Final Exam

Ayush Kumar Shah

2020-10-08

Contents

Introduction

In this report, the `h1b` dataset will be analyzed to generate and communicate findings of different aspects of the data in an effective and presentable manner. The analysis of the `h1b` dataset will help to identify different factors affecting the H-1B Visa acceptance as well as interrelationships between the different variables present in the data.

About the H-1B data

This dataset contains administrative data from employers' Labor Condition Applications (ETA Forms 9035 & 9035E) and the certification determinations processed by the Department's Office of Foreign Labor Certification, Employment and Training Administration where the date of the determination was issued on or after October 1, 2016, and on or before June 30, 2017. All data were extracted from the Office of Foreign Labor Certification's iCERT Visa Portal System, an electronic filing and application processing system of employer requests for H-1B nonimmigrant workers.

The H-1B visa is a temporary or nonimmigrant "specialty occupation" U.S. visa, which means the holder is employed in a position that requires specialized skills or knowledge, for which the employer cannot find a US-based worker. It allows a foreign worker to go to the United States and work for an American company. Hence, it is an important source of information from which useful insights can be formed.

The `h1b` data in this report includes detailed information about the application like submission date, case status, and employer's information like name, address, state, proposed wage, H-1B dependence, and so on. It also includes information about the job name being requested for temporary labor conditions.

Data extract

Let us view an extract of the `h1b` data with only a few columns.

Table 1: Extract of H-1B Visa Data 2016/17 (continued below)

CASE_STATUS	EMPLOYER_NAME	EMPLOYER_STATE	SOC_NAME	WAGE_RATE_OF_PAY_FROM	H-1B_DEPENDENT	WILLFUL_VIOLATOR
CERTIFIEDWITHDRAWN	DISCOVER PRODUCTS INC	IL	ANALYSTS	65811	N	N
CERTIFIEDWITHDRAWN	DFS SERVICES LLC	IL	ANALYSTS	53000	N	N
CERTIFIEDWITHDRAWN	EASTBANC TECHNOLOGIES LLC	DC	ANALYSTS	77000	Y	N
WITHDRAWN	INFO SERVICES LLC	MI	COMPUTER OCCUPATION	102000	Y	N
CERTIFIEDWITHDRAWN	BBandT CORPORATION	NC	ANALYSTS	132500	N	N

Tidying the data

The first step in analyzing any data is converting it into a tidy form by removing any infelicities present in it so that analysis can be performed easily. The given data is very close to tidy, and we perform a few operations to make it perfect for further exploration and analysis. We perform various transformations to the data using the `dplyr` and `tidyverse` package to achieve this.

Checking and removing duplicates

The original data consists of 528134 rows.

We first compute the nonduplicate dataset using `distinct()` which returns only the unique observations from the original `H-1B Visa` data. We then compare the resulting dataset to the original dataset using `all_equal()`.

After removing 62945 duplicate rows, number of remaining observations = 465189

So, we use the nonduplicate dataset for further exploration.

Removing irrelevant columns and Filtering data

Columns with missing values

Let us view the counts of empty rows in each column. We only display the columns which have N.A. values.

Table 3: Columns with NA values

	NA.Count
NAICS_CODE	2
FULL_TIME_POSITION	3
WAGE_UNIT_OF_PAY	4
EMPLOYER_STATE	10
PW_SOURCE_YEAR	31
PW_UNIT_OF_PAY	33
PW_SOURCE	33
H-1B_DEPENDENT	10171
WILLFUL_VIOLATOR	10172

We remove the observations containing missing values corresponding to the variables `EMPLOYER_STATE`, `FULL_TIME_POSITION`.

We also remove the columns `PW_SOURCE_YEAR, NAICS_CODE, WORKSITE_STATE, DECISION_YEAR, DECISION_MONTH, DECISION_DAY, PW_SOURCE_OTHER, WAGE_RATE_OF_PAY_TO, WORKSITE_POSTAL_CODE` since they are not required. We use observations corresponding to Employers of the USA only and hence remove the column `EMPLOYER_COUNTRY` as well.

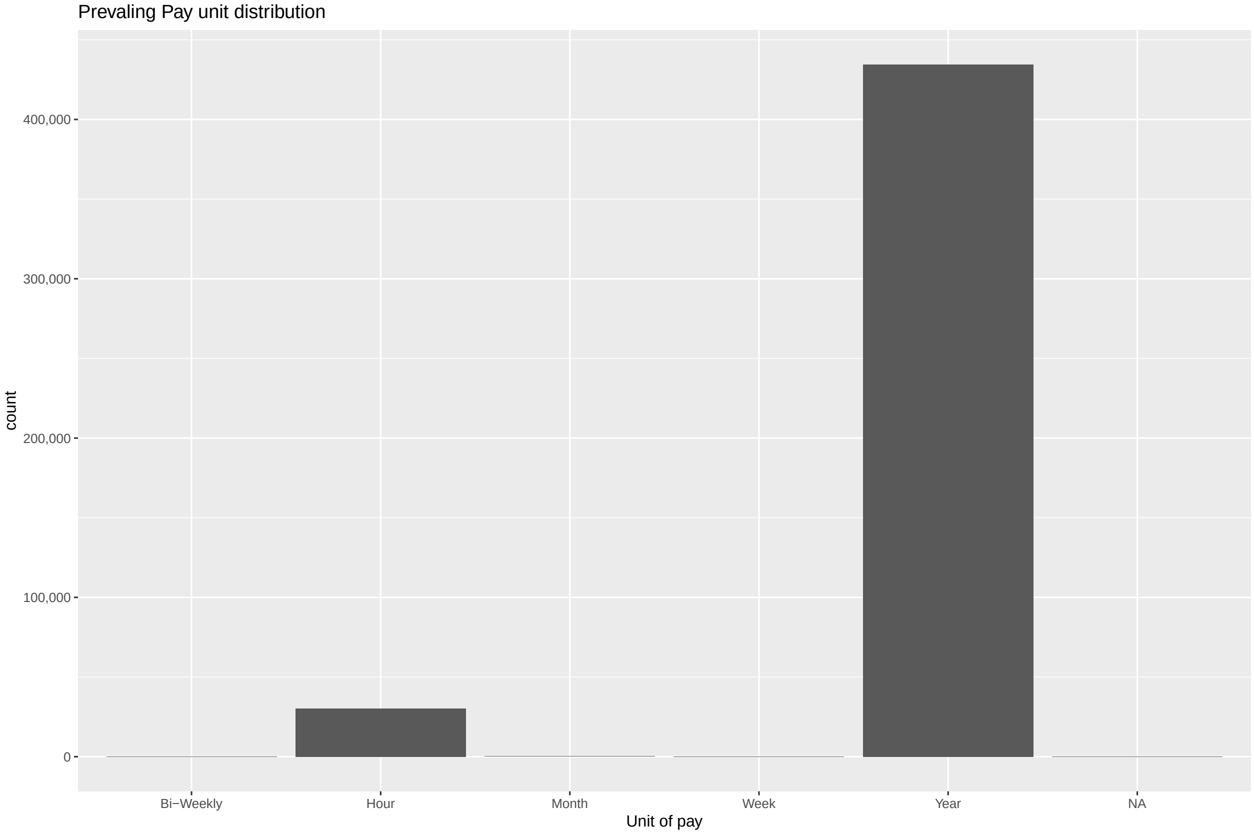
Let us take a look at the year distribution in which the applications were submitted.

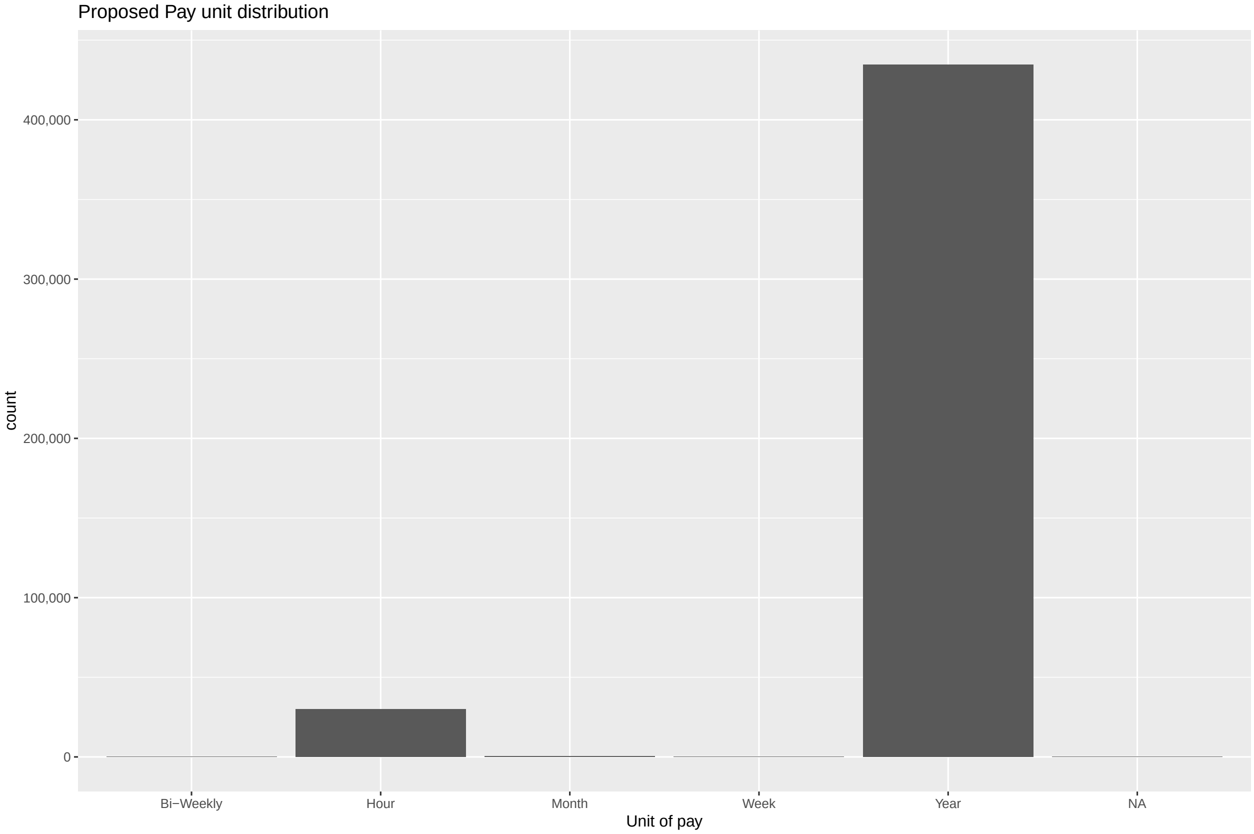
CASE_SUBMITTED_YEAR	count
2017	359510
2016	93962
2015	6910
2014	4516
2013	278
2012	11
2011	2

We can see that the data consists of negligible cases submitted in the years 2011 - 2013. So, we remove observations corresponding to those years.

