H1-B Disclosure Dataset

Rohit Ravishankar - Feb 23, 2018

Index

Introduction

This report reflects the analysis and transformation of data received from the Office of Foreign Labor Certification (OFLC), a division of the U.S. Department of Labor. The main duty of OFLC is to assist the Secretary of Labor to enforce part of the Immigration and Nationality Act (INA), which requires certain labor conditions exist before employers can hire foreign workers. H-1B is a visa category in the United States of America which allows U.S. employers to employ foreign workers.

The data set had 528134 rows and 27 attributes

```
> r <- read.csv("/Users/rohitravishankar/Desktop/Big Data/Assignment 3/h1b-disclosure-dataset/H1B Disclosure Dataset
Files/1. Master H1B Dataset.csv", header = TRUE)
> dim(r)
[1] 528134     27
>
```

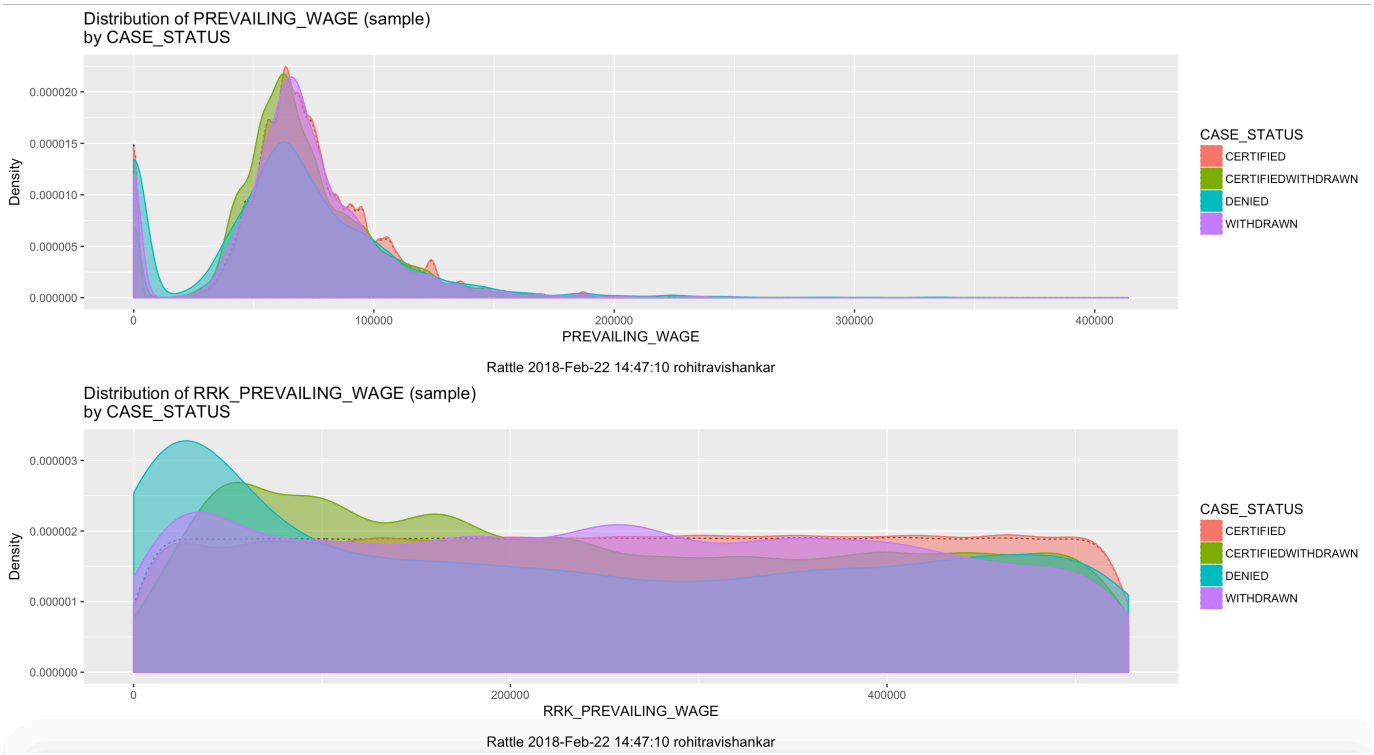
Some of the key attributes used across the data analysis were:-

