**Analysis of H-1B visa data**

Department of Applied Data Science

San Jose State University

DATA 230: Data Visualization

Professor: Andrew H. Bond

May 17, 2022

By

Deep Arvind Bambharoliya

015935777

**Table of Content**

[**1.Introduction**](#_1u3idhpxur7) **3**

[1.1 Purpose of this Document](#_qe2ww22xtkiq) 3

[1.2 Intended Audience](#_2f4sv3cit27g) 3

[1.3 Scope](#_rrv63diwu10l) 3

[1.4 About The dataset](#_pjk7aaco89zk) 4

[1.4.1 Definitions](#_7aaac1x7y35c) 4

[**2. Background and Objectives**](#_pnc807uhz1bn) **5**

[**3. Architecture and High level design**](#_shpnwluyb1cm) **6**

[**4.Organization**](#_34qi4g3c8f1q) **7**

[4.1 Project group](#_2z3ihgu4urli) 7

[**5. Development Process**](#_ncuh55kbwuo7) **7**

[**6. Deliverables**](#_myj05xhmacmo) **10**

[**7. Github**](#_81e7wkfen75b) **11**

[**8. Charts**](#_s9vlkkufl1hg) **11**

[**9. Dashboards**](#_nsmsk0y8dodu) **21**

[**10. References**](#_71zkosy9zlgv) **23**

# 

# **1.Introduction**

## **1.1 Purpose of this Document**

The purpose of this document is to provide a detailed project description of the analysis performed on H-1B data, which is designed to understand H-1B visas for students and foreign workers and provide statistics and trends about it through visualizations. This document

includes details about organization, roles, deliverables, project risks, time plans and financial plans.

## **1.2 Intended Audience**

This document shall be used in all phases of the project as a guideline. Intended audiences of this project are all project stakeholders:

* Professor Andrew Bond
* TA Ritanjali Jena
* Deep Arvind Bambharoliya

## **1.3 Scope**

This document shows the project plan for the analysis of H-1B visa data using Tableau. The document includes the purpose of the project, organization plan for the project, data collection, and applications used. The data for the project includes information about the H-1B visa which is a visa issued by the US government, so the scope of data is limited to the United States and data is collected only for the past 4 years .i.e. 2017 - 2021.

## **1.4 About The dataset**

### **1.4.1 Definitions**

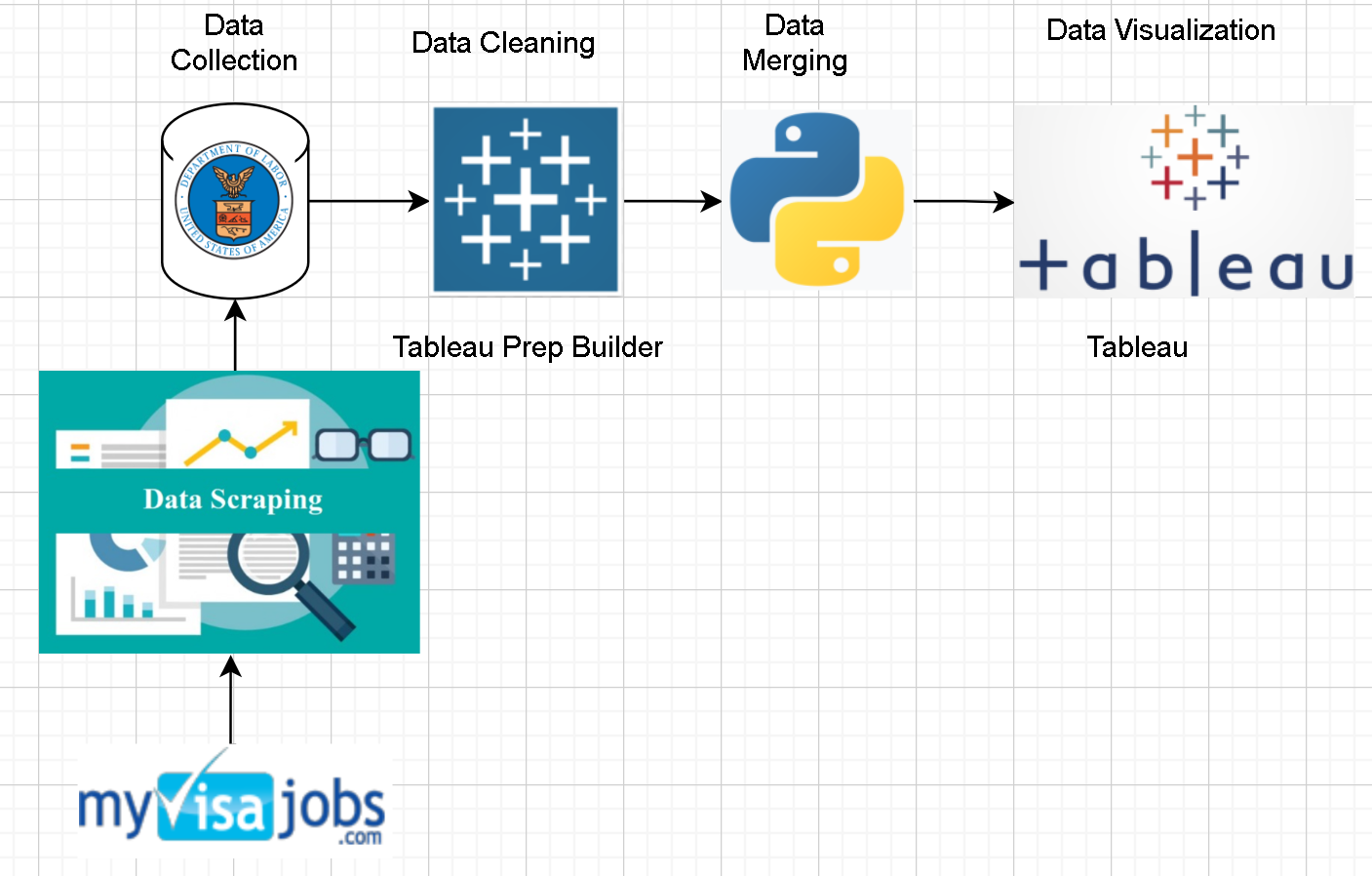
The dataset contains following fields

* **Case Number** - Unique identifier assigned to each application submitted
* **Case Status** - Status associated with the last significant event or decision. Valid values include “Certified”, “Certified-Withdrawn”, Denied”, and “Withdrawn”.
* **Case Submitted** - Date the application was submitted
* **Visa Class** - Indicates the type of temporary application submitted for processing. Values include H-1B, E-3 Australian, H-1B1 Chile, and H1B1 Singapore.
* **Employer Name** - Legal business name of the employer submitting the Application
* **Employer Address -** Employer street address
* **Employer City** - Employer city name
* **Employer State** - Employer state
* **Employer Postal Code** - Employer postal code
* **Job Title** - Title of the job
* **SOC Code** - Occupational code associated with the job as classified by the Standard Occupational Classification (SOC) System.
* **SOC Title** - Occupational title associated with the SOC code
* **NAICS Code** - Industry code associated with the employer, as classified by the North American Industrial Classification System (NAICS).
* **Full Time Position** - Y = Full Time Position; N = Part Time Position.
* **Prevailing Wage** - Prevailing Wage for the job being requested
* **PW Unit of Pay** - Unit of Prevailing Wage Pay. Valid values include “Hour”, “Bi-weekly”, “Week”, “Month”, and “Year”.
* **PW Wage Level** - OES Wage Level. Variables include "I", "II", "III", "IV" or "N/A
* **Wage Rate of Pay From** -Wage paid to nonimmigrant worker
* **Wage Unit of Pay** - Unit of Wage Pay. Valid values include “Hour”, “Bi-weekly”, “Week”, “Month”, and “Year”.
* **H1B Dependent** - Y = Employer is H-1B Dependent; N = Employer is not H-1B Dependent.

# **2. Background and Objectives**

H-1B visas are a category of employment-based, non-immigrant visas for temporary foreign workers in the United States. For a foreign national to apply for an H-1B visa, a US employer must offer them a job and submit a petition for a H-1B visa to the US immigration department. This is also the most common visa status applied for and held by international students once they complete college or higher education and begin working in a full-time position. This can include occupations in fields such as IT, finance, engineering, architecture or more. According to USCIS, foreign workers with H-1B visas can stay in the US for up to six years, as an H-1B visa is initially valid for three years and can be extended for another three years. The Office of Foreign Labor Certification (OFLC) generates program data, including data about H1-B visas. The disclosure data is updated annually and is available online. The dataset contains four years of H-1B petition records from 2018 to 2022 with approximately 2.5 million of records overall.(2,534,222)

# **3. Architecture and High level design**



The data for the project come from two different sources

1. US department of labor - it is an official government website which is updated annually with new data
2. Myvisajobs.com -it is an open source website which contains more detailed data about the H-1B filings which is sourced from the government website itself.

For collecting the data data scraping has been applied on data source 2, to scrape data from the website and store it in csv format.

After the initial data collection, all the data was imported into tableau prep builder and cleaning of the data was done and then it was exported to python in the form of data frame and the data files were merged using python pandas library.