Also, we remove the observations in which the number of workers is 0 or the wages are 0.

Likewise, we keep only the wage values which have Yearly units of pay otherwise, the analysis of wages would not be correct as it would involve mixed units. We can do this since observations with units of pay other than Year are very low.





After tidying the data, number of observations is reduced from 465189 to 433765 and 12 columns have been removed.

Findings

We now explore the data by performing various kinds of transformations and visualizations on the tidied data we have produced. The main objective of the exploration is to find useful insights and findings from the data so that we can answer various questions that we have about the data as well as explore the interesting patterns in the data.

To begin the explorations, we must have a set of questions initially to answer them through the visualizations. A few of the questions might be:

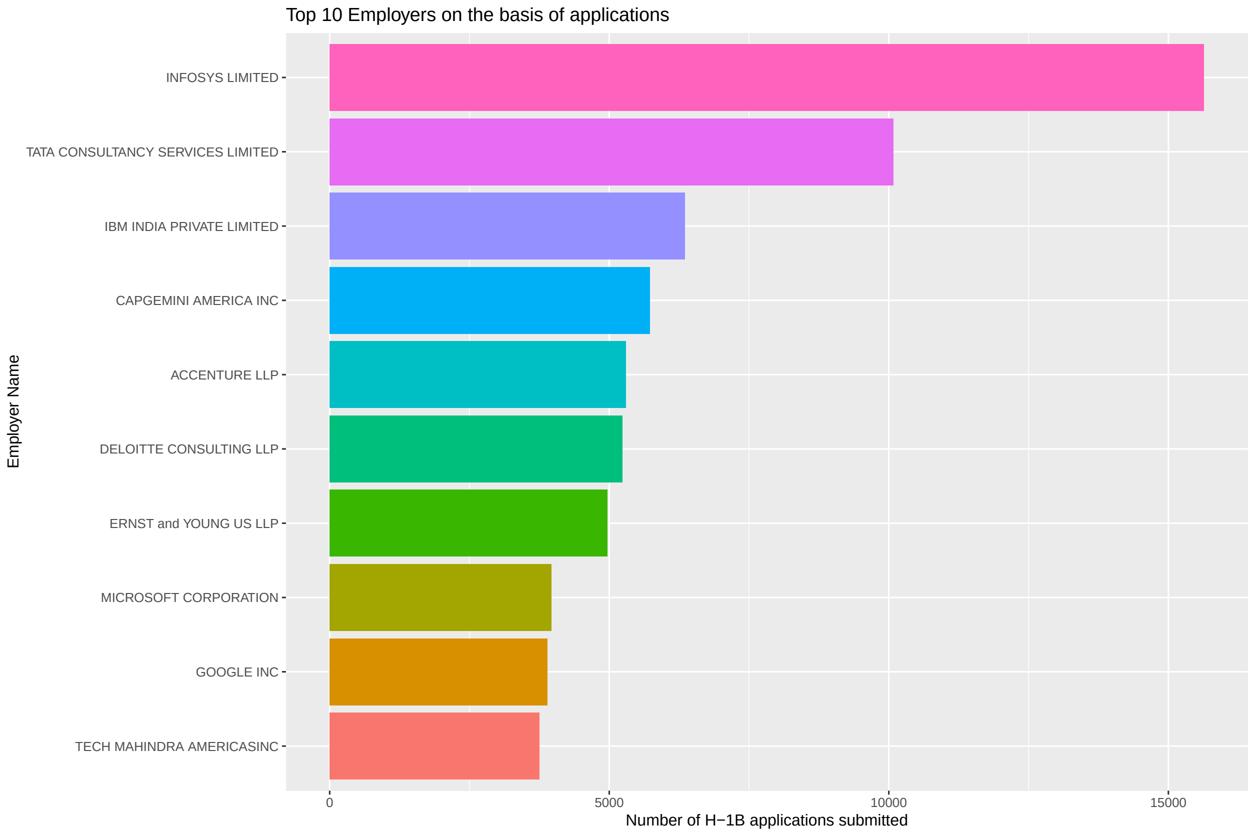
1. What are the top 10 companies submitting the H-1B VISA applications?
2. What are the top 10 companies in terms of proposed wages?
3. How does the monthly trend in submissions of H-1B VISA Applications look like?
4. How does the monthly trend in submissions of H-1B VISA Applications in the top 10 states look like?
5. What is the distribution of case status and the variation with full-time positions and employers' H-1B dependence?
6. How are the Wages and number of workers distributed?
7. How are the Prevailing Wages distributed to its source?
8. Is there a linear relationship between the prevailing wage and the proposed wage?
9. How is the proposed wage distributed among popular jobs?
10. Is there a dependence between the Case status and submission month?
11. How does the Case status co-vary with Job count?
12. How does the monthly trend in popular jobs look like?
13. How are the states and the case status related?

We will follow Wickham's layered grammar of graphics approach using `ggplot2` from CRAN to perform the explorations and create visualizations to answer these questions. The components of the layered grammar allow us to completely and explicitly describe a wide range of graphics in order to explore the data effectively.

1. What are the top 10 companies submitting the H-1B VISA applications?

Table 5: Top 10 Employers on the basis of applications

EMPLOYER_NAME	APPLICATIONS_COUNT	Proportion in %
INFOSYS LIMITED	15639	3.605
TATA CONSULTANCY SERVICES LIMITED	10084	2.325
IBM INDIA PRIVATE LIMITED	6352	1.464
CAPGEMINI AMERICA INC	5727	1.32
ACCENTURE LLP	5301	1.222
DELOITTE CONSULTING LLP	5239	1.208
ERNST and YOUNG US LLP	4967	1.145
MICROSOFT CORPORATION	3962	0.9134
GOOGLE INC	3898	0.8986
TECH MAHINDRA AMERICASINC	3748	0.8641



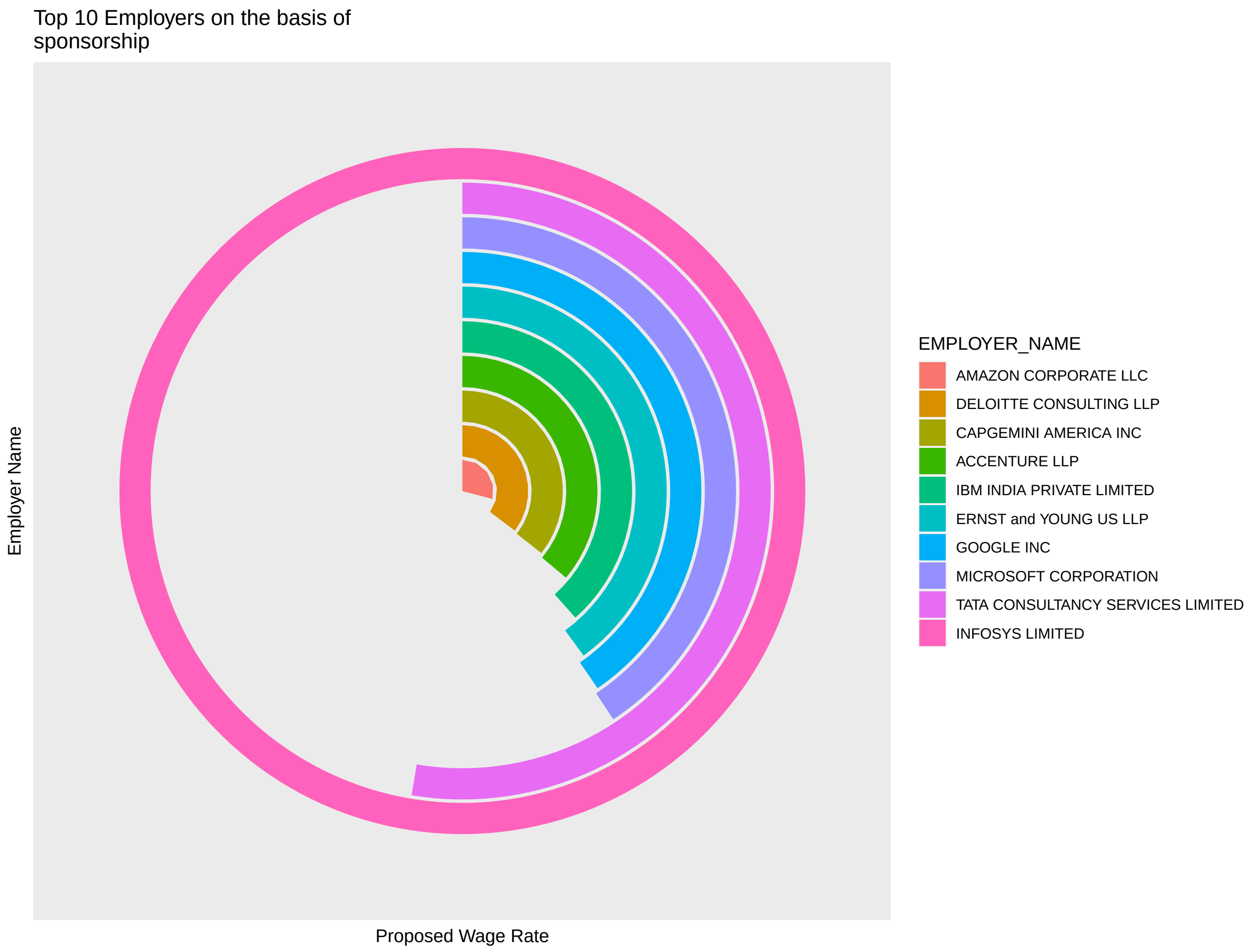
We can see that Infosys Limited has submitted the most number of H-1B VISA Applications, far more than Microsoft and Google.

