Northeastern University

College of Computer and Information Science

DA-5020 : Collecting Storing and Retreiving.

Spring 2017

Exploratory Data Analysis of H-1B Visa Applications.

Sumedh R. Sankhe

NUID : 001799940

emai-id : sankhe.s@husky.neu.edu

Submission Date : 04-23-2017

Contents

Abstract

The H-1B is a non-immigrant visa in the United States under the Immigration and Nationality Act, section 101(a)(17)(H). It allows U.S. employers to temporarily employ foreign workers in specialty occupations. The regulations define a "specialty occupation" as requiring theoretical and practical application of a body of highly specialized knowledge in a field of human endeavor including but not limited to biotechnology, chemistry, architecture, engineering, mathematics, physical sciences, social sciences, medicine and health, education, law, accounting, business specialties, theology, and the arts, and requiring the attainment of a bachelor's degree or its equivalent as a minimum (with the exception of fashion models, who must be "of distinguished merit and ability"). Likewise, the foreign worker must possess at least a bachelor's degree or its equivalent and state licensure, if required to practice in that field. H-1B work-authorization is strictly limited to employment by the sponsoring employer.

The current law limits to **65,000** the number of foreign nationals who may be issued a visa or otherwise provided H-1B status each fiscal year (FY). Laws exempt up to **20,000** foreign nationals holding a **master's or higher degree from U.S. universities** from the cap on H-1B visas. In addition, excluded from the ceiling are all H-1B non-immigrants who work at (but not necessarily for) universities, non-profit research facilities associated with universities, and government research facilities. Universities can employ an unlimited number of foreign workers as cap-exempt. This also means that contractors working at but not directly employed by the institutions may be exempt from the cap as well. **Free Trade Agreements** carve out 1,400 H-1B1 visas for Chilean nationals and 5,400 H-1B1 visas for Singapore nationals. However, if these reserved visas are not used, then they are made available in the next fiscal year to applicants from other countries. Due to these **unlimited exemptions** and roll-overs, the number of H-1B visas issued each year is significantly more than the 65,000 cap, with 117,828 having been issued in FY2010, 129,552 in FY2011, and 135,991 in FY2012

Data Introduction

The H-1B is an employment-based, non-immigrant visa category for temporary foreign workers in the United States. For a foreign national to apply for H1-B visa, an US employer must offer a job and petition for H-1B visa with the US immigration department. This is the most common visa status applied for and held by international students once they complete college/ higher education (Bachelors, Masters, PhD) and work in a full-time position.

The Office of Foreign Labor Certification (OFLC) generates program data that is useful information about the immigration programs including the H1-B visa. The disclosure data updated annually, the H-1B application data can be found under the LCA Program at the link below.

