H-1B Visa Petitions 2011-2016
CS581 : (Final Project Report)

Tathagata Dutta (NetID: td425) **Sai Mounica Pothuru** (NetID: sp1912)
Nikhil Neroor (NetID: nn323)
Chaitanya Sharma Domudala (NetID: cd817)

December, 2020

Introduction

Every human being is constantly taught to pursue work that they are passionate about. For few, it includes working in diverse fields while some like to work in different countries. Most of the countries provides opportunities for such individuals to work by offering work visas. Temporary working visas are provided for immigrants who would like to work in a foreign country over a fixed period. There are multiple types of the visas available and vary according to each country's immigration rules.

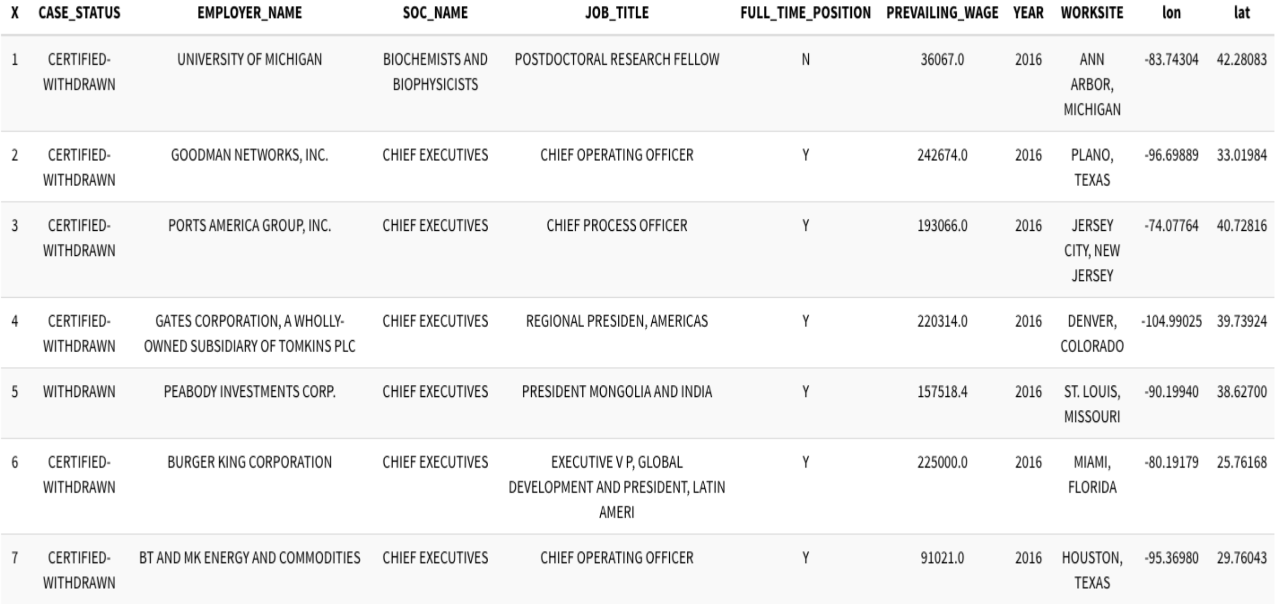
H-1B visas are a classification of business based, non-settler visas for transitory unfamiliar specialists in the United States. For an outside public to apply for H1-B visa,a US employer must extend to them an employment opportunity and present a request for a H-1B visa to the US immigration division. This is additionally the most well-known visa status applied for and held by international students once they complete school or advanced education and start working in a full-time position.

Dataset Description

The H1B data set was collected from United States Department of Labor - Office of Foreign Labor Certification (OFLC). The Office of Foreign Labor Certification (OFLC) generates disclosure data that is useful information about the immigration programs including the H1-B visa.

The dataset being used for this project is H1B Visas from Kaggle and is a time-series data. It contains 5 years of H-1B request information from 2011 to 2016, with around 3 million records.

Link to the dataset: https://www.kaggle.com/nsharan/h-1b-visa



X	CASE_STATUS	EMPLOYER_NAME	SOC_NAME	JOB_TITLE	FULL_TIME_POSITION	PREVAILING_WAGE	YEAR	WORKSITE	lon	lat
1	CERTIFIED-WITHDRAWN	UNIVERSITY OF MICHIGAN	BIOCHEMISTS AND BIOPHYSICISTS	POSTDOCTORAL RESEARCH FELLOW	N	36067.0	2016	ANN ARBOR, MICHIGAN	-83.74304	42.28083
2	CERTIFIED-WITHDRAWN	GOODMAN NETWORKS, INC.	CHIEF EXECUTIVES	CHIEF OPERATING OFFICER	Y	242674.0	2016	PLANO, TEXAS	-96.69889	33.01984
3	CERTIFIED-WITHDRAWN	PORTS AMERICA GROUP, INC.	CHIEF EXECUTIVES	CHIEF PROCESS OFFICER	Y	193066.0	2016	JERSEY CITY, NEW JERSEY	-74.07764	40.72816
4	CERTIFIED-WITHDRAWN	GATES CORPORATION, A WHOLLY-OWNED SUBSIDIARY OF TOMKINS PLC	CHIEF EXECUTIVES	REGIONAL PRESIDEN, AMERICAS	Y	220314.0	2016	DENVER, COLORADO	-104.99025	39.73924
5	WITHDRAWN	PEABODY INVESTMENTS CORP.	CHIEF EXECUTIVES	PRESIDENT MONGOLIA AND INDIA	Y	157518.4	2016	ST. LOUIS, MISSOURI	-90.19940	38.62700
6	CERTIFIED-WITHDRAWN	BURGER KING CORPORATION	CHIEF EXECUTIVES	EXECUTIVE V P, GLOBAL DEVELOPMENT AND PRESIDENT, LATIN AMERI	Y	225000.0	2016	MIAMI, FLORIDA	-80.19179	25.76168
7	CERTIFIED-WITHDRAWN	BT AND MK ENERGY AND COMMODITIES	CHIEF EXECUTIVES	CHIEF OPERATING OFFICER	Y	91021.0	2016	HOUSTON, TEXAS	-95.36980	29.76043

