An Analysis of PERM Labor Certification and Labor Condition Applications from the United States Department of Labor

Arunkumar Ranganathan
Brian Detweiler
Jacques Anthony

October 20, 2016

Abstract

Foreign born workers make up 17% of the United States workforce. In 2014, nearly one million foreign nationals became lawful permanent residents in the United States. Of those one million, 140,000 came through visas which are allocated to employment based residency. Where are these workers, and what do the demographics look like? How does each company's compensation measure up? Here, we use statistical analysis and business analytics to examine visa application data from the U.S. Department of Labor from 2008 to 2016. We intend to create an interactive data product that will make this publicly available information more accessible to the students who are entering the workforce, as well as to US citizens and permanent residents. This will empower them to competitively position themselves in the job market by making more informed decisions.

1 Introduction

The U.S. Department of Labor provides data for Labor Condition Applications and PERM Labor Certifications dating back to 2008. This data contains a wealth of job market information including prevailing wage, and the wages offered by particular companies to individuals with particular qualifications.

Foreign workers can work legally in the U.S. under temporary or immigrant visas. To employ foreign workers legally in the U.S., employers must submit requests to the Office of Foreign Labor Certification. This process keeps borders secure, while protecting foreign workers from unfair treatment by employers. Businesses are required to pay the visa holder the higher of the prevailing wage for the position or the prevailing wage for the occupation in the geographic region of employment.

Here, we look for disparities in wages for domestic and foreign workers if any exist. The H-1B visa program allows businesses to hire skilled professionals with at least a bachelor's degree in areas of specialization.

Each year the federal government grants a maximum of 65,000 applications for the H-1B visa program and additional 20,000 applications for applicants with at least a U.S. master's degree. Permanent applications are limited to 140,000 per year.

Applicants whose petitons are denied for permanent residency must leave the U.S. for one year but are eligible to reapply later on.

The maximum length for the H-1B visa is six years. For permanent residency, an individual under the H-1B program may be sponsored by the employer. Workers with remarkable ability in science, education, arts, business, and education such as professors, researchers, or business executives and managers are eligible to apply for permanent residency without a labor crefication. Some special exceptions to these rules include business investors who invest over half a million dollars in the US, employees of US foreign service posts, retired employees from international organizations, and other classes of foreigners may also petition for permanent residency with no labor certification.

1.1 Document reproducibility

The entirety of this project is reproducible using R (version 3.3 and above) with the `Knitr` package. All code, including this document, is available through GitHub. [**GitHub**]

2 About the data

The Office of Foreign Labor Certification, under the Department of Labor provides data for PERM Labor Certification (LC) applications and Labor Condition Applications (LCA) via XLSX files. Data is available from 2008 onward. The iCERT system was implemented in 2009, so there are two files for 2009. Each file is structured similar to the others, but there are differences which must be addressed.

PERM (Program Electronic Review Management) data is available from 2008 to the present. Years 2015 and 2016 have over 120 attributes while 2008 through 2014 have 27-30 attributes. We plan on using 27 attributes common across 2008-2016, as well as an additional 14 attributes found only in 2015-2016. These may be useful for limited analysis. In this data, we have employer and employee details, job function, salary, university, job city, number of years of experience, country of citizenship, and industry.

2.1 H-1B Data preparation

The H-1B data is about 75% larger than the PERM data, and spreadsheet programs do not handle these well, so the first task was to export these to CSV formatted files so that they could be handled with better tools. When exporting, they were also given more uniform file names in the form of "H-1B_yyyy.csv" where "yyyy" is the year of the data. The one exception here is the 2009 iCERT data, which was named "H-1B_2009_icert.csv". Once in CSV format, we needed to identify common columns across all spreadsheets. The difficulty here, is that the columns do not have the same names across spreadsheets, even though they may be holding the same data.

Using the UNIX tool `head -n 10 *.csv > headers.txt`, we took the first ten rows of each file and put them into a separate file. Each of the CSVs first ten rows were then copied and pasted into another spreadsheet, and we undertook a manual effort to match columns of the same identity. We also discarded some excess information that we deemed to be unnecessary for our purposes.

It is also important to note that there was not always a match for the columns we had selected. For instance, we found some interesting information regarding the attorney used by the employer to file the H-1B application. This was only introduced in the 2015 and 2016 datasets, however, so prior years would have no data for this.

After determining the standard columns, we wrote an import script in R that made use of the function `data.table::fread`. This allowed us to not only quickly read in the file, but also select only the columns of interest, and rename them to the standard naming convention upon read.

Once the data was read into individual data frames, additional cleaning rules were applied. In some years, wage data contained invalid numeric characters such as dollar signs or a range of wages in a single column. To get around this, dollar signs were removed before converting to numeric, and ranges were split into a *from* and a *to* column. Ultimately, all wages were transformed into ranges. If there was no range for the wage, then the wage itself was used as the range.

Another normalization task was the wage unit. Some wages were represented as yearly salary, some as hourly, and others as monthly, weekly, and bi-weekly. There can be subtle differences in each type of pay, but these were normalized according to a yearly salary. Hourly wage was multiplied by 40 hours a week and 52 weeks a year. Monthly wage was multiplied by 12, weekly by 52, and bi-weekly by 26. This allows all wages to be treated roughly on the same scale.