Finally, the cleaned and merged file was imported into Tableau for the data analysis and visualization of the data

# **4.Organization**

## **4.1 Project group**

It is an individual work done by Deep Arvind Bambharoliya (015935777), San Jose State University

# **5. Development Process**

The process for the project starts from collecting the data. For analyzing we needed historical data about H-1B visas. For this purpose, the data was collected from the US department of labor which contained all the data from the government database about the H-1B filings which happened over the past number of years.

Once the data was collected the individual csv files were imported to tableau prep builder. Tableau prep builder is a Tableau product which helps to clean and process the data by using gui and intuitively.

For each of the files, data types for the columns were set according to the data it contained such as changing data type for city column from string to geographical location and many more.

Many rows were removed because it contained multiple null values which made the whole row meaningless. Also few rows were deleted on the basis of the salary range which was set manually to $1,000,000. So if the salary for an instance was greater than the permissible range then the record was considered as outlier and was removed from the dataset.

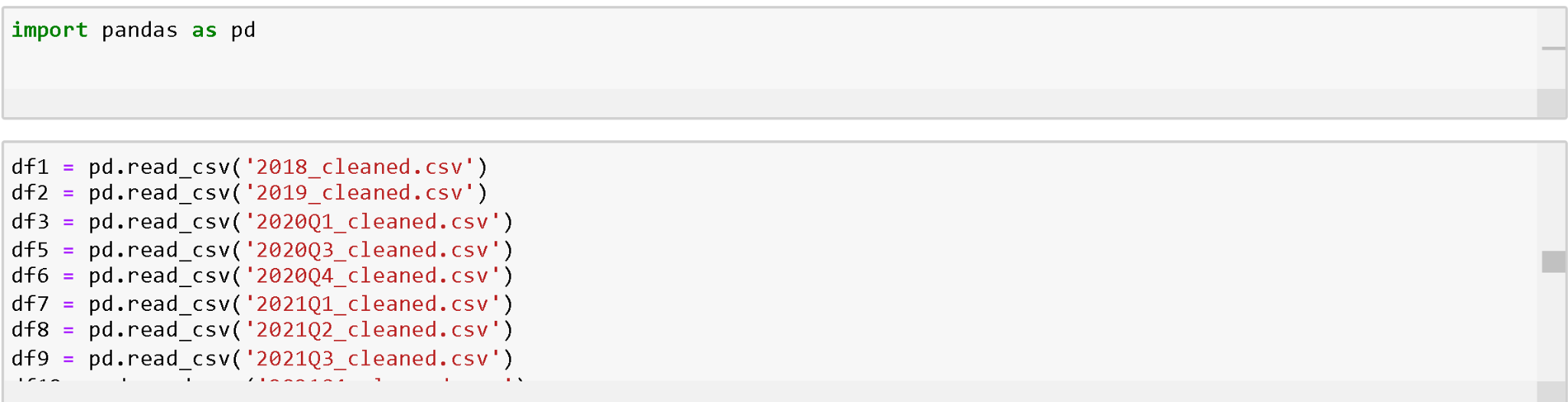


Also a couple of calculated fields were generated to handle the ambiguity in the salary unit.Many cases reported the salaries in hours, weekly, bi-weekly, monthly or yearly. So to make this uniform, calculated fields were created which would convert all the units into a common (yearly) salary unit which would further help us to analyze the data in a better way. Below is an image showing the generated calculated field.

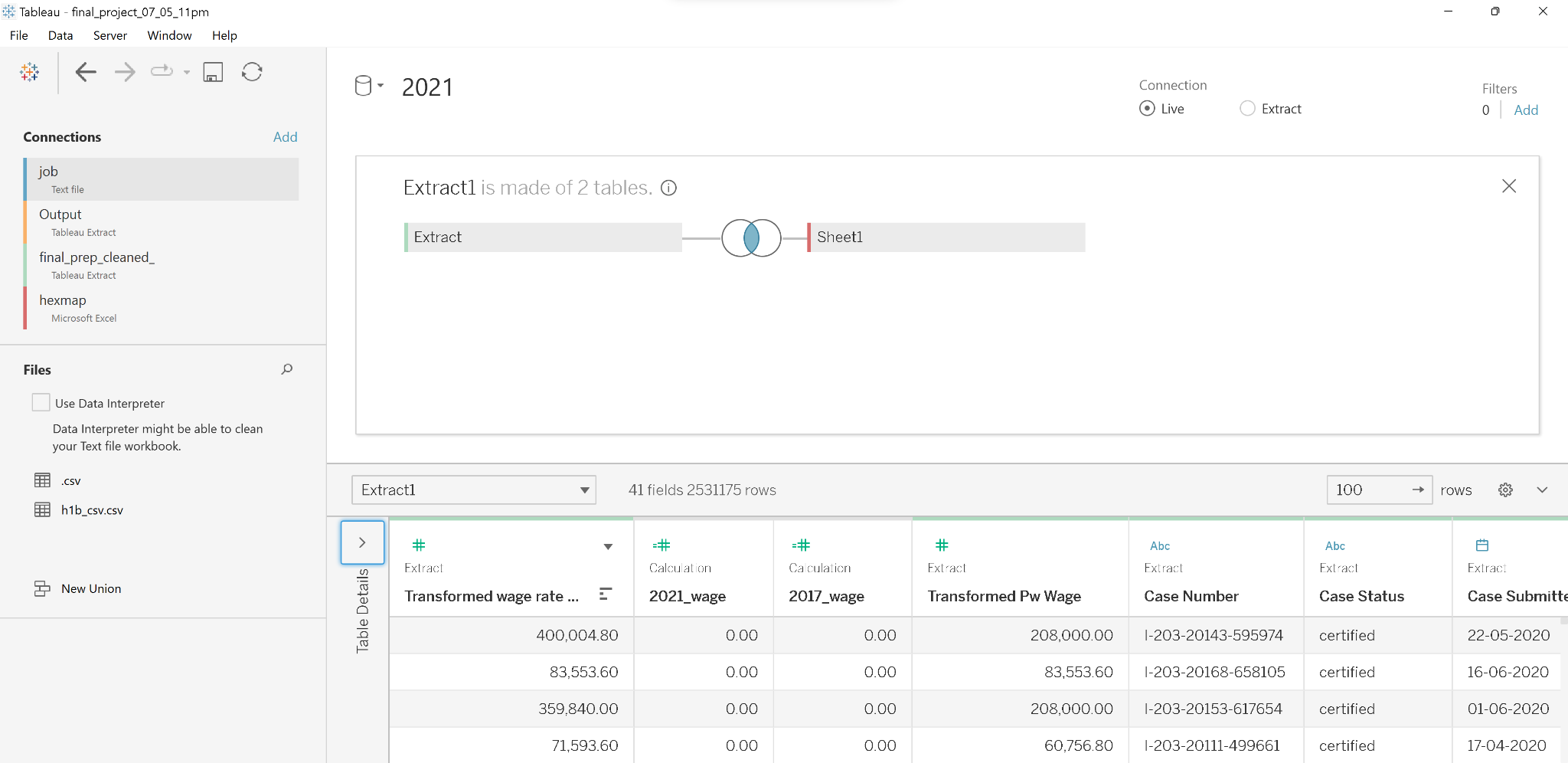


Once the data was cleaned and ready to be merged, all the data files were imported into pandas dataframe using python. The reason for selecting python was because the tableau prep builder was taking a lot of resources and time, which was comparatively faster using python.

The below figure shows the importing of individual csv files into a dataframe. Once the data frames were ready some cleaning was performed like removing duplicate rows, renaming columns to similar column names across all the data frames which is necessary while merging the data



Once all the data files were merged it was time to perform analysis and create visualizations in Tableau. The figure belows shows the sample of data imported in tableau ready for visualization