Figure 1: H1-B Visa Data

Data Dictionary

1. Serial Number

2. CASE_STATUS: This field denotes the status of the application after LCA processing. Certified applications are filed with USCIS for H-1B approval. 'CASE_STATUS: CERTIFIED' does not mean the applicant got his/her H1B visa approved, it just means that he/she is eligible to file an H-1B.

3. EMPLOYER_NAME: Name of the employer submitting labor condition application.

4. SOC_NAME: Occupational name associated with the 'SOC_CODE'. 'SOC_CODE' is the occupational code associated with the job being requested for temporary

labor condition, as classified by the Standard Occupational Classification (SOC) System.

5. JOB_TITLE: Title of the job.

6. FULL_TIME_POSITION: 'Y' = Full Time Position; 'N' = Part Time Position.

7. PREVAILING_WAGE: Prevailing Wage for the job being requested for temporary labor condition. The wage is listed at annual scale in USD. The prevailing wage for a job position is defined as the average wage paid to similarly employed workers in the requested occupation in the area of intended employment. The prevailing wage is based on the employer's minimum requirements for the position.

8. YEAR: Year in which the H-1B visa petition was filed.

9. WORKSITE: City, State

10. lon: Longitude of the location

11. lat: Latitude of the location

Data Exploration

The data set comprises of 3 million records of cases petitioned for H1-B Visas. There are diverse people with various job_titles applied.

A word cloud is generated to look at the most prevalent jobs. From it, we can observe that Software Engineer is the most demanded job for H1-B, followed by Computer Programmers, System Analysts, Software Developers and Business Analysts.

