H1b Visa Petition Data Visualization and Analysis

Pooja Mahajan

College of Business and Economics

California State University East Bay  
Hayward, United States

[pmahajan4@horizon.csueastbay.edu](mailto:pmahajan4@horizon.csueastbay.edu)

Sagar Jethwa

College of Business and Economics

California State University East Bay  
Hayward, United States

[sjethwa@horizon.csueastbay.edu](mailto:sjethwa@horizon.csueastbay.edu)

Karisma Shaik

College of Business and Economics

California State University East Bay  
Hayward, United States

[kshaik4@horizon.csueastbay.edu](mailto:kshaik4@horizon.csueastbay.edu)

Sahana Sureshkumar

College of Business and Economics

California State University East Bay  
Hayward, United States

[ssureshkumar@horizon.csueastbay.edu](mailto:ssureshkumar@horizon.csueastbay.edu)

**PROJECT OVERVIEW AND PROBLEM STATEMENT**

An H1b visa is a business visa issued by the US government to international employees that gives permission to work in the USA. The H-1B visa category is for non-immigrant, employment-based temporary foreign workers in the US. A US company must make a job offer and file a petition for an H-1B visa with the US immigration agency before a foreign national can apply for one. When an international student has completed college or higher education (Master's or Ph.D.) and is employed full-time, they typically apply for and maintain this visa status. I have found this dataset on [H-1B Visas | Kaggle](https://www.kaggle.com/datasets/jmpark746/h1b-visas?select=h1b16.csv). This dataset contains information on all H1b applications filed by US companies for their employees from the year 2011 to the year 2016. This study will use R to apply statistical analytic techniques to this dataset to extract important findings on trends, probabilities, and other statistical insights regarding the acceptance of H1b visas. Data is quite good and already in tidy format but still, it requires some tidying.

**Requirements for H1B:**

• Filing of H1B petition before the deadline – As there are certain limitations on issuing the number of H1B visas each fiscal year, you must timely file your H1B petition.

• Check eligibility for H1B cap-gap extension – If you have filed an H1B visa petition, but your case is still pending, you may become eligible for a cap-gap extension wherein you can remain in the U.S. on F1 visa status until your H1B employment is approved.

**STRATEGY AND EXECUTION PLAN**

To get an in-depth understanding of what attributes of data we would be working on, we will go through the dataset and the available metadata information on the Kaggle website.

**Data Overview**

These sample datasets tend to be revised once a year, barring errors.

**Unnamed**: 0 ID of the row.

**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 the H1-B 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 conditions, as classified by the Standard Occupational Classification (SOC) System.

**JOB\_TITLE**: Title of the job using which we can filter specific job positions e.g., Data Scientist, Data Engineer, etc.

**FULL\_TIME\_POSITION**: Whether the application is for a full-time position or for a part-time position. Y = Full-Time Position; N = Part-Time Position

**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.

**YEAR**: The year in which the H-1B visa petition was filed.

**WORKSITE**: The foreign worker’s intended area of employment. We will explore the relationship between prevailing wages for Data Scientist positions across different locations.

**Lon**: Longitude of the employer worksite.

**Lat:** Latitude of the employer worksite.

**DATA MINING**

We would be collecting the data from different datasets available related to H1b filings on Kaggle throughout the years and merge them to serve as the dataset for our further study. The dataset contains about 3 million records and 11 columns from 2011-2016.

**DATA CLEANING**

**Sampling:**

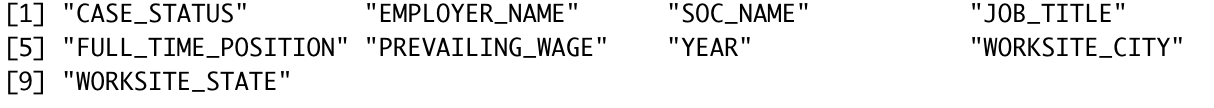
We would be collecting the data from different datasets available related to H1b filings on Kaggle throughout the years and merge them to serve as the dataset for our further study. We extracted 10% of the data with random sampling.

Records after sampling: 300K

**Removal of unused columns and NA values:**

In the visualization of this dataset, we don’t require X, Lon, and lat columns therefore they will be removed from the sampled dataset. Also, the column WORKSITE has information on the city and state which we require to visualize the data in a deeper manner, therefore, the WORKSITE column is split into two columns such as STATE and CITY.