With the exception of exporting data to CSV format, the rest of the steps have been combined into a single script, `csv_manipulation.R`. [**GitHub**]

2.2 PERM Data preparation

PERM data from the Department of Foreign Labor Certification is in XLSX format. We can read and manipulate these using the `xlsx` and `dplyr` packages.

Transforming all character variables to upper case deemed as necessary to avoid duplicity or incorrectness when performing data manipulation.

In total there were 9 files downloaded, with each file ranging from 40,000 to 90,000 records.

Most column names from 2008-2014 were the same with minor differences, while 2015-2016 had a different format, but were similar to each other. Column names are standardized to the 2016 version for posterity

and the assumption that in the short term future, the columns will remain the same. To standardize all columns, we have replaced spaces in column names with underscores and created all column names in upper case. The first step in preparing data was to select all the relevant columns and fill empty columns wth `NA`s.

Once we have all the data available for each year, a final version is created with the raw data without standardizing or further cleaning. This dataset can be used by anyone looking for this PERM data and can standardize based on their needs.

Once we loaded and cleaned the data, we selected only the required columns and exported this as an RDS file. In this process of creating final file, we further reduced the columns and kept only the columns that are essential for our data analysis.

Multiple XLSX files are initially filtered with the selected 41 columns and merged back into one large file with 622,637 records.

Data cleaning specific to this project is also required not just for standardizing data but also to have the data in a specific format.

During the process of creating final file, most columns are read as text and a few columns such as 'decision_date' are read as date, while salary information was read as numeric. Decision date is first transformed to YYYY-MM-DD format for consistency within our database.

Transforming all character variables to upper case deemed as necessary to avoid duplicity or incorrectness when performing data manipulation.

Wage information is marked in unit as weekly, monthly, bi-weekly, and annually. This wage units weren't consistent in the file - so we have to make all the weekly to 'WK', monthly to' MTH' and bi-weekly to 'BIWK' and yearly to 'YR' across data.

Wage information was normalized to create a new column just with annual salary - i.e. weekly, hourly, monthly, and bi-weekly salaries are transformed to yearly. Range of salary is also normalized in this step.

Employer zip codes are trimmed to 5 digits to keep all the zip codes at the same level of depth

Employment state data is both at abbreviated and expanded level. We needed to standardize either as a two digit code or expanded for consistency. For this exercise state names are created at the expanded level for all 50 US states as well as US territories.

Next we needed to create a unique id to tie all the information back together as well as to create index in the future, so row number was created as unique id.

This cleaned data is saved as an RDS file before the next step.

2.3 Geolocation data

Another RDS file (PermEmpMapsdat.rds) is created with summarized employer name address, city, state and zip codes order by number or perm applications processed and by their mean salary. This dataset would give us which employer sponsors most employees for permanent full time employment as well as who pays more. Another use of this dataset is in geocoding.

In order to get geocodes for employer address we decided to go with a distinct employer addresses in the order of most common PERM application employer who also pays the highest wages.

As we are restricted to 2,500 geocode requests from Google Maps in a 24 hour period, we decided to create a program that continuously runs until the rate limit is encountered, sleeps, and wakes up every hour trying to hit the google server to look for the addresses.

This program will collect all the addresses and store them in a temporary file and updat them again based on their index in a main file. This program is designed to run at multiple locations for different index ranges to collect as much address as possible.

2.4 Shortcomings

As mentioned in the previous section, because the data is not homogeneous, there are bound to be disparities. Missing data - columns which are not found across all spreadsheets - is the biggest issue. We can make assumptions when there is sparse data, but it would not be prudent to make assumptions where there is no data. For this reason, we fully disclose the absence data where necessary.

All of the data has been entered by humans at some point, so there are likely many human-generated errors. Some of these can be seen as outliers. Particularly in the 2008 and pre-iCERT 2009 data, the wage

unit is most certainly incorrect in some spots. For example, some wages are listed at $500 per hour, but the intended unit may have been per week. It is not possible to fix this programmatically though, because there are, in fact, some jobs that pay $500 per hour (CEOs, for instance). This data must be dealt with in one of two ways. They can either be corrected by hand inspection of outliers, or outliers can be removed completely. This results in a slight loss of fidelity. Extremely high paying jobs, such as CEO or physician may not be displayed.

Another issue is the switch from U.S. Citizenship and Immigration Services Dictionary of Occupational Titles (DOT) codes in 2008 and pre-iCERT 2009 data, to the North American Industry Classification System (NAICS) codes. The DOT codes are three digits and fairly high-level, where as the NAICS codes are hierarchical, with the first two being the industry, and the specification of the job title narrowing with up to six digits. For this reason, it is difficult to get consistent job titles across years.