United States Department of Labor OLFC Performance Data

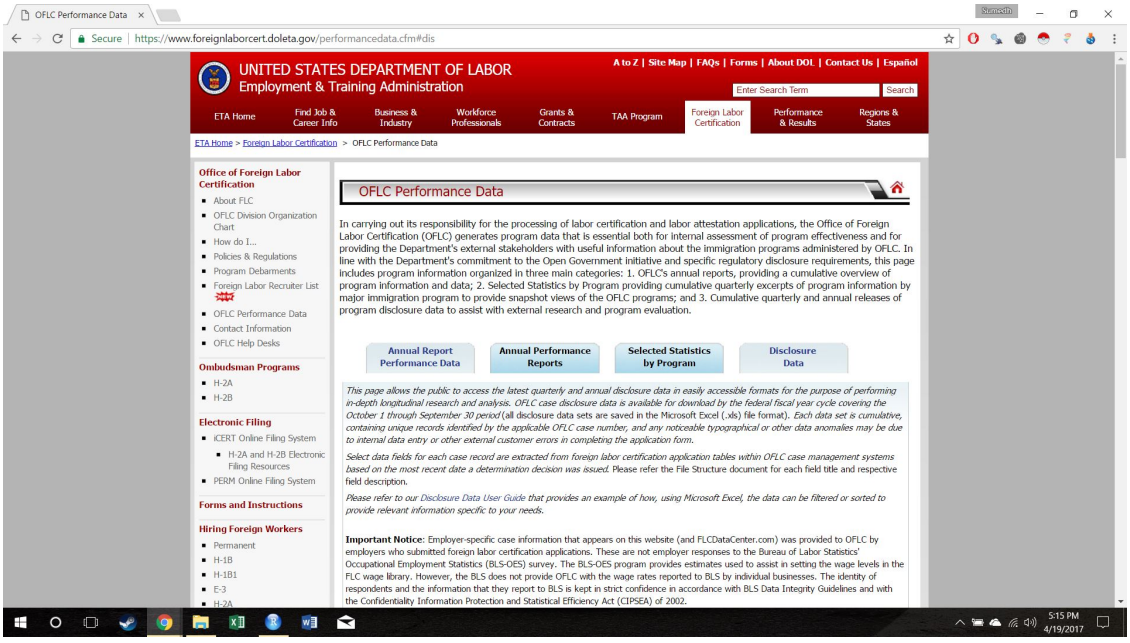


Figure 1: United States Department of Labor Screenshot

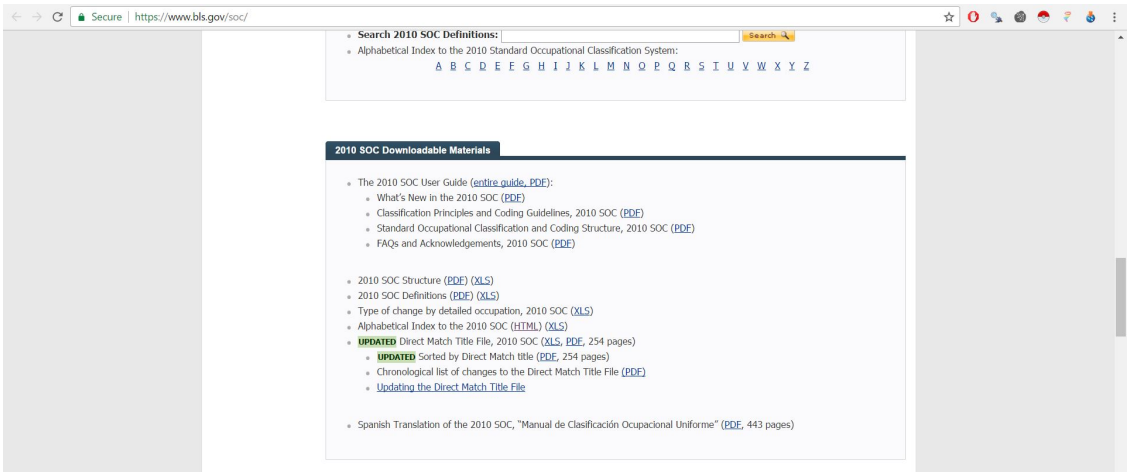


Figure 2: Bureau of Labor Statistics

Dataset Description

Firstly, I shall describes the key columns of this data-set that I have considered in this project. The data-set has 35-40 columns depending on the year of the data. Since the column names have not been the same over the years, my first step was to rename the columns to match the most common and newest specification among all the data. To identify the different years the data belonged to a column called Year is added to the data-sets depicting the year.

CASE_NUMBER : Unique identifier assigned to each application submitted for processing to the Chicago National Processing Center.

CASE_STATUS : Status associated with the last significant event or decision. Valid values include "Certified," "Certified-Withdrawn," Denied," and "Withdrawn". These statuses do not mean that a VISA was granted or denied, they just specify whether a particular application is eligible for VISA application or not.

EMPLOYER_NAME : Name of employer submitting labor condition application(LCA) i.e. H-1B Application.

EMPLOYER_ADDRESS / EMPLOYER_CITY / EMPLOYER_STATE : Contact information of the Employer requesting temporary labor certification.

CASE_SUBMITTED / DECISION_DATE : The dates on which a particular H-1B application was recevied by USCIS and the date on which a decision was given on an application.

JOB_TITLE : Title of the job using which we can filter specific job positions for e.g., Data Scientist, Data Engineer, Industrial Engineer, Process Engineer etc.

SOC_CODE / SOC_NAME : Occupational code associated with the job being requested for temporary labor condition, as classified by the Standard Occupational Classification (SOC) System. SOC_NAME the name associated with SOC_CODE.

NAICS_CODE : Industry code associated with the employer requesting permanent labor condition, as classified by the North American Industrial Classification System (NAICS).

WORKSITE_CITY / WORKSITE_STATE : The foreign worker's intended area of employment. We will explore the relationship between prevailing wage for Data Scientist position across different locations.

PREVAILING_WAGE : 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. This column will be one of the key metrics of the data analysis.

Problems Faced

The data-set showed up the following problems during the exploration.
i) Missing value / NA Values.
ii) Spelling Mistakes.
iii) Human Data Entry Error.