2. What are the top 10 companies in terms of proposed wages?

We find out the top 10 employers based on the sum of the wages proposed by them in all of their H-1B visa applications.

Table 6: Top 10 Employers on the basis of sposnorship

EMPLOYER_NAME	PROPOSED_WAGE	Proportion in %
INFOSYS LIMITED	1.292e+09	3.342
TATA CONSULTANCY SERVICES LIMITED	679924654	1.759
MICROSOFT CORPORATION	525765741	1.36
GOOGLE INC	522274418	1.351
ERNST and YOUNG US LLP	515262536	1.333
IBM INDIA PRIVATE LIMITED	4.96e+08	1.283
ACCENTURE LLP	466105488	1.206
CAPGEMINI AMERICA INC	459789088	1.189
DELOITTE CONSULTING LLP	455915645	1.179
AMAZON CORPORATE LLC	376081980	0.9729

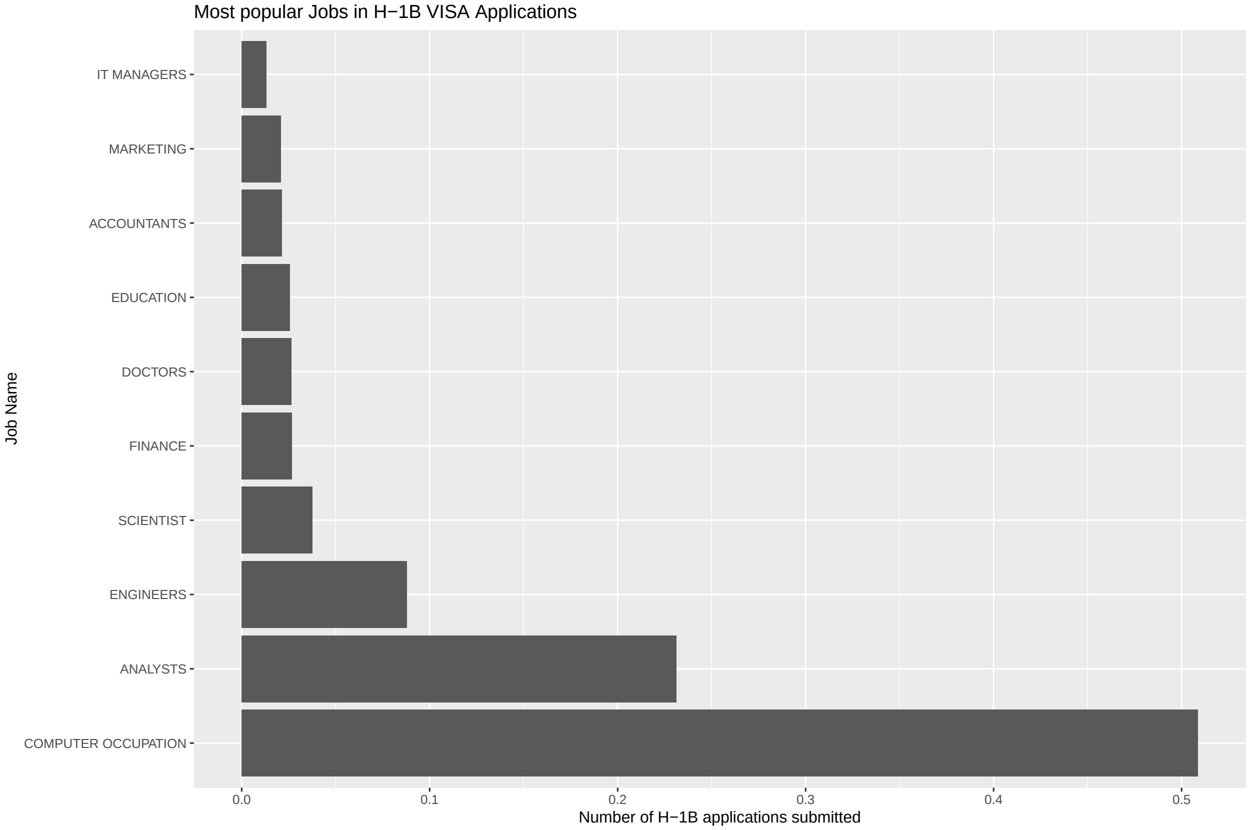


We can see that Infosys limited is the top employer in terms of sponsorship as well. Tata, Google, and Microsoft are just behind Infosys.

3. What are the most popular job titles in H-1B VISA Applications?

Table 7: Most Popular Jobs in Applications

SOC_NAME	APPLICATIONS_COUNT	Proportion in %
COMPUTER OCCUPATION	203151	46.83
ANALYSTS	92311	21.28
ENGINEERS	35136	8.1
SCIENTIST	15056	3.471
FINANCE	10669	2.46
DOCTORS	10611	2.446
EDUCATION	10197	2.351
ACCOUNTANTS	8522	1.965
MARKETING	8388	1.934
IT MANAGERS	5286	1.219

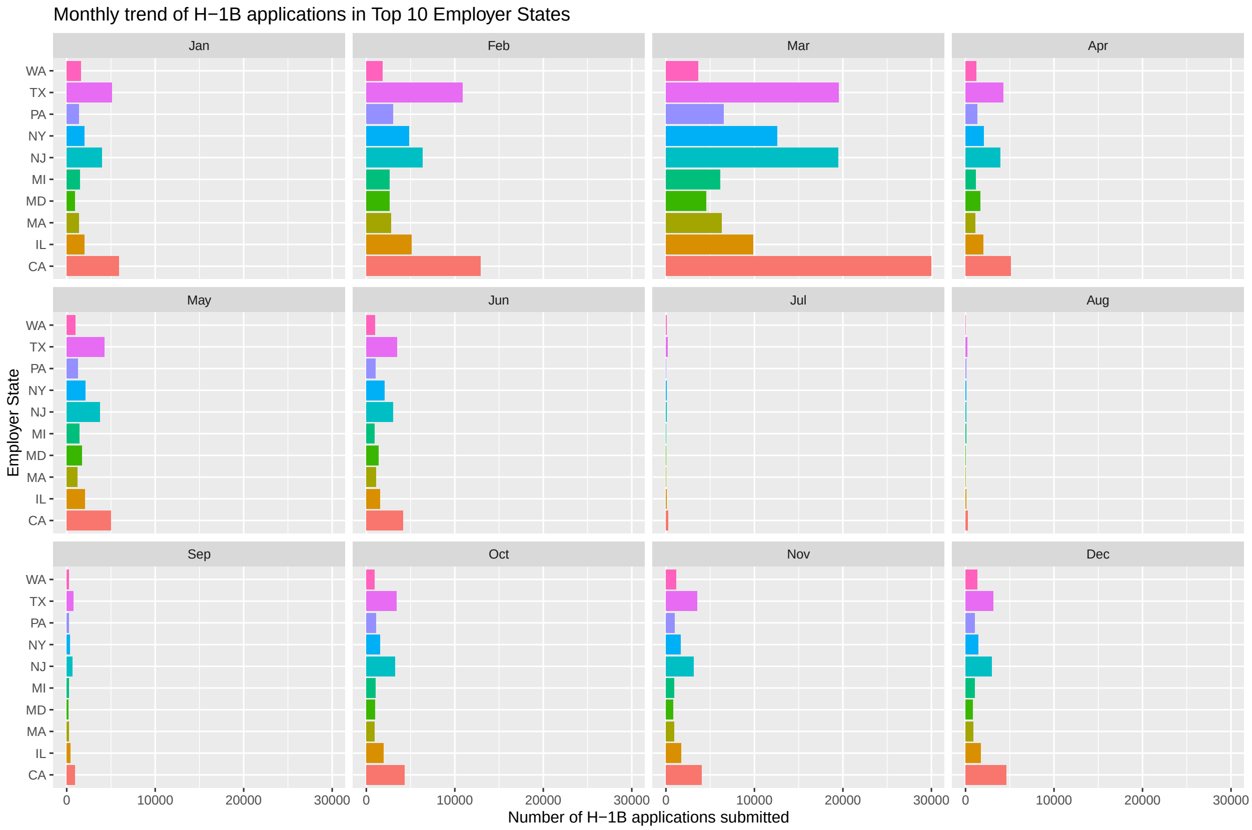


We can see that Computer Occupation occupies more than 50% of the job among the top 10 most popular jobs in VISA applications. This result is in accordance with our expectations since people going to the U.S. to work in the computer field have been increasing rapidly.