PERM data is comprehensive for 2015 and 2016, however the data for each year have inconsistencies. We had to make some educated judgements about some categories and ignore many cases with empty values. Years 2008-2014 have 25 attributes, but most of them do not give us substantial information other than salary and employer name. The nature of this data made it hard to graph plots based of these categorical variables. Here is a sample of the categorical data:

53,011	Job title based on Perm data
46,410	Job Title based on Work
40,431	Study Major
135,955	Employers
213	Countries in the "other" category

3 Methods

The numeric data provided by the H-1B and PERM datasets are mostly in the form of wages, both the wage that the employer is offering and the prevailing wage.[1] Also of interest, are the number of workers an applicant is filing for, and the implicit number of applications faceted by status and year.

3.1 Data product

The data is provided as an interactive `Shiny` application, that allows the user to filter wages by various criteria.

The plots consist of distributions and and heat maps of wages across the United States. Our objective is to have data that can be sliced by users into different factors; by state, city, employer, job major, function, salary. The data products we produce will help our target audience make informed decision about kinds of employment held by high skilled immigrants, what employers are hiring, and what kinds of skills employers are looking for.

First we will look the number of applications processed since 2008. Using these statistics, we can gague an increasing or decreaseing trend in the number of applications processed. Although there are caps on the annual number of employment based immigrations, an increasing trend or maximum use of employment based immigration would provide insight.

Next, we can look for immigration based on geography. In what regions are employers hiring? We can also look at time series data on hiring by job function, job majors and university of education. Box plots, histograms, bar charts and heat maps are some of those plots we will use to draw inferences.

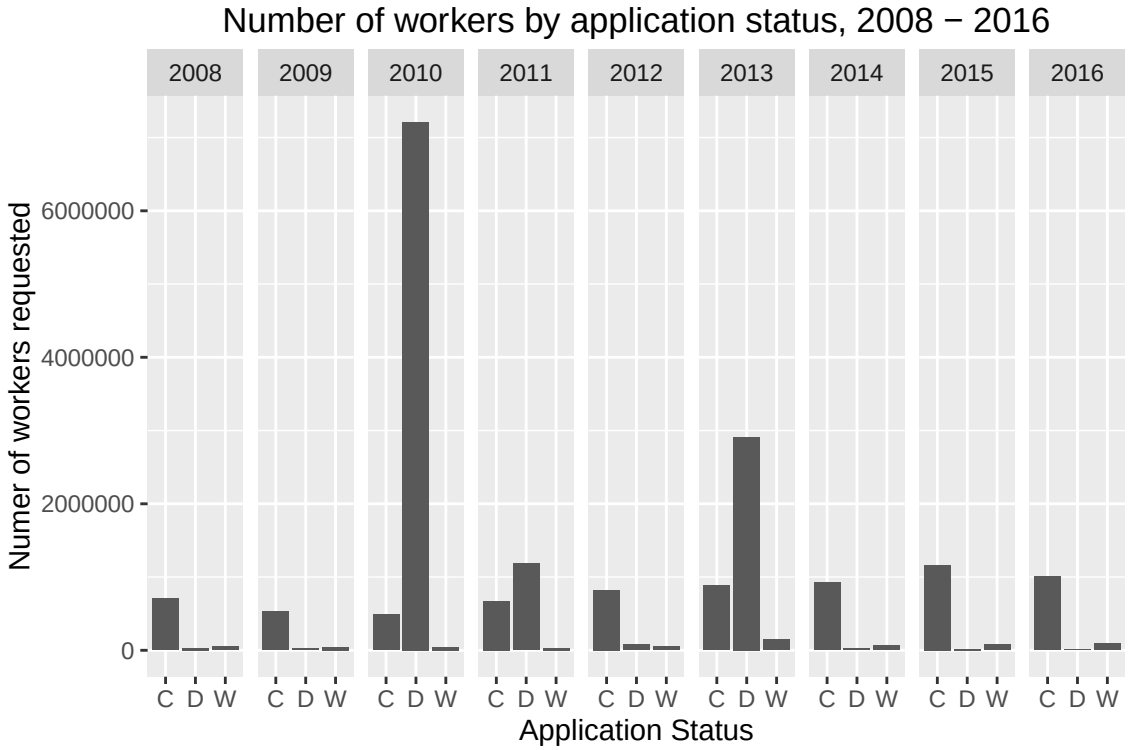
[1] Prevailing wage is defined as the hourly wage, usual benefits and overtime, paid to the majority of workers, laborers, and mechanics within a particular geographic area.

4 Results

4.1 Overview

4.2 Prevailing Wage

It would be interesting to see how many workers are getting approved for H-1B visas over the years. We can see this by plotting the number of workers within each visa status category (Certified, Denied, and Withdrawn).