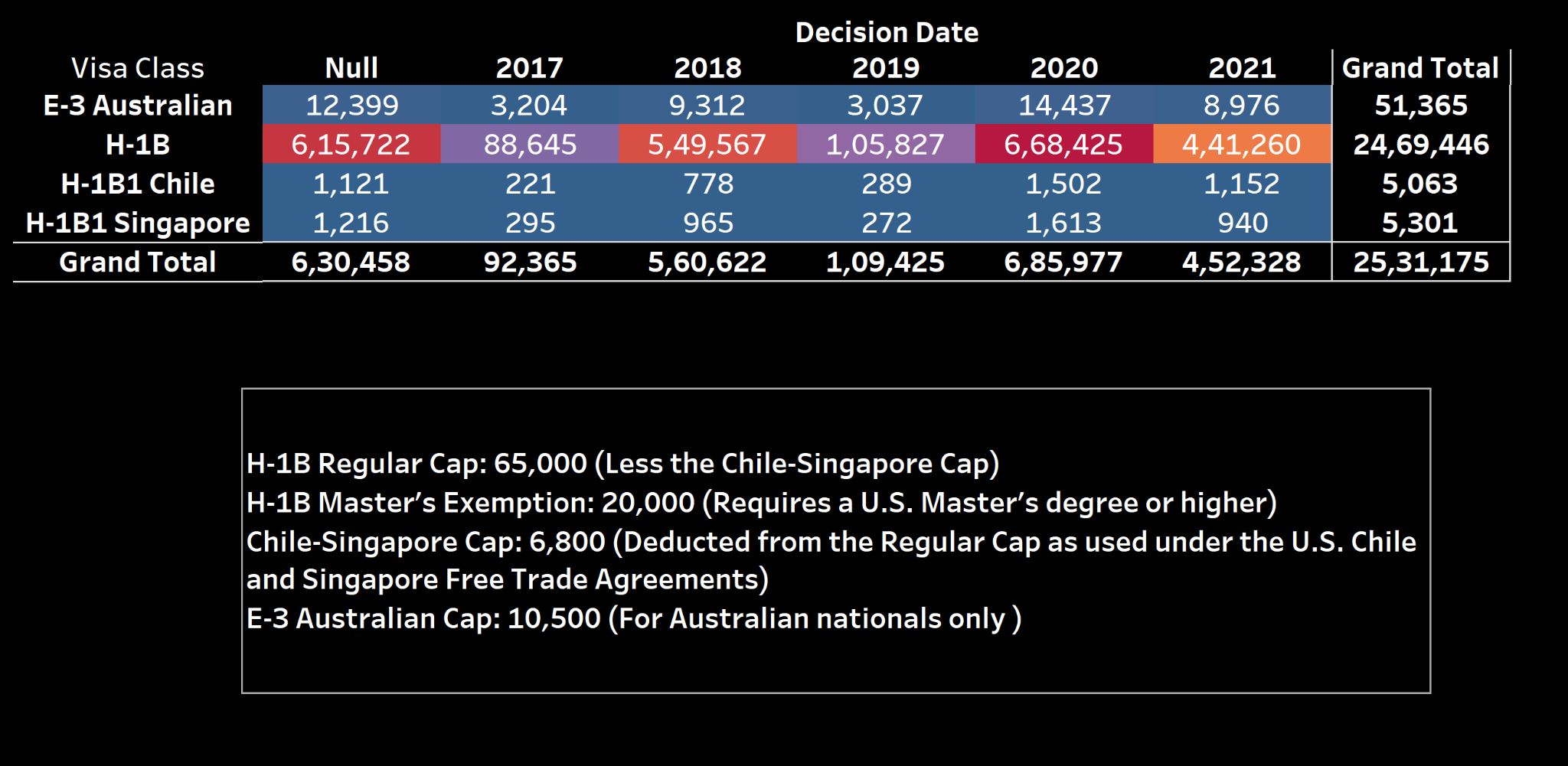
# **6. Deliverables**

The most important deliverable of this project is to provide 2 important dashboards for overall distributed analysis. A project presentation is also given. And the Final Project report will be submitted. All the data processing and visualization codes are to be delivered.

# **7. Github**

<https://github.com/DeepBambharoliya>

# **8. Charts**

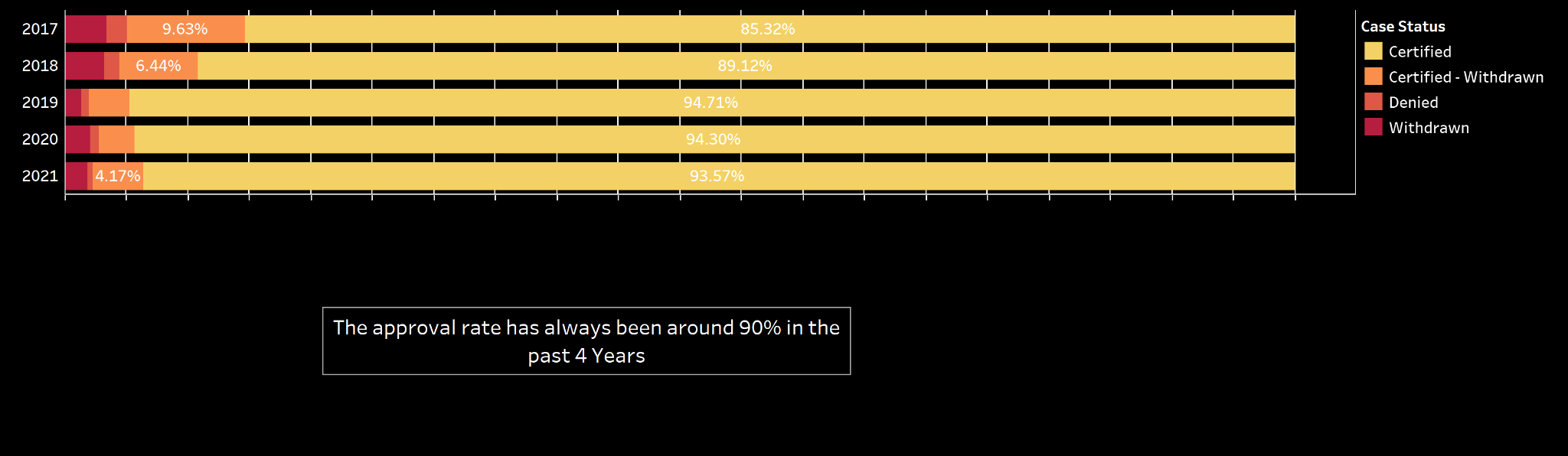


The are 4 sub categories in H1b visa

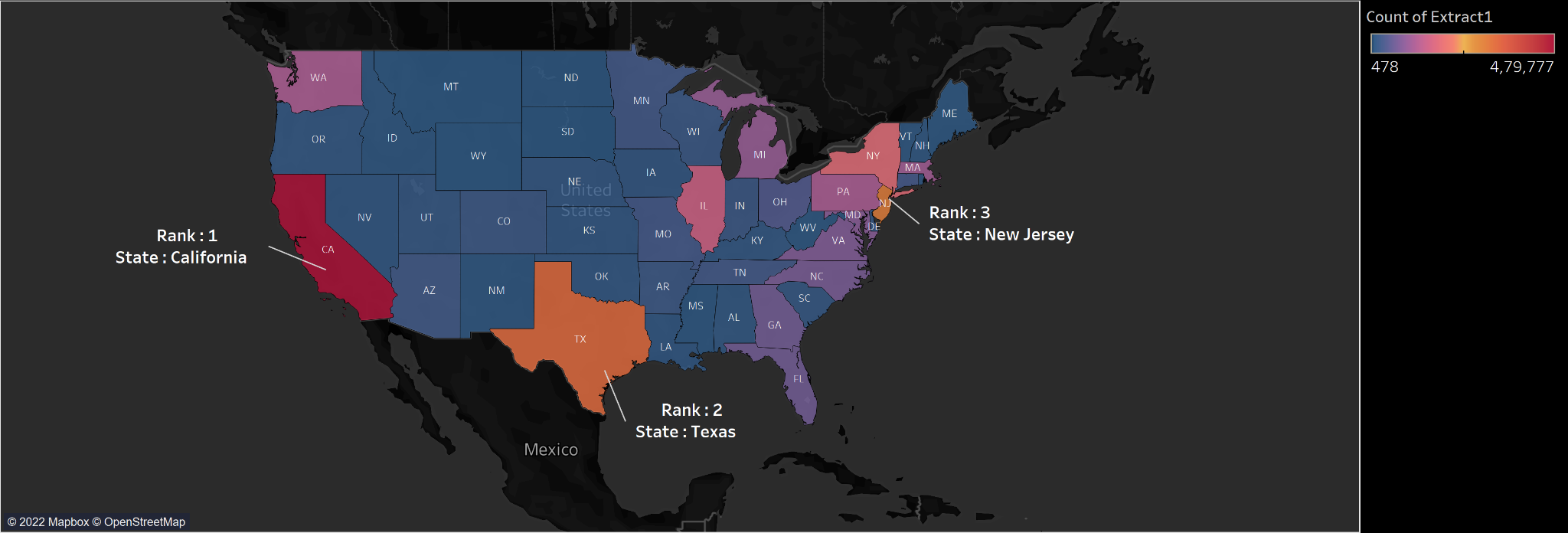
E3 australian which is for australian nationals, H1b1 chile and singapore which are for their nationals and h1b for all other foreign nationals. Yearly only 65000 h1b visas are granted and more 20000 are granted for the individuals who hold a US master's degree or higher.

Majority of applications are made for the H-1B category. A total of 24,69,446 applications of H-1B visas were filed during the last 4 years.

We can see that there are fewer applications during 2017 and 2019 when compared to other years.