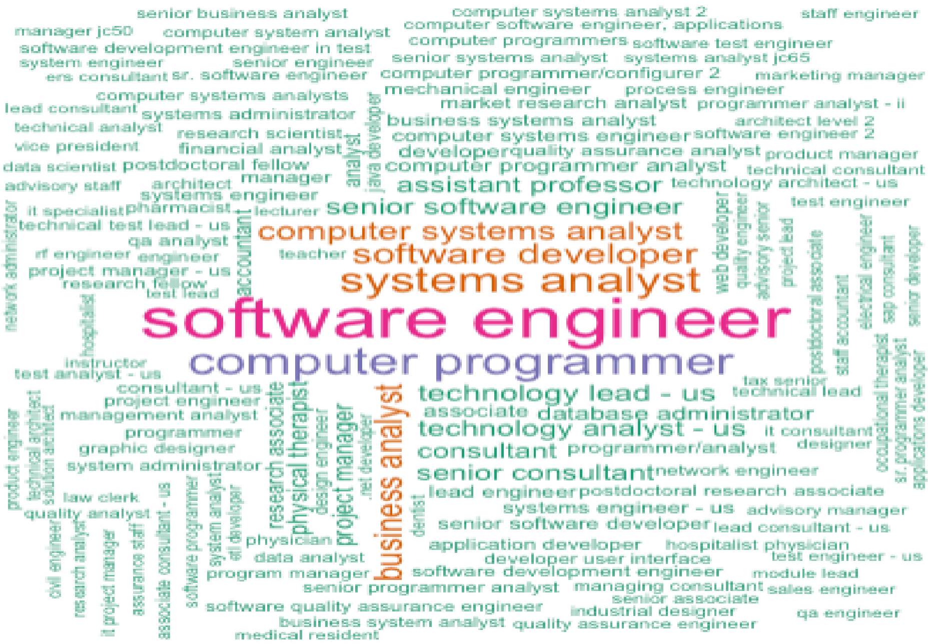
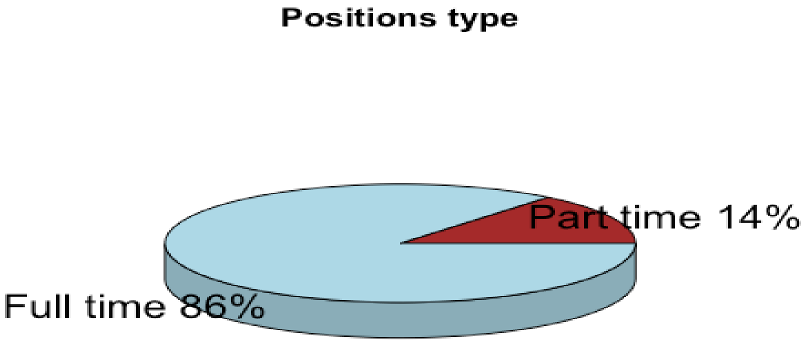
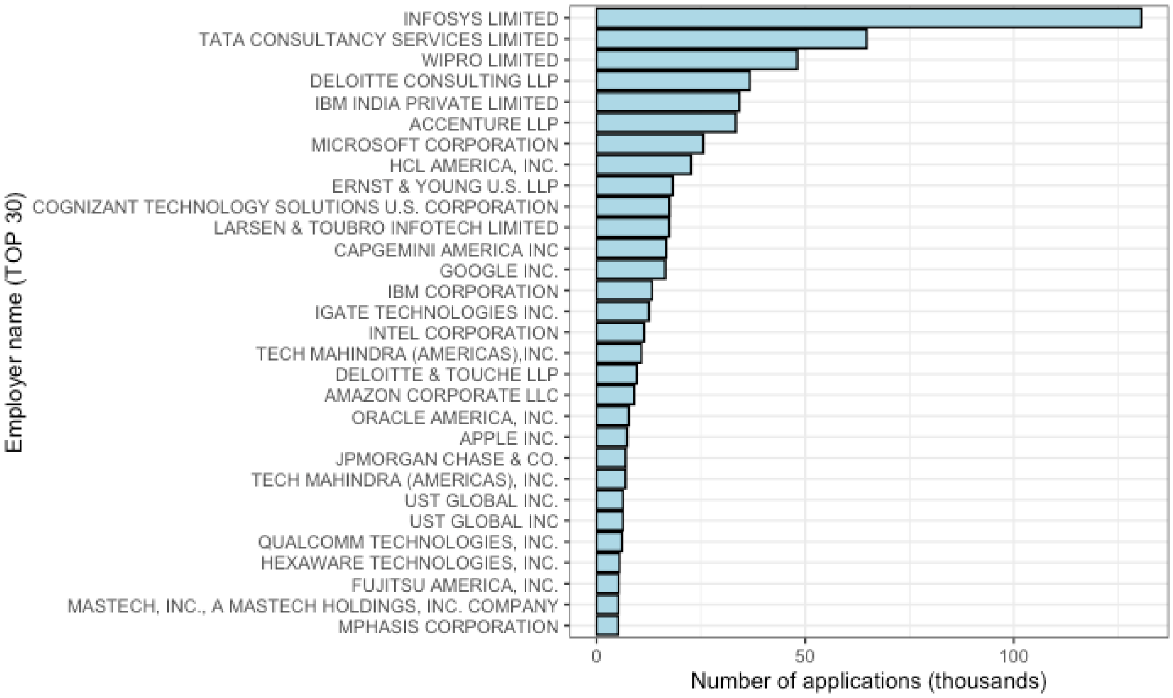


Figure 2: Word Cloud for Jobs

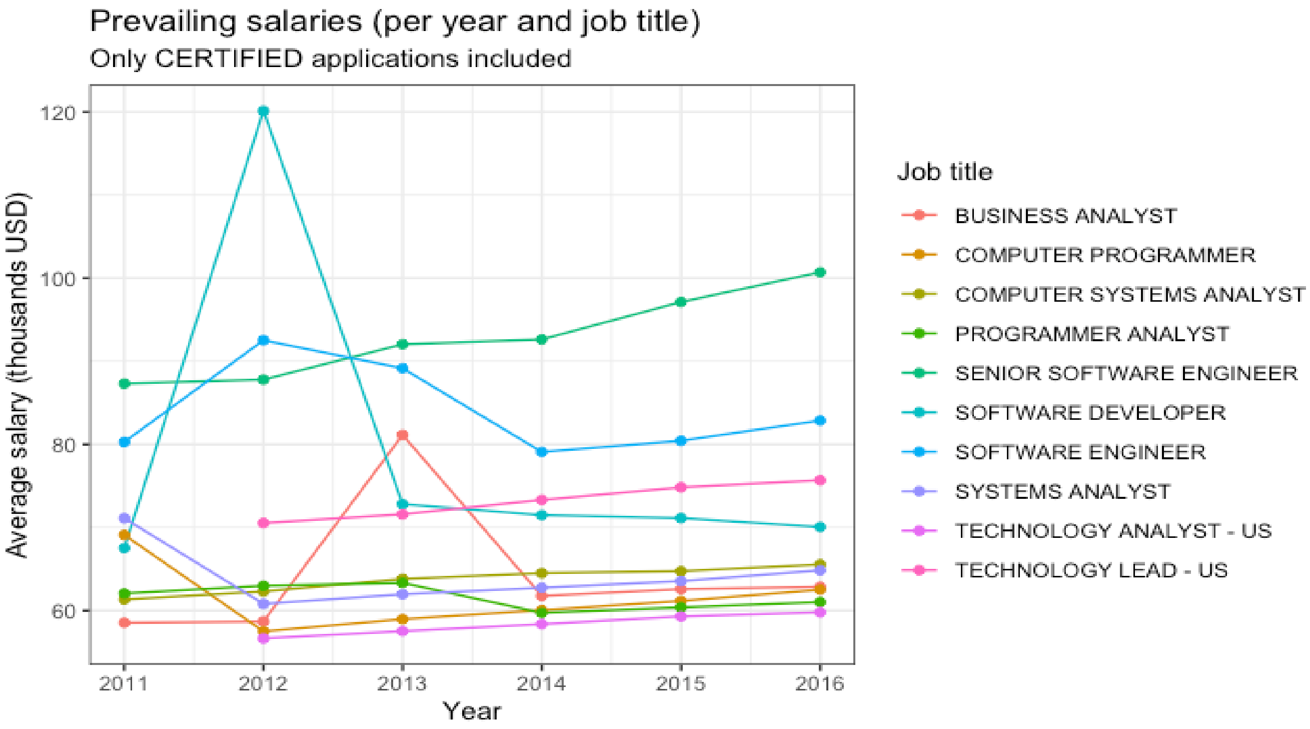
Next analysis was done to look at the type of position that obtained most H1-B's. It resulted in 86% full time positions and 14% part time positions.