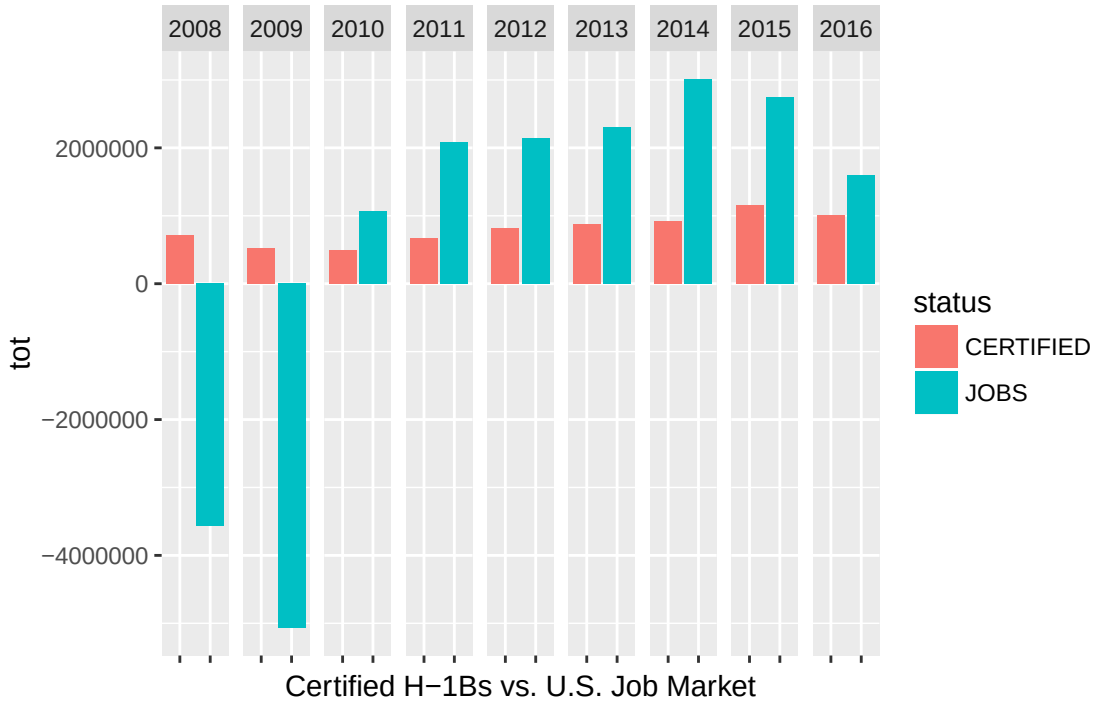
Clearly, we can see huge outliers in 2010 and 2013 that don't seem to fit the data. Upon investigation, it is safe to say that all H-1B applications requesting over 1,000 total workers are either denied or withdrawn. These can be safely ignored.

Once we remove all requests for more than 1,000 workers, we can start to see a pattern. The number of denied and withdrawn applications remains fairly constant, but the number of certified workers shows a steady rise after 2010, most likely due to a strengthening economy and returning jobs. [**BLS**]

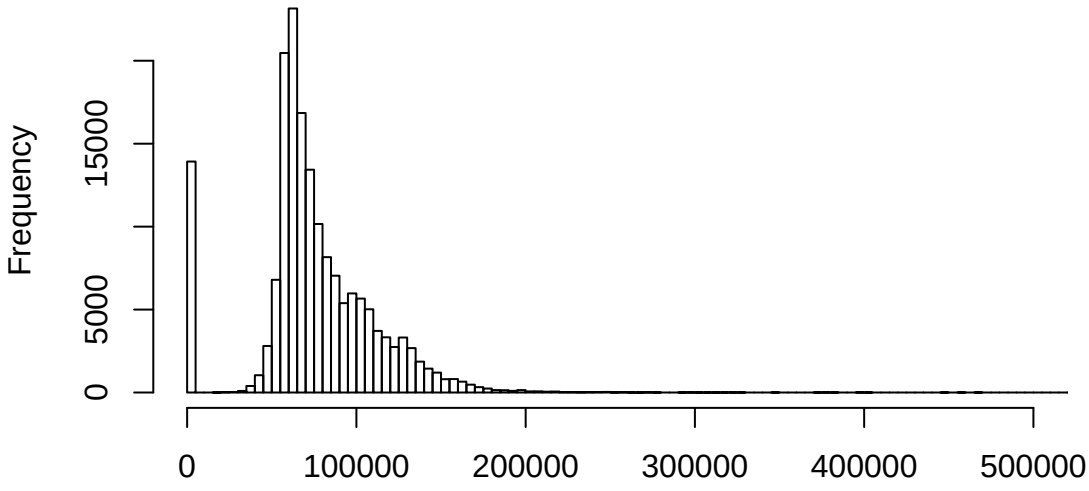
Table 1: Applications requesting over 1000 workers

fy	total_workers	status
2011	2010	DENIED
2012	2011	DENIED
2013	2012	DENIED
2013	2012	DENIED
2014	2013	DENIED
2011	10000	DENIED
2012	43555	DENIED
2010	52000	DENIED
2010	53000	DENIED
2010	53000	DENIED
2013	58500	DENIED
2013	60000	WITHDRAWN
2013	793172	DENIED
2011	1132014	DENIED
2013	2031308	DENIED
2010	7000000	DENIED

