H1B Visa Petitions Analysis

Ramya Bhat

H1B Visa

There is a special kind of visa for non-immigrants of United States who are interested in working for companies situated in USA. 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 (Masters, PhD) and work in a full-time position. Under the H1B visa, any company can employ a foreign worker for up to six years. Filing of a H1B visa is not in hands of the individuals, only the employer is allowed to file the petition for the respective employee.

Dataset Description

The H-1B Dataset selected for this project contains data from employer's Labor Condition Application and the case certification determinations processed by the Office of Foreign Labor Certification (OFLC). The dataset has more than 500,000 records overall. The relevant columns included in our data set are: employer name, job title, case number, visa class, employment start and end dates, case status and so on.

The data was converted from the XLSX to CSV format, due to memory issues. CSV files seem much easier to handle. The data has 410,605 observations of 52 variables. Picking certain variables, we would look at how they are concentrated in the data.

Some of the variables of interest are CASE_STATUS, EMPLOYER_NAME, EMPLOYER_CITY, EMPLOYER_STATE, FULL_TIME_POSITION, PW_WAGE_LEVEL,H-1B_DEPENDENT, SOC_CODE and these are selected out of the data.

The columns in the dataset include:

- **CASE_STATUS:** Status associated with the last significant event or decision. Valid values include “Certified,” “Certified-Withdrawn,” Denied,” and “Withdrawn”.
- **EMPLOYER_NAME**: Name of employer submitting labor condition application.
- **SOC_NAME**: Occupational name associated with the SOC_CODE. **SOC_CODE** is the occupational code associated with the job being requested for temporary labor condition, as classified by the Standard Occupational Classification (SOC) System.
- **JOB_TITLE**: Title of the job
- **FULL_TIME_POSITION:** Y = Full Time Position; N = Part Time Position
- **PREVAILING_WAGE**: Prevailing Wage for the job being requested for temporary labor condition. The wage is listed at annual scale in USD. 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.
- **YEAR:** Year in which the H-1B visa petition was filed
- **WORKSITE:** City and State information of the foreign worker’s intended area of employment

Objective

The objective of this project is to have a deeper insight about the H1B applications filed in United States of America and also to predict the case status of an application submitted by the employer to hire non-immigrant workers under the H-1B visa program.

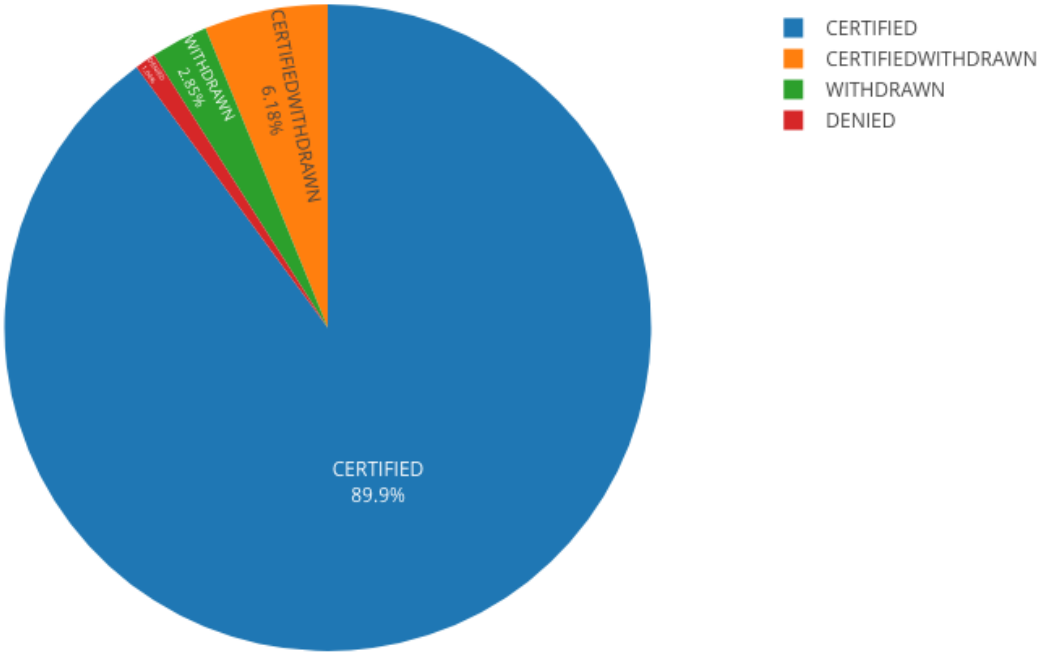
Data Preprocessing

Data preprocesing involves transforming data into a basic form that makes it easy to work with. Our set of data containing more than 500,000 rows must be pre-processed and filtered down to suit the requirements of our project.

The challenge we all face in data preprocessing is to handle null, N/A values and the outliers. All the rows of the class attributes that has N/A or null values in them must been removed and outliers has to be dealt with. Activities done in this step includes detecting the presence of missing (NA) values and outliers, or duplicate data.