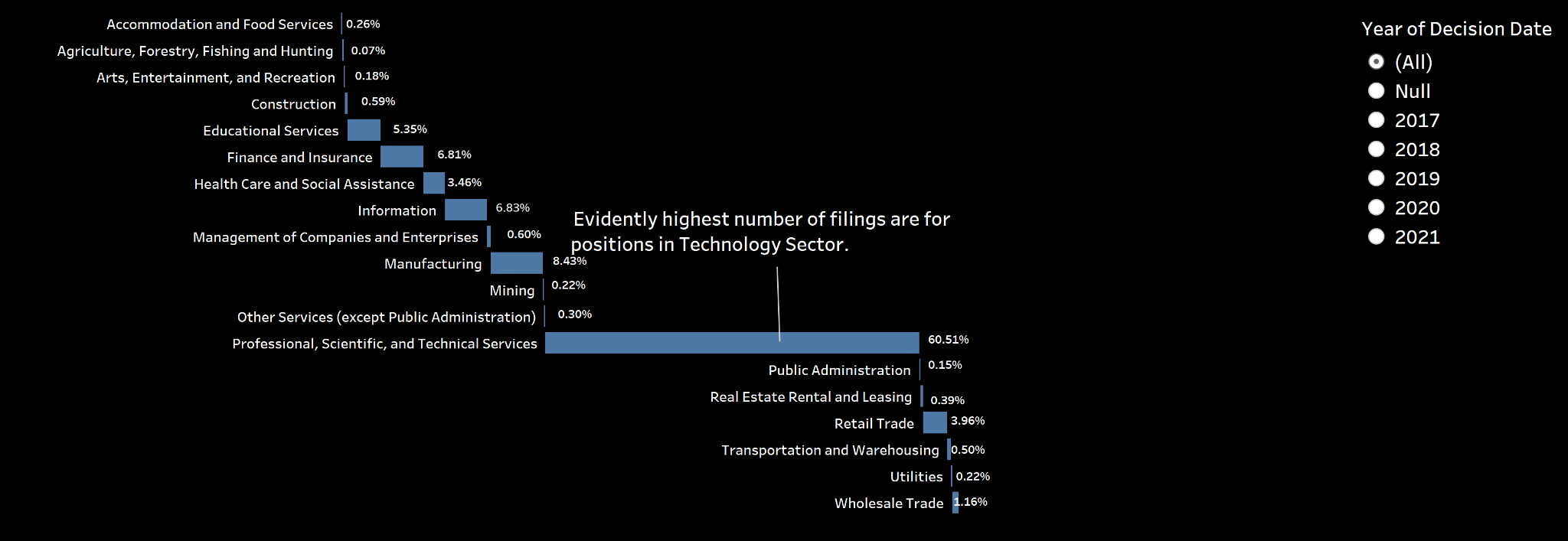


For each case filed there can be four decision certified , certified- withdrawn, denied or withdrawn. Based on the case information and background of the employee and the company filing the application, the government department makes a decision on the case. In the past four year, it can be seen from this bar graph that the approval rate has always been more than 90%. It means out of 100 cases received 90 cases are certified and granted the H-1B visa.



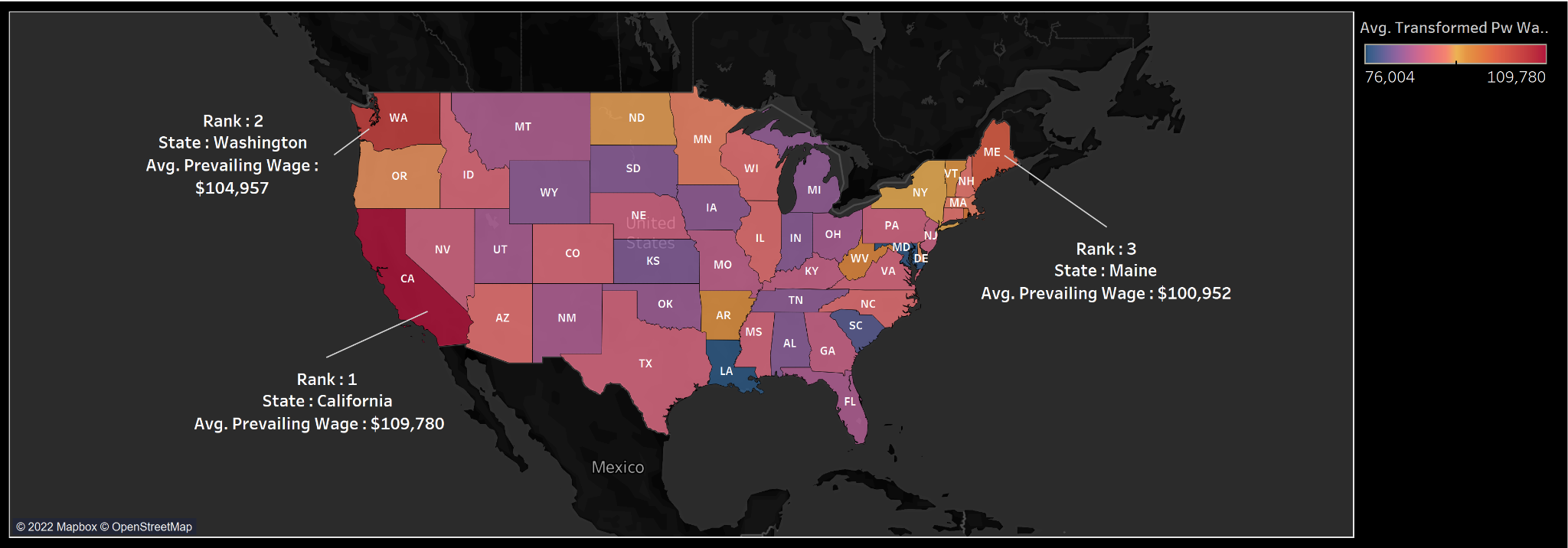
The geographic map shows the distribution of the application US - wide.

It shows the total number of applications made by each state in the whole US. California was the state filing the most number of applications followed by Texas and New Jersey



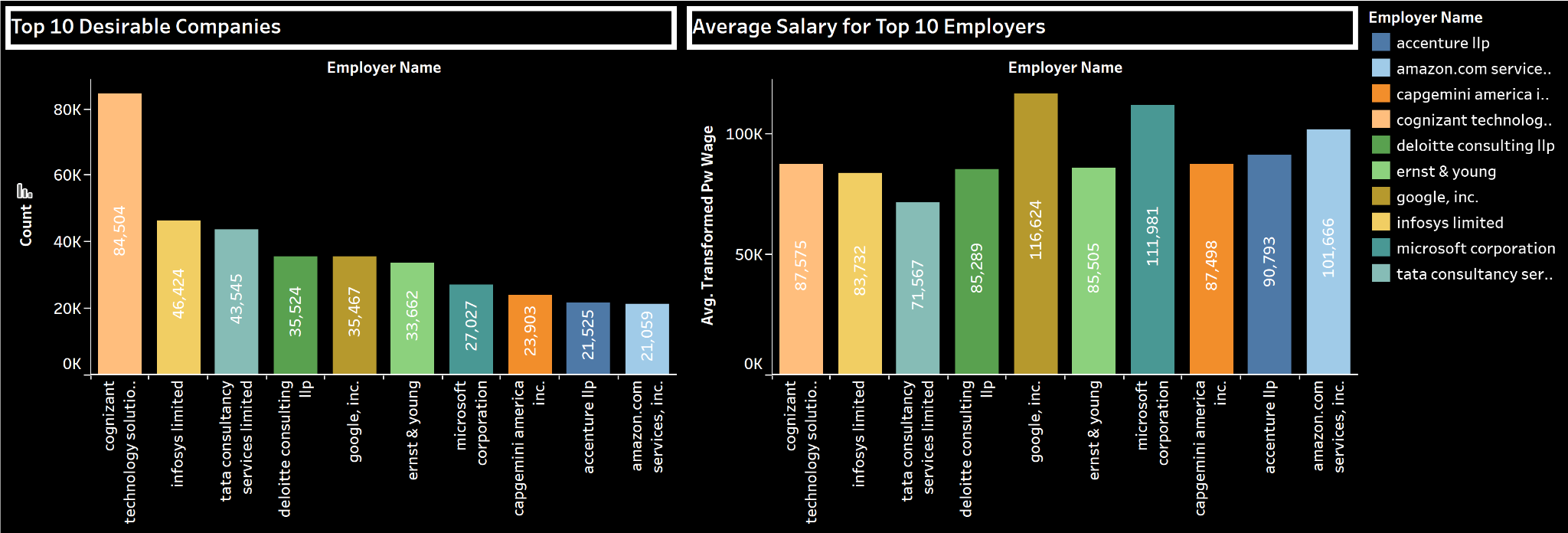
The gantt chart shows the percentage of total cases in a year filed in a particular industry. The government based on the work done by an organization has grouped into an industry which is represented by NAICS code.

From the graph it can be inferred that the most number of foreign workers are hired in the technology sector, that is companies belonging to the Professional, scientific and technical services industry.