- Prevailing Wage - The Wage of the applicant
- NAICS Code - The code the depicts the industry of the applicant company
- H-1B Dependent - Whether the H1-B holder has any dependents
- Willful violator - Indicates whether the H1-B applicant was a willful violator
- Decision Month - Month when the decision of the H1-B applicant was made

Inspiration

How is the wage of an applicant related to the probability of being of certified status? What is the probability that an H1-B applicant has dependents and is a willful violator? What are the months and days an applicant is most likely to get a decision?

Distribution of Prevailing Wage - Re-centering

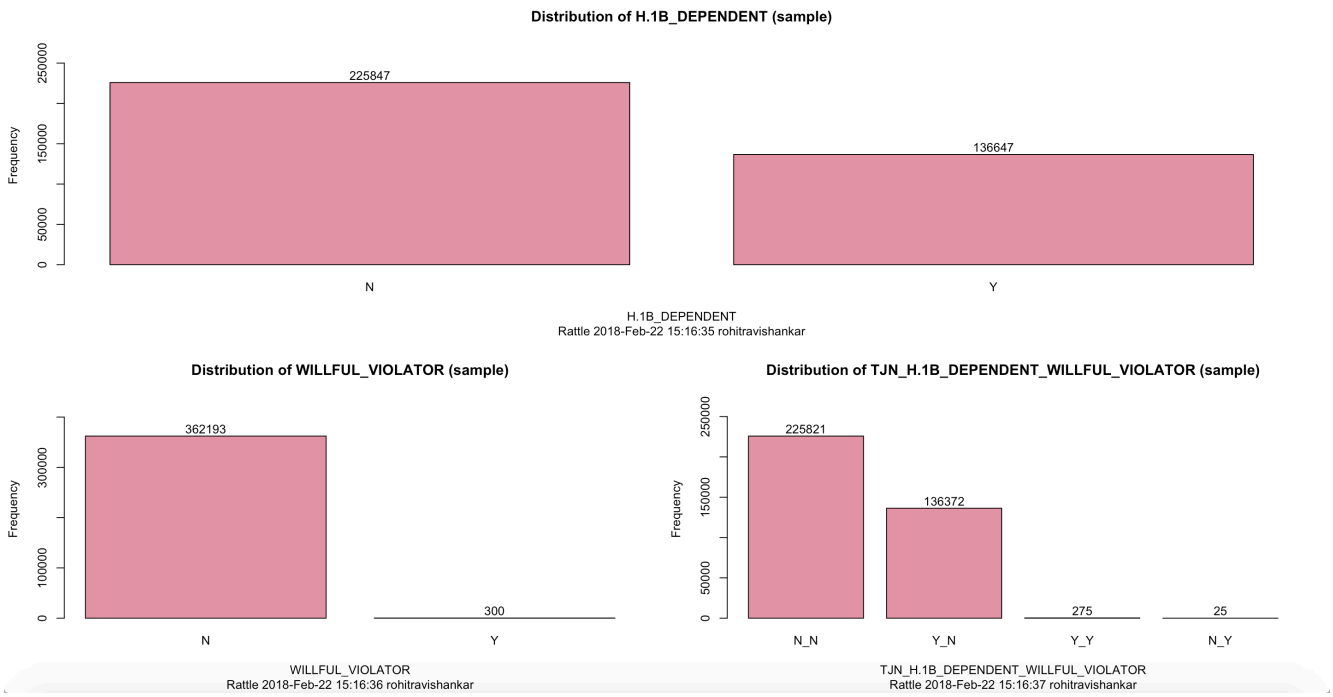


The aforementioned graph, is a histogram plot of the distribution of the prevailing wage and the probability of being of certified H1-B status. The second graph is a **re-centered** rendition of the same graph using **ranking**.