umber of workers by application status vs. U.S. job market, 2008 − 2016



s[which(visas$naics_title == "Computer Systems Design Services



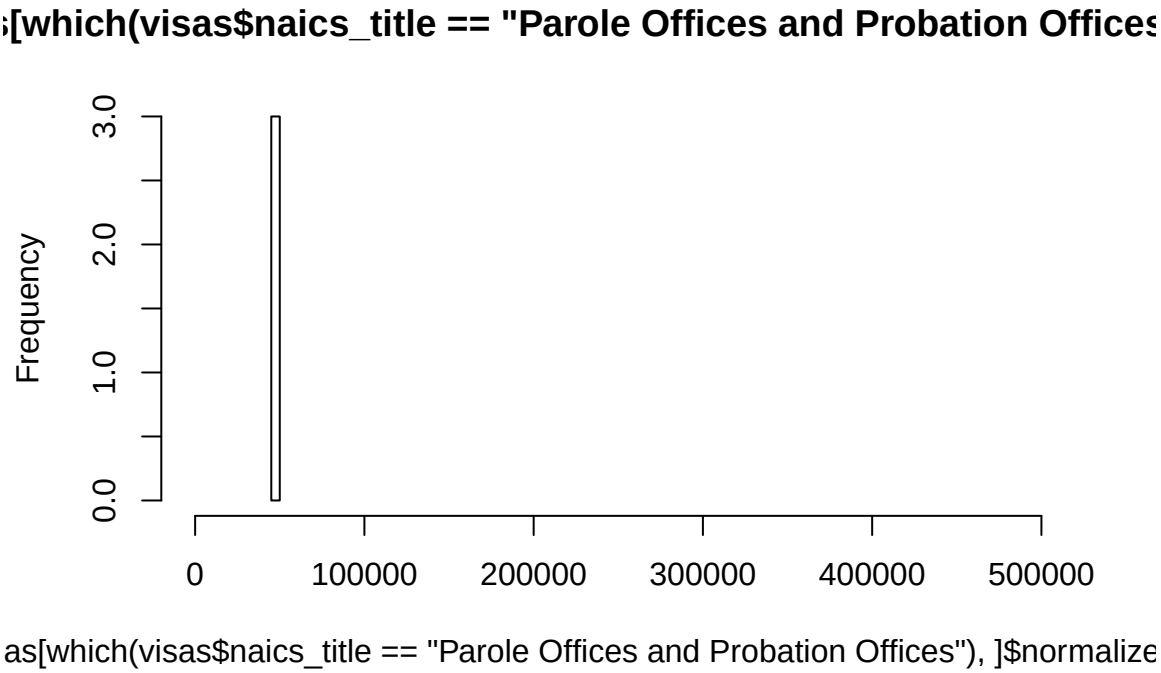
as[which(visas$naics_title == "Computer Systems Design Services"),]$normalize

Table 2: Average Pay by Case Status

YEAR	CERTIFIED	CERTIFIED-EXPIRED	AVG_SALARY
2008	111,437	155,360	136,532
2011	114,053	115,513	115,057
2012	21,865	122,280	122,276
2013	111,551	104,429	107,209
2014	116,833	99,348	109,291
2015	90,882	98,669	94,602
2016	117,454	94,946	111,082