The NA values are not reflected in the dataset and this column is not significant to our analysis.



As there are ‘3002458’ rows before the Data cleaning and finally we got ‘2983619’ rows after the Data cleaning.

**DATA ANALYSIS**

First, we consider the total applications approved over the year and we found that the number of approved applications increases over the year. It’s probably because every year companies file more h1b applications.

**Analysis of all cases of 2011-2016**

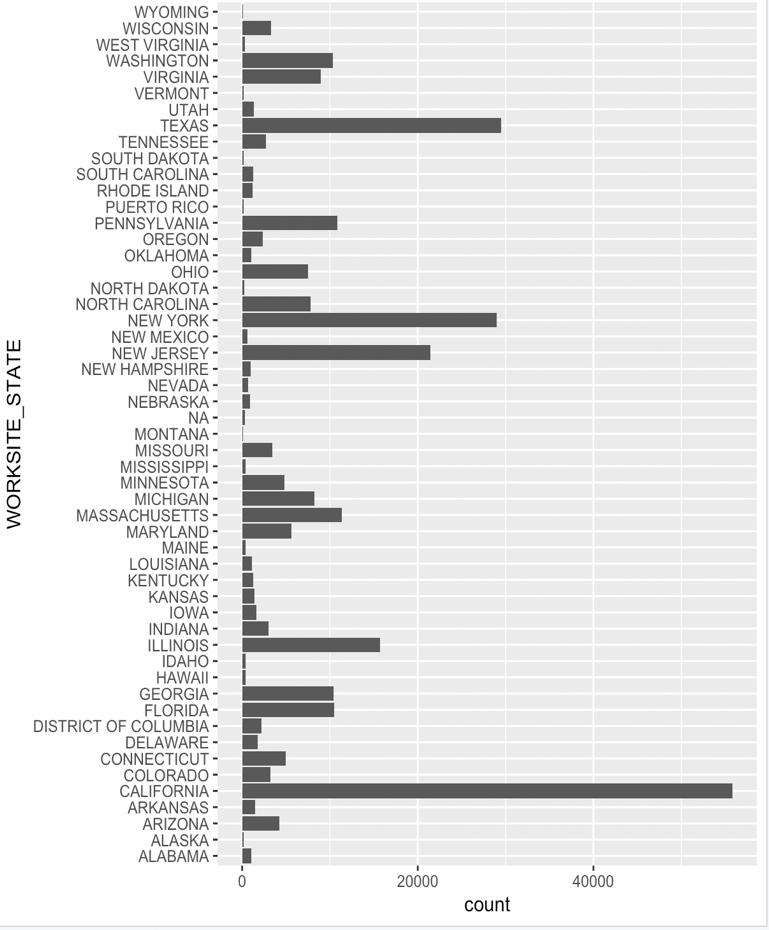
First, we consider the total application approved over the year.

Chart, bar chart

Description automatically generated

From the above graph, we can see that the number of approved applications increases over the year. It’s probably because every year companies file more h1b applications.

To get to know about the dataset and h1b application, we need to have information that which state has the highest number of an application filed:



From the bar chart, it is clearly shown that California has the highest number of h1b applications followed by Texas and New York. The reason behind California being at the top is obvious, SILICON VALLEY and San Francisco, where most of the companies have their offices.

To get more understanding that which are the top 5 states which have the highest number of h1b applications filed:

Chart, bar chart

Description automatically generated

From the faceted graph by year, we can say that California has the highest number of applications, but if we notice Texas and New York from 2011 to 2016, applications from Texas are increasing significantly more than those from New York. The reason behind this, because of the vast number of international students in Texas.

We have come to know about States so far but to get to know specifically which city of the state has the highest number of applications.

Chart

Description automatically generated

We have found that California is the top state but from the above-faceted plot, it is clearly shown that New York City has the highest number of applications. We might be thinking that this could be one of the cities of California, but it’s not. The reason might be that, In California, there are so many regions(cities) where the IT sector is there, not largely in one city. But New York, New York City is the heart of the USA and most of the companies have their offices in NYC only in New York State. That’s why New York City has the highest number of applications.

So far, we have come to know that California is the top state and New York City is the top city for the highest number of applications. The next important factor is the position, for which position most H1b applications are filed.