4. How does the monthly trend in submissions of H-1B VISA Applications in the top 10 states look like?

Table 8: Top 10 Employer States

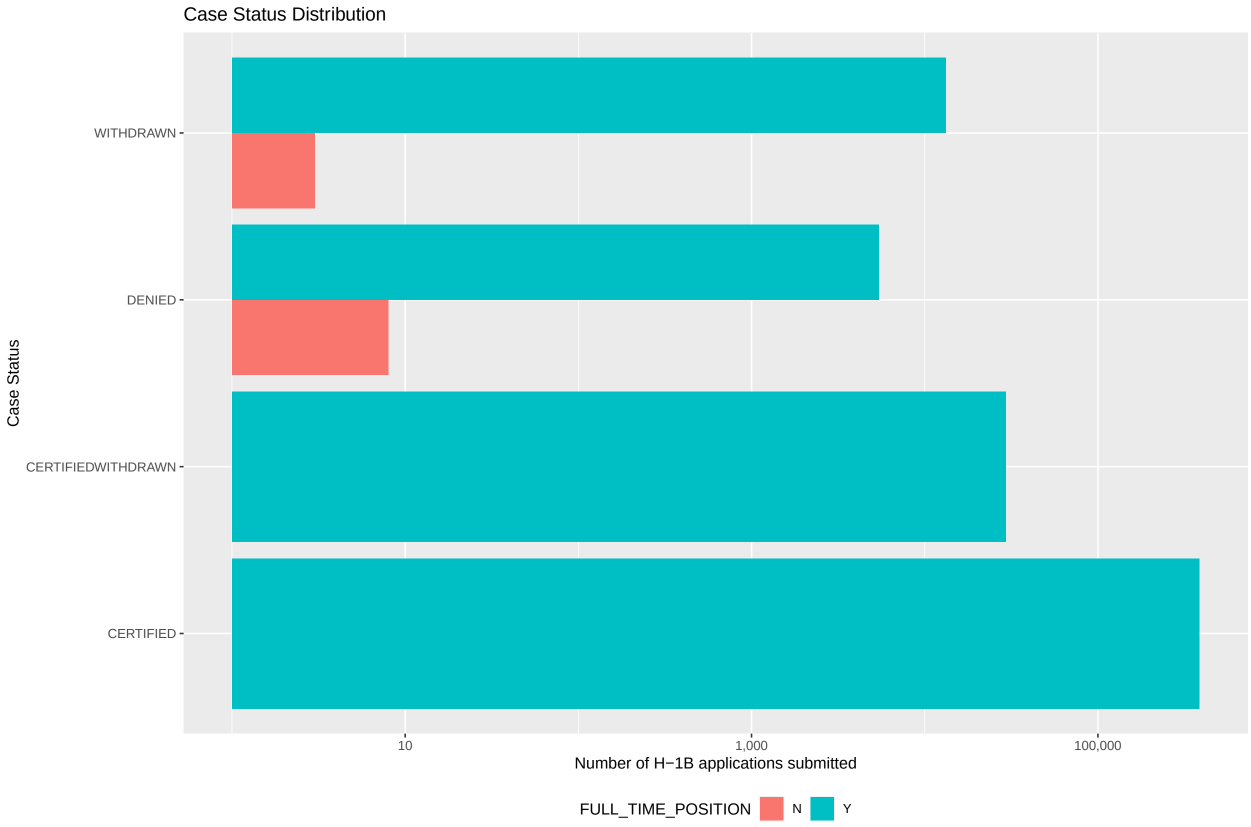
EMPLOYER_STATE	APPLICATIONS_COUNT	Proportion in %
CA	77422	17.85
TX	58872	13.57
NJ	50772	11.7
NY	30844	7.111
IL	28601	6.594
PA	18167	4.188
MI	17217	3.969
MA	16970	3.912
MD	15731	3.627
WA	14041	3.237

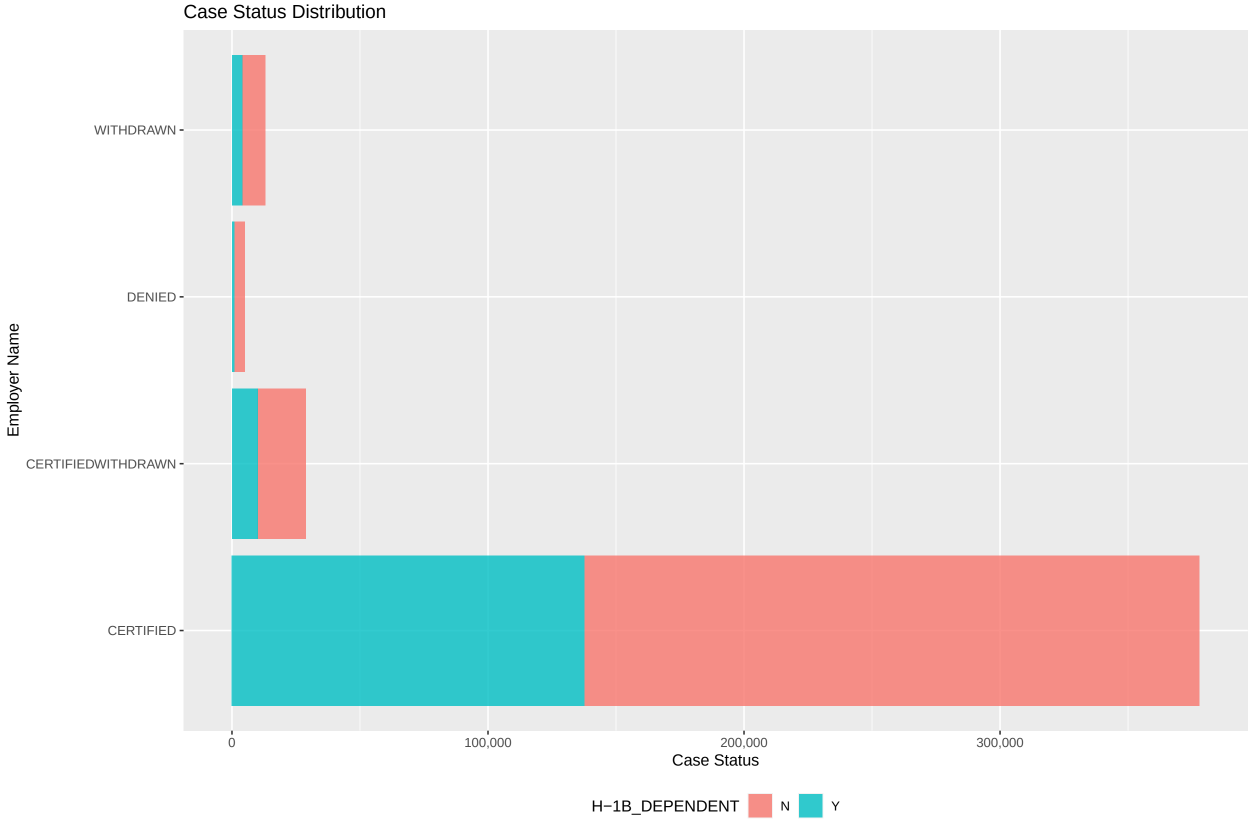


We can observe that H-1B applications submitted in July to September are extremely low compared to other months. Most applications are submitted in March. It might be the case because people are generally busy in those months due to fewer holidays.

Likewise, California has the highest number of application submissions.