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##      7.2  55765.0  68037.0  70605.5  85218.0 414007.0
```

Categorizing the H1B applicants into groups



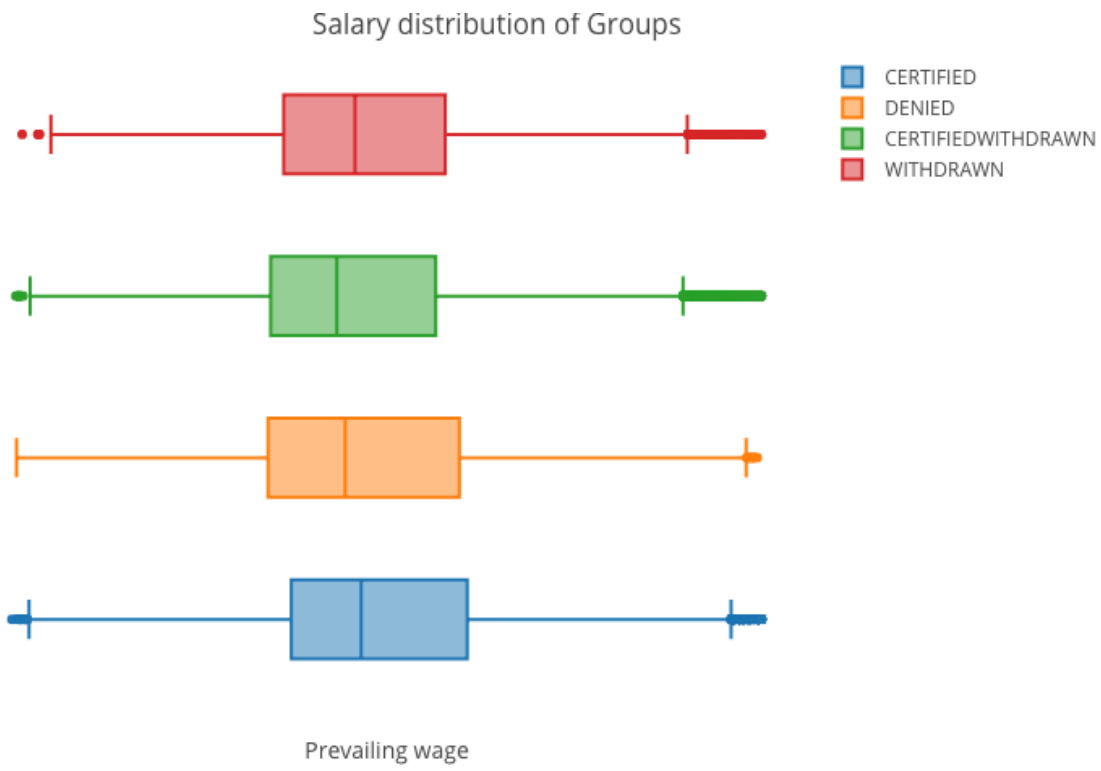
The case status of any application can take on 4 values, which are:
Certified: A certified Labor Condition Application (LCA), is a prerequisite to H-1B approval. So, “certified” means the employer filed the LCA, which was approved by Department of Labour (DOL) and the necessary prerequisite for an H-1B approval is in place.
Certified-Withdrawn: This means the LCA was approved but was later on withdrawn by the employer, for some reason. It could be that the employee worked for some years before the contract was terminated.

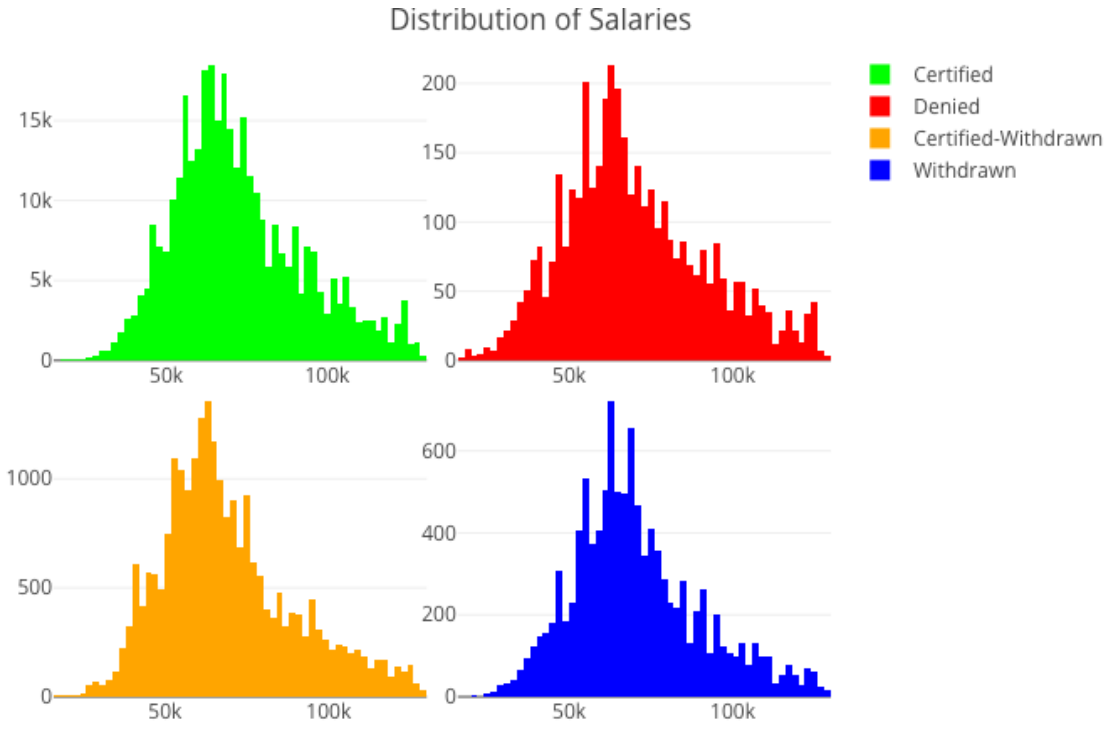
Denied: Means that the LCA was denied and so, the necessary prerequisite for an H-1B approval is not in place.