Chart, bar chart

Description automatically generated

From the above analysis, Programmer Analyst is the most demanding position in the USA for H1b

immigrants followed by Software engineers and Software Developers.

Chart, timeline, surface chart

Description automatically generated

The above graph gives us information about applications by job title over the year. It is clearly shown in the faceted plot that the programmer analyst is at the top. The strange thing is, in 2011, the job title “Technology Lead – US and Technology Analyst – US” is not demanding but over the period of 6 years, in 2016, it increases significantly, the reason may be because over this period, so many new technologies come in the market and companies need leaders to lead their projects or migrating their projects from old to new technology.

The interesting thing to know about the job title vs state is which state demands which role?

Chart, timeline

Description automatically generated

From the above analysis, it gives us the same result that a Programmer Analyst is highly demanding but if we look at California graphs, we will find out that in California, there is a need for a Software Engineer more than a Programmer Analyst. This is because of Silicon Valley and more technology companies in the California region that require employees who can implement their projects flawlessly.

The most crucial factor to determine in this dataset for international immigrants is the list of employers who filed most h1b applications.

Chart, bar chart

Description automatically generated

The above graph describes the relationship between Employers and the applications filed by them. We were all under the impression that big IT giants like Google, Facebook, Microsoft, etc. might file the highest h1b applications, but it’s not, It's Infosys Limited, that fields the highest number of h1b applications followed by Tata Consultancy Service. The reason could be that both are Indian companies, and they require their employees from India here in the USA to provide the best services to their clients.

Chart, pie chart

Description automatically generated

The following Pie chart gives us the information that about 87% of the applications get approved and around 4% of the applications are denied.

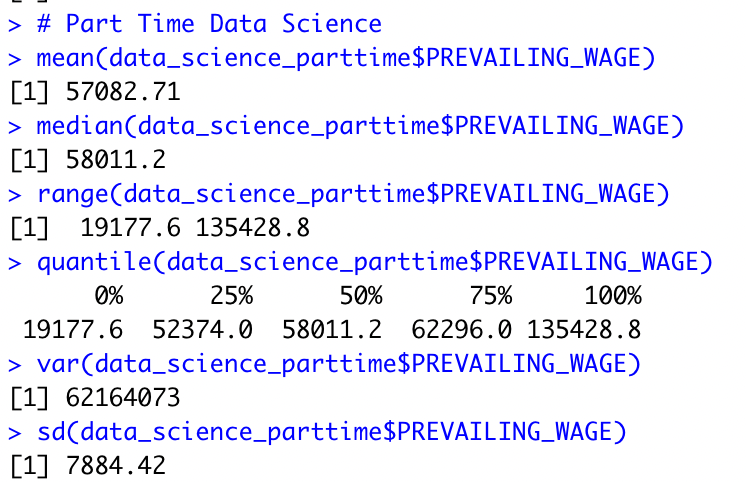
So, from the above analysis, we get the information about which state and city have the highest approval rate and which employer and job title are safer for approval of applications.

**DATA VISUALIZATION**

As Business Analytics students we focused more on the data-related roles and the dataset has been filtered on the basis of the JOB\_TITLE column for identifying Data science full-time and part-time employees. Job Titles like Data Science, Business Analyst, Data Analyst, and Data Engineer are categorized into Data Science jobs for analysis.

The PREVAILING\_WAGE column is used for finding the average wage of Data Science full-time/Part-time employees. Therefore, more analysis is done for median, range, Standard deviation, Quantile, and variance.

Text

Description automatically generated with medium confidence

**The coefficient of variation full-time is 2059.189.**

**The coefficient of variation part-time is 13.81227**

**DATA MODELING**

 Using the statistical methods and libraries in R, we would create models which would be helpful to derive statistical insights about the dataset like future predictions, feature engineering, and statistical analysis (T-Test, Z-Test). Here we are using Linear regression modeling.

* Feature importance
* statistical analysis
* GLM (Generalized Linear Model)

**DATA EXPLORATION**

The coefficient of variance is used for comparing two different datasets. And PREVALING\_WAGE is used for the comparison of differences in wages in these data sets. Hence, the coefficient of variance is not too high, which shows the level of dispersion around the mean is low.