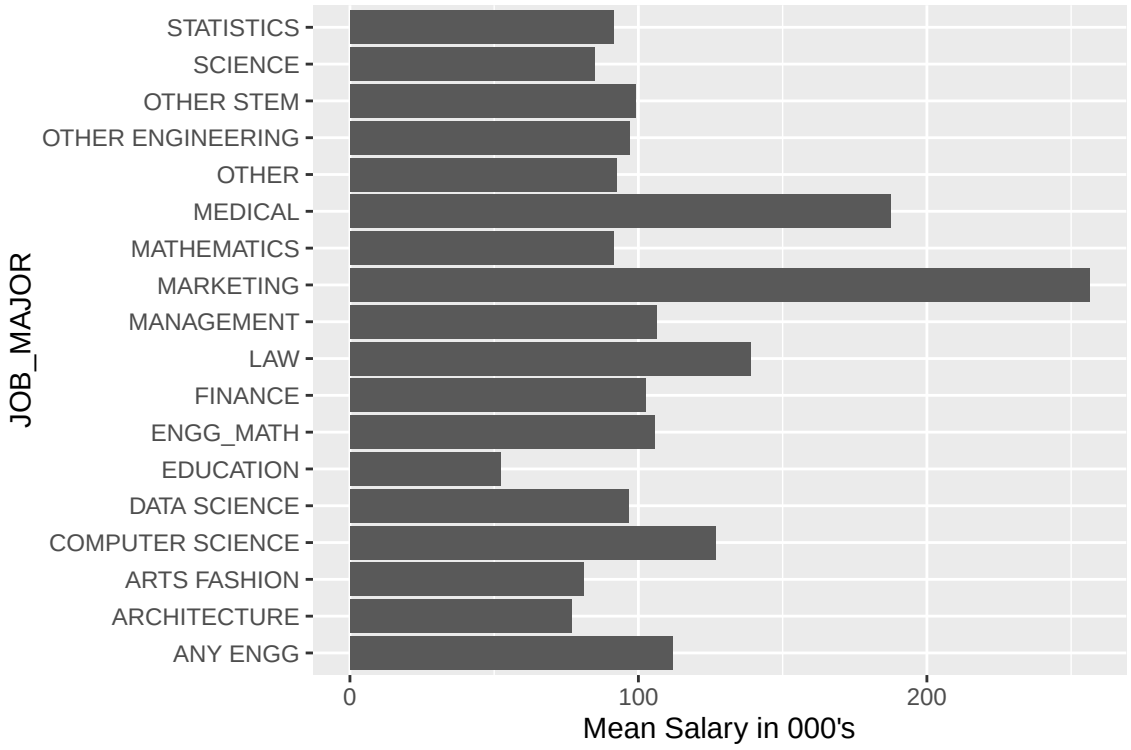
Table 3: Applications by year

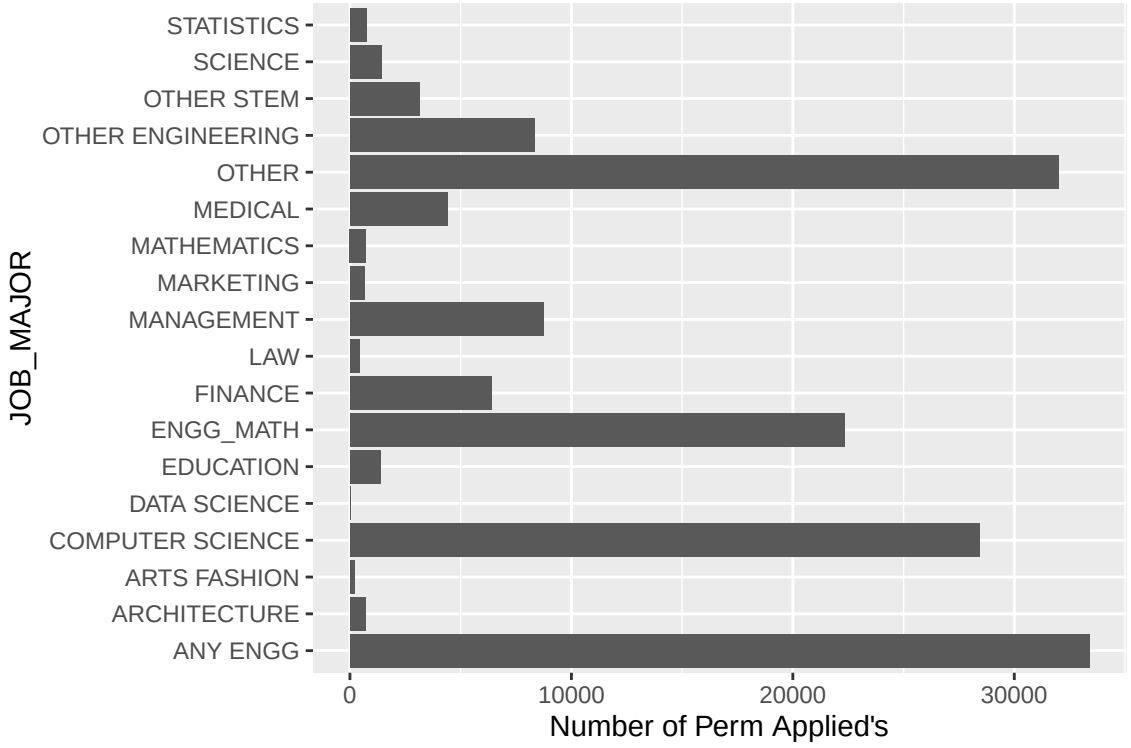
YEAR	CERTIFIED	CERTIFIED-EXPIRED	DENIED	WITHDRAWN	APPLICATIONS
2008	21,092	28,113	10,729	2,063	61,997
2011	18,714	41,149	10,384	2,960	73,207
2012	2	54,579	8,642	3,265	66,488
2013	13,742	21,461	5,874	3,075	44,152
2014	35,615	27,018	4,349	4,016	70,998
2015	41,223	37,715	5,999	4,362	89,299
2016	63,709	25,158	4,254	3,716	96,837



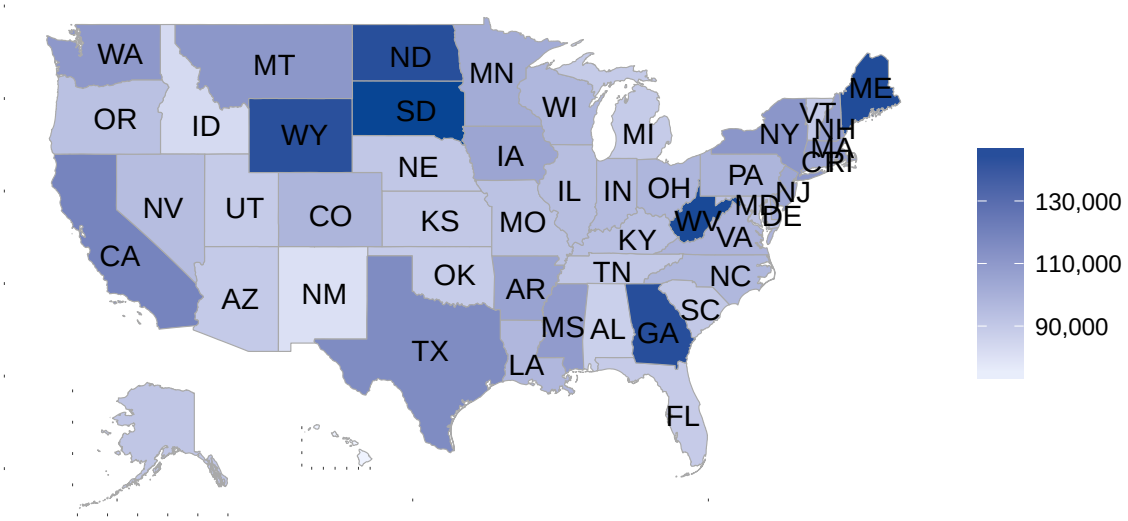
We are seeing a higher than average salary from 2008-2013 and in the last 3 years average salary has dropped down, incidently number of applications also have increased as shows in the next table

```
## Loading required package:  tcltk
```

