ANALYSIS OF WORK VISAS IN THE UNITED STATES

Project 1 in EMIS 7331 - Data Mining.

Nikhila Byreddy & Samarth Suresh Kumar
CSE/EMIS 7331, Project 1

Table of Contents

Executive Summary

In this report, using data analytics, we use the publicly available immigration data for the available work visas in the United states to analyze trends in the visa applications over time.

For a foreign national to work in the United States, an employer must offer the job and file a petition for an H1B visa with the US immigration department. The primary focus of this project is to analyze about the attributes in the dataset that contribute to the final visa status.

We used two different datasets to analyze trends: H1-B Data and US Permanent Work Visa.

For each dataset, before analyzing the data, cleaning is performed. All the duplicate data is removed, appropriate methods are implemented in dealing with the missing data and outliers. Some attributes are selected, and the analysis is performed individually.

This report shows the distribution of top employers, top job positions, Salary trends, top states and dissemination of case status. This gives us the understanding of how these variables are distributed in the dataset.

We have analyzed the relationship between different variables such as prevailing wages from 2011 to 2016 using boxplot, violin plots, prevailing wages for various job positions and the percentage of most successful states with their status as certified and trends in Visa applications each year. We further analyzed and visualized data within a company to understand salaries and case status trend , and distribution of immigrants from various countries.

Many foreign nationals seeking H1B visas are unclear about the visa process and suffer visa denials. This report helps them to analyze them about the chances of getting their status "Certified."

List of Figures:

List of Tables:

Data Understanding:

H1b Visa-Dataset:

H1B visas are a category of employment-based, non-immigrant visas for temporary foreign workers in the United States. Labor Condition Application(LCA) is a mandatory document that H1B Sponsor/employer needs to file with US Department of Labor before they submit an H1B petition with USCIS for any non-immigrant worker. As foreign workers are new to America, certain employers can take advantage and mistreat them regarding wages, benefits, etc. In this context, US Dept. of Labor had mandated LCA, and it is essential for a foreign worker as it protects their fundamental rights.

LCA form has essential information about the offered job position for the foreign worker. The fields in LCA form include Job title of the post provided, duration of job position if the position offered is full time or not, salary submitted for the position, location of job position, the prevailing wage for the same position in that area.

The primary goal of the analyst is to analyze the fields of LCA form that could influence the LCA status (Case status). One can interpret the statistics such as

Which employers send most number of H1B visa applications.

Is the number of petitions with a specified Job title increasing over time.

The Location which has the most number of Data Engineers.

1. Are the jobs concentrated in few specific regions?
2. The employers who file the most petitions per year.

The given dataset has 3,002,458 records. Important variables of data file and type of data is provided in the table below:

Table 1: H1b Visa- Dataset Variable info

Column Names	Type	Description	Statistics
X	Id	ID	
CASE_STATUS	Nominal	Case status of application	
EMPLOYER_NAME	Nominal	Name of the employer	
SOC_NAME	Nominal	Job title	
JOB_TITLE	Nominal	Title of the job	
FULL_TIME_POSITION	Nominal	True or false	
PREVAILING_WAGE	RATIO	Salary of the applicant	MIN: 0 MAX: 6.998e+09 Median: 6.502e+04 Mean: 1.470e+05

YEAR	Interval	Year of the application	
WORKSITE	Nominal	Location of the employer	
lon	Interval	Location - longitude	
lat	Interval	Location - latitude	

EMPLOYER_NAME can be used in comparing the prevailing wages, number of applications for different employers and finding the employers who have status as "Certified."

JOB_TITLE - We can find out Specific job position (E.g. Software Engineer) based on which we can compare the case status for a given job title.

PREVAILING_WAGE - We can explore the relationship between prevailing wage, job title.

WORKSITE - Using which we can compare prevailing wage for a given position (software Engineer) for various work locations.

CASE_STATUS - This helps us to analyze how many H1B Visas are certified by different employers.

US Permanent Visa Dataset:

A permanent labor certification issued by the Department of Labor permits an employer to hire a foreign worker to work permanently in the US. Roughly 140,000 immigrant visas are available each fiscal year for foreigners (and their spouses and children) who seek to immigrate based on their job skills. With the right combination of skills, education, and work experience, one may be able to live permanently in the United States.

An employer can submit an immigration petition to USCIS; the employer must obtain an approved labor certification from the U.S. Department of Labor (DOL). The DOL labor certification verifies the following:

- There are insufficiently available, qualified, and willing U.S. workers to fill the position offered at the prevailing wage
- Hiring a foreign worker will not adversely affect the wages and working conditions of similarly employed U.S. workers

Data covers 2012-2017 and includes information on the employer, position, wage offered, job posting history, employee education and past visa history, associated lawyers, and final decision.

We breakdown the 154 columns available in the US Permanent visa and describe them as follows:

Table 2 : US Permanent Visa Dataset info

Column names	Description
add_these_pw_job_title_9089	Job Title
agent_city	Agent Details.
agent_firm_name	
agent_state	
application_type	Application/Case Details
case_no	
case_number	
case_received_date	
case_status	
class_of_admission	
country_of_citizenship	
country_of_citzenship	
decision_date	
employer_address_1	Employer Details.
employer_address_2	
employer_city	
employer_country	
employer_decl_info_title	
employer_name	
employer_num_employees	
employer_phone	
employer_phone_ext	
employer_postal_code	
employer_state	
employer_yr_estab	
foreign_worker_info_alt_edu_experience	Foreign Worker Details.
foreign_worker_info_birth_country	
foreign_worker_info_city	
foreign_worker_info_education	
foreign_worker_info_education_other	
foreign_worker_info_inst	
foreign_worker_info_major	
foreign_worker_info_postal_code	
foreign_worker_info_rel_occup_exp	
foreign_worker_info_req_experience	
foreign_worker_info_state	
foreign_worker_info_training_comp	
foreign_worker_ownership_interest	
foreign_worker_yr_rel_edu_completed	
fw_info_alt_edu_experience	
fw_info_birth_country	
fw_info_education_other	

fw_info_postal_code	
fw_info_rel_occup_exp	
fw_info_req_experience	
fw_info_training_comp	
fw_info_yr_rel_edu_completed	
fw_ownership_interest	
ji_foreign_worker_live_on_premises	Job Information
ji_fw_live_on_premises	
ji_live_in_dom_svc_contract	
ji_live_in_domestic_service	
ji_offered_to_sec_j_foreign_worker	
ji_offered_to_sec_j_fw	
job_info_alt_cmb_ed_oth_yrs	
job_info_alt_combo_ed	
job_info_alt_combo_ed_exp	
job_info_alt_combo_ed_other	
job_info_alt_field	
job_info_alt_field_name	
job_info_alt_occ	
job_info_alt_occ_job_title	
job_info_alt_occ_num_months	
job_info_combo_occupation	
job_info_education	
job_info_education_other	
job_info_experience	
job_info_experience_num_months	
job_info_foreign_ed	
job_info_foreign_lang_req	
job_info_job_req_normal	
job_info_job_title	
job_info_major	
job_info_training	
job_info_training_field	
job_info_training_num_months	
job_info_work_city	
job_info_work_postal_code	
job_info_work_state	
naics_2007_us_code	North American Industry Classification System
naics_2007_us_title	
naics_code	
naics_title	
naics_us_code	
naics_us_code_2007	
naics_us_title	
naics_us_title_2007	
orig_case_no	Original Case Information

