means that the application is processed by LCA and is eligible for filing a petition at USCIS for H1-B.
- Certified withdrawn: This status means that the application was processed by LCA and is certified for applying to H-1B but is withdrawn by the company or individual, typically there are very less cases pertaining to this case.
- Rejected: This means that the application is rejected by LCA for filing H-1B at USCIS
- Withdrawn: This status means that the application is not processed by LCA and is withdrawn by the individual.

ii Employer Name: This field denotes the Name of the employer who applies for H-1B on behalf of the employee.

iii SOC Name: Based on Standard Occupational Classification systems code, a name is given to the category of job application.

iv Job Title: Employees job title specific to company

v Full Time Position: Y or N is assigned based on employees work status, either full time or part time

vi Prevailing Wage: Average salary for particular position in USD paid by the employer to its employee at the time of application.

vii Year: Visa petition is filed in this year.

viii Worksite: Employers communication address with city and state as its value.

ix LAT: latitude location co-ordinate of Worksite

x LON: longitude location co-ordinate of the Worksite.

6 Data Processing Techniques

Any visualization technique requires proper data preprocessing, as to avoid any results which are misleading, we did data preprocessing for our data to be clean and proper to get good insights, some things like replacing missing values with mean values where done. For the bubble chart, we had to aggregate the values based on the state name by extracting it from Worksite attribute, so we performed aggregation for bubble charts considering the number of H-1B petitions raised from each state. We have written SQL queries to find the top five employers from each state, based on the total number of petitions each employer is filing. We did queries to find the top salary for a position in an area, we also had queries to find the least salary in an area, also we scripted queries to find the top 10 positions each year which are contributing towards the H-1B petitions. After this is done, we have used d3 functions like slice(), nest() etc., for more details.

7 Technical Implementation

We have used MS Access for processing our dataset, D3.js (Data Driven Documents) for visualizing the data. It is a JavaScript Library which gives life to the data by creating beautiful visualizations using SVG elements. The data can be fed into d3.js using CSV, JSON, or TSV and many more. We have used a lot of CSV and Json files as they can be processed and rendered at a higher speed to the screen. Along with D3, HTML and CSS have been used to provide the data with styling elements.

8 Visualization Techniques

Choosing the proper visualization techniques is the main factor for a good visualization, we have chosen

i Geo Visualization: depicts the salary ranges in different parts of the united states

ii Bubble graphs: displays the total number of petitions from each state and shows the top five employers from each state iii Stacked Bar Graph: depicts the number of petitions each positions is contributing from the total number of petitions.

iv Donut Chart: gives a more detailed overview about the top ten positions in the list of H-1B

9 Results and Work Samples

9.1 Geo Map Visualization

This Technique gives us insight about how various visa applicants (ranging from Sales Assistant to HR) filing a petition are scattered throughout the US. Considering data for each year we tried to visualize how each job title and its associated salary is located on the map using longitude and latitude attributes for a worksite. After visualizing year wise data, we found that there is no considerable difference in the distribution of salaries on average. Due to huge size of the dataset and data retrieval from database, d3 tries to render all the objects to this map, but in a slow pace. So, below example results are taken only for year 2016 and found the following insights.