Here we observe that:

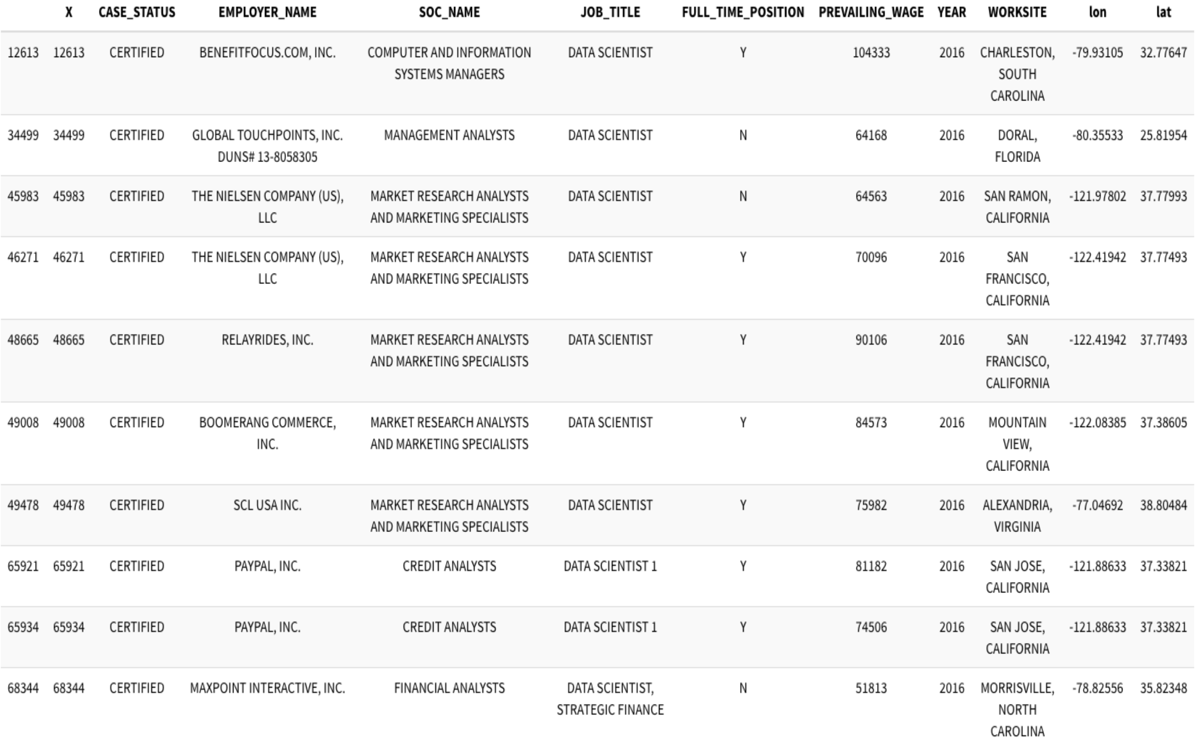
- Infosys has submitted highest H1B Visa Applications in last 5 years.
- Top 5 Companies that submit most H1B Visa Applications are companies based in India associated with software /IT services.
- Some companies may apply for more than what they need to increase the chances of getting their petitions selected in the lottery system.



We can observe the peak of SENIOR SOFTWARE ENGINEER in 2012 and the gradual increase of average prevailing wages for PROGRAMMER ANALYST from 2012, in parallel with a severe drop from 2012 of SOFTWARE ENGINEER. It is obvious that all the high-paid jobs are in the Computer Science related areas. From 2012 we can observe an almost perfectly correlated raise of prevailing wages for most of the top 10 income job titles.