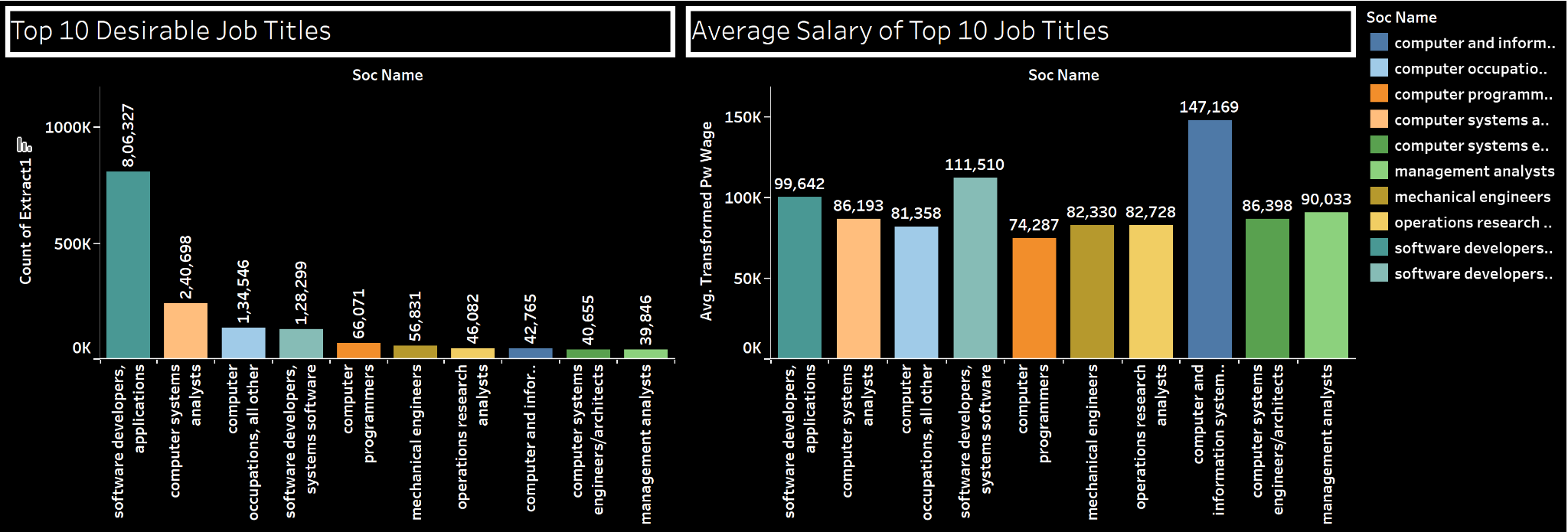
The next analysis was done on the wages provided to the foreign worker in each state. The map shows the distribution of the average prevailing wage paid to the worker by its employer.

The highest average prevailing wage was in California followed by Washington and Maine.



As noted above the majority of the applications were made in the technology sector, next analysis was performed on the employers. This bar graph shows the top 10 employers filing the most number of applications and hiring foreign workers in their companies. The top company was Cognizant technology solution, which is an IT company.

Another analysis was done on the salary for the workers paid by these companies. The leader in paying the highest average prevailing wage was Google followed by Microsoft.

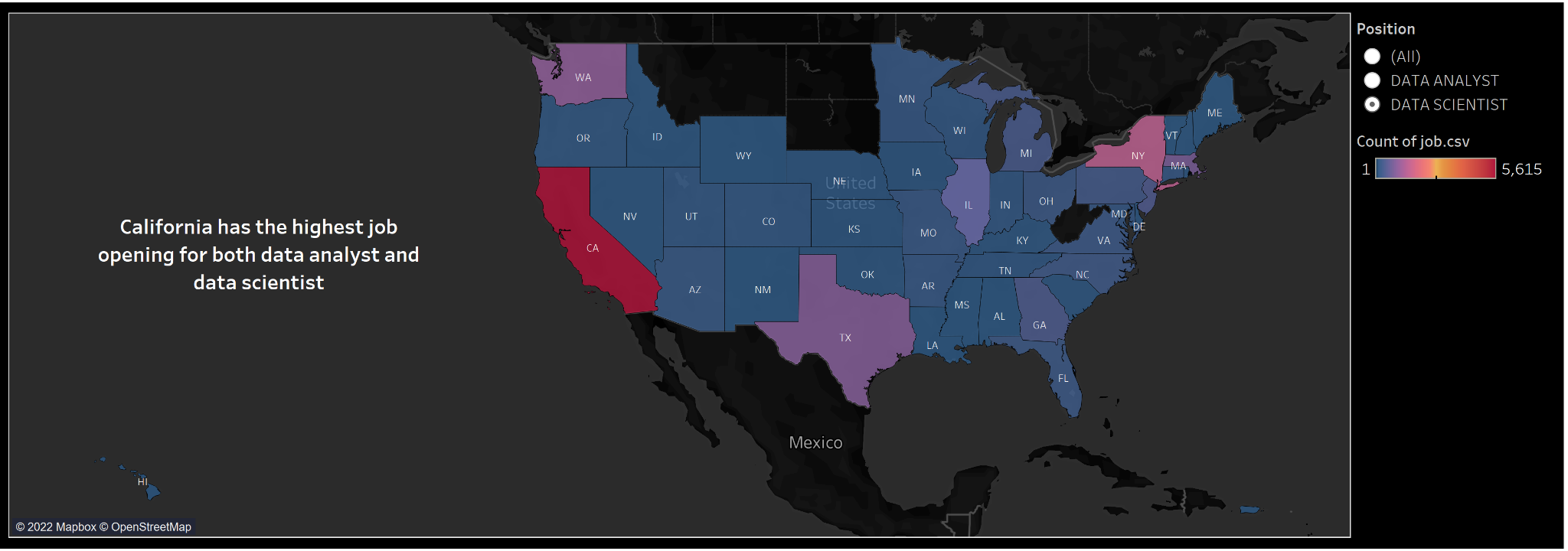


Next the analysis focused on the top 10 job titles. For each applicant the job title is also given in the dataset. So based on this a bar graph was plotted showing the top 10 job titles based on the count. The highest number of cases were having job titles of Software developers.

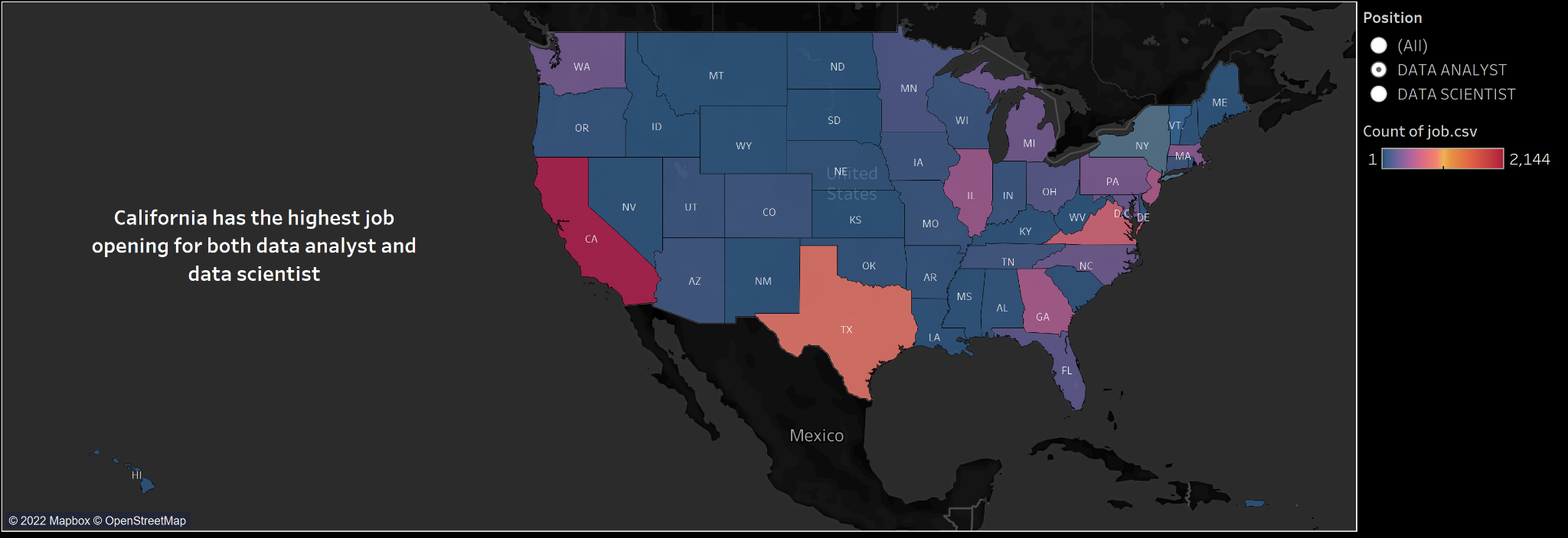
Another similar comparison was made on the salaries paid to these top 10 job titles. From the graph it was noticed that workers working at the position of computer and information systems managers were paid the highest average prevailing wages when compared to all other job titles.