Coefficient of variance = 13.81227

**PMF, CDF, and Expected value for accepted applications from 2011- 2016**

We calculated the PMF (probability mass function), and CDF (cumulative distribution function) for data science professionals, we can see In the PMF graph, 2011 has a Probability of around 0.08 and 2016 has a probability of around 0.25. Therefore data science jobs' acceptance probability has increased over the years as a result of policy changes or a growing demand for that particular position.

For 2011–2016, there is also a positive correlation between the Year and the number of visa applicants.

Chart, bar chart

Description automatically generated

**PMF for Accepted Applications from 2011–2016**

**Chart, histogram

Description automatically generated**

**CDF for Accepted Applications from 2011–2016**

**Number of Applicants per year for Data Science Jobs**

Barack Obama's presidency led to an increase in data science applications every year, but after the 2017 election, applications decreased because of the new policies. amended by the Trump administration. In 2016, around 12000 employees applied for H1-B visas.

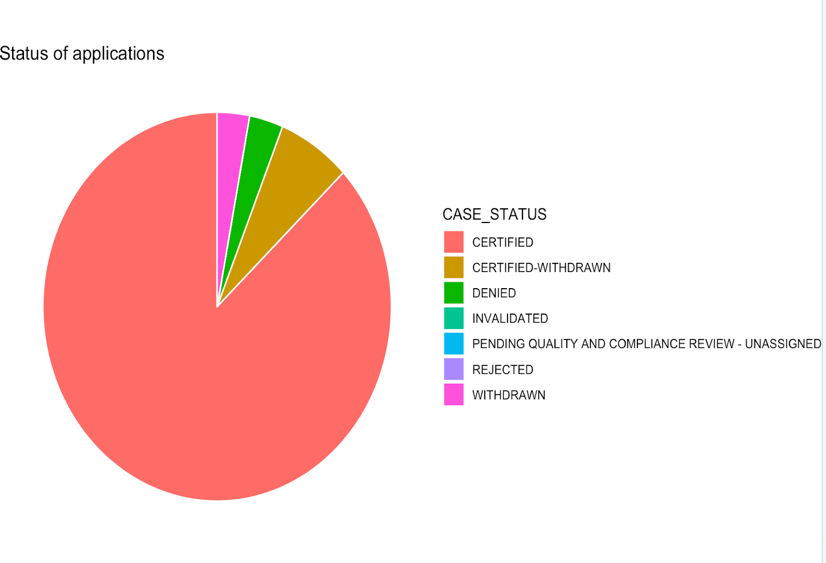
Chart, bar chart

Description automatically generated

**Number of Applications per year for Data Science Job**

**Distribution of H1B Visa Case Status**

The Pie Chart below shows the distribution of H1B visa status. Most cases are "Certified" in this dataset. Thus, we will only analyze CERTIFIED cases, which will provide more accurate insights into this scenario.



**Distribution of Prevailing Wage**

A histogram shows the distribution of the dominant wage vector simplest. Despite this, the dataset contains over 3 million documents, many of which have extreme values. Alternatively, the wage histogram can be displayed by randomly sampling about a tenth of the records and excluding the bottom 10% and top 5%.

Now we have the ideal prevailing wage histogram. The right tail of the distribution shows us that there are fewer foreign workers as the wages increase. This right-skewed distribution indicates that most foreign employees earn between $60,000 and $65,000 a year.

The biggest flaw in this histogram is that we didn’t adjust the wage for inflation. This chart includes all the data from 2011 to 2016.

Chart, histogram

Description automatically generated

**Histogram of Prevailing Wage**

**Prevailing wages of top 10 employees**

Microsoft's median salary is higher than any of the other firms, with a wage range of $0 to $15,000.

Tata Consultancy’s interquartile distribution of prevailing salaries is the smallest and wages for the middle 50% of H1B employees are the least variable at Tata Consultancy.

**Chart, box and whisker chart

Description automatically generated**

Chart, box and whisker chart

Description automatically generated

**Wage distribution for each year**

The distribution of wages for each year is right-skewed. Each year, the bulk of H1B applicants have salaries between $50,000 and $70,000.