Withdrawn: Means that the LCA was withdrawn before approval or denial. So, no decision was taken before the employer withdrew the application.

Determining the distribution of salary of the applicants based on the status of their job

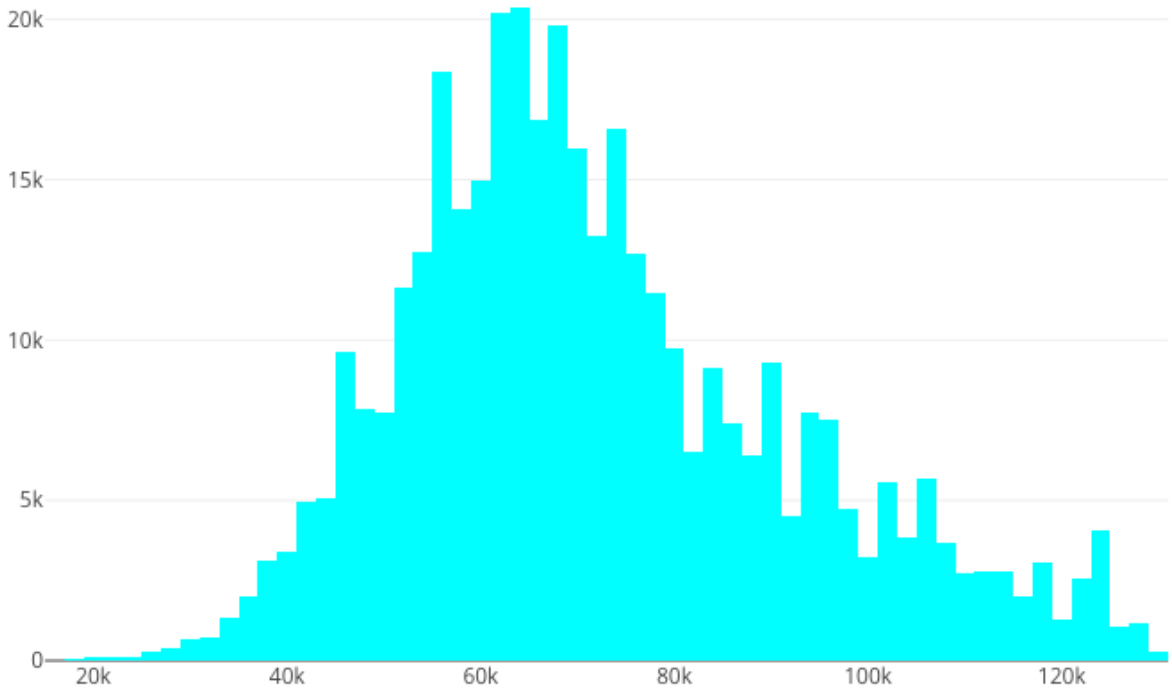
Below is a box plot and histogram of the prevailing wage distribution for each group of H1B applicants.

To investigate the prevailing_wage variable, histogram would be the best visualization to see the distribution. In general, the prevailing wages of denied H1B cases have more extreme values than those of certified H1B cases. The 1st quartile, median and 3rd quartile of prevailing wages for certified cases are greater than those of denied cases respectively.


Salary distribution of Groups
CERTIFIED
DENIED
CERTIFIEDWITHDRAWN
WITHDRAWN
Prevailing wage


Distribution of Salaries
Certified
Denied
Certified-Withdrawn
Withdrawn
15k
10k
5k