orig_file_date	
preparer_info_emp_completed	Preparer Information
preparer_info_title	
pw_amount_9089	Prevailing Wage Information.
pw_determ_date	
pw_expire_date	
pw_job_title_908	
pw_job_title_9089	
pw_level_9089	
pw_soc_code	
pw_soc_title	
pw_source_name_9089	
pw_source_name_other_9089	
pw_track_num	
pw_unit_of_pay_9089	
rec_info_barg_rep_notified	Recruitment Information
recr_info_barg_rep_notified	
recr_info_coll_teach_comp_proc	
recr_info_coll_univ_teacher	
recr_info_employer_rec_payment	
recr_info_first_ad_start	
recr_info_job_fair_from	
recr_info_job_fair_to	
recr_info_on_campus_recr_from	
recr_info_on_campus_recr_to	
recr_info_pro_org_advert_from	
recr_info_pro_org_advert_to	
recr_info_prof_org_advert_from	
recr_info_prof_org_advert_to	
recr_info_professional_occ	
recr_info_radio_tv_ad_from	
recr_info_radio_tv_ad_to	
recr_info_second_ad_start	
recr_info_sunday_newspaper	
recr_info_swa_job_order_end	
recr_info_swa_job_order_start	
refile	
ri_1st_ad_newspaper_name	
ri_2nd_ad_newspaper_name	
ri_2nd_ad_newspaper_or_journal	
ri_campus_placement_from	
ri_campus_placement_to	
ri_coll_tch_basic_process	
ri_coll_teach_pro_jnl	
ri_coll_teach_select_date	
ri_employee_referral_prog_from	

ri_employee_referral_prog_to	
ri_employer_web_post_from	
ri_employer_web_post_to	
ri_job_search_website_from	
ri_job_search_website_to	
ri_layoff_in_past_six_months	
ri_local_ethnic_paper_from	
ri_local_ethnic_paper_to	
ri_posted_notice_at_worksite	
ri_pvt_employment_firm_from	
ri_pvt_employment_firm_to	
ri_us_workers_considered	
schd_a_sheepherder	
us_economic_sector	
wage_offer_from_9089	Wage Information.
wage_offer_to_9089	
wage_offer_unit_of_pay_9089	
wage_offered_from_9089	
wage_offered_to_9089	
wage_offered_unit_of_pay_9089	

Out of the 154 columns, the following variables were of most significant which had valid data:

Table 3 : PERM visa description.

Column Name	Type	Description	Statistics
CASE_STATUS	Nominal	Status of the application	
PW_SOC_TITLE	Nominal	information of the Status of an application.	
EMPLOYER_NAME	Nominal	Name of the employer	
PW_AMOUNT_9089	RATIO	Salary of the applicant	MIN : 6 MAX : 13528320 MEAN : 85350 MEDIAN : 85675
COUNTRY_OF_CITIZENSHIP	Nominal	Country of citizenship of the applicant	
EMPLOYER_CITY	Nominal	City of the employer	
LOCATION	Ordinal	Added Later (Lat and Lon from Employer City)	

Data Quality:

H1b Visa-Dataset:

The first column of H1B visa dataset is just a row count. Below is the analysis on the important variables in our dataset:

CASE_STATUS:

There are 13 NA s, one blank (") for CASE_STATUS. However, same CASE_STATUS occurs more than once. We should check if there is a necessity to remove them. Checking the status reveals that the possible status of any employee should be "CERTIFIED," "CERTIFIED-WITHDRAWN," "WITHDRAWN," "DENIED," INVALIDATED," REJECTED," UNASSIGNED." Checking the status shows that no mistakes occurred because different employees may have the same case status.

EMPLOYER_NAME:

There are missing values for EMPLOYER_NAME. The same EMPLOYER_NAME existed in the dataset. There is no mistake in the repetition of EMPLOYER_NAME because the employer may be filing LCA form for different job titles.

SOC_NAME:

There are eleven missing values in the dataset we chose (one "N/A" and ten "NA") for SOC_NAME. Duplicate data is not a problem here because different employers may file a petition for the same case. As SOC_NAME is nominal variable, all the missing values are replaced by "NA" for consistency. These missing values may affect our analysis.

JOB_TITLE:

There is missing data for JOB_TITLE. It is a nominal variable and replaces it with "NA." Also, there are duplicate data for the job title. But this should not be a problem because an employer may need to file a petition for different employees with different job titles.

FULL_TIME:

There is one missing value for the employer "Four seasons heating and air conditioning," for which the CASE_STATUS is "Denied." As there is only one lost value in the entire dataset, this missing value should not make much difference in our analysis. The ratio of missing values and the existing values are minimal. Also, there are duplicate values but that should not be a problem because the possible answers if the employee is full time or not is Y(Yes) or N(No) or a missing value.

PREVAILING_WAGE:

There are some missing values in the prevailing wage. There are some wages which are in billions. One may replace all the salaries which are greater than one million(1e6) with NA. Also, the minimum wage is zero which is again exceptional. Prevailing wage cannot be zero because a foreign professional must receive at least some fixed prevailing wage to be eligible to file for an H1B visa. One may replace this with the median. Also, duplicate salaries exist. But this should not be a problem because different employers may have same prevailing wages.

YEAR:

There are no missing values in YEAR. Duplicate values do exist, but this should not be a problem because the given dataset consists of all the petitions for H1B by the employers in the years 2011 to 2016.

WORKSITE:

There are some missing locations in the dataset. Although data about the state or country is present, the exact location is missing. Duplicate locations exist but that should not be a problem because there may be different employers in the same area.

LON:

There are many of missing values in longitudes. Duplicate longitudes exist but that should not be a problem because there may be different employers in the same location.

LAT:

There are a lot of missing values in latitudes. Duplicate latitudes exist but that should not be a problem because there may be different employers in the same location.

Here the number of missing values of both latitudes and longitudes are same.

After cleaning and removing the duplicate entries, we have 2068225 records.

US Permanent Visa Dataset:

There were no complete duplicate entries found and the no. of unique rows were 374,362

0 Complete cases found in the entire dataset.

CASE_STATUS: 0 NA Values.

PW_SOC_TITLE:

- 2336 data points had whitespace values which were discarded.
- 0 NA Values.

EMPLOYER_NAME:

- 12 data points had whitespace values which were discarded.
- 0 NA Values.

PW_AMOUNT_9089:

- Minimum = 6, Max = 13528320
- 2216 NA Values.
- Rows with value greater than 500000 = 46.
- Rows with value less than 60,000 (Minimum wage to attain H1b) = 74592

COUNTRY_OF_CITIZENSHIP:

- 0 NA Values.
- 20633 data points had whitespace values which were discarded.
- Duplicate column with name typo, not considered for analysis.

EMPLOYER_CITY:

- 0 NA Values.
- 12 data points had whitespace values which were discarded.

LOCATION: - Added later using geocode to fetch geolocation of the employer city.

LAT: - Ordinal - Latitude

LON: - Ordinal – Longitude

Single Variable Visualizations:

With data representing our dataset, we obtained the following plots using R and ggplot:

H1b Visa-Dataset:

Case status distribution using CASE_STATUS:

CASE_STATUS is a Nominal variable which is visualized in the bar plot given below. We notice that majority of applicants have their status as "Certified"(around 1.75 million).