Analysis of Data Scientist Jobs

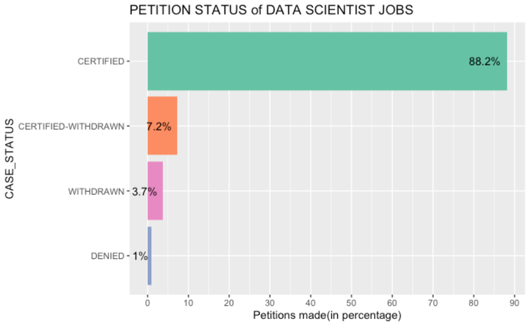
Filtered the data and extracted subset of it related to Job Title :'Data Scientist' & performed exploratory analysis.

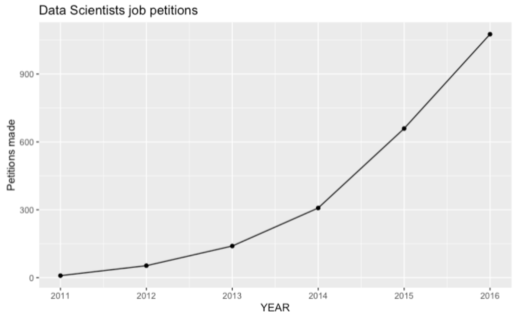


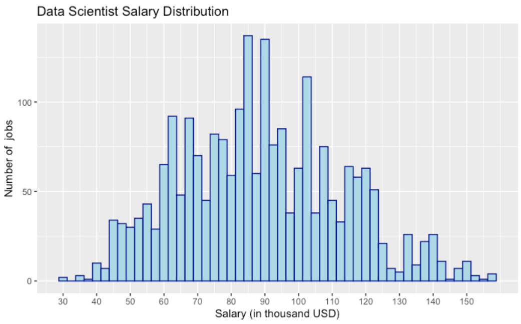
	X	CASE_STATUS	EMPLOYER_NAME	SOC_NAME	JOB_TITLE	FULL_TIME_POSITION	PREVAILING_WAGE	YEAR	WORKSITE	lon	lat
12613	12613	CERTIFIED	BENEFITFOCUS.COM, INC.	COMPUTER AND INFORMATION SYSTEMS MANAGERS	DATA SCIENTIST	Y	104333	2016	CHARLESTON, SOUTH CAROLINA	-79.93105	32.77647
34499	34499	CERTIFIED	GLOBAL TOUCHPOINTS, INC. DUNS# 13-8058305	MANAGEMENT ANALYSTS	DATA SCIENTIST	N	64168	2016	DORAL, FLORIDA	-80.35533	25.81954
45983	45983	CERTIFIED	THE NIELSEN COMPANY (US), LLC	MARKET RESEARCH ANALYSTS AND MARKETING SPECIALISTS	DATA SCIENTIST	N	64563	2016	SAN RAMON, CALIFORNIA	-121.97802	37.77993
46271	46271	CERTIFIED	THE NIELSEN COMPANY (US), LLC	MARKET RESEARCH ANALYSTS AND MARKETING SPECIALISTS	DATA SCIENTIST	Y	70096	2016	SAN FRANCISCO, CALIFORNIA	-122.41942	37.77493
48665	48665	CERTIFIED	RELAYRIDES, INC.	MARKET RESEARCH ANALYSTS AND MARKETING SPECIALISTS	DATA SCIENTIST	Y	90106	2016	SAN FRANCISCO, CALIFORNIA	-122.41942	37.77493
49008	49008	CERTIFIED	BOOMERANG COMMERCE, INC.	MARKET RESEARCH ANALYSTS AND MARKETING SPECIALISTS	DATA SCIENTIST	Y	84573	2016	MOUNTAIN VIEW, CALIFORNIA	-122.08385	37.38605
49478	49478	CERTIFIED	SCL USA INC.	MARKET RESEARCH ANALYSTS AND MARKETING SPECIALISTS	DATA SCIENTIST	Y	75982	2016	ALEXANDRIA, VIRGINIA	-77.04692	38.80484
65921	65921	CERTIFIED	PAYPAL, INC.	CREDIT ANALYSTS	DATA SCIENTIST 1	Y	81182	2016	SAN JOSE, CALIFORNIA	-121.88633	37.33821
65934	65934	CERTIFIED	PAYPAL, INC.	CREDIT ANALYSTS	DATA SCIENTIST 1	Y	74506	2016	SAN JOSE, CALIFORNIA	-121.88633	37.33821
68344	68344	CERTIFIED	MAXPOINT INTERACTIVE, INC.	FINANCIAL ANALYSTS	DATA SCIENTIST, STRATEGIC FINANCE	N	51813	2016	MORRISVILLE, NORTH CAROLINA	-78.82556	35.82348