0
50k
100k
200
150
100
50
1000
500
600
400

Distribution of applicants by measure of Salary.

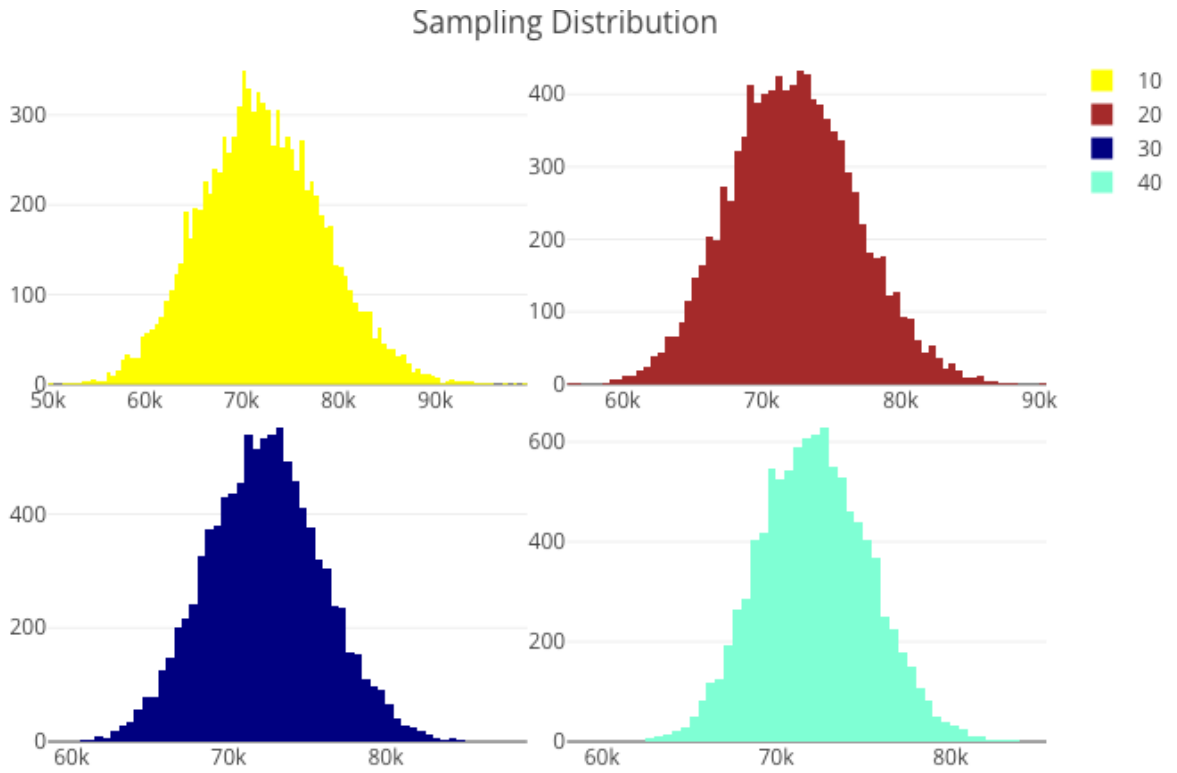


Central Limit Theorem

The central limit theorem states that the distribution of sample means, taken from independent random sample sizes, follows a normal distribution even if the original population is not normally distributed. As a result we can apply statistical techniques that assume normality even when the population is non normal.
In this data set the applicability of the central limit theorem can be shown by using the Prevailing_wage attribute . As we can see from the above histogram, the salary distribution of all applicants have a normal distribution hence Prevailing_wage is used here for central limit theorem. Below are histograms showing the sample means of 1000 random samples of sample size 10, 20, 30, and 40 following a normal distribution.