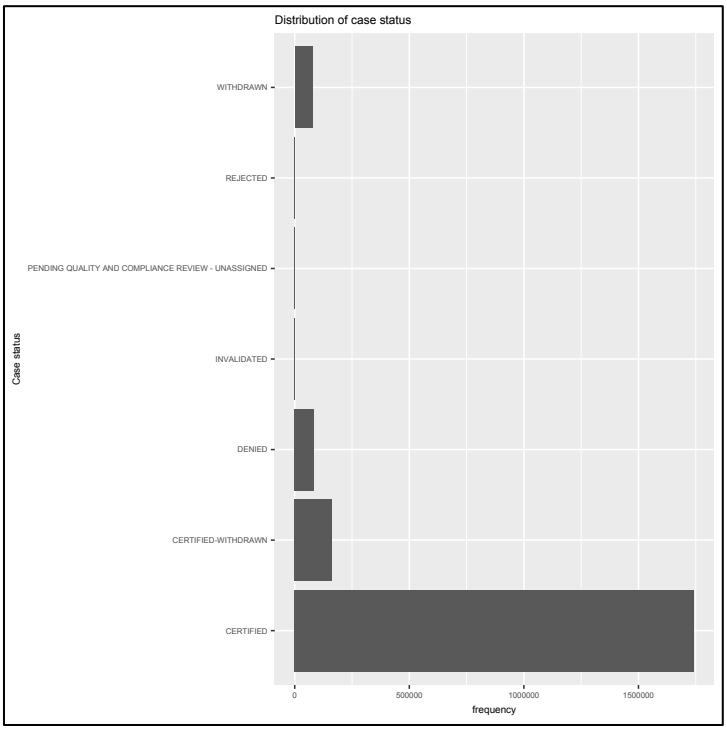

Figure 1 : Case Status Distribution H1b

There is smaller number of applicants with status "CERTIFIED-WITHDRAWN"(163,228), "DENIED"(84505). Number of applicants with status "COMPLIANCE REVIEW - UNASSIGNED"(15), "REJECTED"(2) and "INVALIDATED"(1) are in even smaller number.

Top 20 Employers using EMPLOYER_NAME:

It gives us the information about the frequency of the Employers in the dataset from which we may analyze about the employers who filed H1B visas.

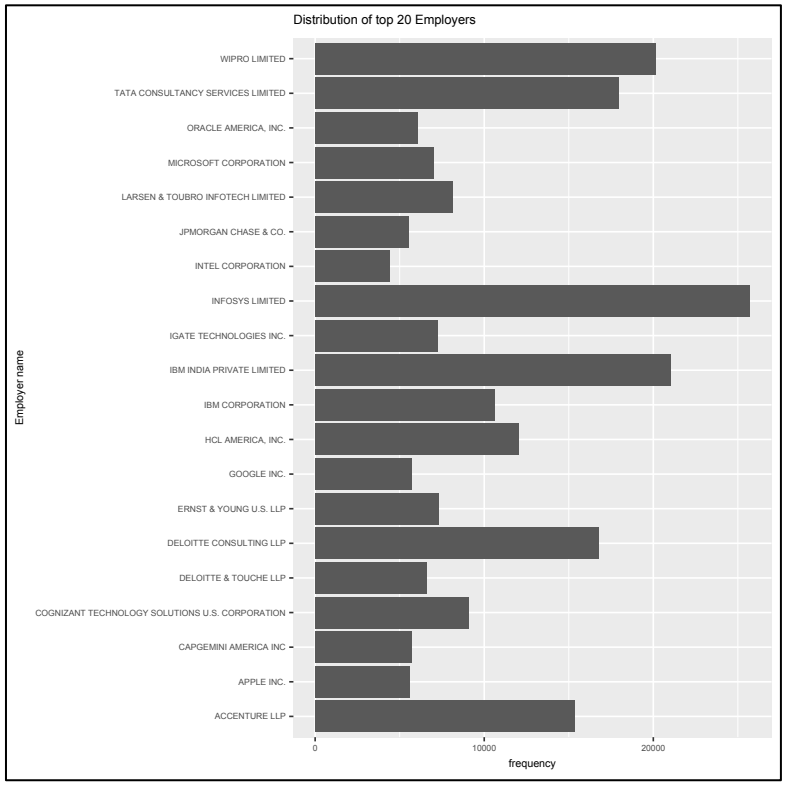

Figure 2: Distribution of top 20 H1b Employers

Represented top 20 employers based on their occurrences in the H1B dataset using a bar plot. Employer name is Nominal, and for plotting the categorical data, bar plot is selected. Also, to illustrate the distribution of employers for the given dataset.

We see that most frequent applications are for PROGRAMMER ANALYST (132381), SOFTWARE ENGINEER (66416), SYSTEMS ANALYST (34833).

Top 20 Jobs using JOB TITLE:

JOB_TITLE is Nominal variable. For better understanding, top 20 job titles based on their occurrences in the H1B dataset is plotted using histogram.

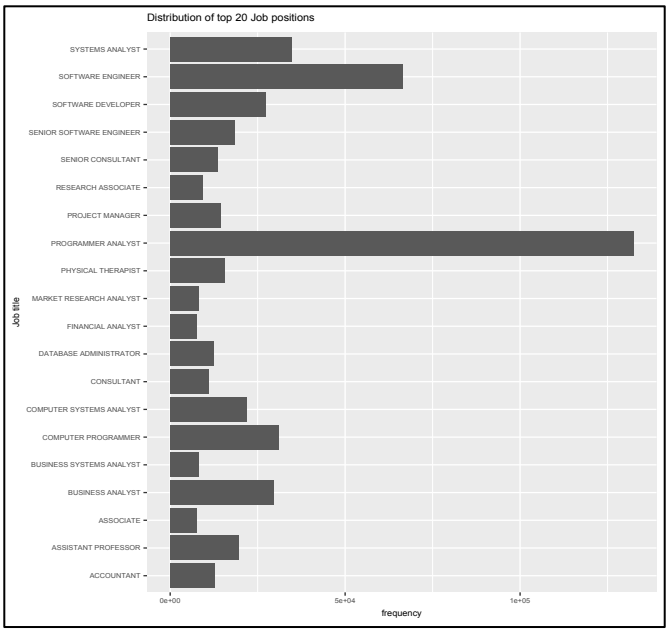

Figure 3:Distribution of top 20 H1b positions

Bar plot gives us the better understanding of the job positions of the employees, for which H1B visas are filed. We may analyze that more significant amount of H1B visas are submitted for Programmer Analysts.

Salary densities using PREVAILING_WAGE:

As the prevailing wage is an Ordinal data type, a histogram is selected for visualizing the salaries in the H1B dataset. In this case, to understand about prevailing wages in our dataset histograms are a better option.

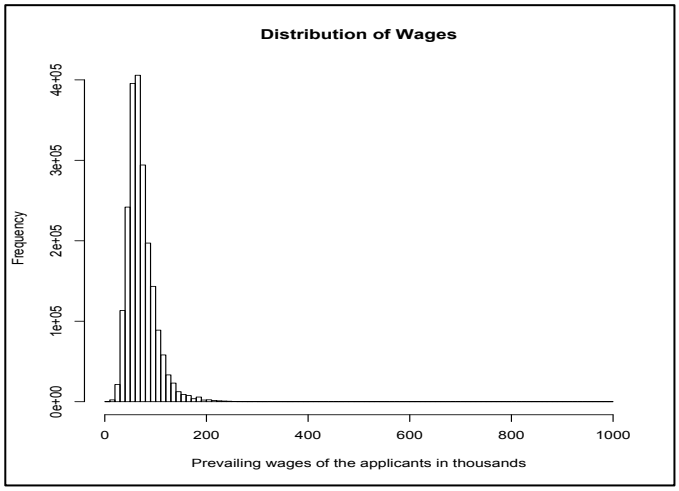

Figure 4 : Salary densities for H1b Applicants

After cleaning the data, i.e., after removing all the exceptional wages, we now have all the salaries below 1,000,000 USD. Most of the prevailing wages are below 200,000 USD. We see that most common wages are 60,000 USD (4732 times), 62566(3028), 55245(2891).