The first graph is not a clear indication of how the wage relates to case status, i.e., certified, withdrawn, denied or certified but withdrawn. Using the second graph, we can conclude that applicants with lower wages are more likely to have their H1-B visas denied as compared to applicants with higher wages, and applicants with higher wages are more likely to be certified.

Looking at the graphs we can also conclude that a large majority of the H1-B applicants are earning higher than the median wage in United States. Also, looking at the prevailing wage reported of H1-B applicants, we could infer that that majority of the H1-B applicants are probably highly skilled workers.

H1-B Dependents and Willful Violators - Recoding

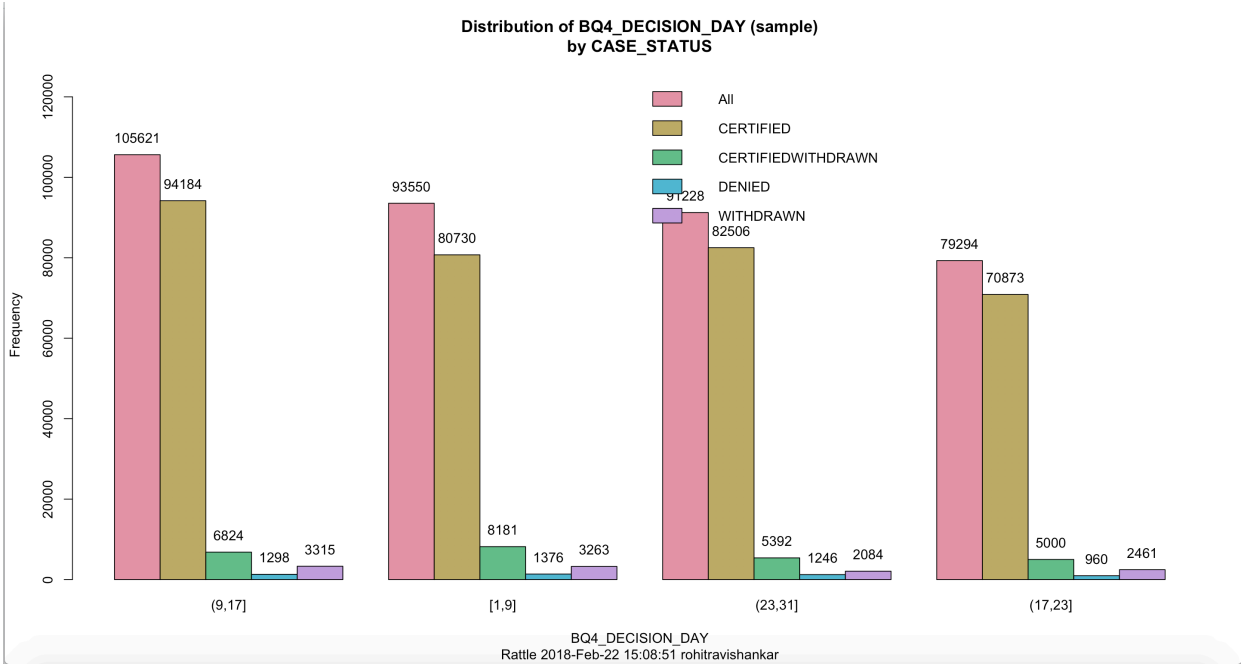


The figure above has 3 bar charts of which 2 of them describe the H1-B applicants having dependents and H1-B applicants who are willful violators. These 2 graphs per se do not provide much insight into the data.