```
## Population , Mean = 72246.11 , SD =  20328.86
```

```
## Sampe Size 10, Mean = 72243.72 , SD = 6428.549
## Sampe Size 20, Mean = 72303.86 , SD = 4545.671
## Sampe Size 30, Mean = 72286.91 , SD = 3711.525
## Sampe Size 40, Mean = 72235.47 , SD = 3214.275
```



Sampling

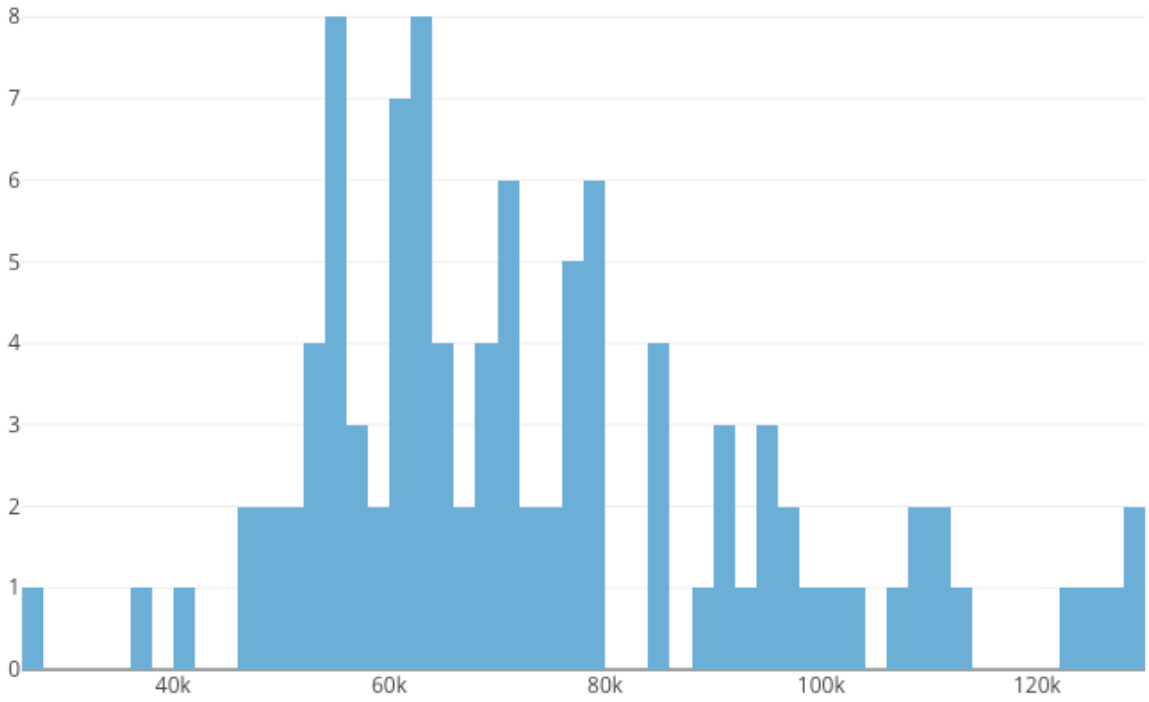
Sampling is a technique to select a representative portion of the population to perform a study on. There are many different sampling techniques ->

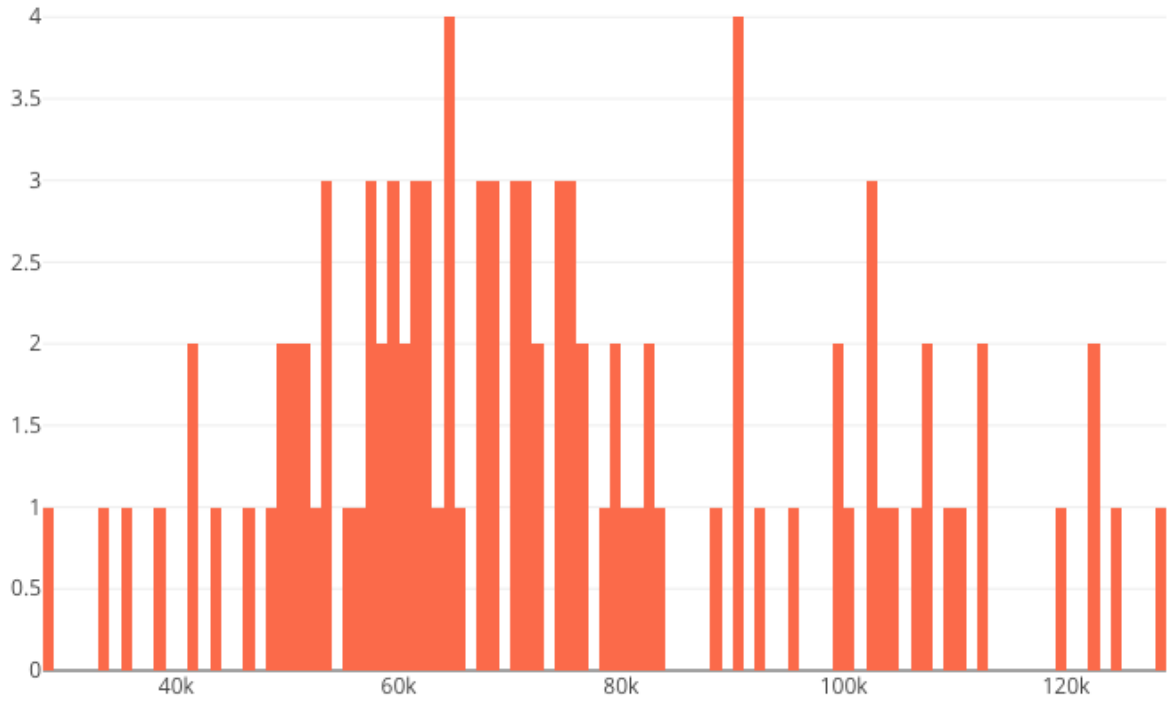
1) Simple random sampling - a basic sampling technique where individual subjects are selected from a larger group. In this case, every sample has the same chance of getting picked.

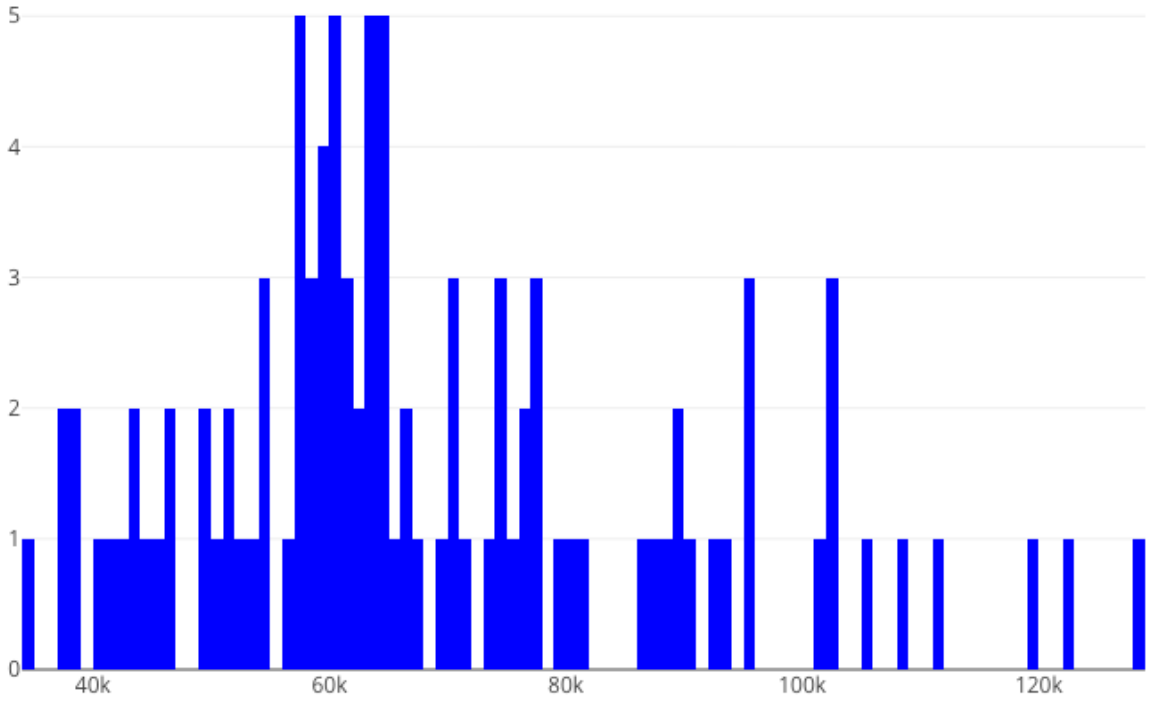