STATE (newly create column for easier analysis):

For the given dataset, Worksite has both city and state. Performed analysis of the state which is made a separate column for better understanding. In this case, State is a Nominal variable and to analyze the occurrences of the state in the dataset;

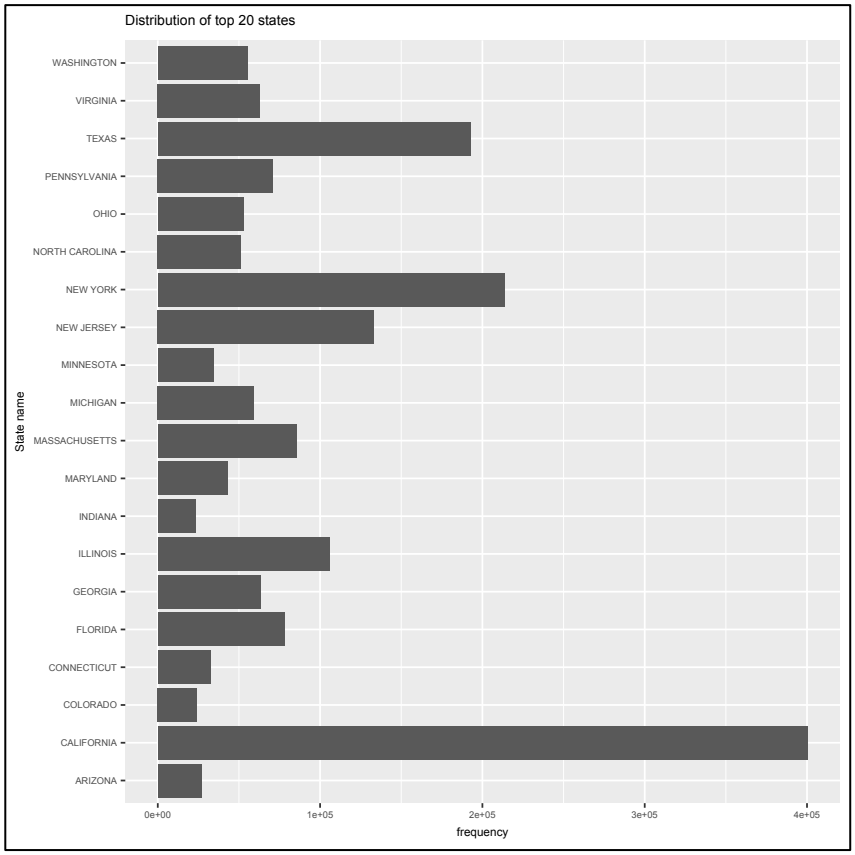

Figure 5 : Distribution of top 20 states with H1b

More in-depth analysis can be done if we separately investigate the applicants based on their state. California, New York and Texas are top states for H1B Visa Applications.

US Permanent Visa Dataset:

Case status distribution using CASE_STATUS:

CASE_STATUS is a Nominal variable whose frequencies are visualized in the bar plot given below:

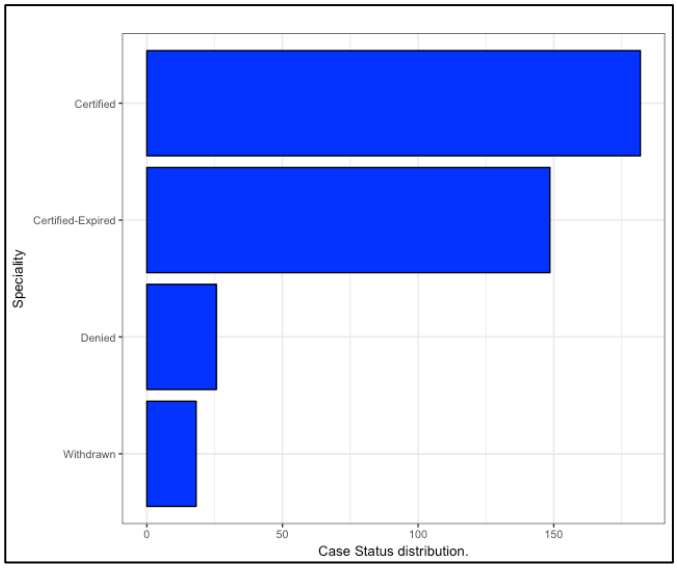

Figure 6 : Case Status distribution (PERM)

There are 150 thousand certified applicants which account for a high percentage of the total no. of applications for Permanent Visas. There is also a high no. Of certified-expired applicants which are applicants whose employers fail to file petition an alien to work in the US within the required timeframe.

Top 10 Job Title distribution using PW_SOC_TITLE:

PW_SOC_TITLE is a Nominal variable whose frequencies are visualized in the histogram given below:

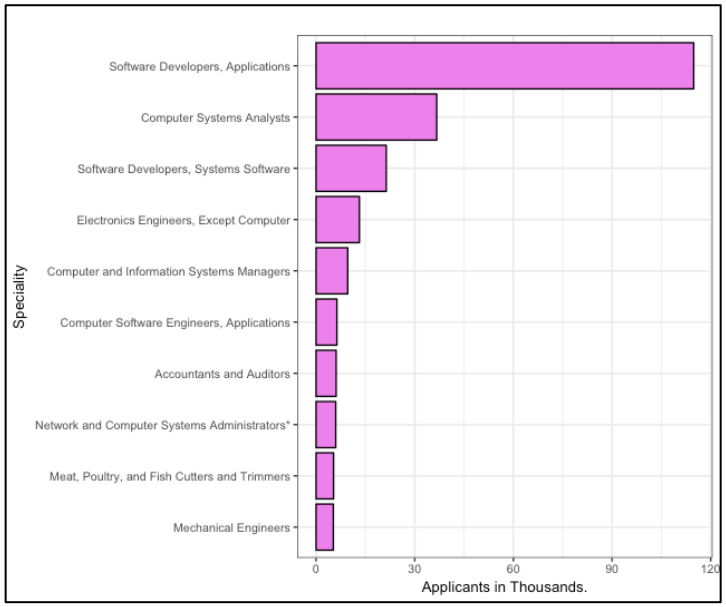

Figure 7 : Top 10 Job Titles (PERM)

Observation: Software Developers is one of the topmost applicants for Permanent Visas regarding frequencies.

Based on the criteria for US Department of Labor criteria, we can also interpret by assuming there is an evident deficit in US Citizens who are skilled Software Developers.

