H1B Visa Research Proposal May, 2018

Mark Ferguson

**Introduction**

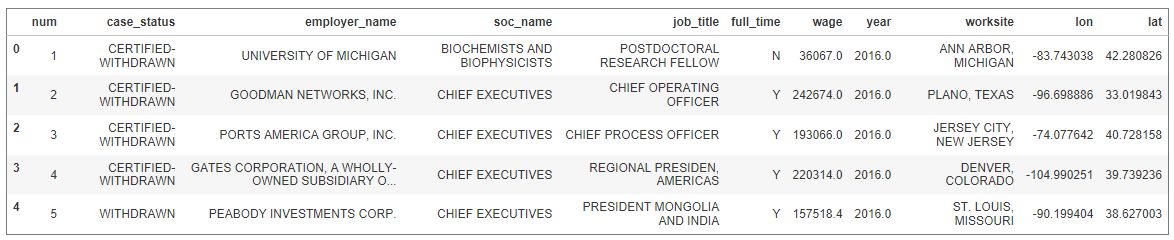
The H1B is a visa type granted by the United States Citizenship and Immigration Services (USCIS), that is intended for professional workers with degrees. A complete description of the H1B program can be found at the following link:

https://en.wikipedia.org/wiki/H-1B\_visa

A dataset was found on the Kaggle website containing details of some three million H1B visa petitions, spanning the period from 2011 to 2016. These are categorized according to approximate job categorization, sponsoring employer, job title, location, salary, and geographical coordinates. The source of the dataset is given below:

https://www.kaggle.com/nsharan/h-1b-visa

Below is a screenshot from the original import of the data:



From a first look at the data, it was apparent that a number of important questions could be posed and answered by a further analysis. These questions are examined in the form of hypotheses, which this study will endeavor to validate:

* Executive-level petitions are less likely to be denied than non-executive petitions.
* Some professions, job titles, and sectors are more likely to be denied than others.
* More generally, average salaries for positions in major metropolitan areas are higher than for positions outside such areas.

Before deciding on an investigative strategy, it is important to consider for a moment what is not known, what we are assuming, and what are the possible implications of these.

As mentioned above, the data are drawn from a kaggle dataset which in turn was taken from a USCIS web site. There is no specific reason to suspect that anything untoward was done to the raw data. However, as we don’t have access to these raw data we are subject to the judgement (including any potential errors in that judgement) that the kaggle user employed in cleaning the data.

**Examining the Data**

We now take a closer look into the data and outline what we understand by each of the heading titles:

Case Status – Denotes the current status of the petition. This can take any of the following values:

CERTIFIED-WITHDRAWN, CERTIFIED, WITHDRAWN, or DENIED. Note that the designation 'CERTIFIED' does not necessarily mean that the petition has been approved, only that it has passed the USCIS' initial screening process. It is not known from the available data what ultimately happened to a petition initially designated ‘CERTIFIED.’ In this study we will effectively lump all the 'CERTIFIED' designations into one basket, since we are concerned with whether the petition was denied or not.

Employer Name – The name of the sponsoring company or institution.

SOC (Standard Occupational Classification) Name – The broad professional category into which the job falls.

Job Title – The job title as specified by the sponsor.

Full Time – A binary Y/N position based on whether the job is full time or not.

Prevailing Wage – The yearly wage paid for the job role.

Year

Worksite – The city and state – presumably this denotes the place where the job will be carried out, as distinct from the location of the sponsor.

Lon, lat – The geographical coordinates of the worksite; a random few of these locations have been verified.