5. What is the distribution of case status and the variation with full-time positions and H-1B dependence of employers?

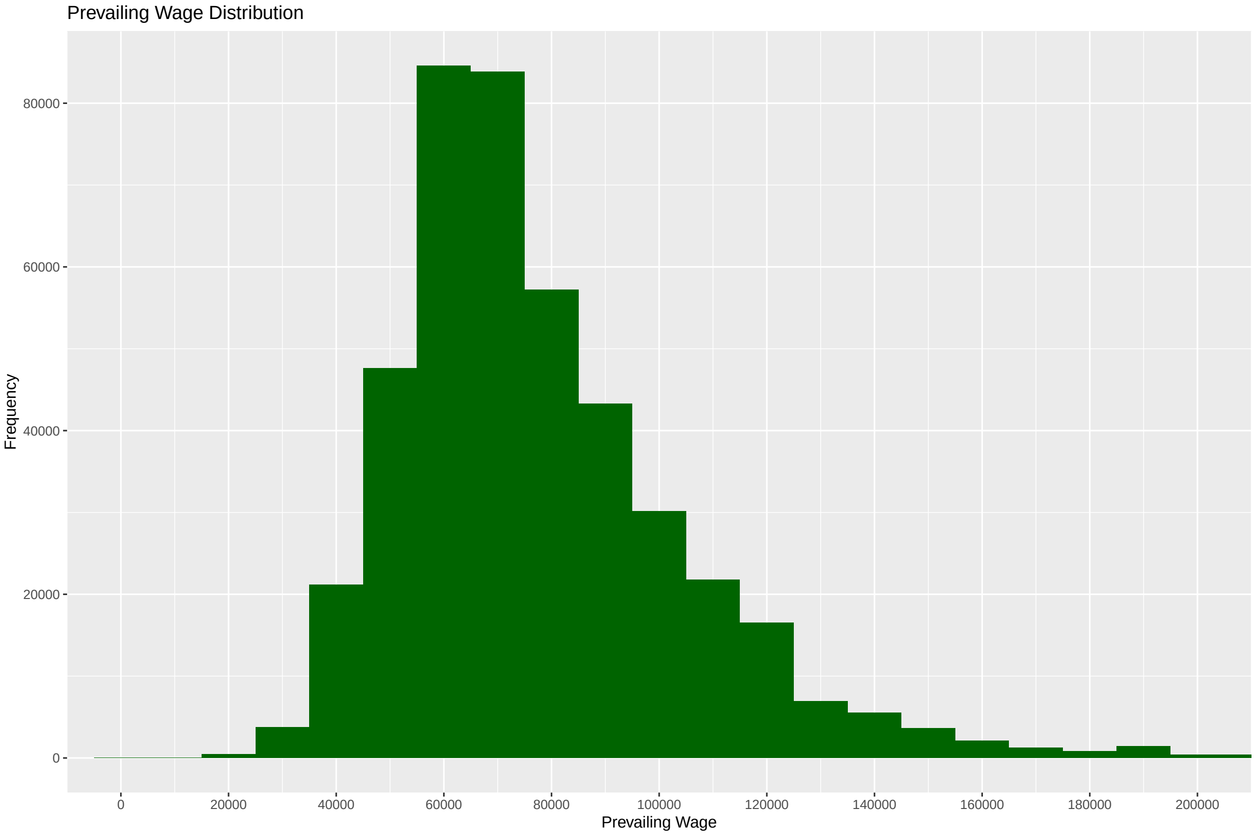




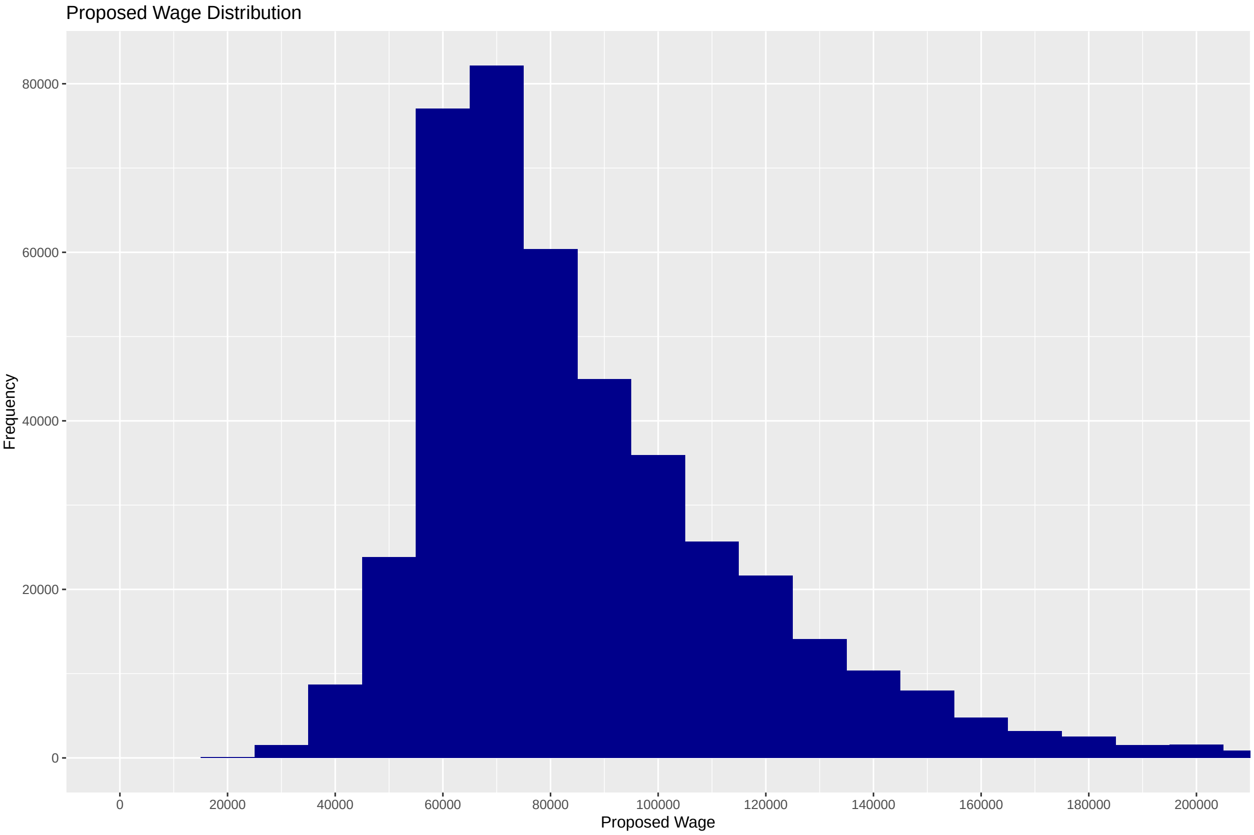
We can see that among the certified cases, a greater proportion of the employers was not declared as H-1B Dependent, which suggests that being declared as H-1B dependent might reduce certification chances.

6. How are the Wages and number of workers distributed?

Prevailing Wage Distribution

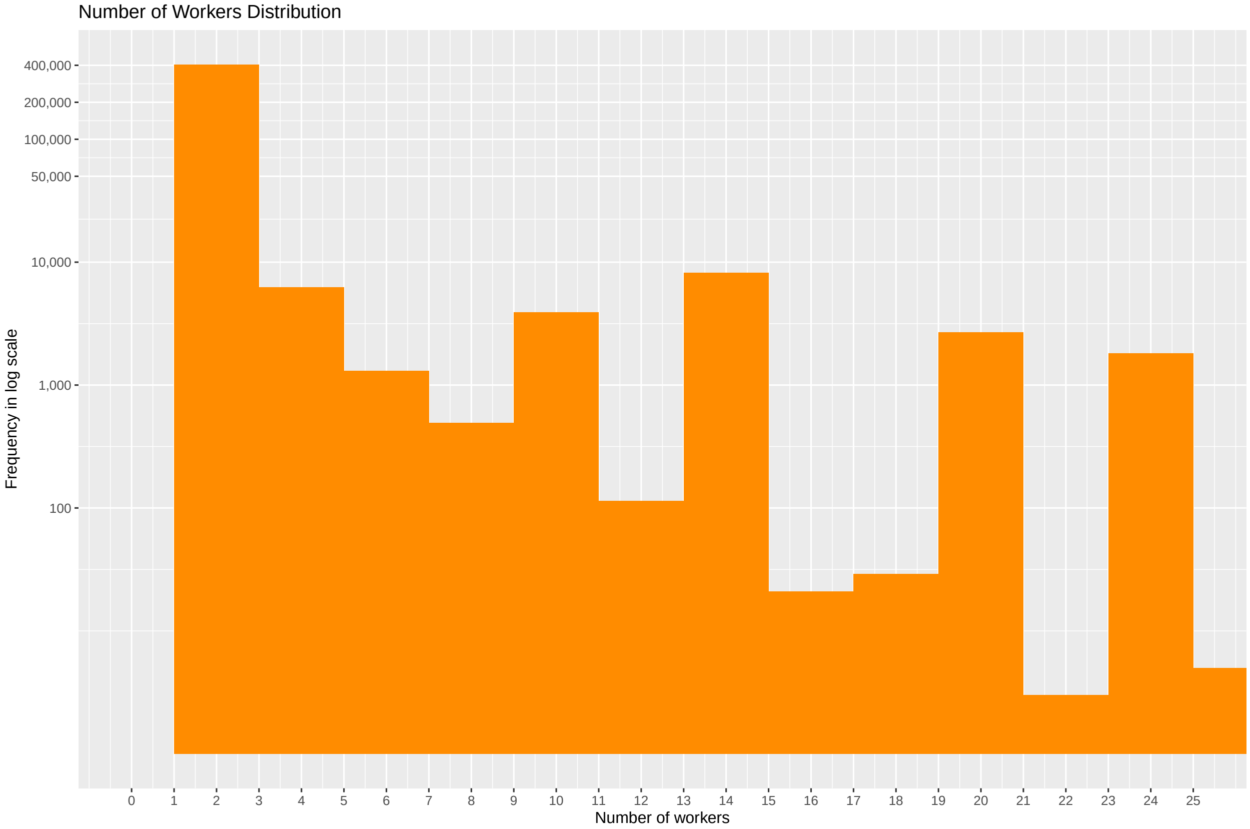


Proposed Wage Distribution



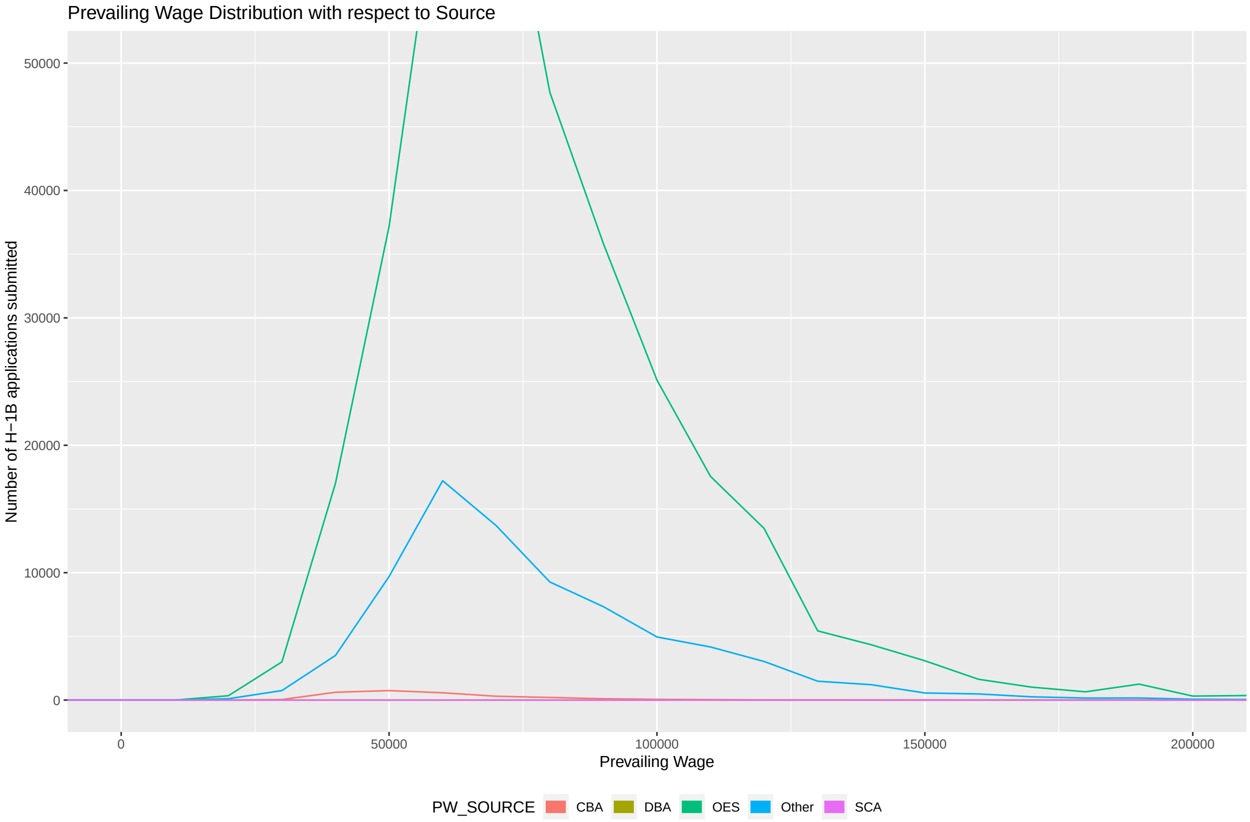
We can observe that both types of wages approximately follow the normal distribution, as suggested by the bell shaped curves. Both are centered around at the wage value of 60000 to 80000, corresponding to the average salary.