Case status distribution using EMPLOYER_NAME:

EMPLOYER_NAME is a Nominal variable whose frequencies by “Certified” and “Denied” - CASE_STATUS is visualized in the histogram:

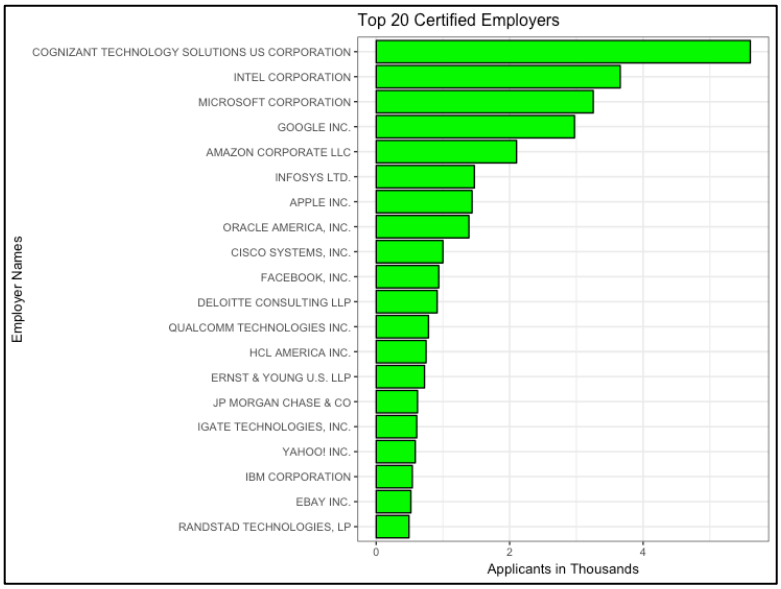

Figure 9 : Top 20 Certified Employers (PERM)

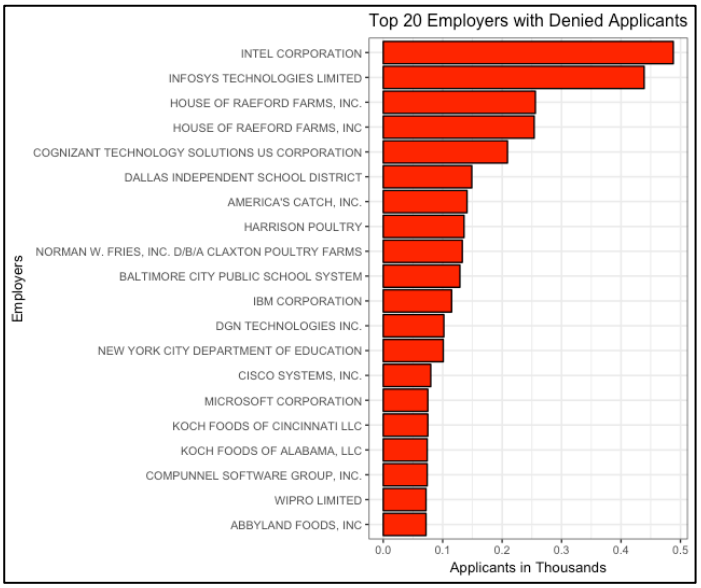

Figure 8 :Top 20 Denied Employers (PERM)

Analyzing the top 20 companies with certified status applicants shows Cognizant has the highest number of applicants with certified applicants followed by Intel.

For an immigrant employee who is interested in being a permanent alien can consider applying in these top companies.

Analyzing the top 20 companies with denied status applicants shows Cognizant has the high number of applicants with certified applicants followed by Infosys.

It's interesting to see Intel in the top 3 employers with both Certified and Denied status companies.

Relationships Visualizations:

H1b Visa-Dataset:

Boxplot of Salaries per year using PREVAILING_WAGE VS YEAR:

For the year 2016 Prevailing wage is higher, year 2012 has the highest wage (1,000,000 USD). There are many outliers in the plot, for better analysis, limit the prevailing wage from 1 million to 150,000 USD.

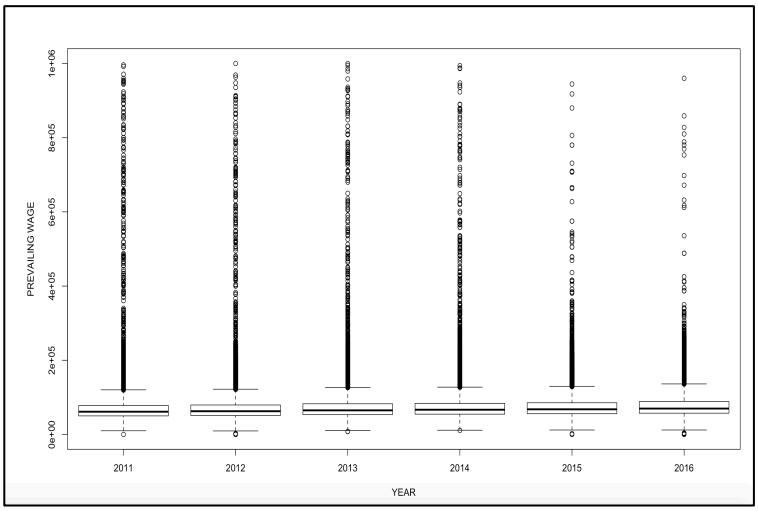

Figure 10 : Box Plot for salaries vs year (H1b)

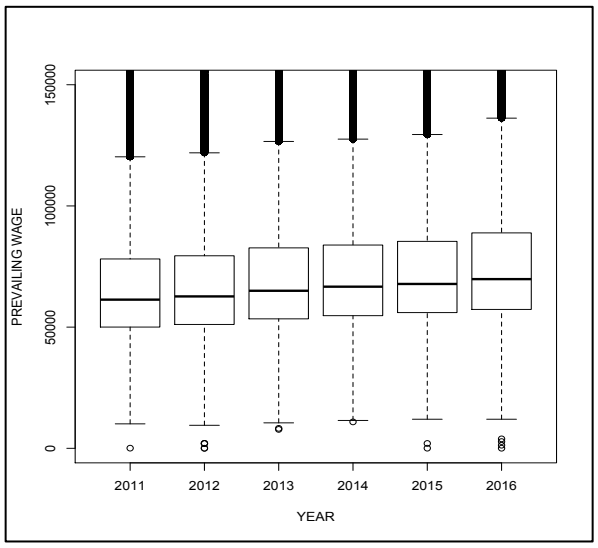

Figure 11: Box Plot for salaries vs year after Limits (H1B)

For the above box plot, the prevailing wage is limited to give a better understanding of the current salaries every year. PREVAILING_WAGE has increased linearly from the year 2011 to 2016. This may be because of increment in the wages or also because of the increase in the number of applicants for an

H1B visa every year. The median weights of the groups of cereal boxes are similar. That means the median of prevailing wages for all the years from 2011 to 2016 is towards the lower Quartile. Also, most of the salaries are in the range of 50k to 80k USD. For the years 2013 and 2014 the outliers (consider for minimum side) are closer to the smallest of a box plot, from which we may analyze that the minimum wages for the years 2013 and 2014 are higher when compared to other years.