Missing Values : The missing values and NA values are prevalent in a huge number in the data-set. After reducing the number of rows to 25 the total number of missing values did reduce, but the number of missing values where to large to ignore or omit out of the data-set. Hence the following approach was considered taking into view the kind of analysis that I was going to perform on the data. The physical locations i.e. the STATE and CITY columns where filled with "UNKNOWN". If a postal code was found to be unknown the observation was omitted. A major issue was the for the data that was available for the year 2016. 100% of the data for the columns of FULL_TIME_POSITION was unknown "NA", hence a median of prevailing wages was considered and the observations that fell on or above the median salary where considered to be full-time positions. Other missing value where not taken into consideration as they were not to be part of the data analysis down the line. At any give point thought there are around 1.2 million missing value spread across the entire data-set, hence simply omitting or ignoring them will not give us the complete picture.

```
sum(is.na(H1b_df_new))
```

```
## [1] 1266963
```

9422	NEW YORK++	1
9421	NEW YORK,NY USA	2
9420	NEW YORK,NEW YORK	3
9419	NEW YORK, NY 10001	2
9418	NEW YORK, NY -	1
9417	NEW YORK, NY	6
9416	NEW YORK, NEW YORK	22
9415	NEW YORK, 10003	1
9414	NEW YORK,	142
9413	NEW YORK NY	1
9412	NEW YORK MILLS	9
9411	NEW YORK DOWNTOWN	1
9410	NEW YORK CITY,	10
9409	NEW YORK CITY	1472
9408	NEW YORK ,	2
9407	NEW YORK CITY	1
9406	NEW YORK	141452
9405	NEW YORJK	1
9404	NEW YOR CITY	1
9403	NEW YOORK	2
9402	NEW YOKR	2
9401	NEW YOK	6
9400	NEW YOEK	1

Figure 3: "Errors in Spellings"

Spelling Mistakes : Spelling mistakes are very rampant across this data-set, having a detailed look at the data gives you a fair idea as to the amount of spelling mistakes that are there to name a few, *New York* was misspelled *New Yrok, New Yok*, *San Francisco* misspelled *San Fransisco* and *Sunnyvale* misspelled *Suunyvale*. The above spelling mistake were just computed for the Worksite Cities. Upon closer observation it was found a lot of Employer names, addresses are misspelled. A package from R called "HUNSPELL" checks and corrects the different words in the data-set. Although the Library as such works perfectly fine it does not complete the job of correcting all the words in the data-set. Further searching and browsing through

stack-exchange and kaggle. I came up on a **Spell Checker** that was made by *Peter Norvig Director of Research at Google* although it is written in Python it provides us with a fair idea about correcting spelling by implementing a probabilistic method. After reading through the .py code and the related articles I decided to make a probabilistic spell checker in R. To describe briefly, this spell corrector finds out every possible transformation to a given word by 1-edit distance including deleting a letter, interchanging of two adjacent letters, inserting a new letter, replacing a letter with another letter from the English dictionary. The spell checker is current work in progress and is not included in the code for this report.

Human Error : The data-set provides classic example of human errors in data entry some can even be attributed to improperly filled applications by the applicant. Some of the example are as listed below, A few organizations have their name noted in varying pattern, 1010data for instance has been recorded as 10100DataServices 7 times, The Column Worksite_City has the address of the worksite instead of the city name in over 45 observations. Most of the employer names are easily fixed by using regular expressions and removing ill-used punctuation marks, and using substitutions for known abbreviations. Wrong entries in the columns have been disregarded and removed from the final data-frame. After basic transformation of the data to match case / remove punctuation / remove white spaces or tabs in the fields. Close to 80% of the errors were eliminated. The errors that still existed in the data-set were kept in the data-frame and dealt with while creating the database system for storing the data.

	WORKSITE_CITY	WORKSITE_STATE
1	<U+FFFD>ATLANTA	GA
2	<U+FFFD>BOTHELL	WA
3	<U+FFFD>BROOKLYN	OH
4	<U+FFFD>BURBANK	CA
5	<U+FFFD>CHICAGO	IL
6	<U+FFFD>COLUMBUS	OH
7	<U+FFFD>DELAND	FL
8	<U+FFFD>EL SEGUNDO	CA
9	<U+FFFD>ELKRIDGE	MD
10	<U+FFFD>FORT<U+FFFD>WASHINGTON	PA
11	<U+FFFD>HARTSVILLE	SC
12	<U+FFFD>JERSEY CITY	NJ

Figure 4: “Parsing Errors”

Data Wrangling

To work on comparing the wages we needed to first verify if all the wages provided have the same units i.e. yearly/hourly. On inspecting the data-set it was found that around **6.9%** of the data is **hourly wage**, less than 1% in weekly and monthly wage rates, majority of the wages occur in yearly pattern. Hence a function was created in converting the prevailing wage to a standard yearly wage pattern. Hence clearing the doubt about the different types of wage patterns in the data. Again values with "NA" were ignored. It is a common belief that H-1B visa applications are only approved if a company offers you a full time position, contrary to belief it is surprising to find that 1.91% of the data-set were not full time applicants and around 27.9% of the applications had "NA" associated with them.