**Data Analyst vs Data Scientist**

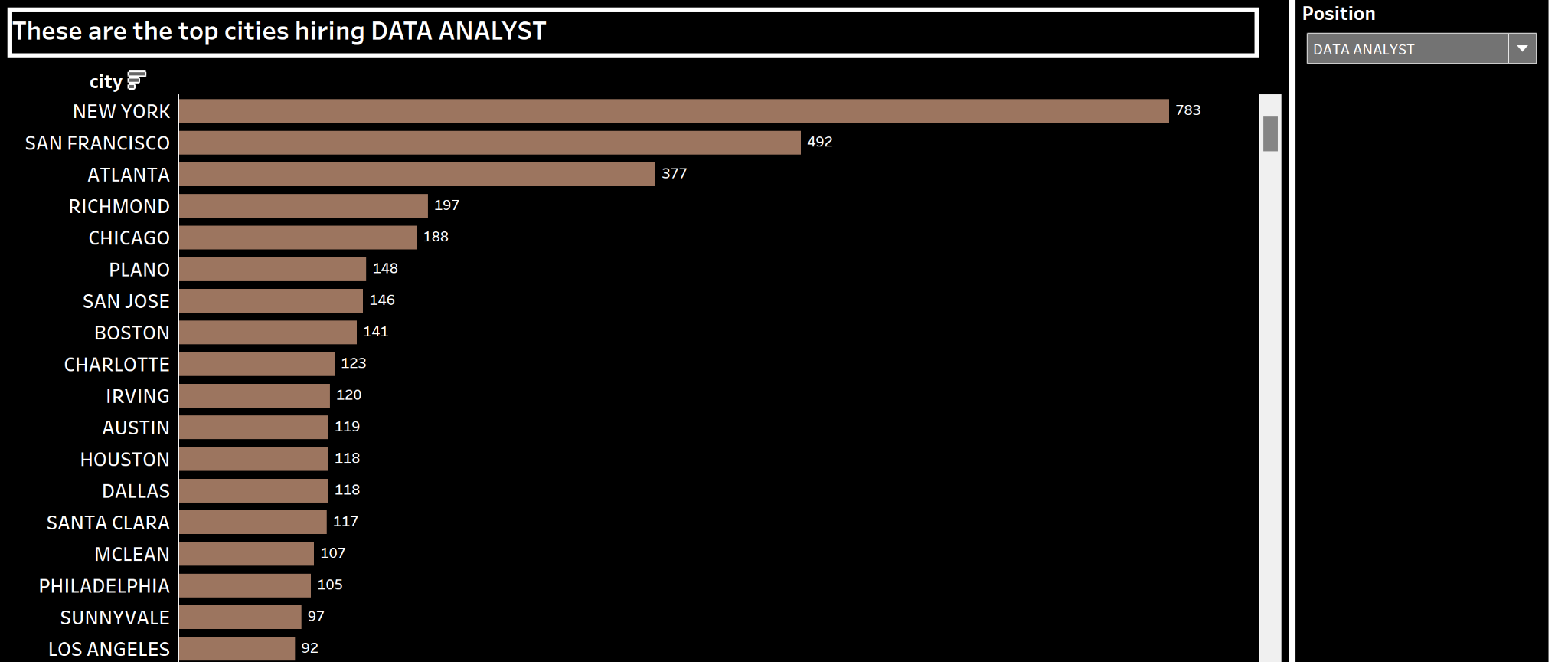
Next phase of the analysis focused on two particular job titles; Data Analyst and Data Scientist



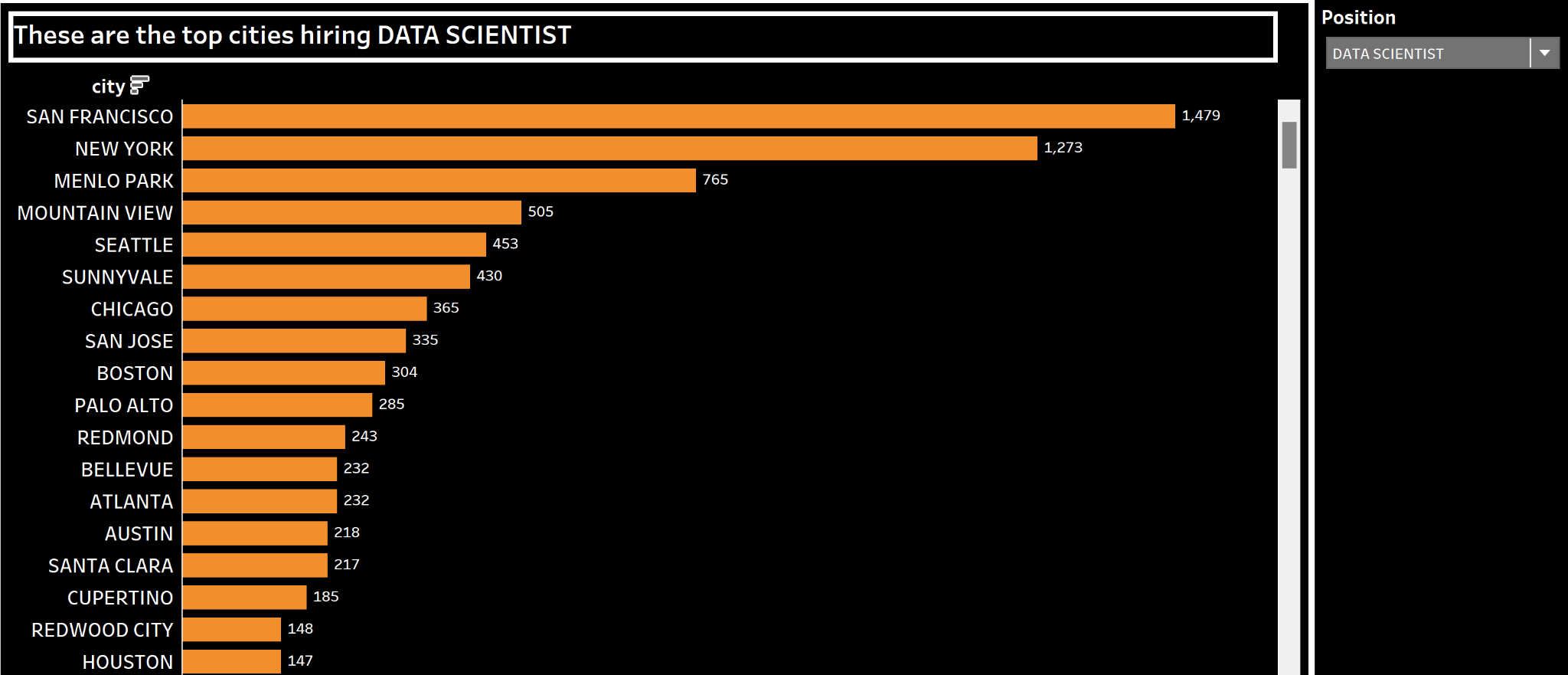
This graph shows the total number of positions available for data scientists in each state of the US. It is observed that the state of California had the highest number of job openings for the position of data scientist and similarly the below graph shows the job openings for the position of data analyst in the US states. Here the same trend was followed and California was leading with the highest number of openings.

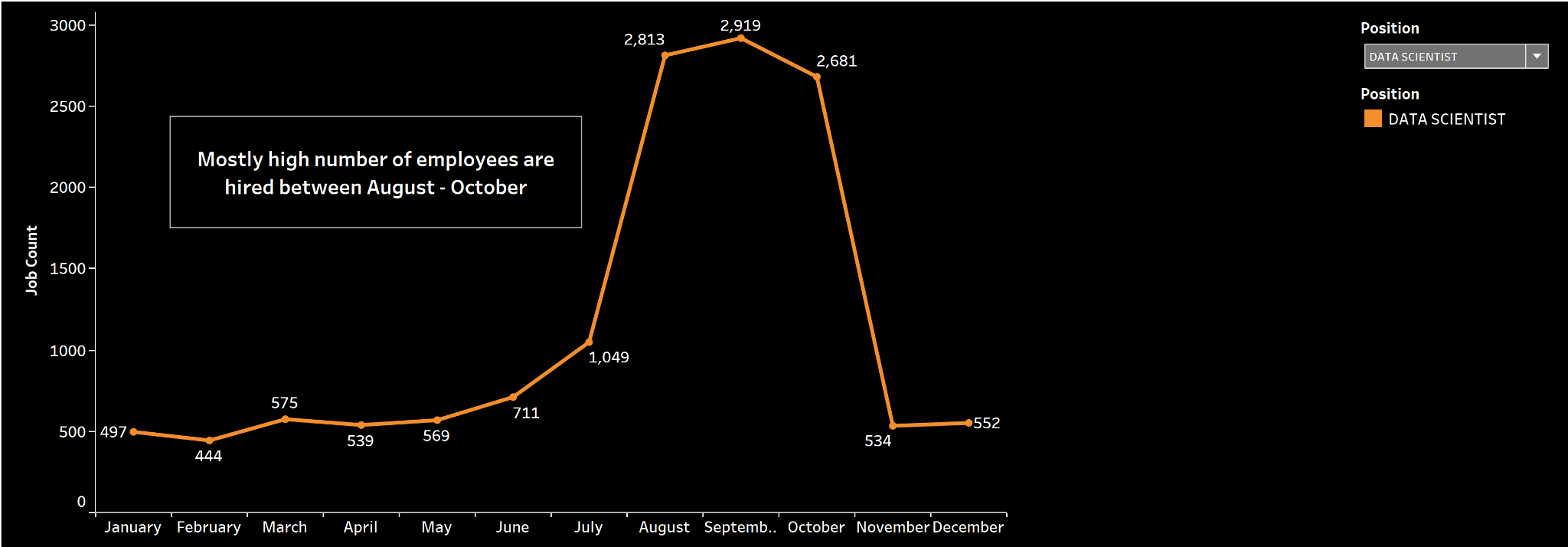


When both the maps are compared, overall the number of opportunities for data scientist are way more than positions for data analyst.