Box plots of salaries of top 5 Job title using JOB_TITLE Vs PREVAILING_WAGE:

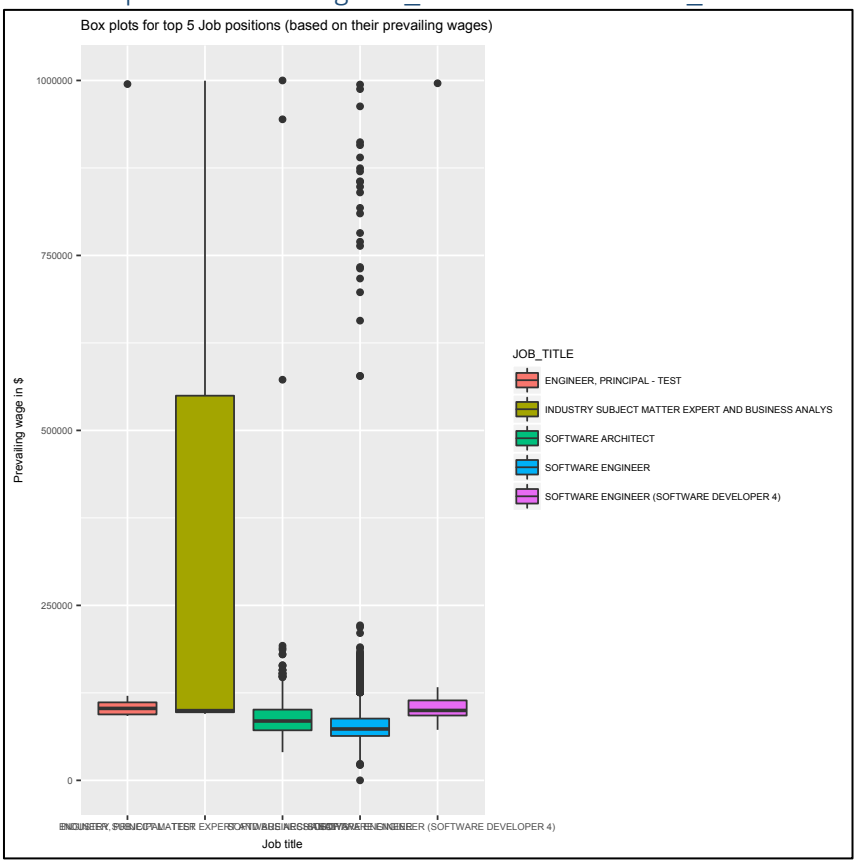

Figure 12 : Box Plots for salaries for Top 5 Job titles by frequency.

For the above box plot, the prevailing wage is limited to give a better understanding of the current salaries every year. PREVAILING_WAGE has increased linearly from the year 2011 to 2016. This may be because of increment in the wages or also because of the increase in the number of applicants for an H1B visa every year. The median weights of the groups of cereal boxes are similar. That means the median of prevailing wages for all the years from 2011 to 2016 is towards the lower Quartile. Also, most of the salaries are in the range of 50k to 80k USD. For the years 2013 and 2014 the outliers (consider for minimum side) are closer to the smallest of a box plot, from which we may analyze that the minimum wages for the years 2013 and 2014 are higher when compared to other years.

Temporal data of case status per year:

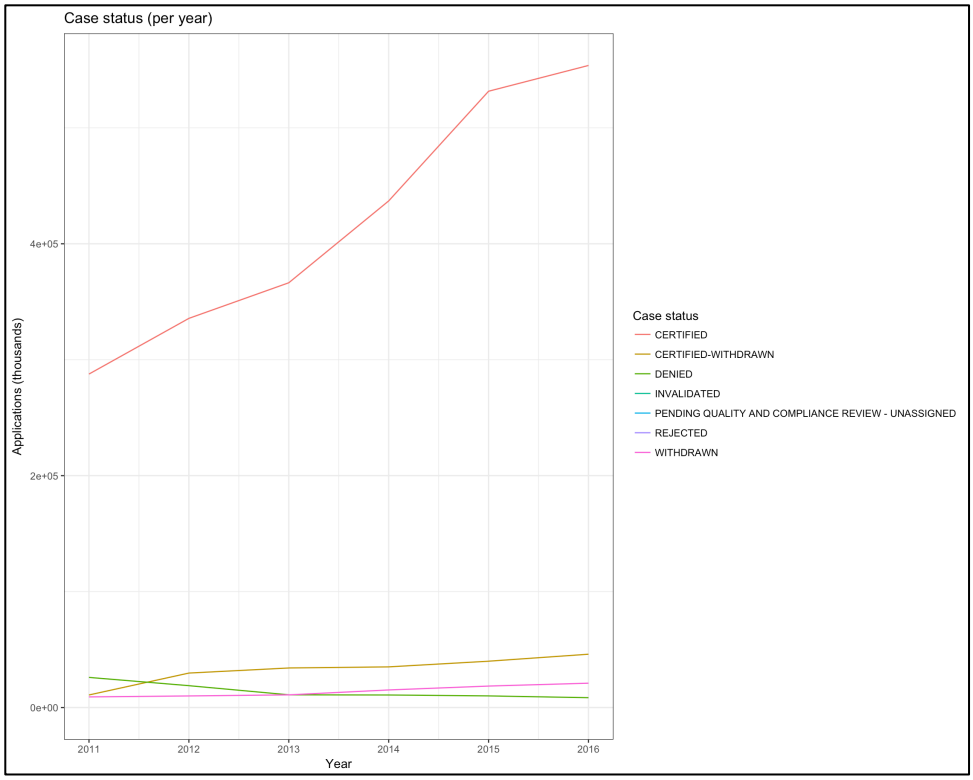

Figure 13 : Case status by year (H1b)

Plot shows a steady year on year growth of certified applicants. No. of applicants getting rejected are reducing year on year too which is great of interested immigrants.

US Permanent Visa Dataset:

Map plot using employer_cities:

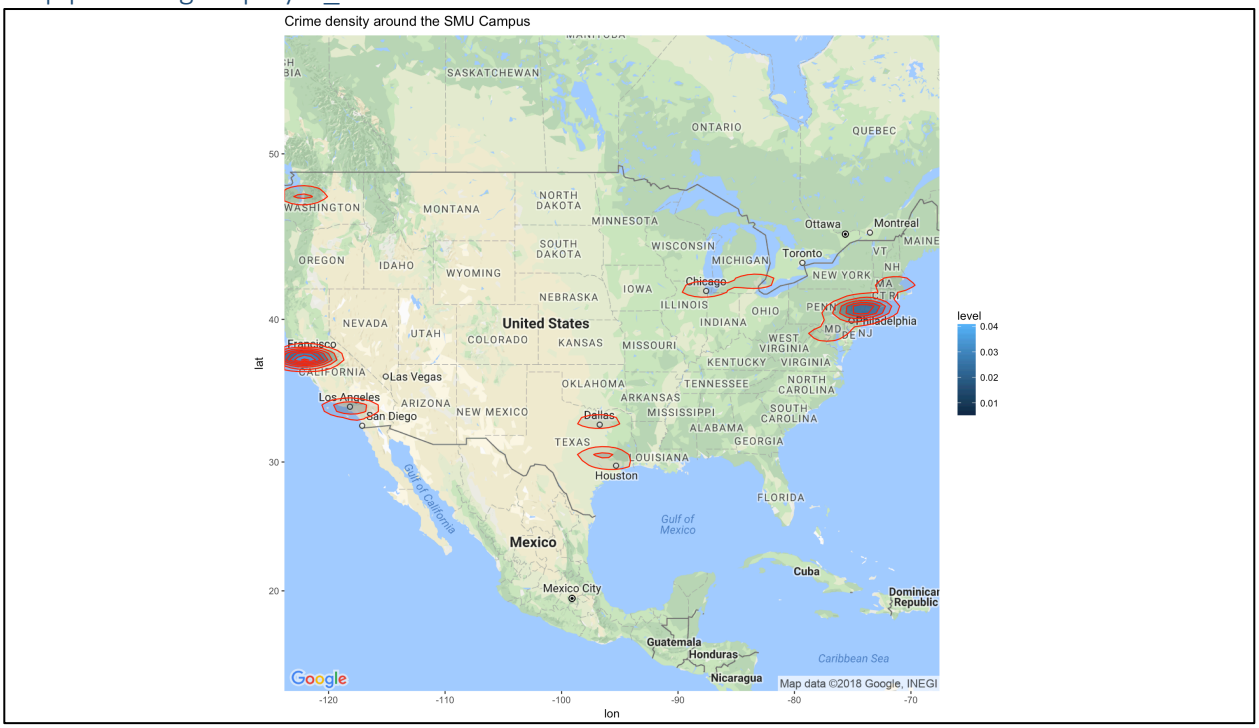

Figure 14 : Map data of Densities of applicants by city. (PERM)

After fetching latitudes and longitudes for the employer_cities data, we plot the densities in the applicants per city. We see that California and New York are top hubs for immigrant applicants along with a handful of cities with comparatively low frequencies.