Attributes used: Prevailing Wage, Job Title, LON and LAT of the Worksite

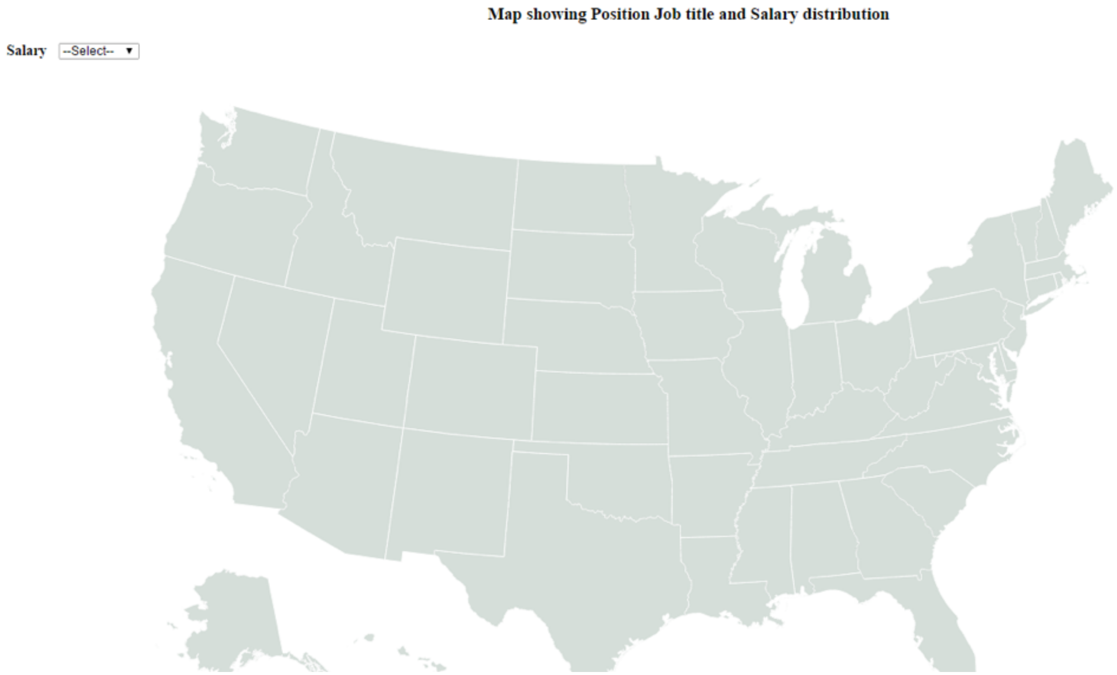
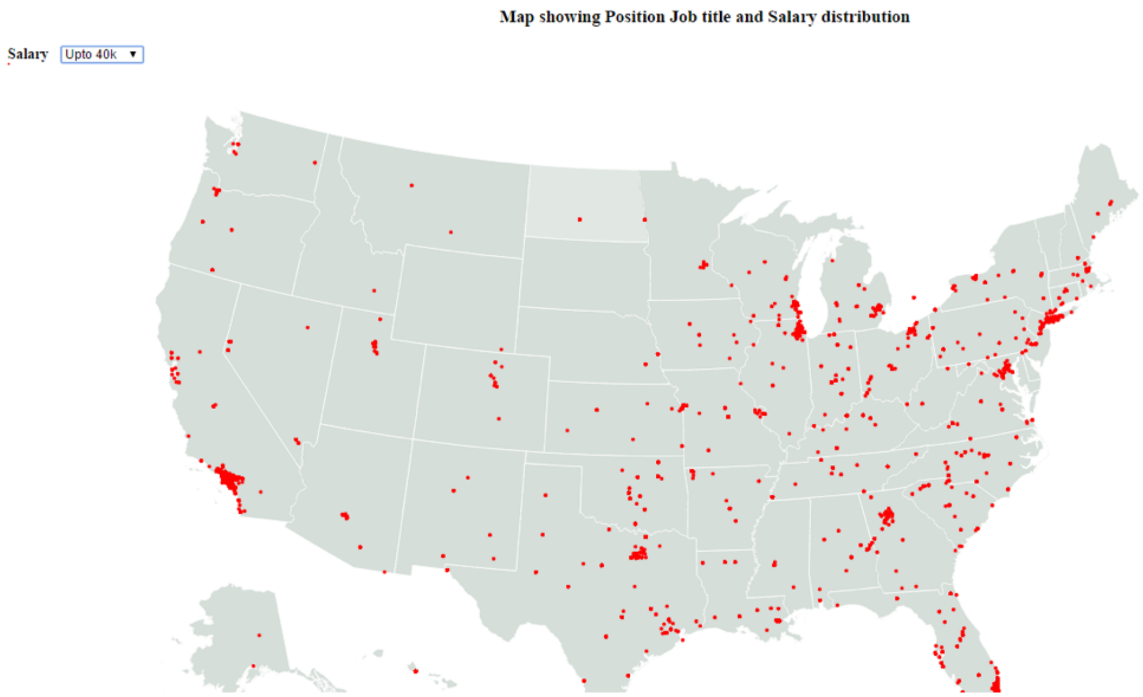


Figure 1: Shows the initial map

The map above shows the initial position of visualization. User can interact with this visualization using Dropdown provided to the left of the screen, by zooming in or out the states and by hovering on each of the location coordinates.