Number of Workers distribution



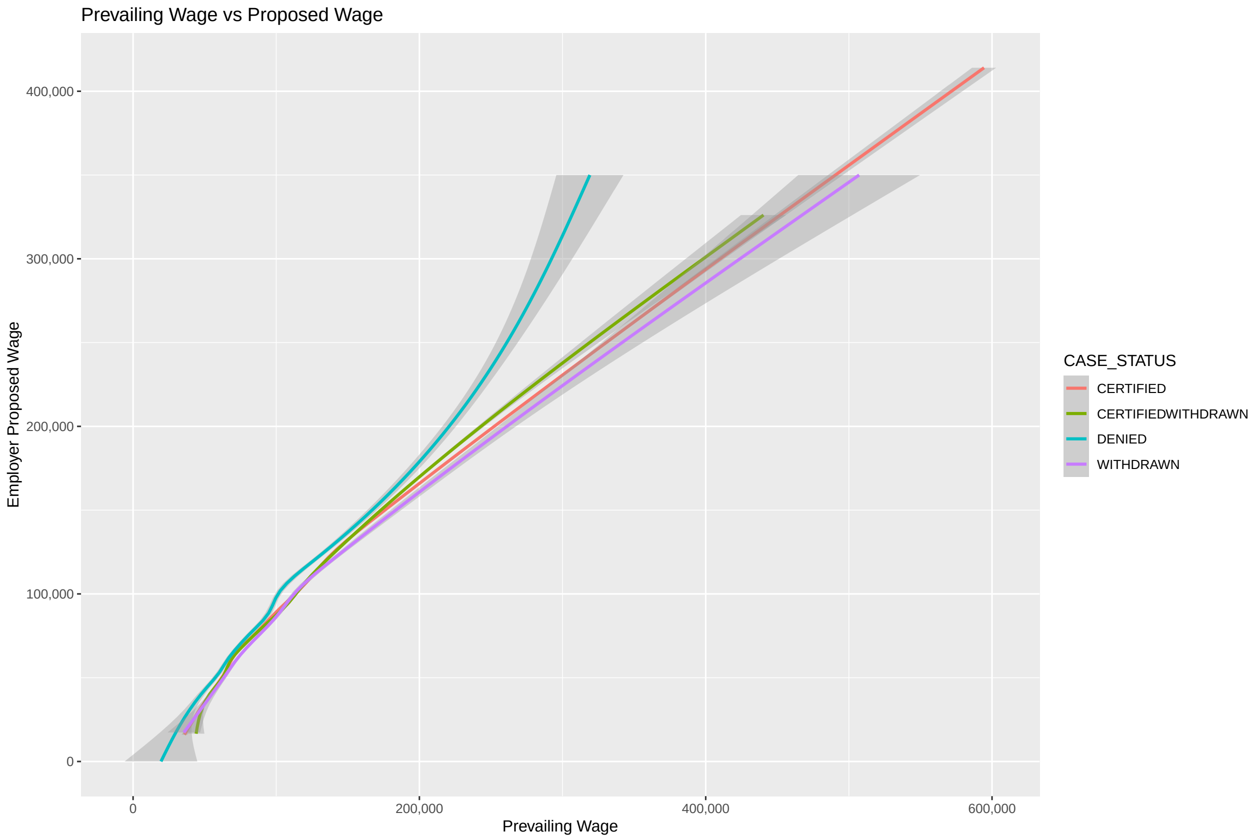
We can see that the distribution of the number of workers does not show any specific standard distribution. The frequency of applications decreases with the increase in the number of workers, which is natural as most companies request only a few workers.

7. How are the Prevailing Wages distributed with respect to its source?



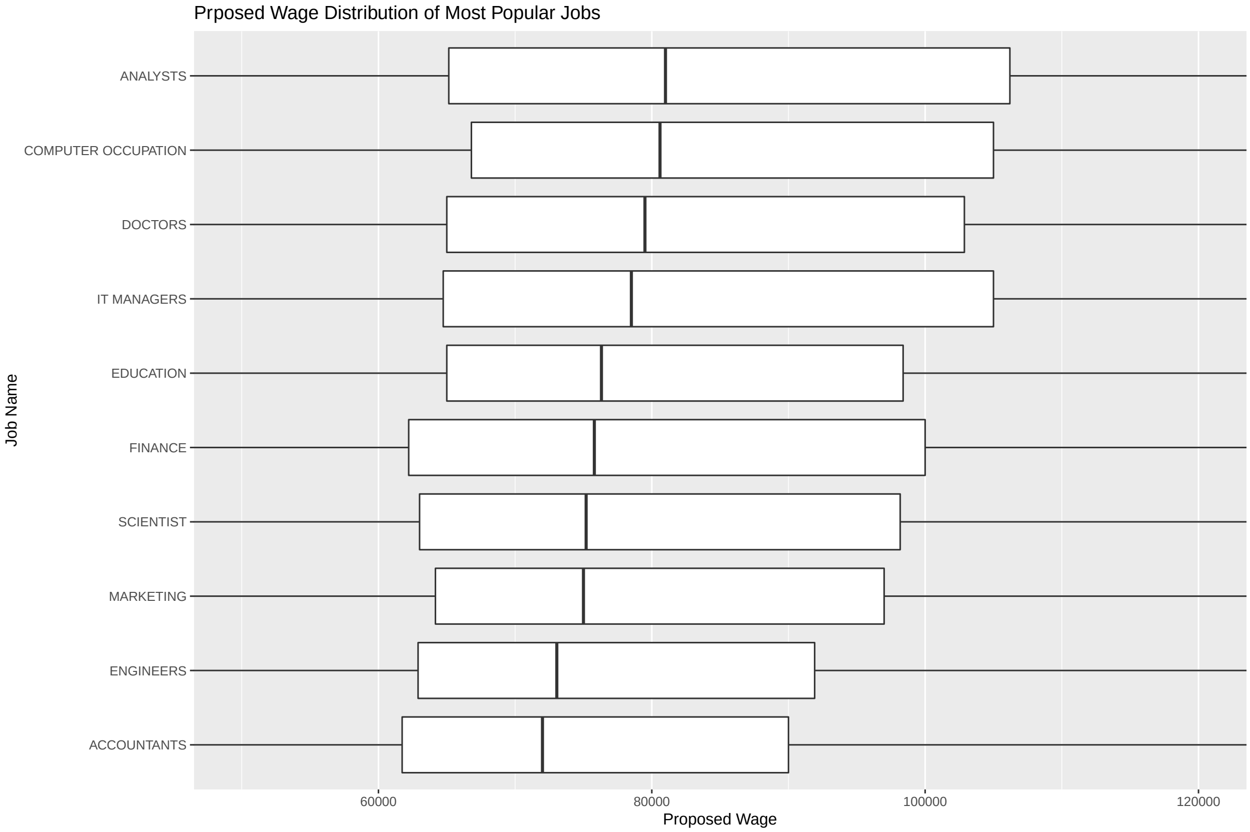
We can see that the Prevailing Wage source OES has higher wages compared to other sources.

8. Is there a linear relationship between the prevailing wage and the proposed wage?



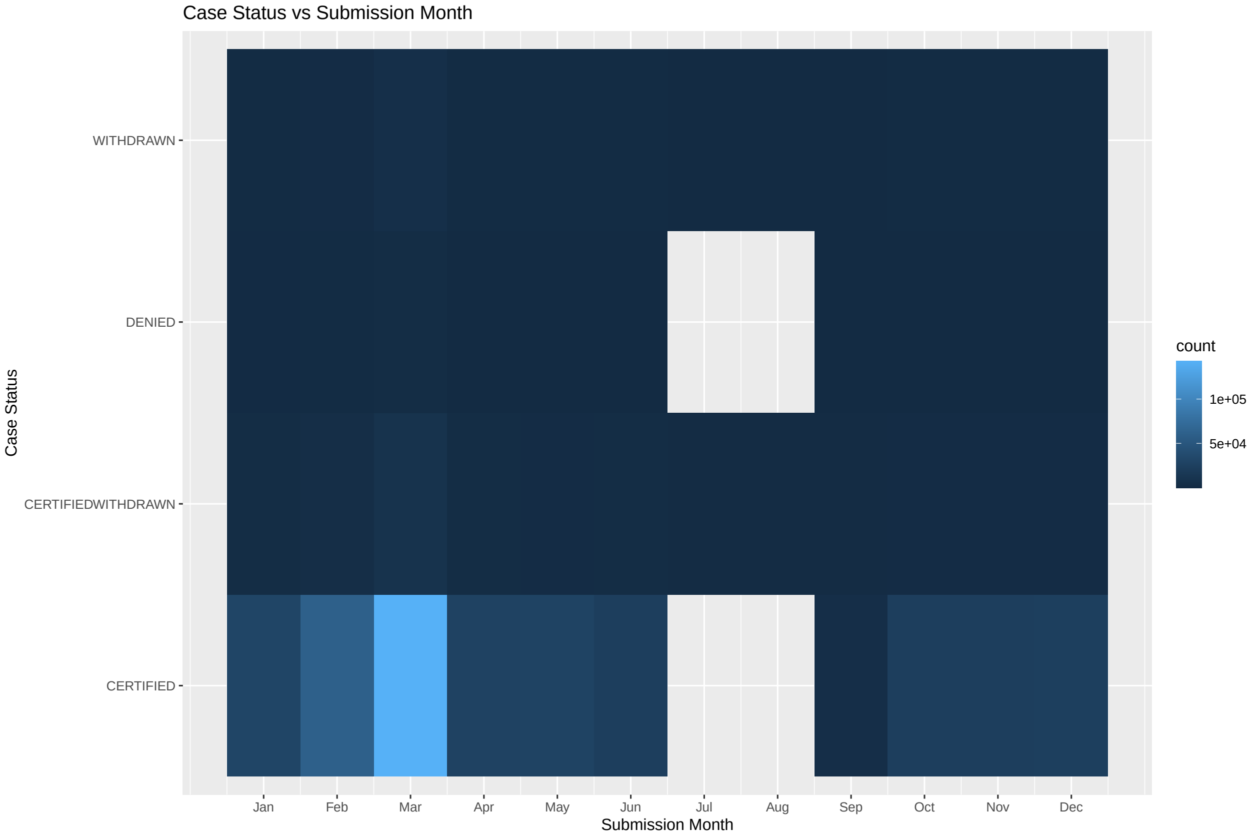
As expected, there is a positive linear relationship between the prevailing and the proposed wages. In the case of denied VISA applications, the proposed wage is slightly higher than the prevailing wage after 200,000, which might be considered the cause of denial.