Table 1: Full Time Position Distribution

FULL_TIME_POSITION	count
N	44445
UNKNOWN	647764
Y	1628783

If you have a look at the year wise distribution, there has been a steady increase in the number of full time applications over the last 5 years the same trend can be seen in part time positions. If you look at the below, you may notice that for the year 2016, 100% of the values are NA we do not have any information for that year whether the applications are made for full time or part time positions. Hence there is a need to determine a cut-off based on the yearly salaries computed before.

Table 2: Full Time Position Distribution across Years

FULL_TIME_POSITION	Year	n
N	2013	13590
N	2014	14209
N	2015	15107
N	2017	1539
UNKNOWN	2013	1
UNKNOWN	2014	2
UNKNOWN	2016	647761
Y	2013	428303
Y	2014	505029
Y	2015	603465
Y	2017	91986

On comparing the median wages of the applications, we see that full time positions have higher median wages compared to part time ones, even the positions that are reported as NA have a higher median wage. Hence it is imperative that we create a baseline on the prevailing wages to make a more informed decision. Considering the median wages that are shown below we will consider $65,000 as a baseline cut-off to do our analysis.

Table 3: Median Wage across Positions

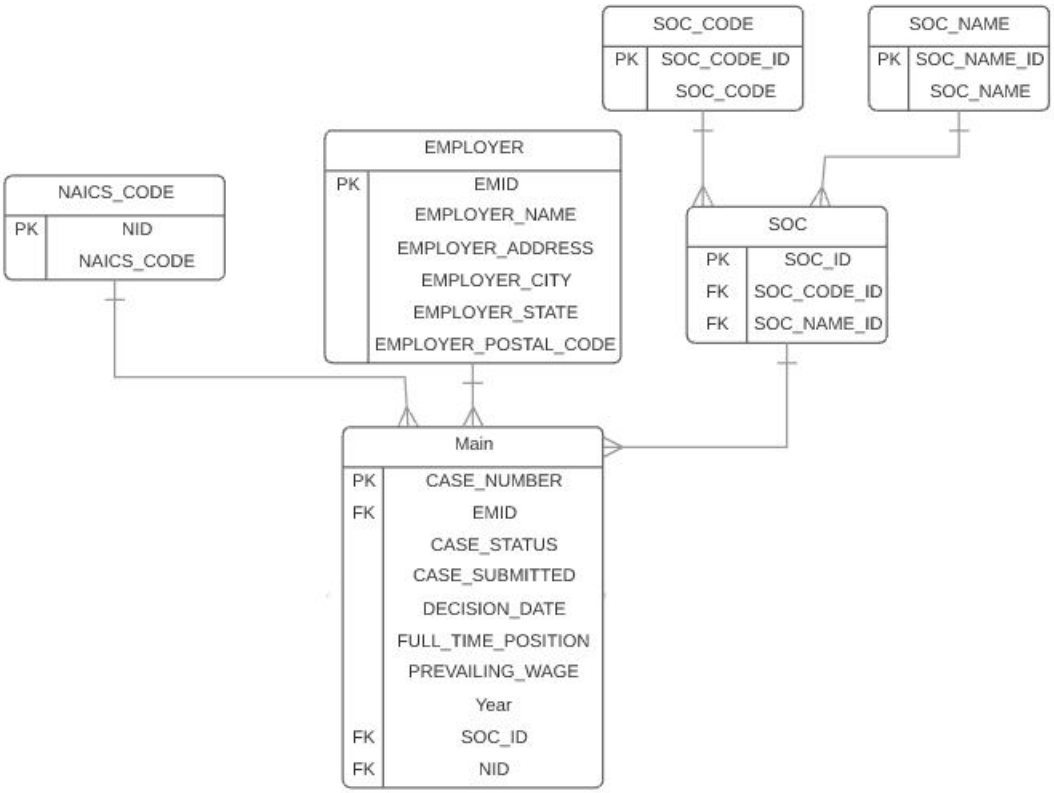
FULL_TIME_POSITION	Median Wage
N	55702.4
UNKNOWN	68245.0
Y	66082.0

Data Storage

The data provided to us is structured and the structure stays consistent over time barring a few column name that may change in years, or addition of some columns in the more recent version. The basic structure stays the same with fixed data-types of characters, integers and dates. Hence a relational database system is the best fit for my need. **SQL-Lite** was the choice of database used and **dplyr** the tool used for extracting the data from the database.