Dropdown contains different salary ranges like less than 40k, 40k to 50k, 50k to 54k, 54k to 58k, 58k to 62k and Above 62k. Figure 2 shows when user selects less than 40k option from the salary dropdown. More dense area can be seen in the Boston, Southern California, Illinois. All these are entry or basic level positions like House

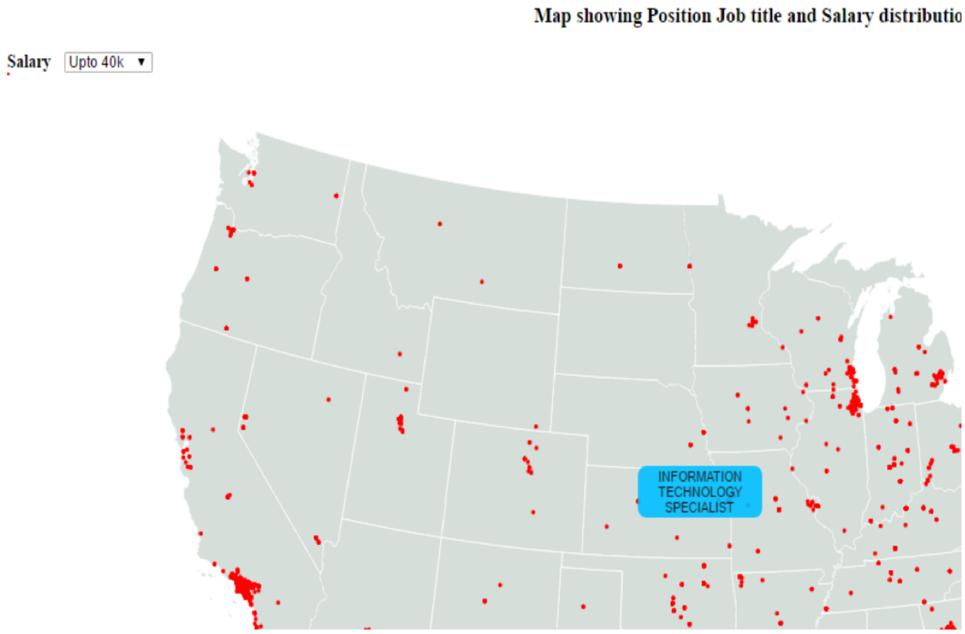


Salary Upto 40k ▼

Figure 2: Shows the salary range from 1-40k

Keeping, Clerk, Assistants etc., Also the east and mid-eastern part of US have good chances of filing a petition for such entry level jobs.

Figure 3 shows the implementation of of zoom functionality and mouse over functionality. For example, users can interact with this map belonging to a particular region or state by clicking the mouse which triggers mouse on click event. This function zooms in or out based on the users operation. To check which Job Title is located at those co-ordinates, user can simply interact by hovering mouse pointer. This triggers OnMouse event and retrieves respective Job title and displays it on the hover box.


Map showing Position Job title and Salary distributio
INFORMATION TECHNOLOGY SPECIALIST

Salary Upto 40k ▼

Figure 3: Implementation of zoom functionality and mouse over functionality

Figure 4. gives two different results on selecting Above 63k option from the drop down. Red bubbles give information about Job title with less than 75k and Black bubbles provide information about Job title with salary greater than 75k. Out of all the salary ranges, people with salary range of 65 - 75k are high in number as we can see that color red dominates the black color in the graph below. All these positions are mostly related to positions like Senior Level Programmer, Human Resource Manager, Business Analyst etc.