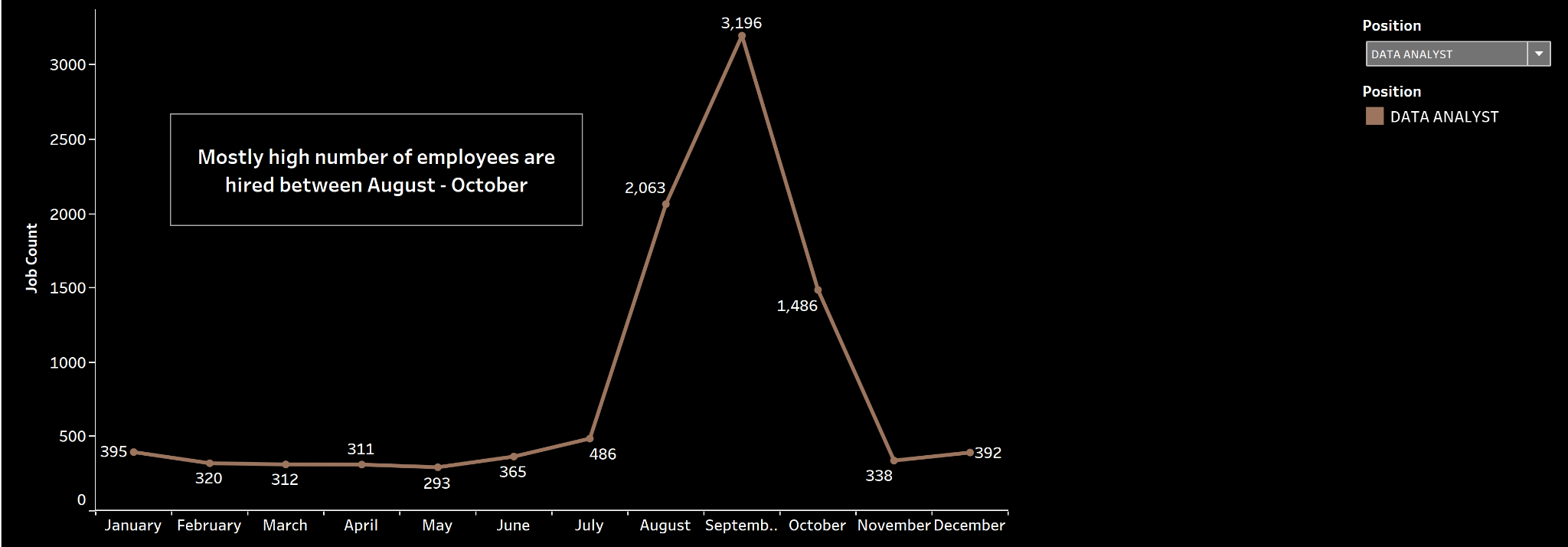


This horizontal bar graph shows the number of openings available in each city for data analysts. The highest number of openings were noted in the city of New York and followed by San Francisco.

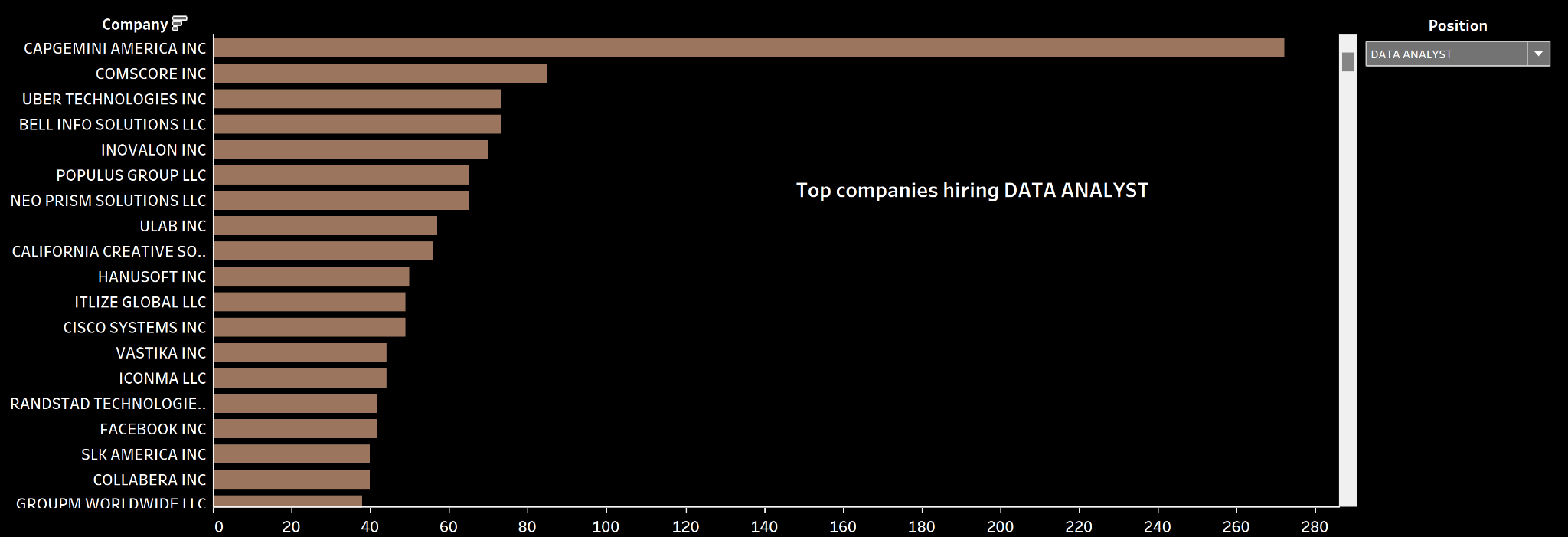


This horizontal bar graph shows the number of openings available in each city for data scientists. The highest number of openings were noted in the city of San Francisco and followed by New York.

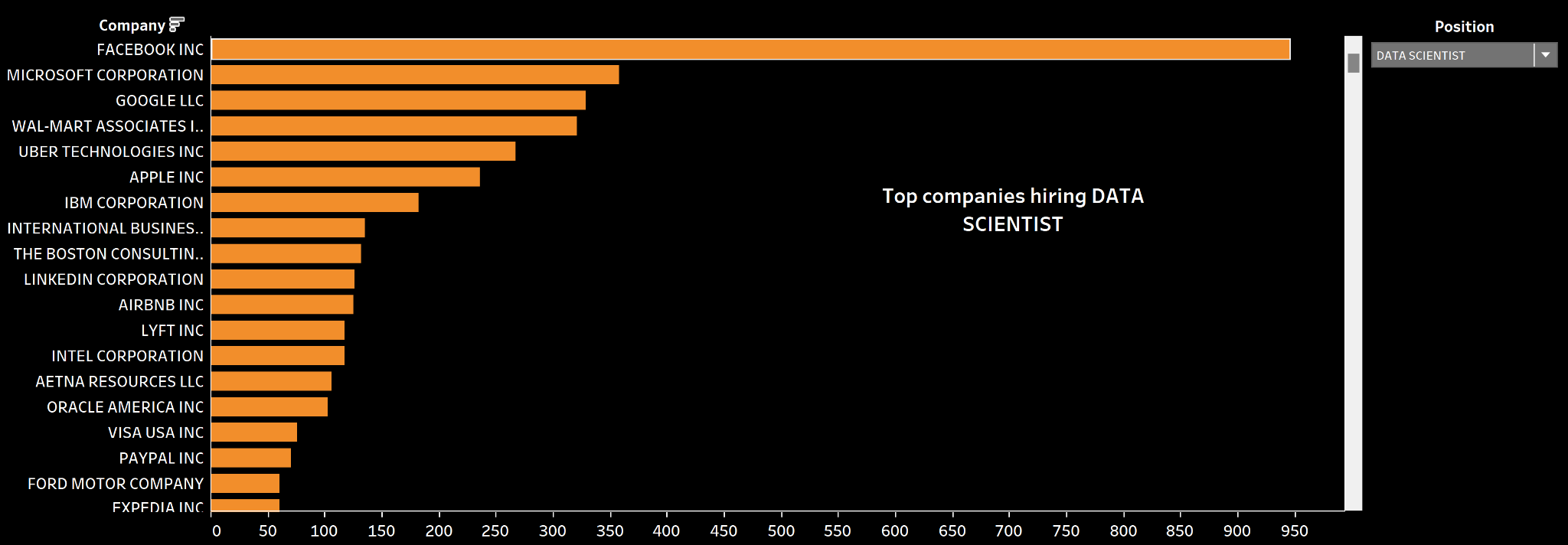
The line graph shows the number of openings available throughout the year for each month for the position of data scientist. There is a significant rise in opening during the months of July - October when compared to the rest of the year.



