4/21/2022

Analysis of Visa Work Petitions to the US Job Market From 2017-2022

Roland Locke

1. Summary

The H-1B, H-1B1, E-3 Vias Petitions 2017-2022 data set, contains data on US work Visa petitions and was compiled by the Office of Foreign Labor Certification (OFLC). Posted by Jishnu on Kaggle, this data set is updated quarterly as new reports from the OFLC are published. The data set contains information about the job the Visa applicant will or would have traveled for, such as employer, job title, and salary. The data set only contains Visa petitions, no information on whether the Visas were granted is included. Data from 2017 through 2022 had to be combined and some extraneous data filtered out. Employers based outside of the United States were also filtered out as well as any rows with NA fields this kept the data focused on US employers. As this analysis will be focused on potential employees or employers looking to be hired or hire outside of the US.

1.1 Visualization of Visa Data

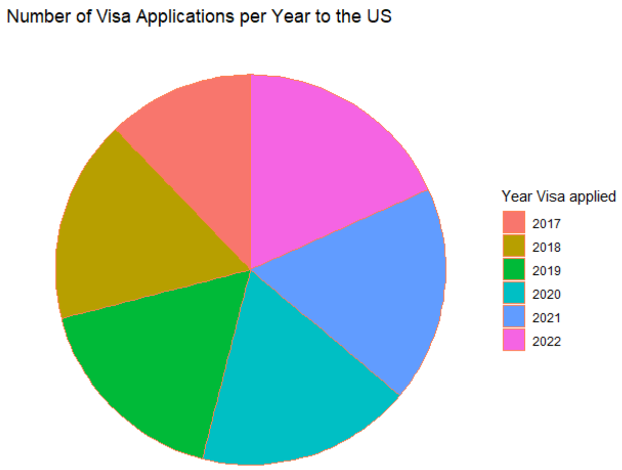
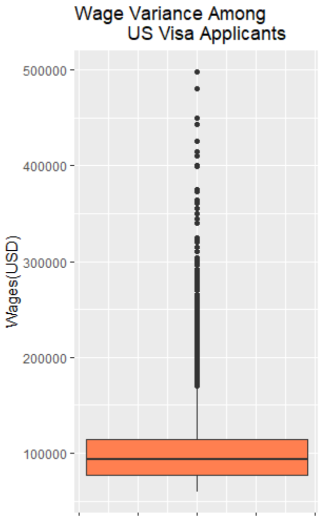
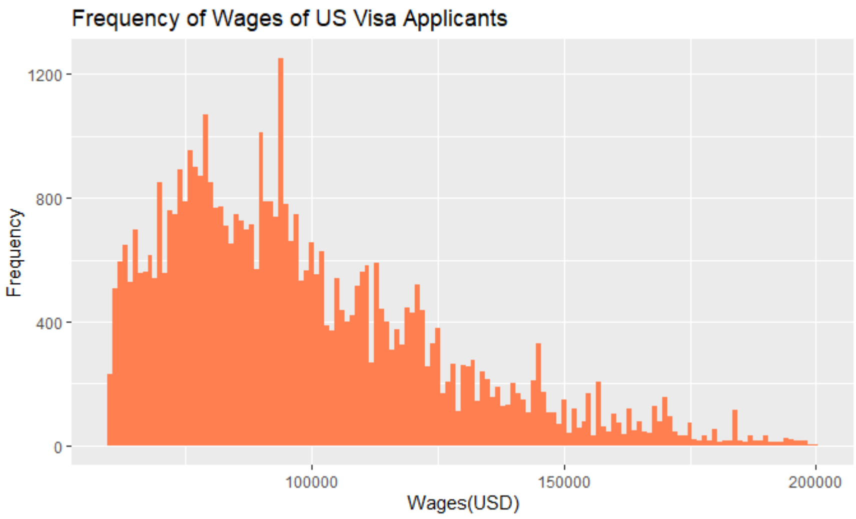


Figure 1: Title: Number of Visa Applications per Year to the US

Firstly, the data shown in years will put into perspective the spread of data available. Figure 1 shows, in percentage, what year in which each application was submitted. Because of the size of the data (over 3 million fields) Equal samples of 10,000 applicants were taken from each of the years. There are relatively equal amounts of applications over all the years with the exception of 2022 and 2021. 2022 data only includes quarter 1 and in 2021 the number of applications increases substantially. There are three Visa types included in this data set not labeled on this graph, however the vast a majority are H-1B Visas. The H-1B Visa are for applicants from 45 different countries while the E-3 Visas are for Australian applicants and the H-1B1 Visas are for Chilean/Singaporean applicants. Because applicants can manually enter their position each one is unique to the applicant. Without further specification stats on each job type cannot be obtained, this will be explored in the modeling section.

1.2 Visualization of Wage Data





(left) Figure 2: Title: Wage Variance Among US Visa Applicants

(right) Figure 3: Title: Frequency of Wages of US Visa Applicants

Data was filtered of applicants whose wage was greater than 60,000, and less than 500,000. However, the data was skewed by various wages above 200,000, as seen in figure 2. Thus, wages above 200,000 were filtered out of the data. From the filtered data Figure 3 was created. Figure 3 shows that most wages were just below $95,000 with two tall peaks around $80,000 and $95,000. There is a very sharp jump in wages at $62,000, just above the minimum value a Visa applicant can apply for. While most jobs are between $60,000 and $100,000 there are is a high number of jobs between $100,000 and $150,000.