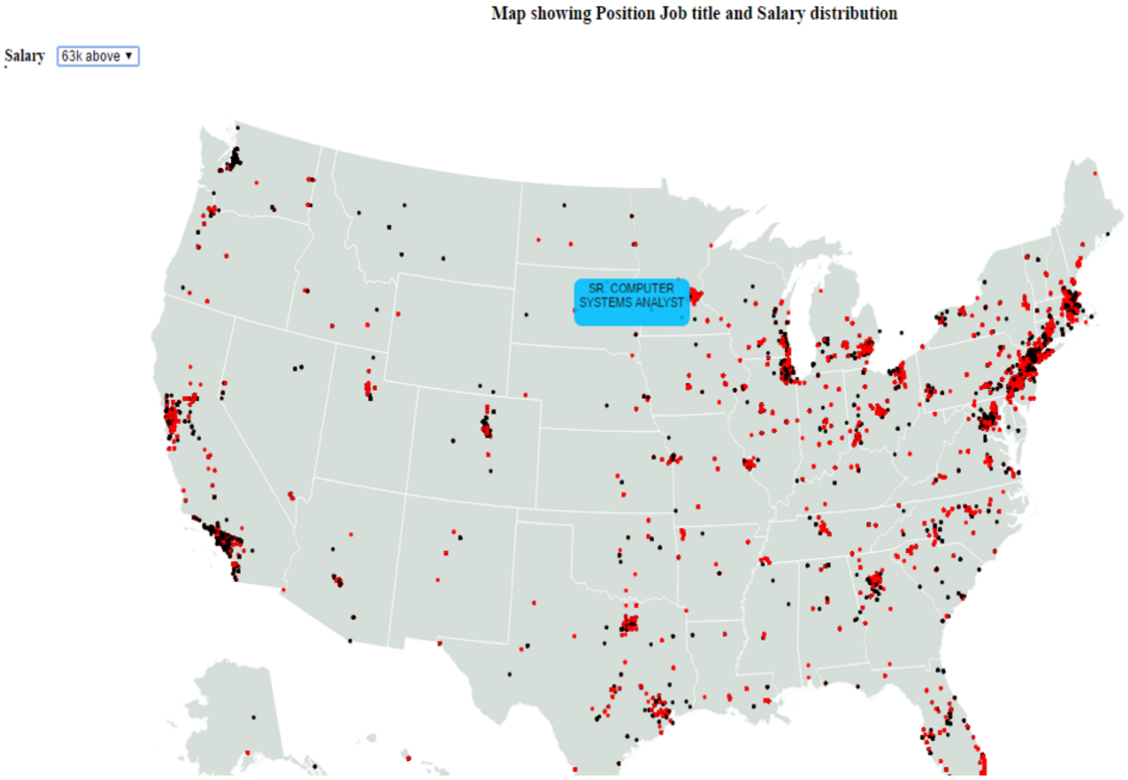

Map showing Position Job title and Salary distribution
Salary
63k above ▾
SR. COMPUTER SYSTEMS ANALYST

Figure 4: salary ranges from 63 and above with different colors

9.2 Force directed Bubble Graph

What if a user wants to interact with a visualization chart and know some information about number of visa petitions and top employers in each state. Bubble layout is chosen to visualize current and past year trends in each state of US.

Attributes Used: Worksite, Year, Count of H1B Petitions

Figure 5 shows us information about the top five employers in each state and the number of application or petitions raised from each state, the color in the bubbles represents different states and the size of the bubbles represent the number of petitions raised from that state, for user interaction we have added MouseOver options, when a user hovers the mouse over a bubble, it displays the information about the total number of petitions raised from that state, and by looking at the size of the bubble we can easily tell the difference between the number of applications or petitions raised from that state.

For further interaction with this graph, we have added some buttons and mouse click functions. The buttons are for changing the years, by which a user can interact with the graph and see the comparison between different