2) Systematic sampling - Systematic sampling is a method where samples are selected via a fixed periodic interval. The interval is calculated by dividing the whole population sample by the desired sample size.

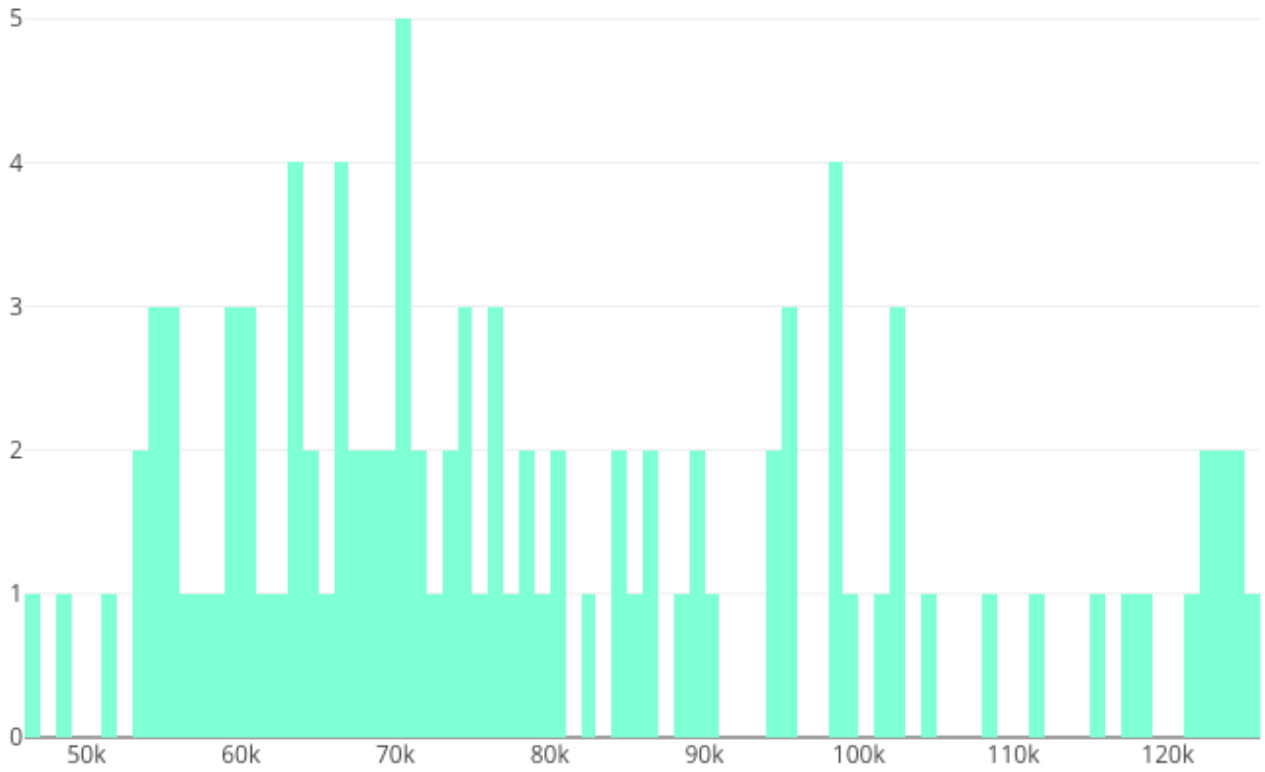
3) Stratified sampling.

For this project - Simple random sampling with replacement, simple random sampling without replacement, systematic sampling, and unequal probability sampling has been be used as sampling methods.




4
3.5
3
2.5
2
1.5
1
0.5
0
40k
60k
80k
100k
120k


5
4
3
2
1
0
40k
60k
80k
100k
120k



Confidence Intervals

Confidence intervals are used to indicate how accurate a calculated statistic is likely to be. Confidence intervals can be calculated for a variety of statistics, such as the mean, median, or slope of a linear regression.