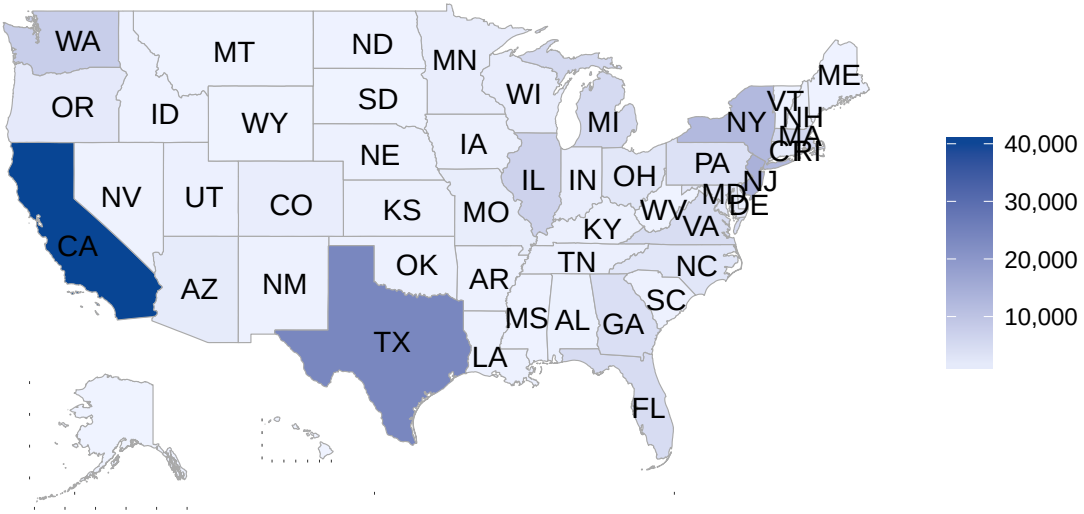


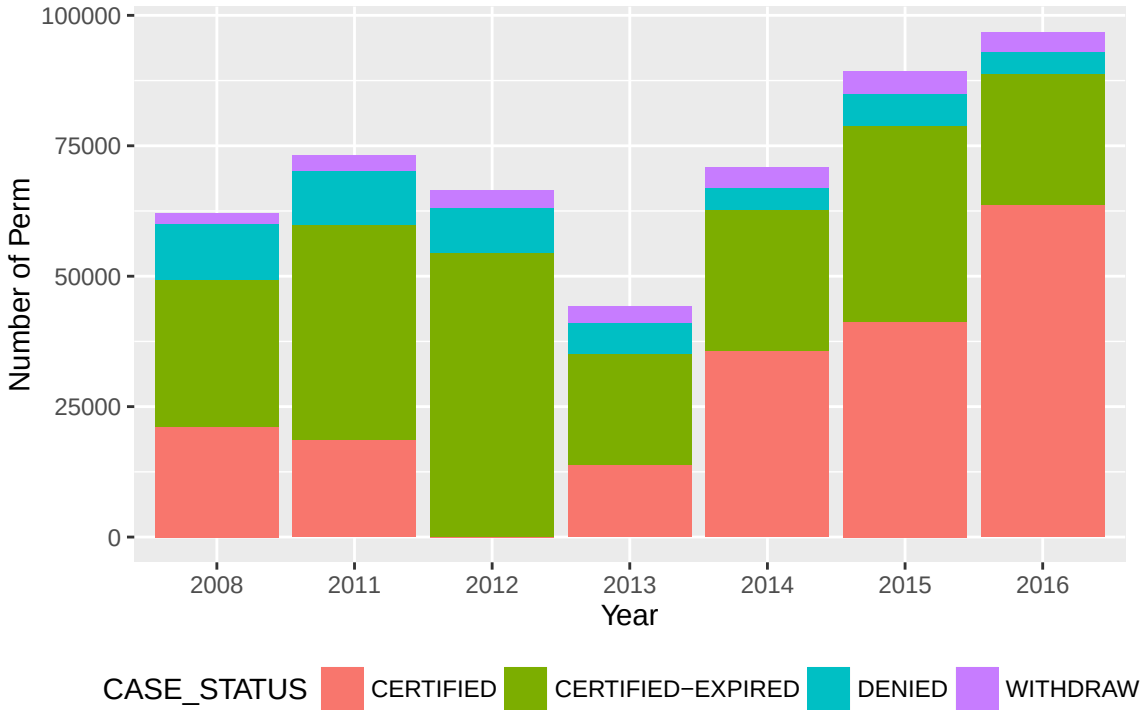


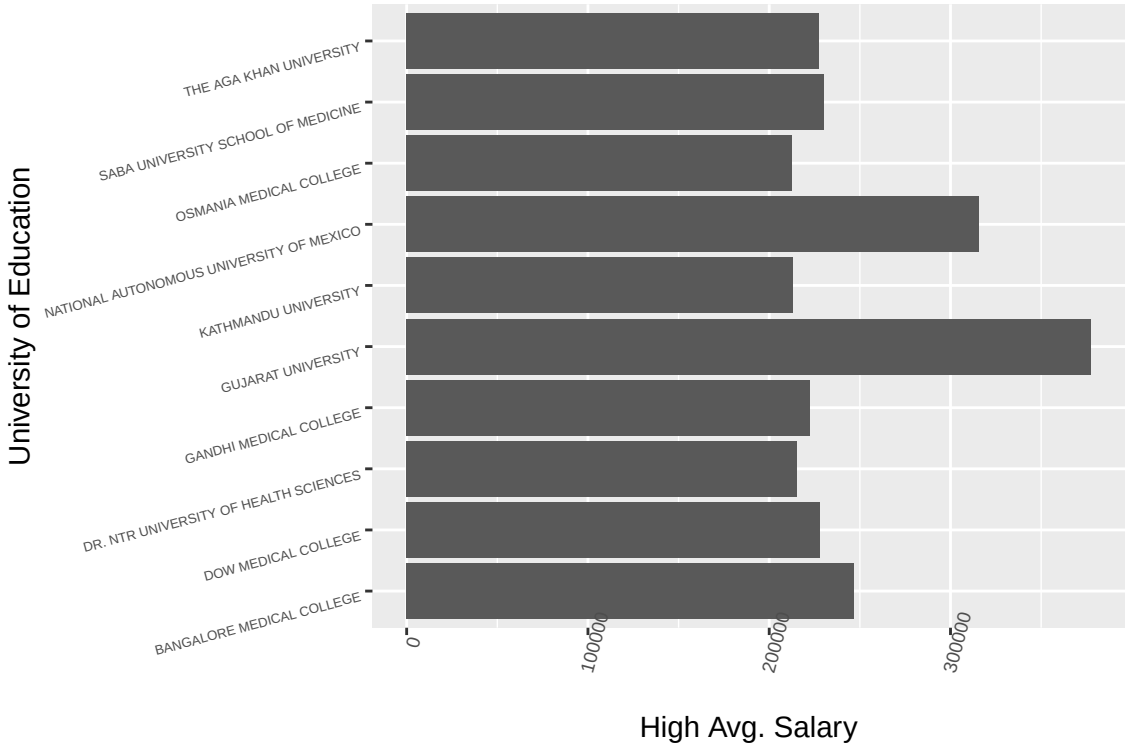
Mean Salary by State



Number of Perm by State







4.3 Offered Wage

We are seeing outliers in the wage offered both in H1B and Perm. How we deal with these outliers would change the way we come up with statistical inferences. Right now Perm data doesnt exclude outliers, however showing data as such clearly skews the results in favor of one or the other. One way to mitigate this is to

include only approved applications to remove any error from denial cases. We can also consider excluding outliers that are above 2 standard deviations from the mean.

4.4 H-1B vs. PERM

5 Conclusion

```
#visas <- readRDS('H1BVisas.rds')
#perm <- readRDS('PermData.rds')
#perm.map <- readRDS('PermEmpMapsdat.rds')

#hist(visas[which(visas$normalized_wage < 250000),]$normalized_wage, breaks=500)

#findmode <- function(x, na.rm = TRUE) {

  #if(na.rm){
    #x = x[!is.na(x)]
  #}

  #ux <- unique(x)
  #return(ux[which.max(tabulate(match(x, ux)))])
#}

#wage.mode <- findmode(visas$normalized_wage)
#wage.mode
#abline(v=wage.mode + 500, col="red")

#hist(visas[which(visas$normalized_prevailing_wage < 250000),]$normalized_prevailing_wage, breaks=500)
#pw.mode <- findmode(visas$normalized_prevailing_wage)
#pw.mode
#abline(v=wage.mode + 500, col="red")
```

References

[1] U.S. Department of Labor, Office of Foreign Labor Certification Disclosure Data, https://www.foreignlaborcert.doleta.gov/performancedata.cfm

[2] GitHub, stat-8416-final-project, Brian Detweiler, https://github.com/bdetweiler/stat-8416-final-project

[3] Bureau of Labor Statistics http://data.bls.gov/timeseries/CES0000000001?output_view=net_1mth