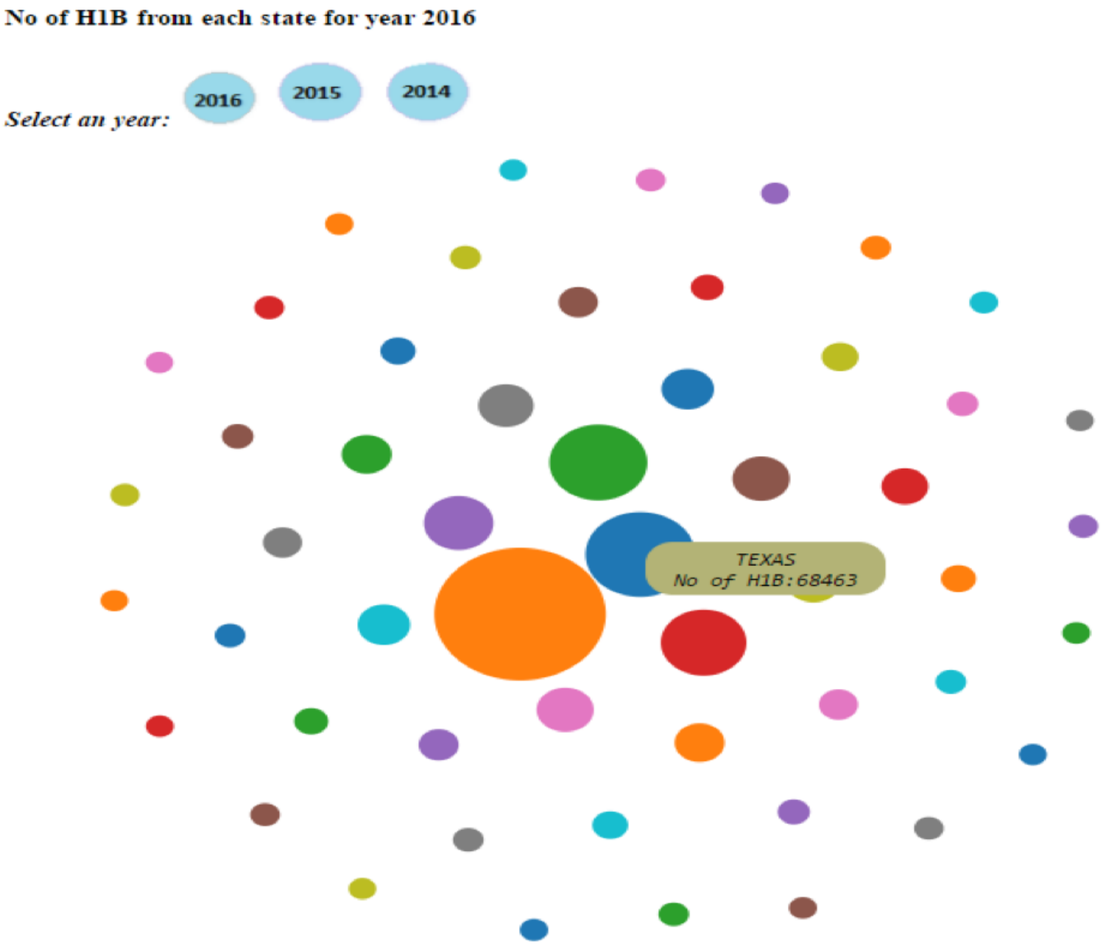


Figure 5: shows the mouse over functions

years, ultimately visualization is best done by comparing, if a user presses the button at the top of the graph, then he/she can see the data from different years. The bubble size changes accordingly to the values. We can easily interpret the results from this visualization.

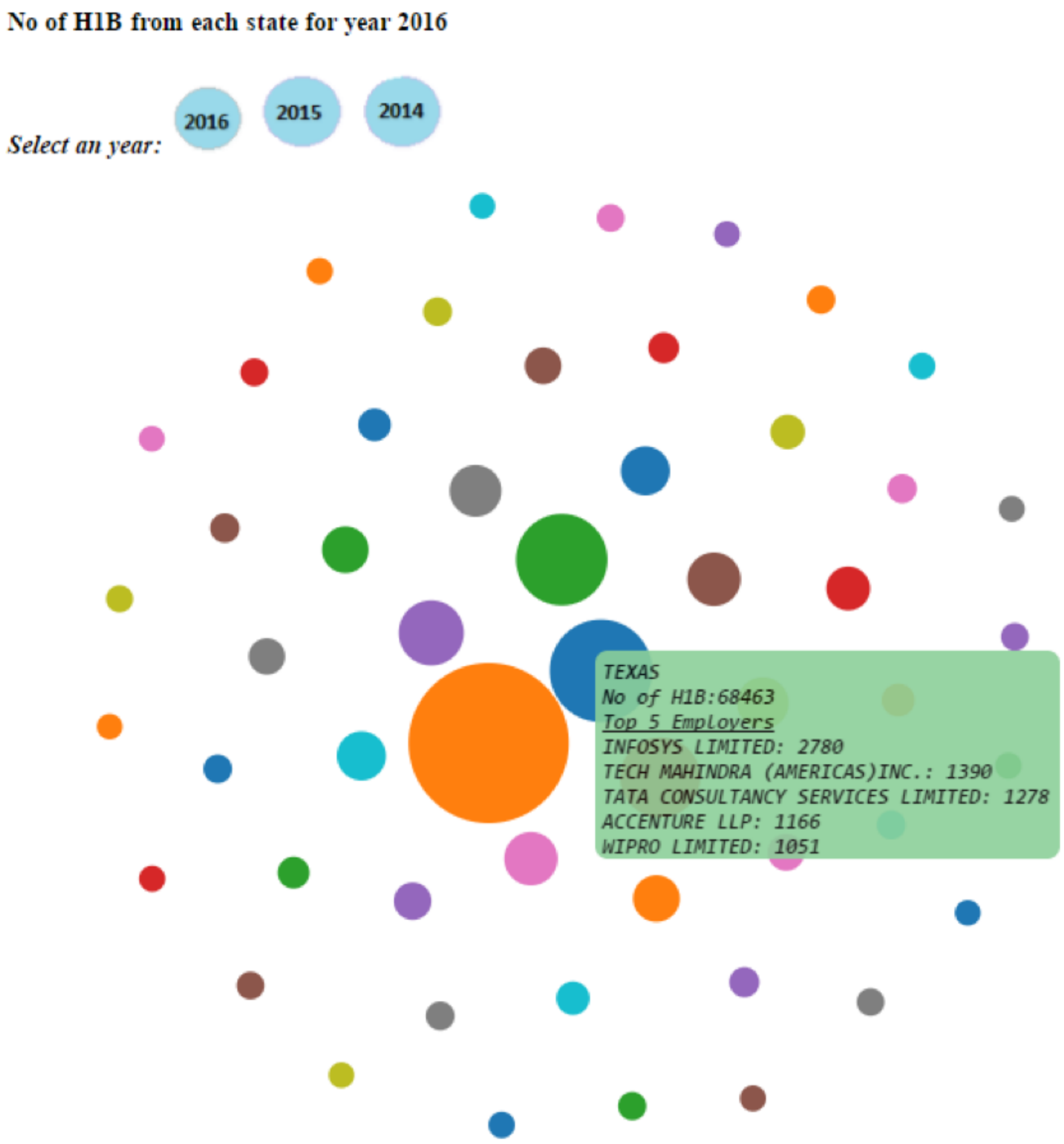
The mouse click function provides the user with mouse click functionalities such as when the user clicks on a bubble, the bubble pops up and a legend is displayed which retrieves details such as the top 5 employers in a state and the number of petitions they have filed each year as in Figure 6.