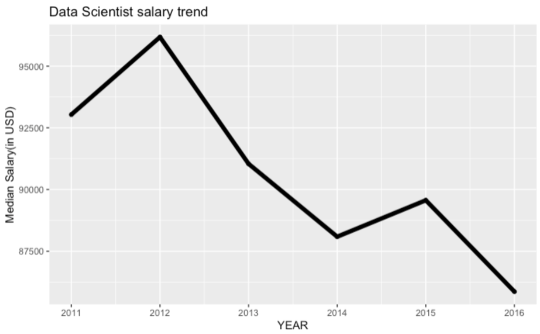
Figure 3: Data Scientist Table

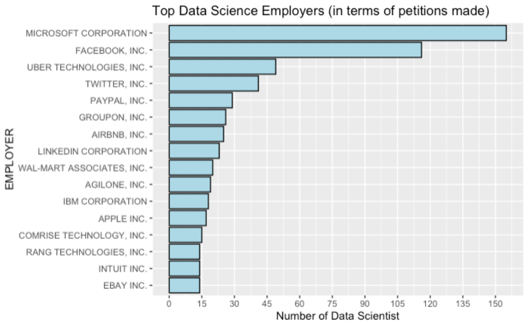
Charted out following plots depicting the various scenarios of H1-B petition status under Data Scientist role.

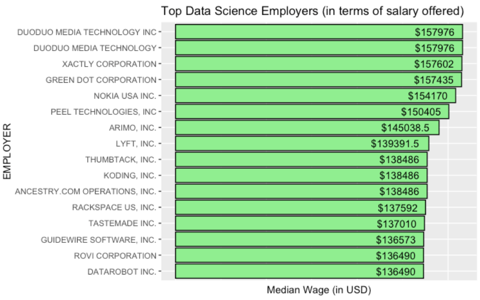


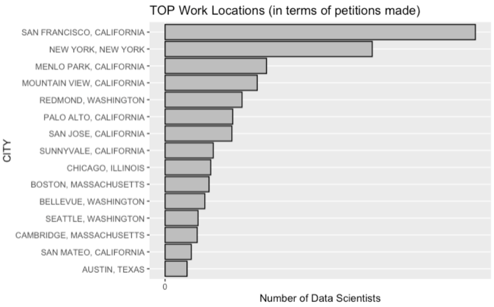


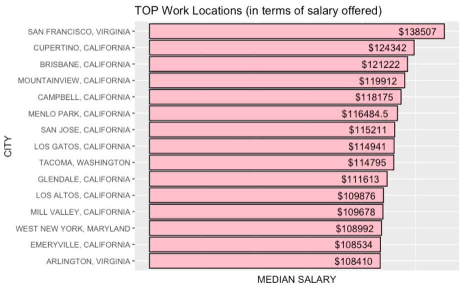

Data Scientist Salary Distribution


Data Scientist salary trend


Top Data Science Employers (in terms of petitions made)


Top Data Science Employers (in terms of salary offered)


TOP Work Locations (in terms of petitions made)


TOP Work Locations (in terms of salary offered)