```
## Prevailing wage : mean = 72246.11  and sd = 20328.86
```

```
## 80% Conf Level (c_alpha = 0.20), CI = 46193.63 - 98298.59
## 90% Conf Level (c_alpha = 0.10), CI = 38808.12 - 105684.11
```

```
## SRSWR : mean = 73974.64  and sd = 2032.886
```

```
## 80% Conf Level (c_alpha = 0.20), CI = 71369.39 - 76579.89
## 90% Conf Level (c_alpha = 0.10), CI = 70630.84 - 77318.44
```

```
## SRSWOR : mean = 71998.53  and sd = 2032.886
```

```
## 80% Conf Level (c_alpha = 0.20), CI = 69393.28 - 74603.78
## 90% Conf Level (c_alpha = 0.10), CI = 68654.73 - 75342.33
```

```
## SRSWOR : mean = 68915.04  and sd = 2032.886
```

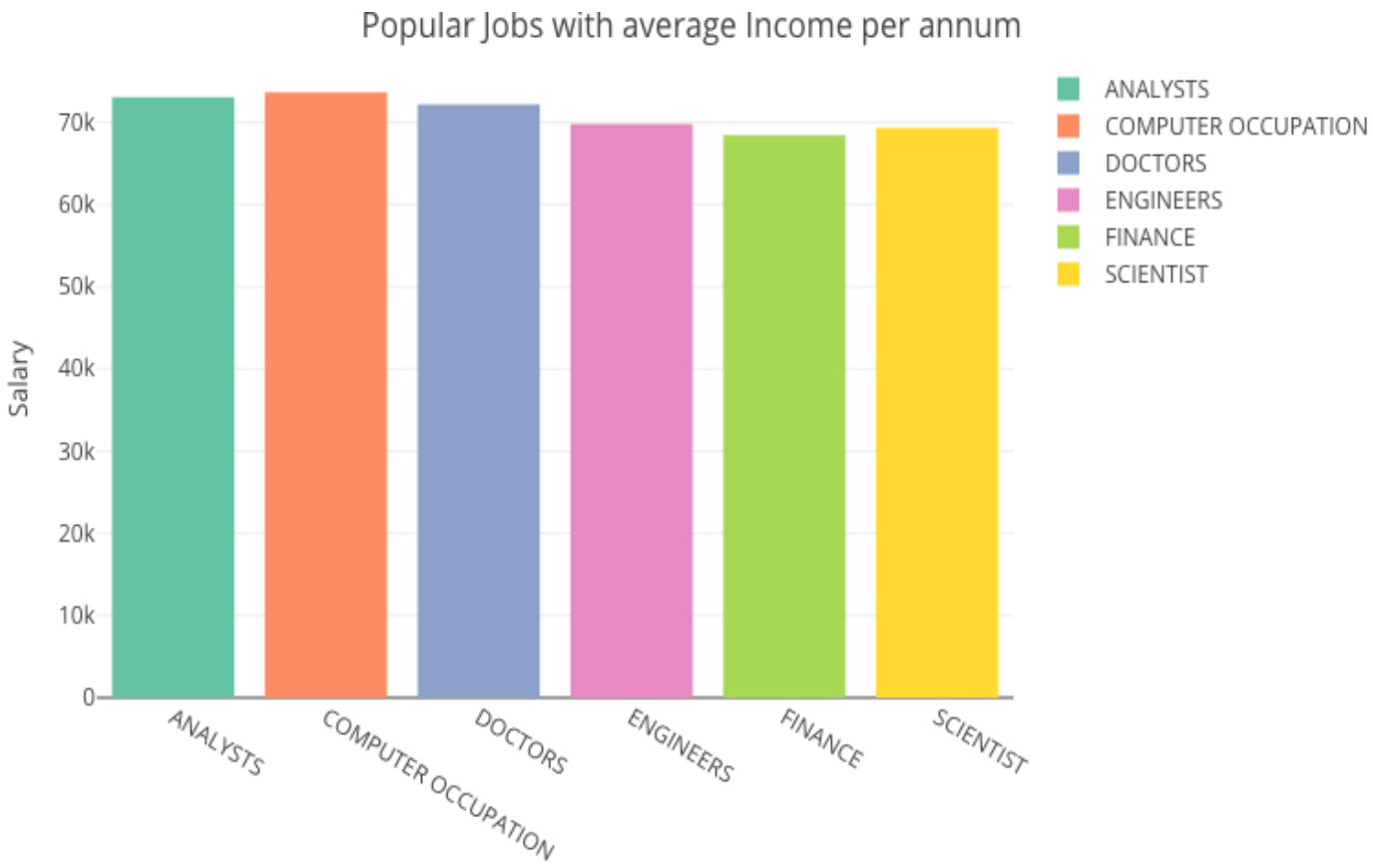
```
## 80% Conf Level (c_alpha = 0.20), CI = 66309.79 - 71520.29
## 90% Conf Level (c_alpha = 0.10), CI = 65571.24 - 72258.84
```