9.3 Stacked Bar graph and Donut Graph

Geo Visualization and Bubble visualization gave us an overview of the trends in data across years. But what if an user is interested in knowing details about a specific Soc_Name and Full Time Position. So this Multi stacked bar and donut chart provides us with some knowledge on such information.

Attributes used: Soc_Name, Full_Time_Position

Figure 7 shows the screenshot from the project which displays a stacked bar graph, each bar represents a different year, the colors in the bar represents different job roles(Soc_Name), all the colors are represented with the legend next to it. The x axis defines the years and the y axis defines the count of the job roles, so the above graph represents the top 6 and least 6 job positions each year and we can see that the Software Developers are



TEXAS
No of H1B:68463
Top 5 Employers
INFOSYS LIMITED: 2780
TECH MAHINDRA (AMERICAS)INC.: 1390
TATA CONSULTANCY SERVICES LIMITED: 1278
ACCENTURE LLP: 1166
WIPRO LIMITED: 1051

Figure 6: shows interaction about the mouse click events and buttons

the ones with highest number of petitions each year. For further interaction, we have added some effects to this graph which are represented in the next figure.

9.4 Stacked Bar graph with Donut

Considering the screen space, we have done this visualization. In figure.8. Whenever a user clicks the mouse on a bar the Donut pops up and displays the share of each job position towards the total number of petitions, Full time or part time for that year. For every bar here, it displays a different donut and we can see the actual number of applicants form this donut when we hover over the donut. Everything in this graph can be used to interact with the user to get some good insights, we have added some mouse click functions to the donut as well, when a user click on the donut it displays the count of those job roles in that year and it highlights the positions of that role.