9. How is the proposed wage distributed among popular jobs?



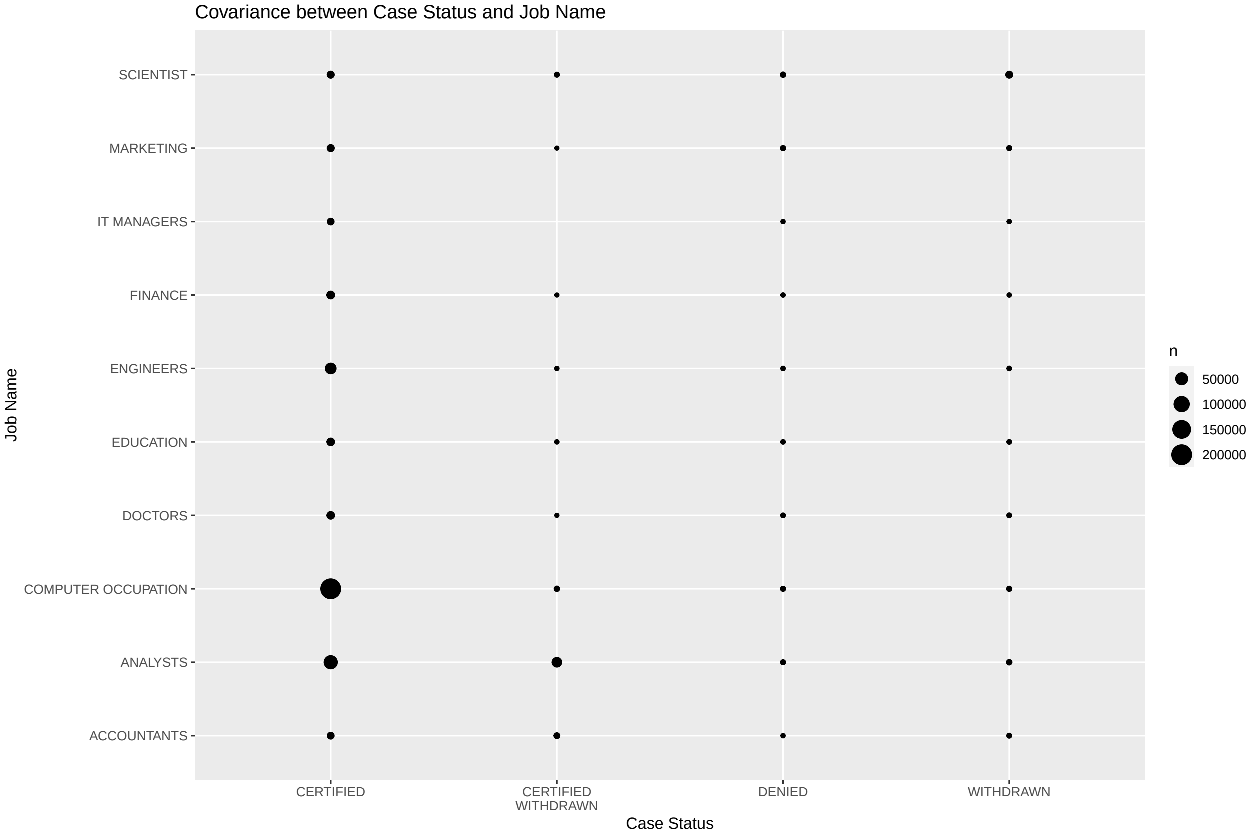
We can see that although Computer Occupation is the most popular jobs, Analysts have a higher median of proposed wage and also has the most variance in wage.

10. Is there a dependence between the Case status and submission month?



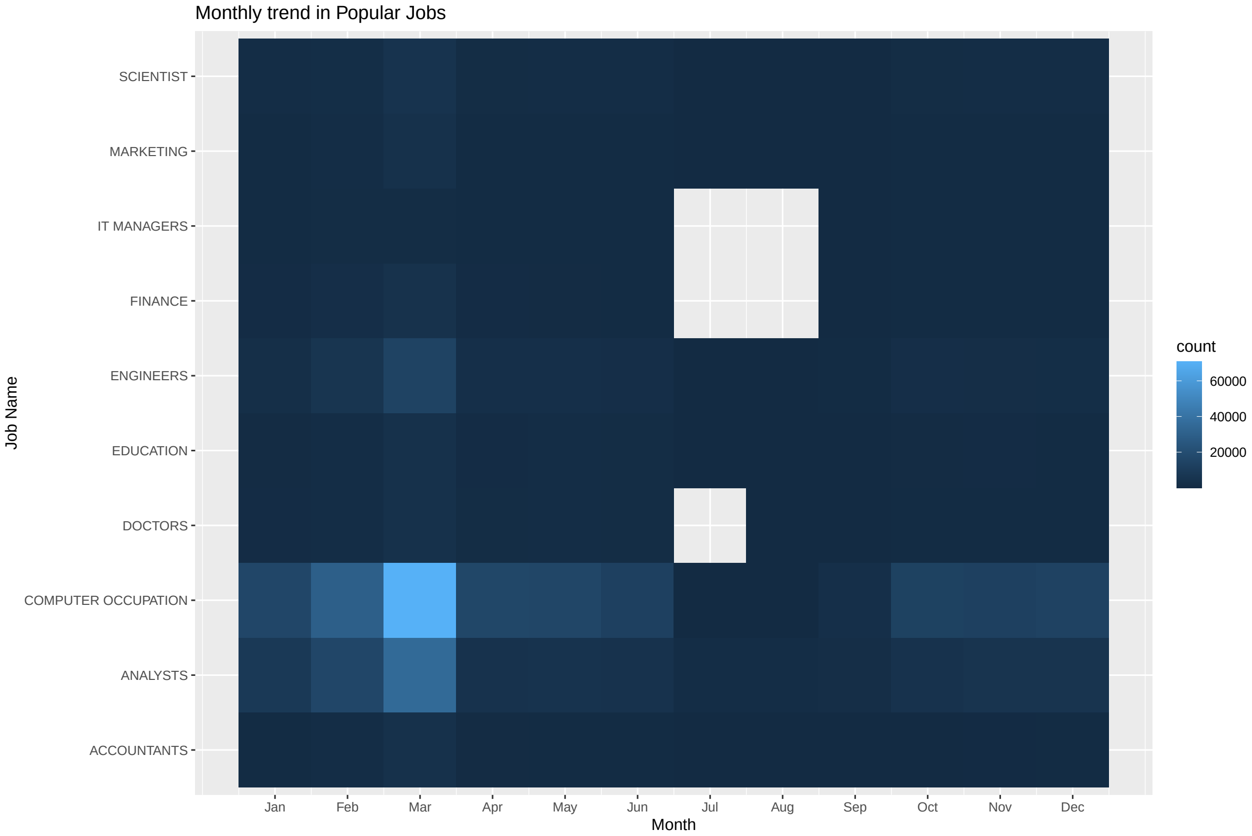
We can observe that March has the most number of certified VISA applications. However, we cannot see a definite relationship between the months and the case status.