```
## UPSystematic : mean = 79946.01  and sd = 2032.886
```

```
## 80% Conf Level (c_alpha = 0.20), CI = 77340.76 - 82551.26
## 90% Conf Level (c_alpha = 0.10), CI = 76602.21 - 83289.81
```

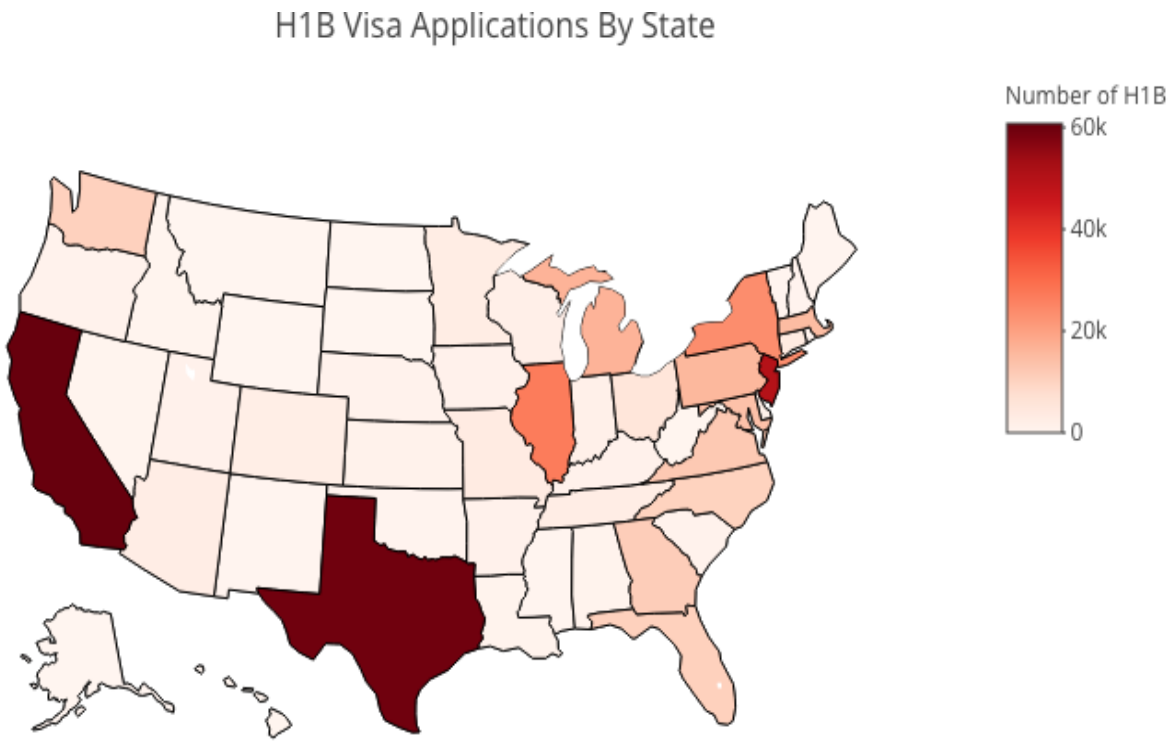
Jobs that are of more demand in market

Below bar chart depicts the top six job designations that opt for H1B visa and their mean salary over the years. Generally, wages for computer and mathematical occupations are higher than other occupations. The median wage level for H1B workers having computer or mathematical jobs is going up every year, with more and more people earning higher than $120,000 annually. During the recent years, computer technology is developing much faster than ever before and big data is booming as well, thus it’s not surprising to see the high demand for talents in the computer and mathematical areas. The bar chart depicts the top six job designations that opt for H1B visa and their mean salary.



Analysis of the applications received by state

The geo plot gives us the information about number of H1B visa applications received from individual states in USA over the years.


H1B Visa Applications By State
Number of H1B
60k
40k
20k
0