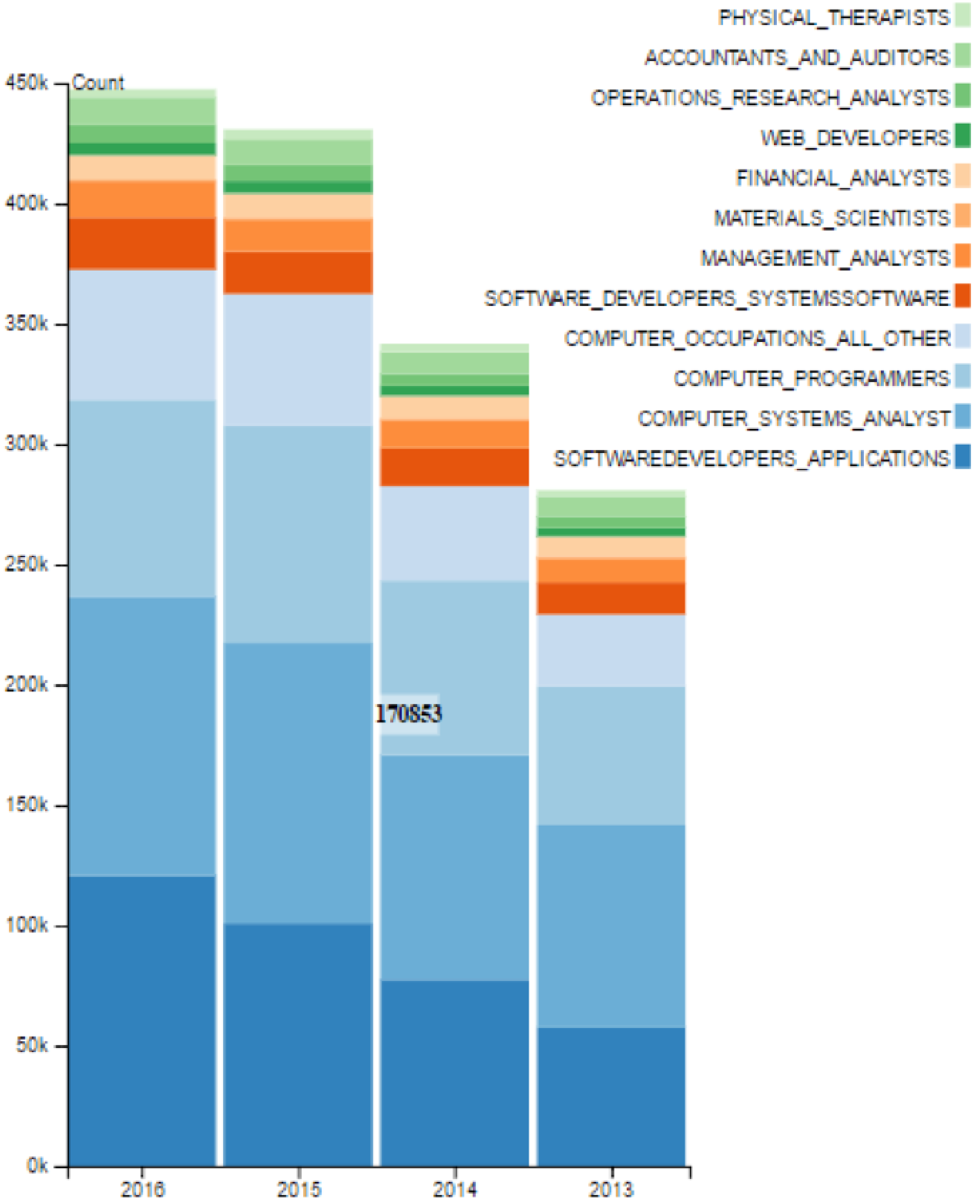


Figure 7: shows the stacked bar graph

9.5 Stacked Bar graph with color highlighting

To see overall year wise comparison in the same graph for a particular position this visualization is done. Figure 9 shows the mouse click options which we have created for the legend, whenever a user wants to interact with the legend or he wants to know the number of positions with a job role he can easily do that by clicking the legend or the name of the position, this will automatically highlight the positions in the bar graph and shows the number of positions pertaining to that position, from which a user can easily interpret the results and get some insights.