The database schema for the H-1B applications has been a challenging task. All the columns in the data-set have inherent importance and cannot be neglected without concrete explanation. Some of the original table had columns which dictated the wage pay ranges "from - to" but a majority of the years had no data available for this set of columns hence these columns were removed from the database. For the past year the data about the Agent or Attorney for a company has been reported although the data is not regular it may provide some useful insight once more data is accumulated over the coming years.
The SOC codes and names are an important classification given by Standard Occupational Classification . The data present us a unique challenge where in quite a few "JOB_TITLES" are associated with a SOC code by their corresponding name varies according to the nature of the job, hence we have multiple names associated with a code and the same is true the other way around. Hence creating a intermediate table between the two to resolve the many to many relationships was implemented.
The NAICS codes or the North American Industrial Classification systems code classify a job according to the industry the job is based in, the department of labor provides us with a list for all the NAICS code along with NAICS Names associated with it. After comparing the the list provided by the government and the data that we have it is seen that there no overlap between the two data-sets, which seems pretty odd since both the data-sets are provided by the department of labor. After going over the Department of Labor's website and finding no other solution to the issue. Creating separate Table for the NAICS names was dropped and only the NAICS Codes were preserved.

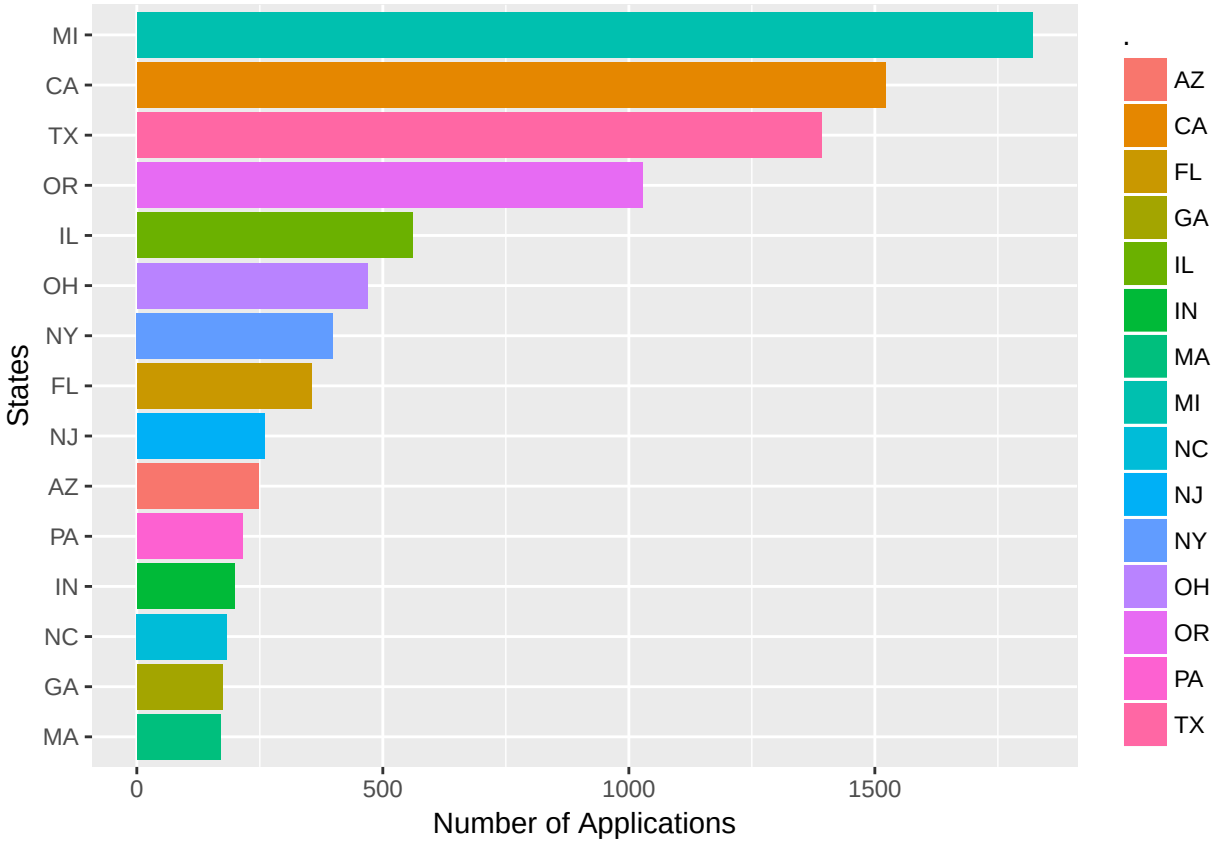
The Employer table has been the most difficult one to make and bring to the third normal form, On account of the number spelling mistakes and the multiple variations in the street addresses used to depict a single location, making the observations unique in this particular table had been challenging, various imputations were tried trying to subdivide the locations by city/state/postal code. But a solution that was unique was not particularly found. Instead of breaking them down to different tables, a single table with a composite key was created with all the variable taking part in the formation of the key and hence making it unique and workable.