**Key Assumptions and Sources of Error**

At this point it is worth considering what assumptions are inherent in this study, and what those assumptions would have on the validity of the analysis if wrong. The main caveats are listed as follows:

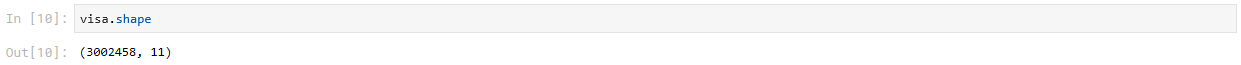
* The data essentially fall into two categories: denied, and certified. An important unknown is which petitions that are initially certified, will ultimately go on to be rejected. Determining this would require data that are simply not available to us.
* The salary data are what is referred to as the “prevailing wage” for the position. It is not known how accurate these data are, as these numbers are left to the sponsoring body to report. This would presumably, but not necessarily, be the same as or similar to, the amount that organization pays or intends to pay that employee once approved. However, different companies may have access to differing wage data. Also, if the sponsor anticipates some potential bias on the part of the USCIS in favor higher-paying roles, they may be even inflate their stated salary figure, in hopes of exaggerating the 'importance' of the position and increasing the chance of approval.
* We aim to use H1B petition data to generalize salary data for professional positions across the US. This may or may not be a valid extrapolation, and in practice could be an important source of contextual bias. For example, a scarcity of workers in one area, say, oil and gas workers in Texas, may cause average Texas wages to appear artificially high if there is a large demand from overseas for such workers; in actual fact, the mean Texas wage may be a lot lower when national (non-H1B) data are examined. Accounting for this potential bias would require comparison with some external dataset(s).

**Preliminary Data Analysis**

We now take a more detailed look at the data, in order to determine exactly the dimensions and characteristics of the dataset:

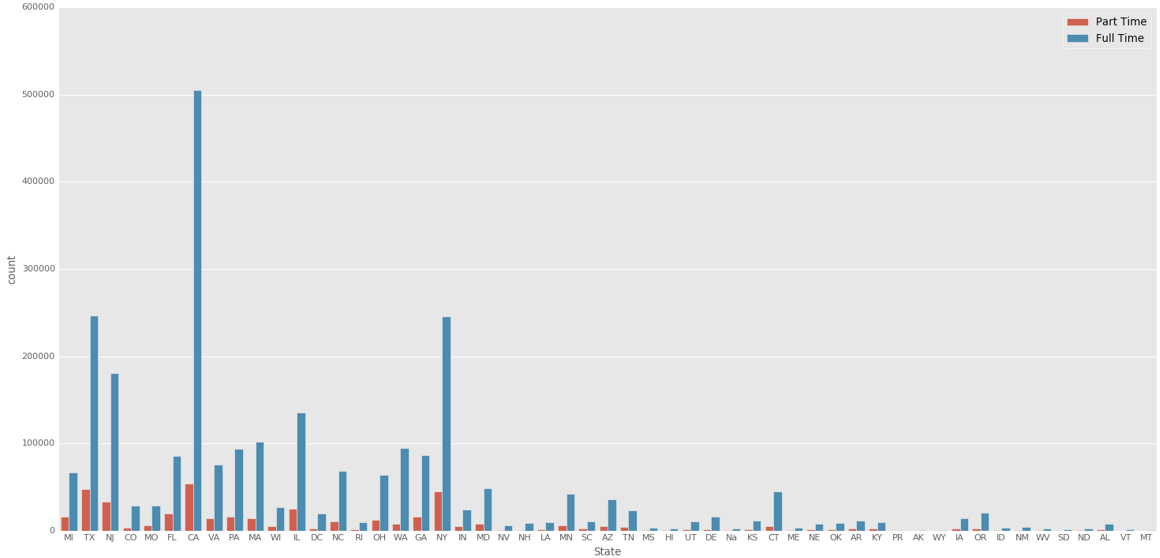
Number of records

The 'visa' dataset has approximately 3 million records:



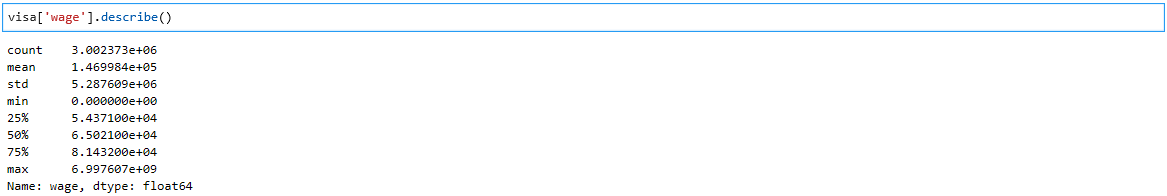
Most common geographical areas

We see from the below that California, followed by New York and Texas are the most common states from which H1B applications are received:



Salary statistics

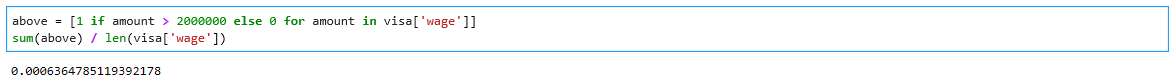
Running a describe on the 'wage' column shows a maximum value of just under $7 billion. One feels well justified in eliminating this as bad data. We also see that around 75 % of the wages in the original dataset are below about $81,000.



Having removed the rogue salary value, we obtain the following boxplot visualizations:

|  |  |
| --- | --- |
|  |  |

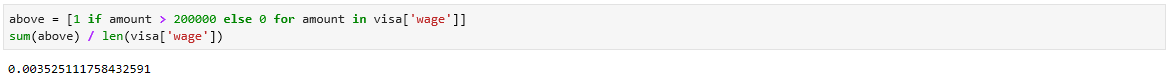
The visualizations show a clear split in the data at around 0.2E7, or $2million. In order to determine what proportion of the data are above this value, we do the following:



Thus only 0.06 % of H1B visa applications relate to positions having a wage above $2MM. This seems to be a logical cutoff point in order to eliminate bad data. However, looking at some boxplots and histograms, we get the following, showing clearly that the bulk of the data lie well below $200,000:

|  |  |
| --- | --- |
|  |  |
|  |  |

Repeating the above list comprehension shows that 99.7 % of the salary values are below the $200,000 mark:



Thus an upper threshold of $200,000 seems even more suitable and appropriate for this dataset, and that is what will be used.

**Outline Research Plan**

Given the considerable size of the dataset, processing times may be very long. If this proves to be a serious drain on the time available to complete the project, there exists the option to modify the method such that random samples are taken from the dataset. Additionally, the study will be narrowed to consider only full-time positions to remove unnecessary variability of the data.

Taking each of the stated projected objectives in turn:

1. Executive-level petitions are less likely to be denied than non-executive petitions.

The job titles will be grouped into 'Executive' and 'Non-executive', based on keywords found in the job category column.

The five status categories - CERTIFIED-WITHDRAWN, CERTIFIED, WITHDRAWN, or DENIED will be split into two broader categories: CERTIFIED, consisting of CERTIFIED-WITHDRAWN, and CERTIFIED; and DENIED. The WITHDRAWN designation will be discarded in this part of the analysis.

As the status will be reduced to a simple 0 or 1 binary classification, means will be calculated for applications in each of the 'Executive' and 'Non-executive' categories, and a t-test can be carried out to compare the mean values of the two ad hoc job categories.

From this we can infer whether a higher proportion of applications for executive-level positions are denied; a p-value below 0.05 will be taken as the threshold such statistical significance.

Having done this on the entire US dataset, we will then repeat this analysis on a number of cities (say, Houston, Los Angeles, New York and Chicago) in order to normalize the results and see whether the national findings are repeated when regional variations are removed.

2. Some professions, job titles, and sectors are more likely to be denied than others.

In a similar way to that described above, the petitions will be split into an appropriate number of groups, this time by profession, then a similar t-test analysis done to determine whether this criterion in any way influences the likelihood of denial.

3. Average salaries for positions in major metropolitan areas are higher than for positions outside such areas.

The data will again be categorized, this time according to whether the worksite location is in one of the top 20 US cities by population. Visualizations will be used to show any correlation between the location of the worksite (specifically whether or not it is in one of the major cities), and the average salaries of positions in those areas. This could be carried out using A/B testing as described earlier.

The findings will then be checked against national wage data obtained from another source, in order to assess, explain and adjust for any bias.