11. How does the Case status co-vary with Job count?



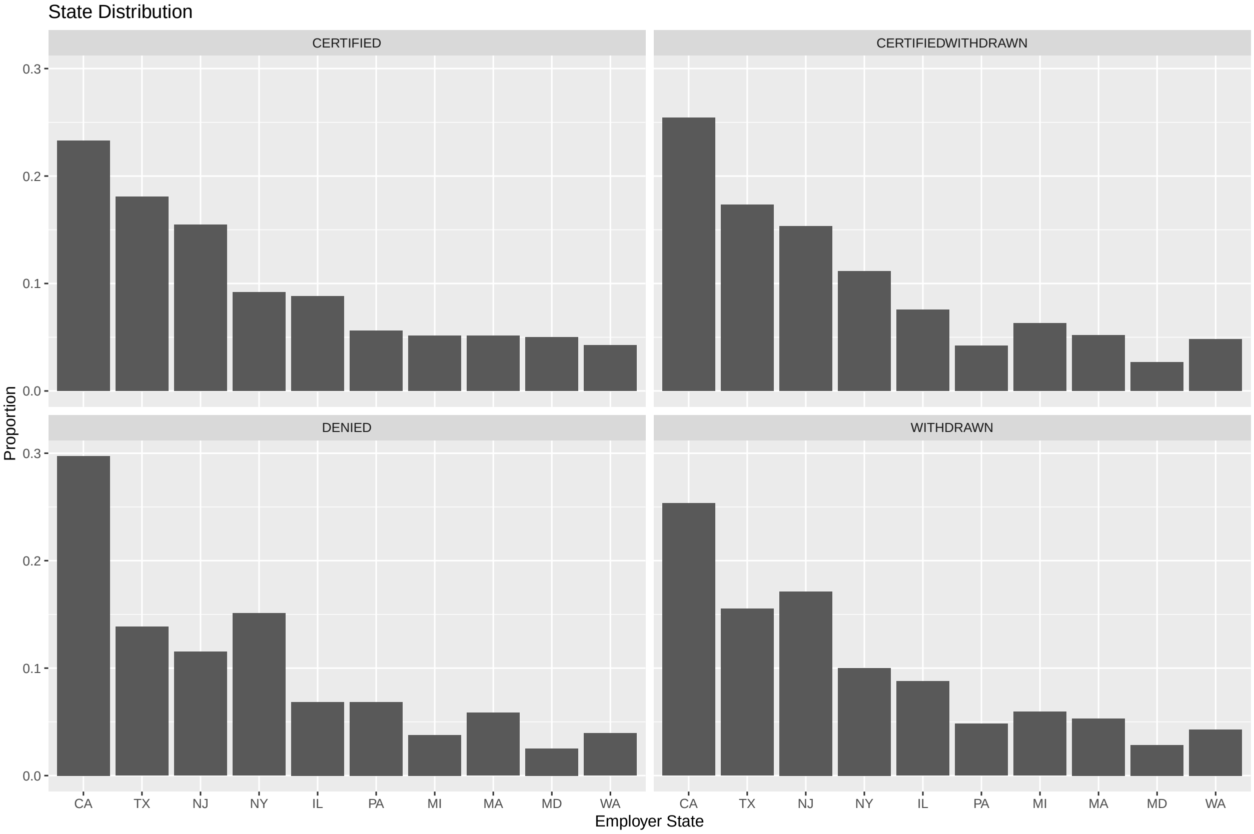
We found out earlier that Computer occupation is the most popular job requested for H-1B VISA. Here, we can observe that it is also the most certified job.

12. How does the monthly trend in popular jobs look like?

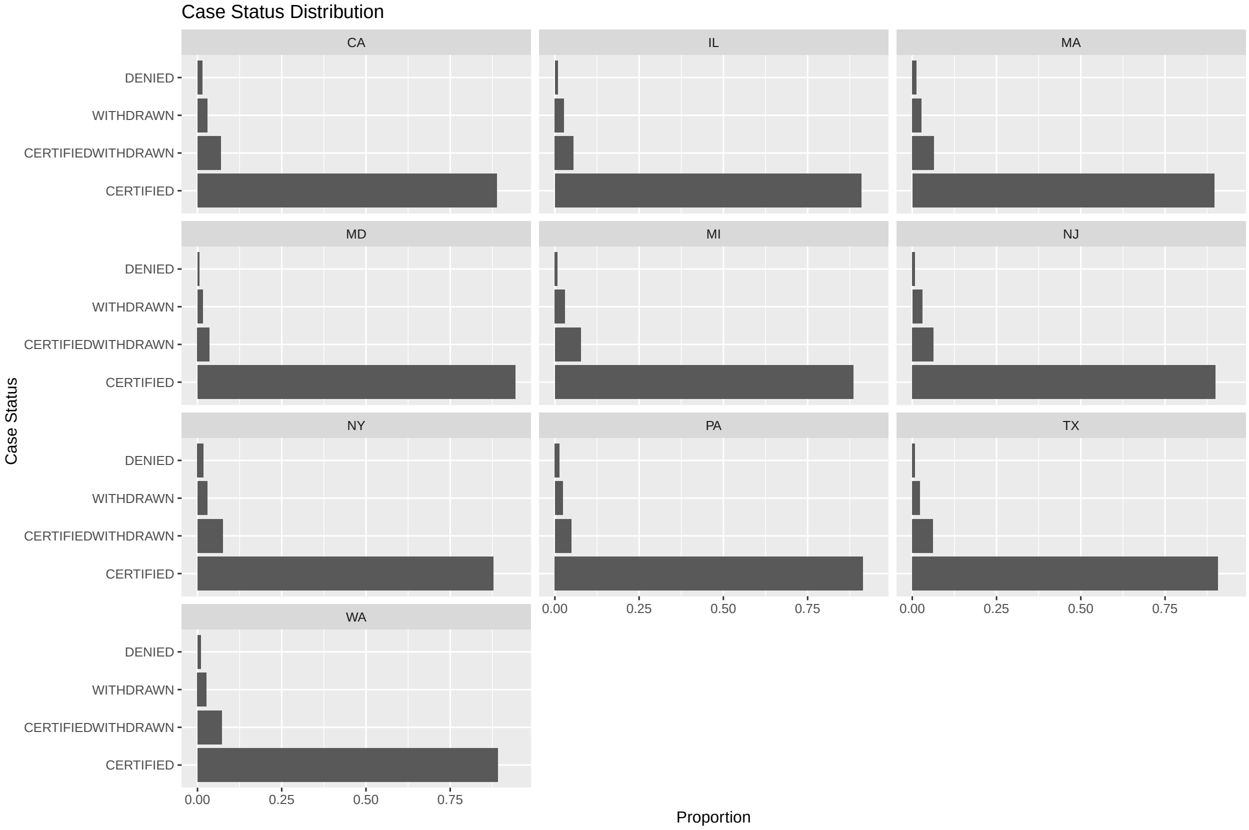


We can observe that the months from January to March receive many applications requesting jobs Computer Occupation and Analysts.

13. How are the states and the case status related?



We can see that California leads other states in both certification and denial. This is because the number of applications in California is high.



We can see that most of the applications are certified and is not dependent on the Employer state. So, we can conclude that there is no strong relationship between the case status and the employer state.

Conclusions

Hence, several transformations and explorations were done on the `h1b` data by initially converting it into appropriate form (tidying, grouping and removal of irrelevant columns).

We applied exploratory data analysis using various data transformation and visualization techniques. We were able to answer several questions about the data. These key findings are summarized below:

1. Infosys Limited has submitted the most number of H-1B VISA Applications and the top wage sponsor.
2. Computer Occupation occupies more than 50% of the job among the top 10 most popular VISA applications.
3. H-1B applications submitted in July to September are extremely low compared to other months. Most applications are submitted in March.
4. Being declared as H-1B dependent might reduce the chances of certification.
5. Both types of wages approximately follow the normal distribution with mean at around 60000 to 80000.
6. Most companies request only a few numbers of workers (1-3).
7. OES has higher wages compared to other sources.
8. There is a positive linear relationship between the prevailing and the proposed wages.
9. Although Computer Occupation is the most popular job, Analysts have a higher median of the proposed wage and have the most variance in wages.
10. March has the most number of certified VISA applications. However, we cannot see a definite relationship between the months and the case status.
11. Computer occupation is the most popular and the most certified job in terms of H-1B Applications.
12. The months January to March receive many applications requesting for jobs Computer Occupation and Analysts.

13. There is not any strong relationship between the case status and the employer state.

Appendices

Data Dictionary

FIELD NAME	DESCRIPTION	FIELD TYPE
CASE_STATUS	Status associated with the last significant event or decision. Valid values include “Certified,” “Certified-Withdrawn,” Denied,” and “Withdrawn”.	Text
CASE_SUBMITTED_YEAR	Year the application was submitted.	Number
CASE_SUBMITTED_MONTH	Month the application was submitted.	Number
CASE_SUBMITTED_DAY	Day the application was submitted.	Number
VISA_CLASS	Indicates the type of temporary application submitted for processing. R = H-1B; A = E-3 Australian; C = H-1B1 Chile; S = H-1B1 Singapore. Also referred to as “Program” in prior years.	Text
EMPLOYER_NAME	Name of employer submitting labor condition application.	Text
EMPLOYER_COUNTRY	Country of the Employer requesting temporary labor certification	Text
EMPLOYER_STATE	State of the Employer requesting temporary labor certification	Text
SOC_NAME	Occupational name associated with the SOC_CODE	Text
DECISION_YEAR	Year on which the last significant event or decision was recorded by the Chicago National Processing Center.	Number
DECISION_MONTH	Month on which the last significant event or decision was recorded by the Chicago National Processing Center.	Number
DECISION_DAY	Day on which the last significant event or decision was recorded by the Chicago National Processing Center.	Number
TOTAL_WORKERS	Total number of foreign workers requested by the Employer(s)	Number
FULL_TIME_POSITION	Y = Full Time Position; N = Part Time Position	Text
PREVAILING_WAGE	Prevailing Wage for the job being requested for temporary labor condition.	Number
PW_UNIT_OF_PAY	Unit of Pay. Valid values include “Daily (DAI),” “Hourly (HR),” “Bi-weekly (BI),” “Weekly (WK),” “Monthly (MTH),” and “Yearly (YR)”	Text
PW_SOURCE	Variables include “OES”, “CBA”, “DBA”, “SCA” or “Other”	Text
PW_SOURCE_YEAR	Year the Prevailing Wage Source was Issued	Number
PW_SOURCE_OTHER	If “Other Wage Source”, provide the source of wage	Text
WAGE_RATE_OF_PAY_FROM	Employer’s proposed wage rate	Number
WAGE_RATE_OF_PAY_TO	Maximum proposed wage rate	Number
WAGE_UNIT_OF_PAY	Unit of pay. Valid values include “Hour”, “Week”, “Bi-Weekly”, “Month”, or “Year”	Text

FIELD NAME	DESCRIPTION	FIELD TYPE
H-1B_DEPENDENT	Y = Employer is H-1B Dependent; N = Employer is not H-1B Dependent	Text
WILLFUL_VIOLATOR	Y = Employer has been previously found to be a Willful Violator; N = Employer has not been considered a Willful Violator	Text
WORKSITE_STATE	State information of the foreign worker's intended area of employment	Text
NAICS_CODE	Industry code associated with the employer requesting permanent labor condition, as classified by the North American Industrial Classification System (NAICS)	Number
WORKSITE_POSTAL_CODE	Zip Code information of the foreign worker's intended area of employment	Number