**Chart, histogram

Description automatically generated**

**Statistical Analysis**

**One Sample Z-test**

X = R.V. of wage of an employee

Thus, our null and alternate hypotheses are:

H0: µ = 176240

H1: µ ≠ 176240

Z-Value = 0.1174008

The z value lies within (-1.96,1.96), we conclude that there is no significant difference between the sample mean wage and population mean wage. Hence, we fail to reject the null hypothesis

**One Sample t-test**

The sample t-test is for the Microsoft Prevailing wage compared to the Mean population salary

Thus, our null and alternate hypotheses are:

X = R.V. of wage of an employee

H0: µ = 84538.73

H1: µ ≠ 84538.73

The p-value is ≤ 0.05, we conclude that there is no significant difference between the sample mean of Microsoft wage and population mean wage. Hence, we reject the null hypothesis.

Text, letter

Description automatically generated

**Left tail Test**

Text

Description automatically generated

**Right tail test**

Text, letter

Description automatically generated

**Two tail tests**

Text, letter

Description automatically generated

**Two sample Z-test**

X1 = R.V. of wage of an employee from the first sample

X2 = R.V. of wage of an employee from the second sample

H0: µ1 − µ2 = 0

H1: µ1 − µ2 ≠ 0

Z-Value = -1.719223

The Significance level of alpha = 0.05. The z value lies within (-1.96, 1.96), we conclude that there is no significant difference between the two samples. Hence, we fail to reject the null hypothesis.

**Two Sample t-test**

X1 = R.V. of wage of a person from the first sample

X2 = R.V. of wage of a person from the second sample

H0: µ1 — µ2 = 0

H1: µ1 — µ2 ≠ 0

The p-value is ≥ 0.05, we conclude that there is no significant difference between the two samples. Hence, we fail to reject the null hypothesis.

Text

Description automatically generated

**Advanced Analytics and Conclusion: -**

In statistical analysis, logistic regression uses prior observations of a data set to predict a binary outcome, such as yes or no. In logistic regression, a dependent variable is predicted based on the relationship between several independent variables.

We have performed Logistic regression for predicting the acceptance of Visa applications based on Job Title, Full-time position, Year, and Wage of an employee.

Text

Description automatically generated with medium confidence

From this understanding, the P-value of Job Title and Year are more responsible in predicting the status of a visa application.

From this model Analysis, we determined the accuracy of the model on test data is 88%.

**Role of H1b Immigrants in the US Economy:**

According to my research on the internet, I have found out that h1b Immigrants play a crucial role in

growth of the US economy. It can be understood by the following figures:

• GDP of USA in 2011 - $15.52 trillion

• GDP of USA in 2016 - $18.56 trillion

• Total H1b applications approved – 2615623

• % change in the economy – 19.587% ▲

This figure shows that H1b immigrants play a crucial role in the growth of the US GDP with a 19.587% increase in 5 years.

**President Trump's policy proposal for H1b visa applications:**

President Trump has proposed a policy for H1B applications that companies can apply for only if employees have a yearly income of more than 130k. Data from the last 5 years (2011-2016) shows that, out of 3M+ applications, only 87095 applicants have yearly wages over 130K. If this proposal is approved, only 2.903% of applicants will be eligible for H1B.

**REFERENCES**

1) R for data science - <http://r4ds.had.co.nz/>

2) GDP data per capita over the years in the United

States:

[https://www.google.com/publicdata/explore?ds=d5bncppj](https://www.google.com/publicdata/explore?ds=d5bncppj%20%20of8f9_&amp;met_y=ny_gdp_pcap_cd&amp;idi%20%20m=country:%20USA:RUS:GBR&amp;hl=en&amp;dl=en)

[of8f9\_&amp;met\_y=ny\_gdp\_pcap\_cd&amp;idi](https://www.google.com/publicdata/explore?ds=d5bncppj%20%20of8f9_&amp;met_y=ny_gdp_pcap_cd&amp;idi%20%20m=country:%20USA:RUS:GBR&amp;hl=en&amp;dl=en)

[m=country: USA:RUS:GBR&amp;hl=en&amp;dl=en](https://www.google.com/publicdata/explore?ds=d5bncppj%20%20of8f9_&amp;met_y=ny_gdp_pcap_cd&amp;idi%20%20m=country:%20USA:RUS:GBR&amp;hl=en&amp;dl=en)

3) Kaggle dataset for years 2011-2016: H[H-1B Visa](file:///Users/sahana/Downloads/H-1B Visa  Petitions 2011-2016 | Kaggle)

[Petitions 2011-2016 | Kaggle](file:///Users/sahana/Downloads/H-1B Visa  Petitions 2011-2016 | Kaggle)