The line graph shows the number of openings available throughout the year for each month for the position of data analyst. There is a significant rise in opening during the months of July - October when compared to the rest of the year.



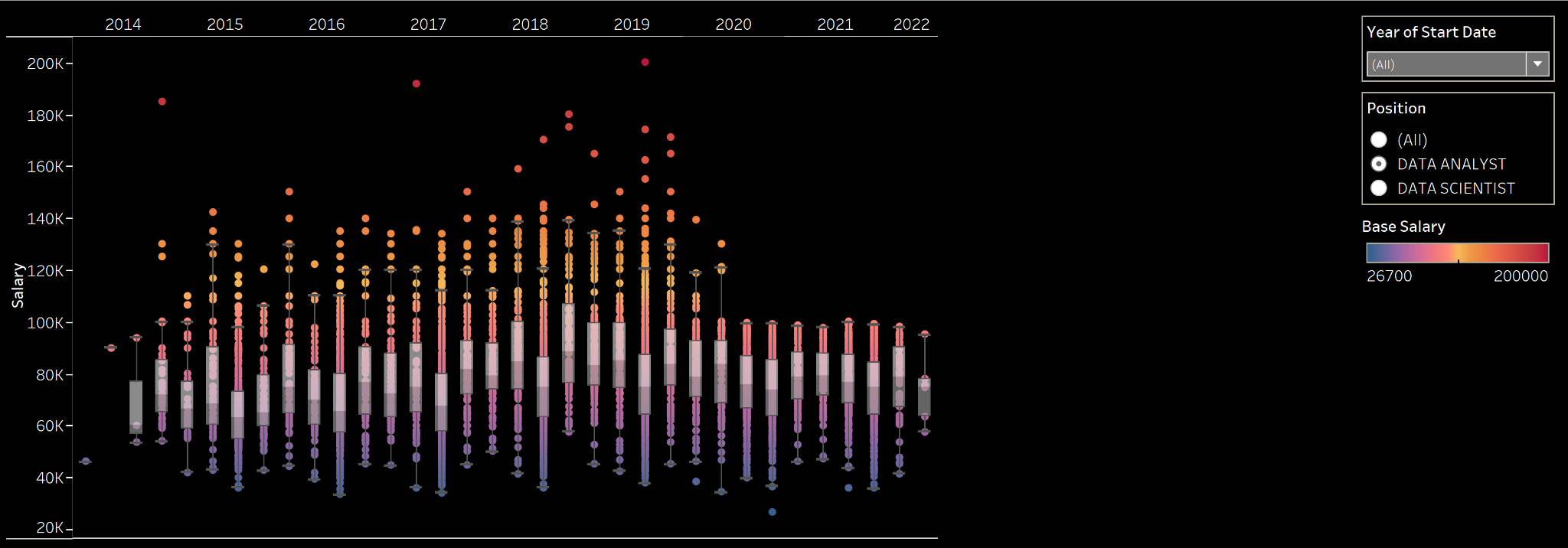
The vertical bar chart shows the top employers for data analysts in the US. Most data analysts are hired by Capgemini America INC followed by Comscore INC.



For the position of data scientists the most number of workers are hired by companies like Facebook and Microsoft and these are all big companies which shows that workers for the position of data scientist are hired in big tech companies.



The line graph shows the salary comparison for data scientists and data analysts over the years. Overall data scientists are paid more when compared to data analysts. The orange line shows the trend for data scientists and the brown line represents the salary trend for data analysts. There is a huge gap in the average salary between both the positions.



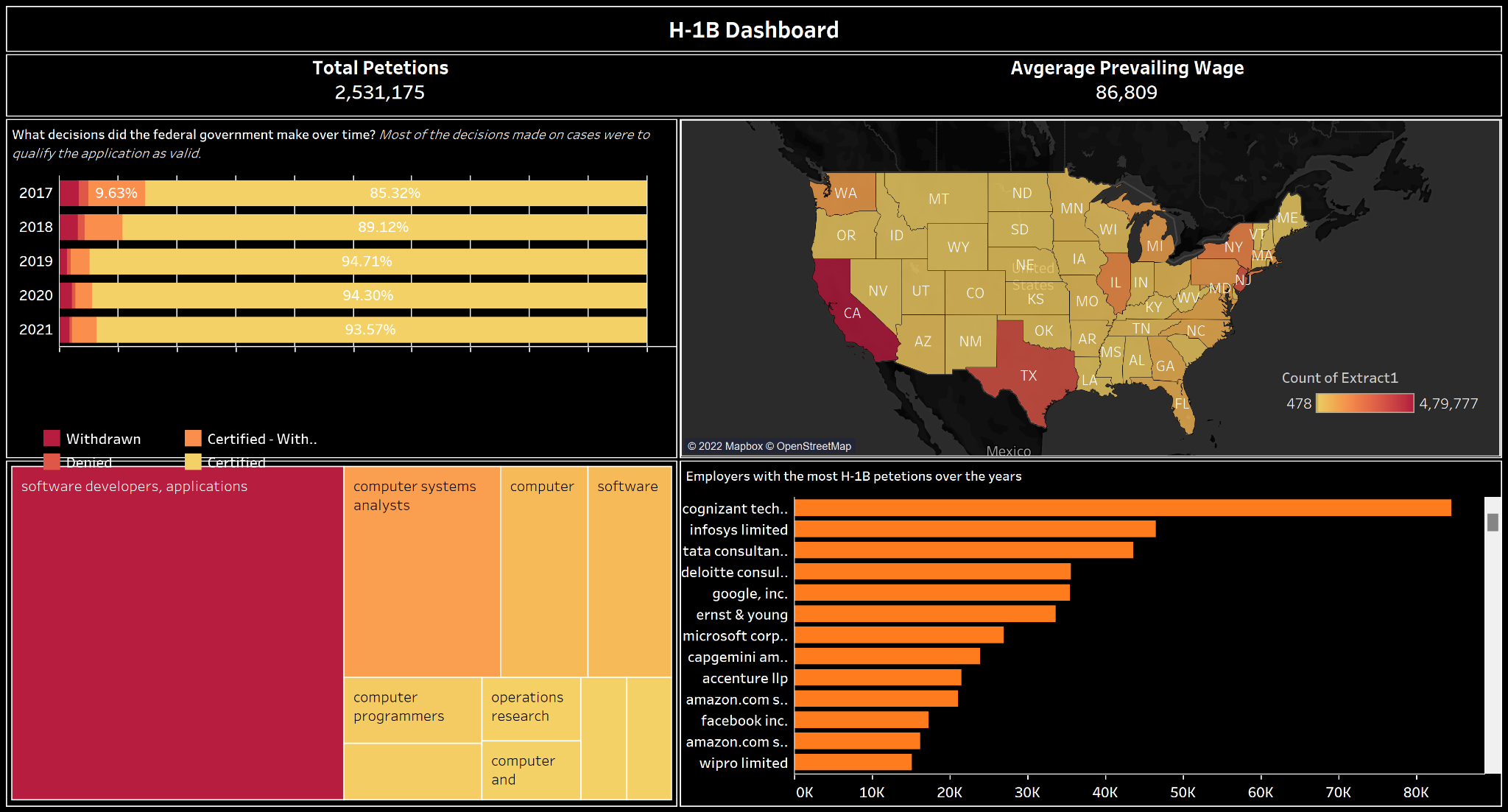
These box plots were plotted for the in depth analysis of salary for the position of data analyst. Each box represents the salary in a particular quarter of the year starting from 2014 and ending in 2022. We can see that some data analysts are paid as high as $200,000 as annual salaries but when compared to the highest salary for data scientists it is still less.



This box plots the salary analysis for the position of data scientists. Each box represents a quarter in each year. We can see that the overall average salary is high when compared to average salary for data analysts and also the highest salary noted was $300,000 which was also higher than the highest salary noted for data analysts.

# **9. Dashboards**

This dashboard was created using the charts created for the comparison of data scientists and data analysts. It shows the overall information for data scientists and data analysts such as the salary trend over the years, number of openings throughout the year, trending employers and trending states and cities.Also a dropdown is provided for the dashboard which allows to select the position between data scientists and data analysts.



This dashboard shows the overall information about the H-1B visas. It contains an information panel on top which shows the total number of applications and the average prevailing wage which can be changed according to the state selected in the map of the US given below the information bar. The dashboard also shows the case status over the years and the top companies and top job titles in the form of horizontal bar chart and tree chart respectively.

# **10. References**

<https://www.dol.gov/agencies/eta/foreign-labor/performance>

<https://www.glassdoor.com/Salaries/us-data-analyst-salary-SRCH_IL.0,2_IN1_KO3,15.htm>

<https://www.glassdoor.com/Salaries/us-data-scientist-salary-SRCH_IL.0,2_IN1_KO3,17.htm>

<https://h1bdata.info/index.php?em=&job=data+analyst&city=&year=All+Years>

<https://h1bdata.info/index.php?em=&job=Data+Scientist&city=&year=All+Years>