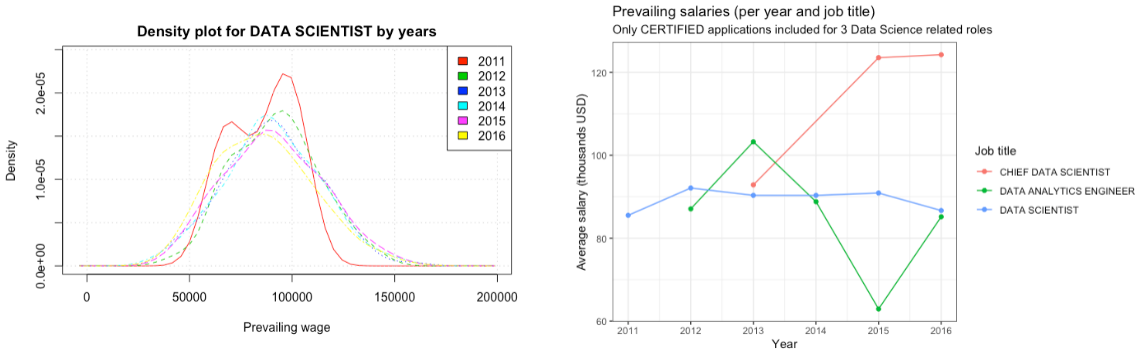


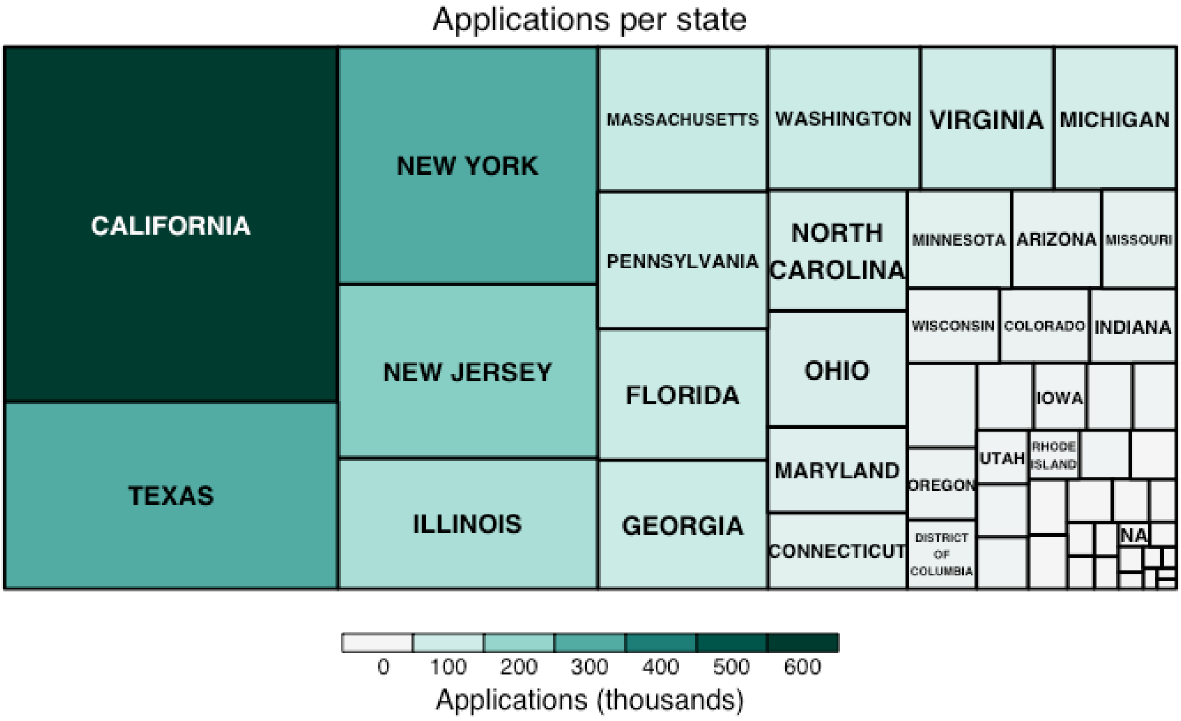
Figure 4: **Data Scientist Job Title - Observations**

Following conclusions were made :

- Microsoft and Facebook are way ahead in filing petitions for Data Scientist jobs. On the other hand some not so familiar companies are paying the highest salaries in the field.
- The average salary reported on an H1B application for last 5 years was around 89k USD for Data Science related jobs. The figure (salary Distribution) tells 25% above than the average salary reported for H1B application for overall jobs
- The median salary stays around USD 90k with a little decrease over the years. But it is still around USD 90k. However, a clear upward trend can be seen in number of petitions made each year.
- For 2011 there is an interesting double peak density plot profile, with highest peak at K$95 and lower peak at K$65. The values span is increasing in 2012 with a longer queue to the upper values, up to K$150 and above (also) and follow almost the same profile until 2016. The prevailing wages interval increased although the averages values did not moved drastically from the initial average.
- We can observe that DATA SCIENTIST shows quite a constant variation during the period 2011-2016, with peak value for 2012.

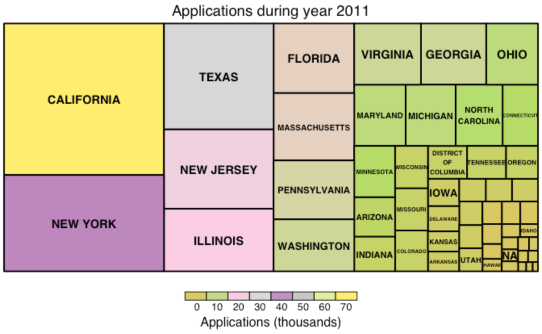
Applications by Worksite and Year

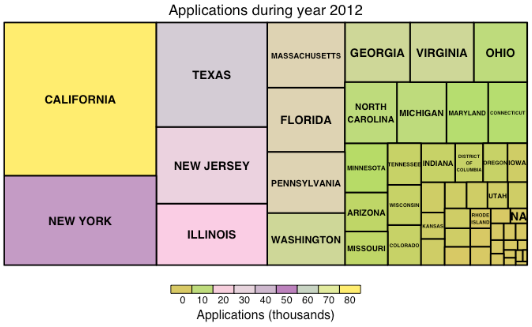
We extracted the state information from the WORKSITE field. Then grouped the applications by state.

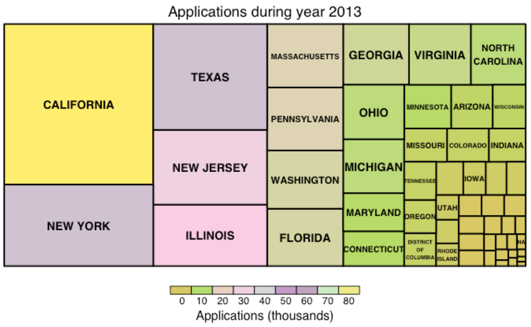


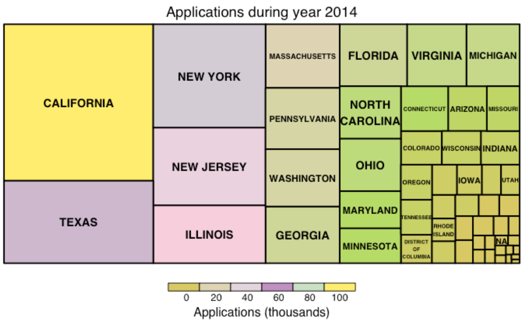
The States with most of the Applications are California, New York, New Jersey, Massachusetts, Illinois, Pennsylvania, Texas.