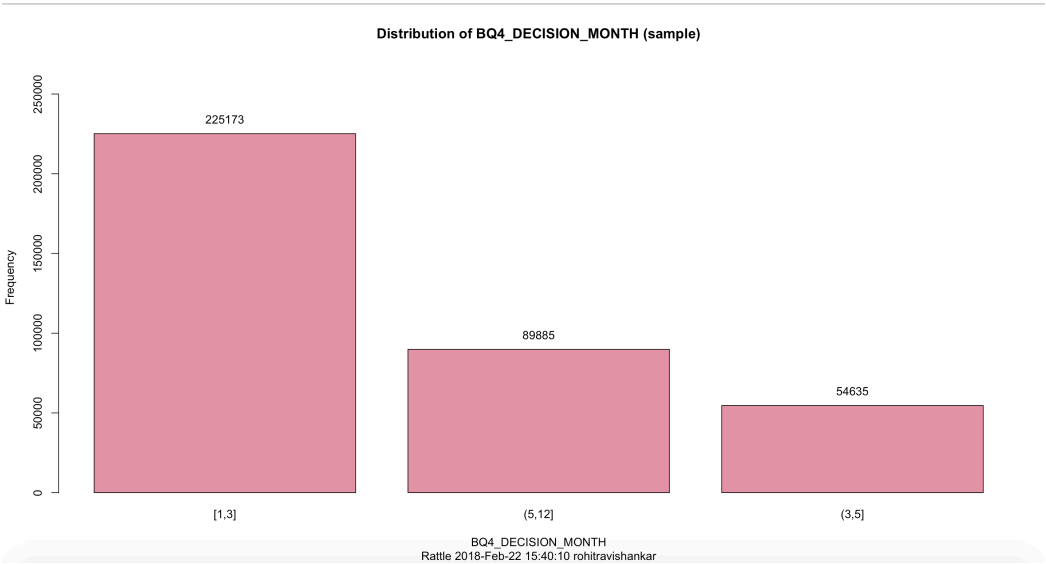
However, on **joining the categories** the results of the distributions we see that number of H1-B applicants with dependents, and H1-B applicants who are not willful violators far exceed the number of H1-B applicants with dependents and H1-B applicants who are willful violators.

Further, from the skewness in the data we can speculate that H1-B applicants with dependents are likely to not be willful violators. We can say that H1-B applicants are mostly not willful violators because they are likely to be well paid and highly skilled workers. (using the conclusion of the previous graph's data on prevalent wages).

H1-B Decision Status by Month and Day - Recoding



The above histogram plot describes the decision for H1-B applicants **recoded** and **binned** by dates. From the above plot we can observe that H1-B applications were likely to be processed and decisions to be made between the 9th (not inclusive) and 17th(inclusive) of any month followed closely by the period between 1st (inclusive) to 9th(inclusive) of the month.



The above histogram plot describes the decision for H1-B applicants **recoded** and **binned** by months of the year. From the above plot we can observe that H1-B applications were likely to be processed and decisions to be made in the first quarter of the year as compared to other periods of the year.

Armed with the knowledge of the two aforementioned results we can conclude that most H1-B applications are processed in the January to March period and between 9th to 17th of these months.

Inserting Missing values - Imputing

When when we take Median/MAD for Total workers we end up with 484363 missing. It is a categorical field type and hence, we insert 'Missing' for all these missing values so that it is easier to perform other transformations such as re-centering & recoding that may be required at any future stage.

Deleting NAICS Code - Cleanup

The North American Industry Classification System (NAICS) classifies business establishments for the purpose of collecting, analyzing, and publishing statistical data related to the U.S. economy. The NAICS industry codes define establishments based on the activities in which they are primarily engaged. NAICS codes are also used for administrative, contracting, and tax purposes.

The data set also contains the names of the companies which can be used to define the establishments based on the activities that are performed by them rather than having a separate column to describe the type of establishment of the company. Hence, the belief that NAICS code need not exist in the transformed data set.

Conclusion

We can summarize our conclusion from the given data down to the following:-

- A large majority of the H1-B applicants are earning higher than the median wage in United States
- H1-B applicants with dependents are likely to not be willful violators.
- Most H1-B applications are processed in the January to March period and between 9th to 17th of these months.
- Not all fields need to remain after transforming the data, such as, NAICS code in this case.

With this knowledge, we can conclude that a large majority of H1-B workers are highly skilled since they earn higher than median salary and most of them are likely to earn their living by honest means.

References

- Kaggle
  https://www.kaggle.com/trivedicharmi/h1b-disclosure-dataset

- SBA
  https://www.sba.gov/contracting/getting-started-contractor/determine-your-naics-code