The **"Main"** table is the main working table of the database, that is the table from which we shall pull all the data required for our analysis using joins. The data that is pulled from this table using dplyr will be collected into data-frames and filtered according to the requirement of the analysis.
This database is designed to give us the result quickly without having to go through multiple joins, a future scope for the database is to use the date attributes to store applications received on a particular day, or get the number of decisions that are given out by the USCIS with regards to the applications.

Analysis

For the purpose of analysis I shall be using the Job titles associated with the Industrial Engineering Field and come up with graphs and observations from the database that I have created using SQL-Lite and using dplyr for retrieving and ggplot2 and graphics for plotting the observations.



From the plot above it is observed that in the last 5 year period, the state of Michigan has the maximum number of applications for list of job titles selected, close to 1540 applications, while the state of California comes second. I have only considered the top 15 fifteen states with the maximum number of applications for this plot, **Massachusetts ranks 15th in this**.

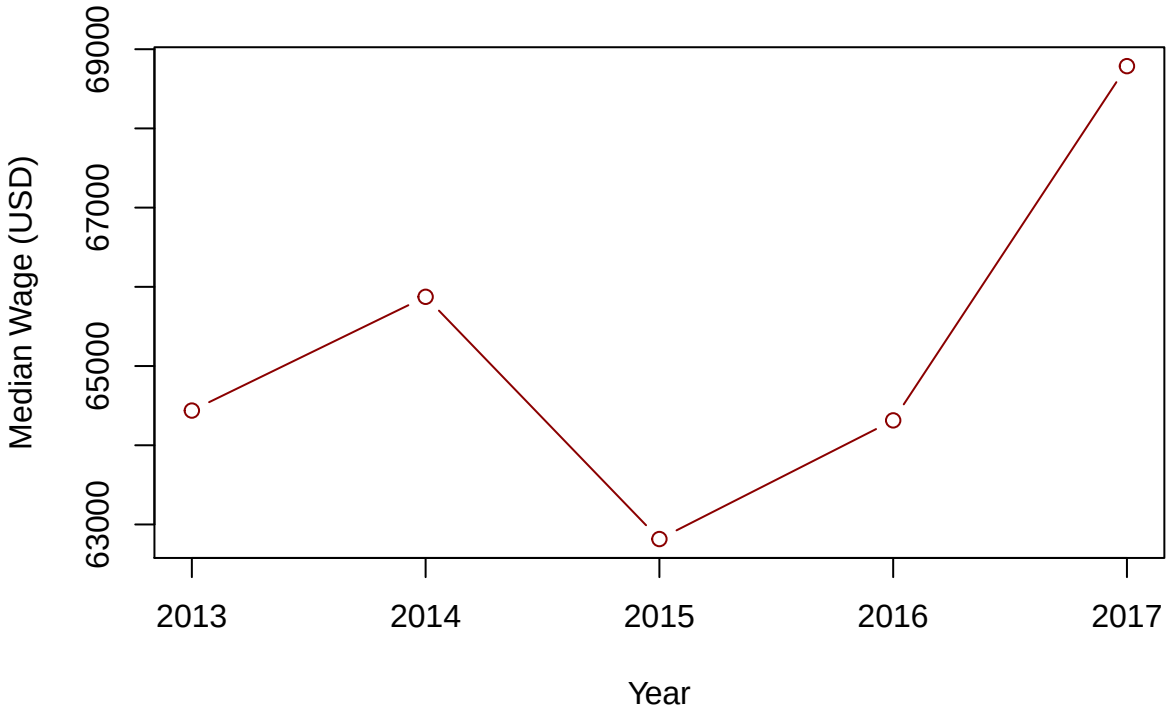
The following table gives us a broad pictures as to which Employers had the maximum successful applications in the last 5 year period. It is curious to note that Intel Corporation had the maximum number. Apple and Amazon are the go to employers for the selected job titles.