Company - Intel's Salary Analysis

We can further investigate the data by taking a subset of the dataset by company. We chose Intel Corporation because tops amongst both Certified and Denied applications. We find useful insights about most occurring jobs, salaries and country of applicants.

Plot below of sorted salaries shows us a range of salaries within intel starting from below $50k to above $200k. The points below and above extreme cases are very few; these data points can be considered as outliers.

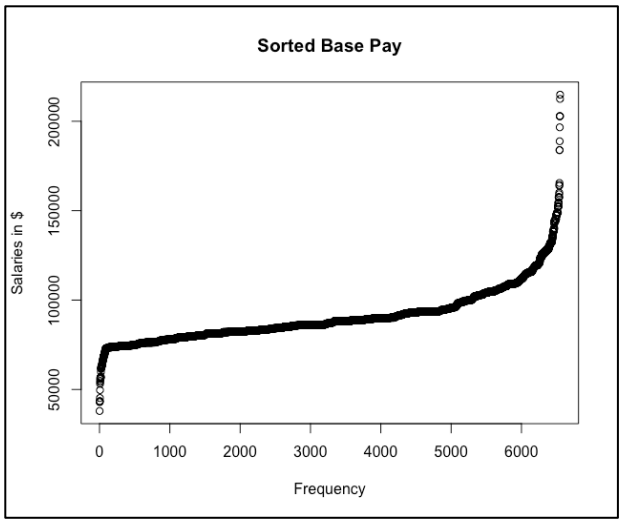

Figure 15 : Scatter plot for sorted salaries in Intel (PERM)

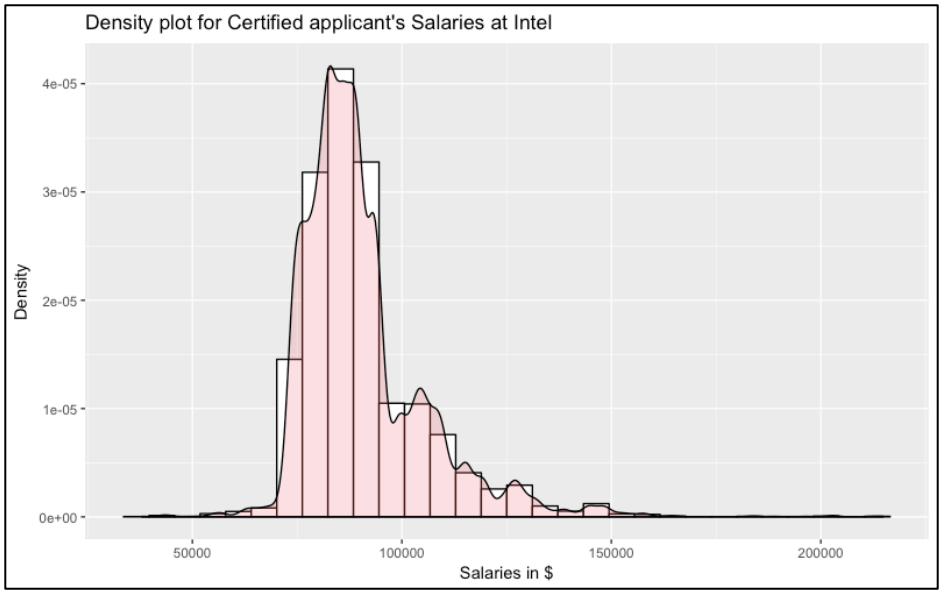

Figure 16 : Salary densities at Intel

Analyzing densities of salaries at Intel, we see a very high density of applicant's pay between $60k to $100k. As an analyst at Intel, with just a glance we can find out the average pay an immigrant applicant.

Violin plots of salaries of Case Statuses using PW_SOC_TITLE Vs PW_AMOUNT_9089 :

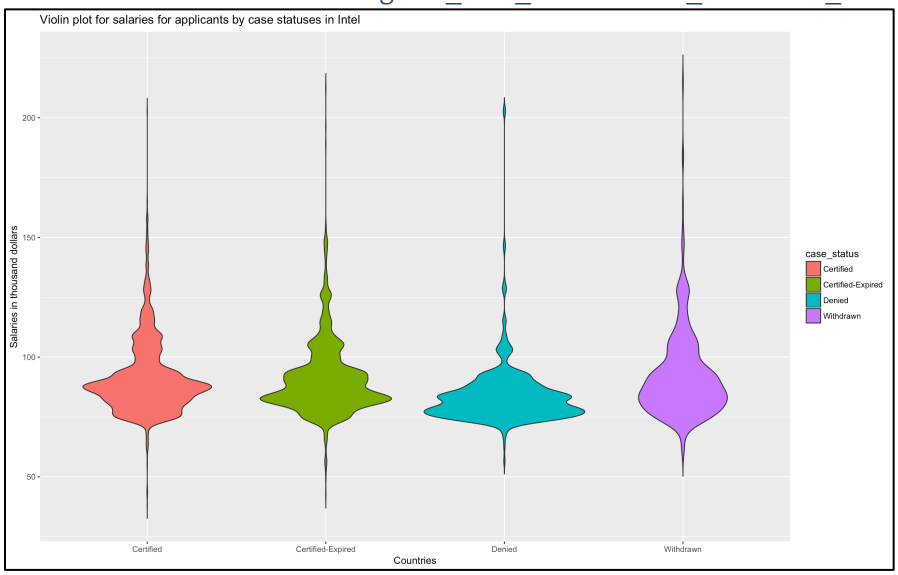

Figure 17 : Violin Plots for Case Status Vs Salaries (PERM)

During analysis of salaries by case status using a violin plot, we see similar shapes to these plots with Certified and certified - expired applications having similar densities but unique shapes for both Denied and withdrawn. By looking at the denied violin plot, we can interpret that the salaries are dense at around $75k.