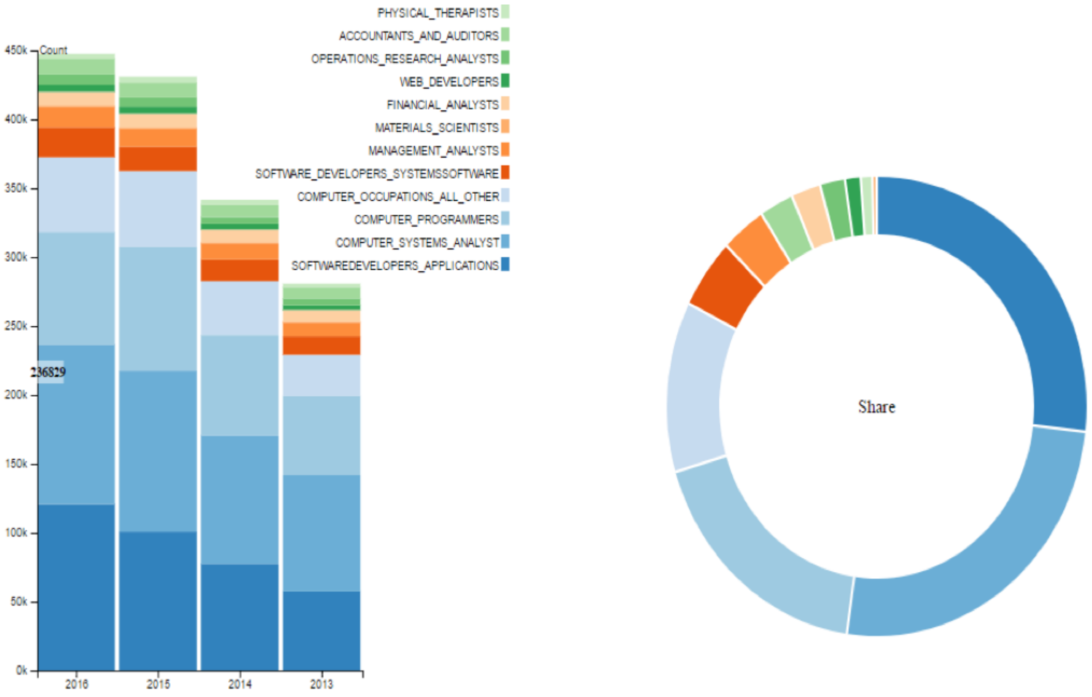


Figure 8: Stack and donut interaction

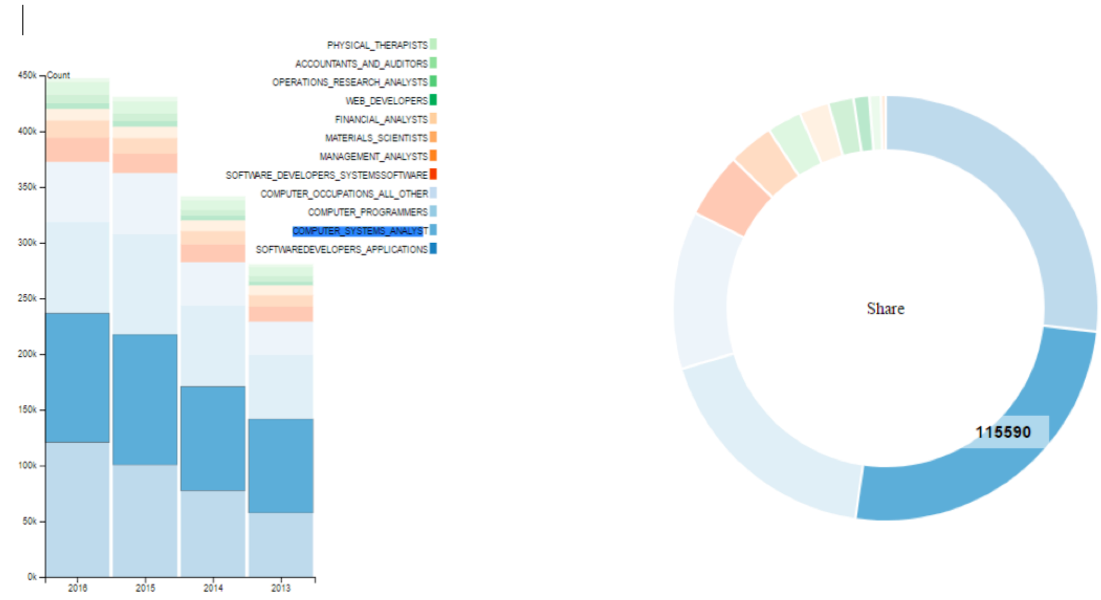


Figure 9: Shows how the legend interactions work

10 Challenges faced

The first and foremost challenge was to figure out the right dataset, after finding the dataset we thought it would be easy to process it and get the results, but the dataset we had contained almost 30 lakh records which seemed to be a nightmare to be processed, we then had to filter the null values, it became very challenging when it came to Geo visualization as we did not find the proper Geo JSON file, we had to generate it on our own using the shape files, we were very new to this, so it took a lot of time to get this done.

11 Learning from the project

The concepts we learned from this visualization are: creating beautiful and meaningful geo visualizations, how to use different inbuilt d3 functions such as d3.geo.path( ). Learned how to create stacked bar graphs, usage of different layouts present in d3.js such as d3.layout.pie( ) so on and so forth. All in all, we gained a lot of useful knowledge by executing this project.

12 Future Enhancements

There are some concepts which we couldnt implement in this project, we will be working on them in the future to achieve all the things which we thought could be good, we will be creating a full stack web interface for our project where the data will be updated automatically, and we will add many more user interactions to our project, where the user himself can select the particular region and add filters according to his requirements.

13 Conclusion

By doing this project we got some great insights about the H-1B process, not only about the facts and figures but also about how to apply and when to apply, we were able to find the details about the top 5 employers in the past five years and bottom 5 employers in the past five years, we also found very interesting insight, which is people with salary range 65k - 75k were high in number almost 60% of the total applicants were people with salary in this range. We found that new jersey and California were the places with high density of software jobs and Data analyst jobs.

14 References

1. http://bl.ocks.org/NPashaP/a74faf20b492ad377312

2. https://bl.ocks.org/mbostock/3886208

3. https://www.kaggle.com/nsharan/h-1b-visa