Table 4: Number of Successfull Applications per Employer

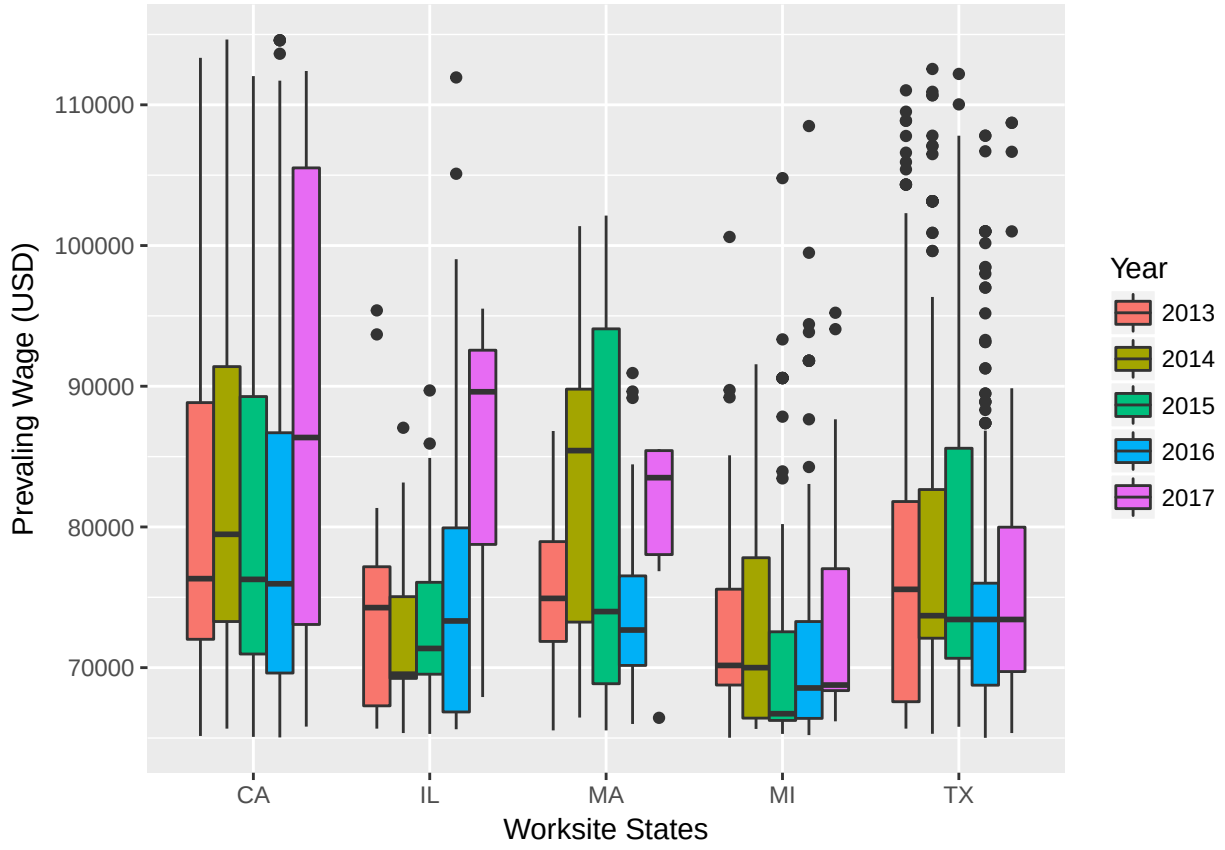
Employer Name	Freq
INTEL CORPORATION	1119
LAM RESEARCH CORPORATION	136
GEOMETRIC AMERICAS INC	82
APPLIED MATERIALS INC	68
ENGINEERING TECHNOLOGY ASSOCIATES INC	65
SCHLUMBERGER TECHNOLOGY CORPORATION	65
TESLA MOTORS INC	58
OPTIMAL CAE INC	56

Employer Name	Freq
FEV NORTH AMERICA INC	43
RACAR INTERNATIONAL DE LLC	29
TECHNIP USA INC	28
DETROIT ENGINEERED PRODUCTS INC	25
CUMMINS INC	24
DPR CONSTRUCTION A GENERAL PARTNERSHIP	23
FORMOSA PLASTICS CORPORATION TEXAS	23

Looking at the trend of the wages for the past 5 years, one can see from the graph below that there has been no major change till 2016, but for the current data for the first quarter of 2017 there is a sharp spike in the prevailing wage, this could be an anomaly since we just have 3 months worth data for 2017, having a look at this graph at the end of the might give us a true picture of the prevalent wages.