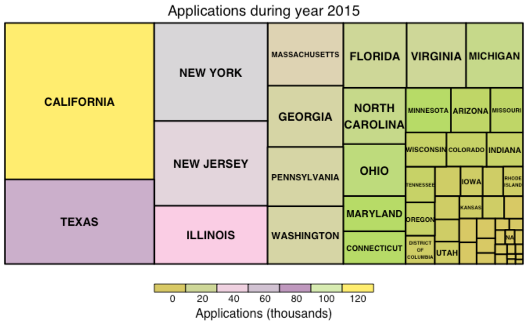
The below plots indicate the H1-B Applications petitioned each year from 2011-2016 grouped by state.

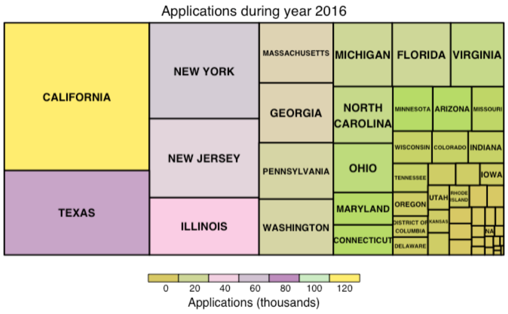












Data Modelling

Since there are multiple discrete and categorical variables , we tried models Logistic Regression, Random Forest and Boosting. Our plan is to model a classifier to predict whether an applicant gets his H1B certified or denied. Hence, we dropped cases where the feature values of the target are otherwise. We dropped the features CITY, latitude and longitude but we kept the feature STATE as they all describe the location. The other categorical variables are transformed as factors and random data points are taken as the dataset is quite large. Our response variable being categorical, we labeled 1 and 0 accordingly.

Logistics Regression

Logistic regression is considered as the response or outcome variable is discrete. Apart from the data pre-processing done above, we took just the cases of latest year to build the model as the dataset is large enough(3 million) to not fit in RAM. All the categorical variables are factorized and VIF is applied to remove the predictors that are deflected the variance. The dataset is then split into train and test datasets in the ration of 79:21. The model is trained using glm function in R and prediction is done on test data. An accuracy of 76.5% is observed for the model.

Logistic Regression		
	Actual	
Predicted	1	0
1	12830	252
0	3343	141

TP	12830
FP	252
FN	3343
TN	141
Accuracy	0.78298926
Misclassification	0.21701074
Precision	0.98073689
Sensitivity/Recall	0.79329747
Specificity	0.35877863

Boosting

The same data set is applied on the gradient boosting model. gbm function in R is applied to train the model and the results proved that the gradient model is a better choice. It resulted in 97% of accuracy.

Gradient Boosting		
	Actual	
Predicted	1	0
1	17051	0
0	426	9

Sensitivity	0.012723
Specificity	0.999753
Pos Pred Value	0.555556
Neg Pred Value	0.976566

Random Forest

We did some further pre-processing on the data. We introduced a system of taking 'n' random datapoints from the dataset, otherwise we were running out of RAM (since the dataset has more than 2.6 million datapoints) instead of a particular year.

We used one hot encoding for the features FULL_TIME_POSITION and YEAR for better training. We used Label Encoder for STATE as using one hot encoder would create too many features. There are three other features which are categorical and have very high cardinality: EMPLOYER_NAME, SOC_NAME, JOB_TITLE. First we used Label Encoding on them and trained our model but didn't get a good enough result. Then we dropped these features and got a fairly decent result.

We used 80% of the dataset for training and the rest for testing. After predicting on the testing data, we generated the confusion matrix and calculated different metrics from it which are shown below:

Random Forest		
	Actual	
Predicted	1	0
1	19311	0
0	663	25

Sensitivity	0.966477
Specificity	1
Pos Pred Value	1
Neg Pred Value	0.0335

Conclusion

- The technical jobs like Computer Programmers, SystemAnalysts, Software Developers and Business Analysts are the ones that are mostly filled by foreign workers proving that they have high demand. So emphasis should be driven on promoting coding in schools across U.S. such that more technical labor force is available for the jobs in the field. Only 40% of the schools in the United States teach computer programming.[Source: Code.org]
- Random Forest gives better set of results to these kind of datasets. However, we need more set of features to improve the accuracy.

References

[1] https://en.wikipedia.org/wiki/H-1B_visa

[2] https://www.analyticsvidhya.com/blog/2016/01/xgboost-algorithm-easy-steps/

[3] "H-1B Fiscal Year (FY) 2018 Cap Season," USCIS. [Online]. Available: https://www. uscis.gov/working-united-states/temporary-workers/h-1b-specialty-occupationsand-fashion-models/h-1b-fiscal-year-fy-2018-cap-season. [Accessed: 20-Oct-2017].

[4] "Predicting Case Status of H-1B Visa Petitions." [Online]. Available: https://cseweb. ucsd.edu/classes/wi17/cse258-a/reports/a054.pdf.