Box plots of salaries of top 5 Job title using PW_SOC_TITLE Vs PW_AMOUNT_9089E:

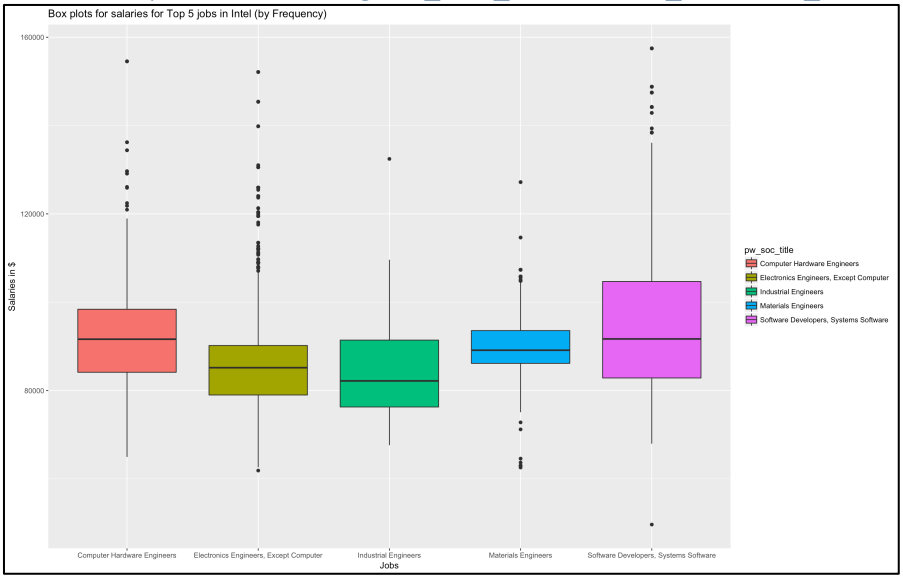

Figure 18 : Box plots of top 5 Jobs at Intel vs Salaries

We further analyze the salaries for top 5 no. of applicants per job type in Intel. We see a median salary between 80k and 120k for these jobs. A high number of intel applicants who are Software Developers with a median salary close to $90k.When analyzing further, we see a few outliers with applicants who earn more than 140k.

Box plots of salaries of top 5 countries using COUNTRY_OF_CITIZENSHIP Vs PW_AMOUNT_9089E:

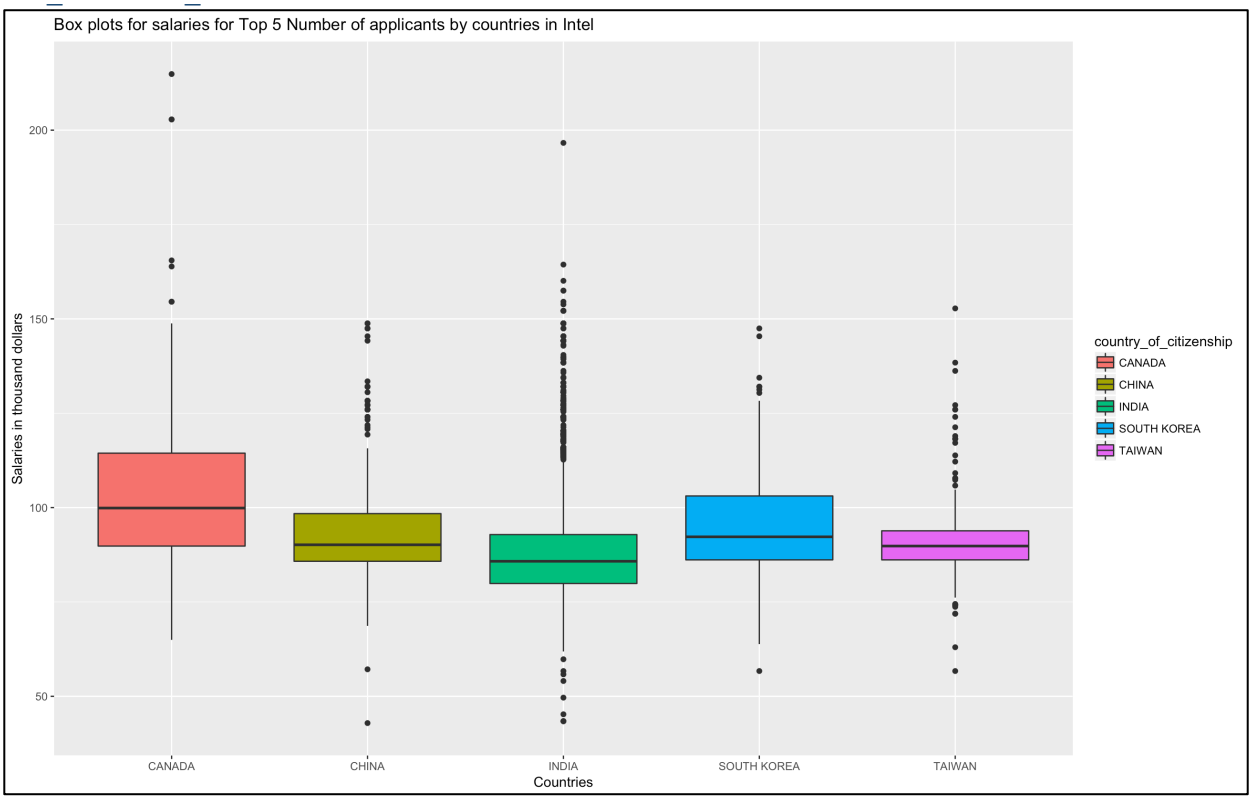

Figure 19 : Box plots of salaries of top 5 countries using COUNTRY_OF_CITIZENSHIP Vs PW_AMOUNT_9089E

We further analyze the salaries for top 5 no. of applicants per Country in Intel. We see median salaries between $80k and $100k. Applicants from India have the lowest median salary with many exceptional cases of outliers. Canadians have the highest earnings compared to other countries.

Scatter plot using case_status Vs decision_year at Intel:

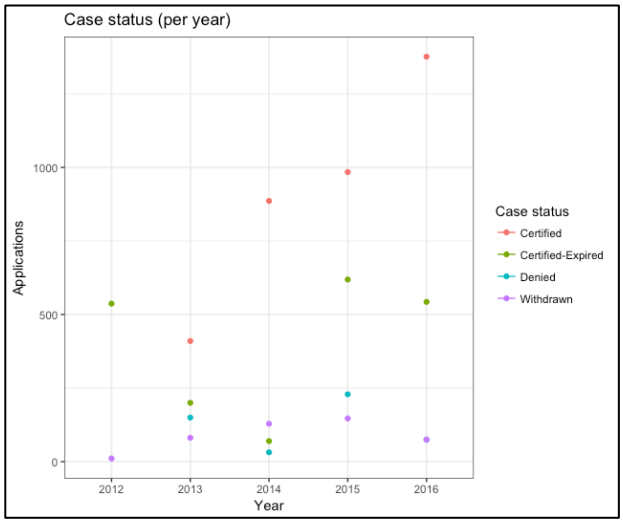

Figure 20 : Case Status per year (PERM)

We are able to see a steady growth in the number of certified application indicating a high change of getting a permeant visa certified. No. of applications withdrawn have reduced in 2016.

Conclusion

The data of the H1B and Permanent Visas highlights useful variables which influence the final decisions.

The following conclusion is made from mining the data in this project:

1. Number of Certified applications are significantly higher than other case statuses, Increasing year on year.
2. Average Salary in order to get a certified visa ranges from $70-$100.
3. Applicants concentrate from states like California, New York, Texas.
4. Top Certified jobs are Software Developers – Which may indicate the US has a clear deficit in Computer Science Professionals.
5. Top employers who petitions for H1b Visas vary from the top employers who petition for permeant status.
6. Permanent visa data within Intel was analyzed to see salaries, jobs and distribution of immigrants from various countries, this process can easily be repeated for any given employer.

References

1. https://www.foreignlaborcert.doleta.gov/docs/Performance_Data/Disclosure/FY16Q2/PERM_FY16_Record_Layout.pdf
2. https://www.uscis.gov/working-united-states/permanent-workers
3. http://www.cookbook-r.com/Graphs/Bar_and_line_graphs_(ggplot2)/
4. https://www.kaggle.com/nsharan/h-1b-visa
5. https://stackoverflow.com/questions/36568070/extract-year-from-date
6. https://www.kaggle.com/jboysen/us-perm-visas