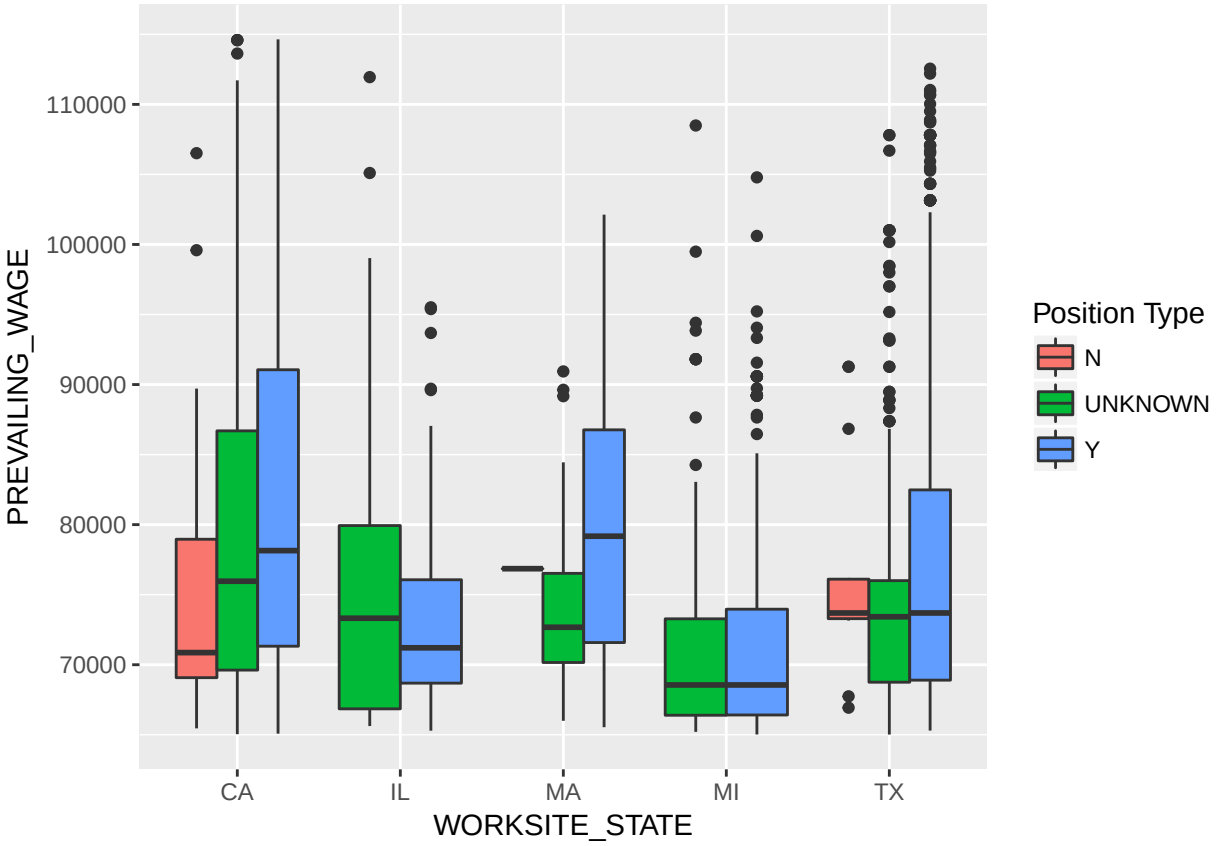
The following plot compares the prevailing wages of the 5 states of Michigan, California, Texas, Illinois and Massachusetts for 5 years starting 2013 to 2017.



Thus we see that the state of California consistently out-performs the other states in terms of wage, although we see Massachusetts wages on par for the year 2015 but all and out California tops the charts and with Michigan and Illinois coming in last.

We can see that majority of the positions that belong to this current wage bracket of $65000 to $114650 fall under the full time position type category in the states that we have selected, why did we choose such a filter for the wages? the states of California and Massachusetts had a very abnormally high wage for 2-3 observations which skewed the plots and made them un-readable hence this bracket was chosen for the prevailing wage.

In the plot below we see the distribution of the types of poistions that are prevalent in the 5 states chosen. We see there is an abnormally large number of positions that can be seen as "UNKNOWN" which can be attributed to the fact that in the Year 2016, 100% of the data reported the full time position to have a **"NA"** value. Otherwise we see maximum number of part-time positions are offered in California followed by Texas whereas Illinois, Massachusetts and Michigan have the least offerings for part-time positions. Whereas full-time positions are offered majorly in California and Massachusetts.



Conclusion

Thus from the observations above we can infer that, the state of Michigan has the maximum number of applications. Although Michigan has the maximum number of applications the wage in California have been consistently greater than the other states, this might be incorporated due to the higher cost of living in the state of California.
The number of full-time and part-time position are the highest in the state of California. The prevailing wages for the field suggest there has been no major improvement in the wages in the past 4 years, for the current year "2017" there has been a spike in the prevailing wages but this could be due to the partial data that we have for the year 2017.

Future Scope

1) Develop a spell checker that would probabilistically give out a list of correct words to choose from.
2) Create a R-shiny application to make the dataset interactive so that different plot and other relevant information for multiple filters can be aesthetically visualized.
3) Collaborate with a few other project from the DA5020 course to justify the differing prevailing wages. Namely the Crime Data Analysis and the University Ranks list as well as Apartment Rental analysis.
4) The ultimate achievement beyond this project would be to create a predictor model based on the data, that help us correctly predict if an application can be certified or not, Although a few more important data-points might be required for that.

References

1) Stack Overflow Question Answers : stackoverflow.com
2) Bureau of Labor Statistics : bls.gov
3) Bureau of Labor Statistics : olfcperformancedata
4) Norvig Spell Checker : norvig.com/spell-correct.html
5) R Cheat sheets for ggplot2 / R Markdown / dplyr.
6) R for Data Science - Garrett Grolemund /Hadley Wickham : R for Data Science