1.3 Data Discussion

Further investigation into the data will include clustering. Clustering the data based on the year can provide data on the average change in wage of Visa applicants each year, this data could be useful when comparing to domestic wages. This could tell job searchers how companies value of employees has changed each year. Or where in the US demand for jobs has changed.

Clustering could also help a Visa applicant decide where they want to reside during their stay on a Visa, based on the average wages in each location. This could also be useful in telling prospective applicants what fields companies in the US job market are (or were) looking for talent. This illuminates how companies, regions, and jobs change in demand as wages fluctuate each year.

2. Modeling

The data set still being so large, a subset of demanded jobs were selected for modeling clustering and regression. Ten job titles were selected for: Software Developer, Data Scientist, UI Developer, Full Stack Engineer, Cloud Engineer, UX Designer, Systems Architect, and Database Administrator. These job titles were picked from Indeed.com's list of 20 In-Demand Information Technology Jobs that Pay Well (Indeed Editorial Team, 2022). Focusing on the most demanded in the field, allows job searchers a better understanding of what talent companies will be hiring now and in the near future. Modeling these jobs can help people looking for careers in the industry understand what wages each combination of job factors may offer. By using Indeed.com, a widely used job search site, data better reflects the real-world demand of the job market.

2.1 Clustering

Clustering the data provides clusters or groups that have similar attributes, this helps determine the conditions of any one scenario based on a similar one. This data set was grouped into 6 clusters based on the job title, Visa class, location of work in the US, wage, and year applied.

2.2 Clustering results

The results of cluster 2 shows that a Software Developer applying to work in Findlay Ohio most likely had a wage of $73,000 and a Visa application type of H-1B. Cluster 5 had the same job title and Visa application type with a salary of $117,000 and applicants were more than likely to work in Atlanta Georgia. The job title Software Developer appeared in 3 of the 6 clusters and had likely wages ranging from $73,000-$117,000. However, the highest wage jobs were most likely to appear in Cluster 6 with a wage at $139,000. Cluster 6 was also most likely to have applicants that worked in San Francisco, as a Data Scientist, and were applying with an H-1B Visa.

A separate group of clustering results are displayed Figures 4 and 5. In cluster 5 (Figure 5) we can see jobs ranged in wage from $123,000 to $153,000, with data scientist and Software developer ranging across all wages. In this specific cluster data scientists, software developers, and database administrators had a majority of wages in the same range. However, data scientist, software developers had outliers with salaries spanning much higher than others. These two positions are also the sole job titles in cluster 4 (Figure 4) which had the highest average wages of all the clusters, ranging from $165,000 to $190,000. Software developer was the position with the highest wage in this cluster at $190,000 with data scientist maxing out at $174,000.

2.3 Clustering discussion

Importantly, the clusters do not represent exactly what each position will be, nor do they represent averages. The results are only providing groups of similar applicants. An applicant can use each cluster to find the cluster that best fits their situation and to know what the general conditions are in each scenario.

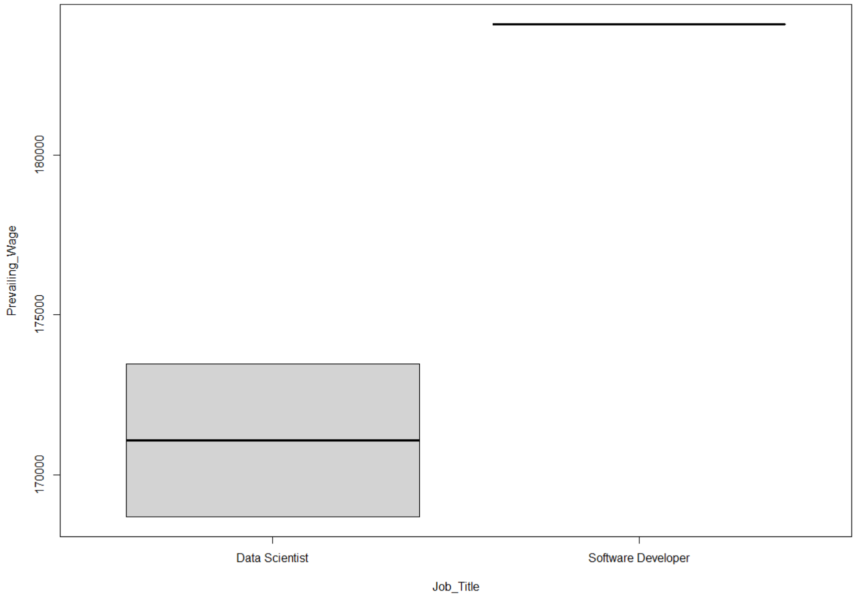


Figure 4: Title: Wage Variance Among the US Visa Applicants in Cluster 4

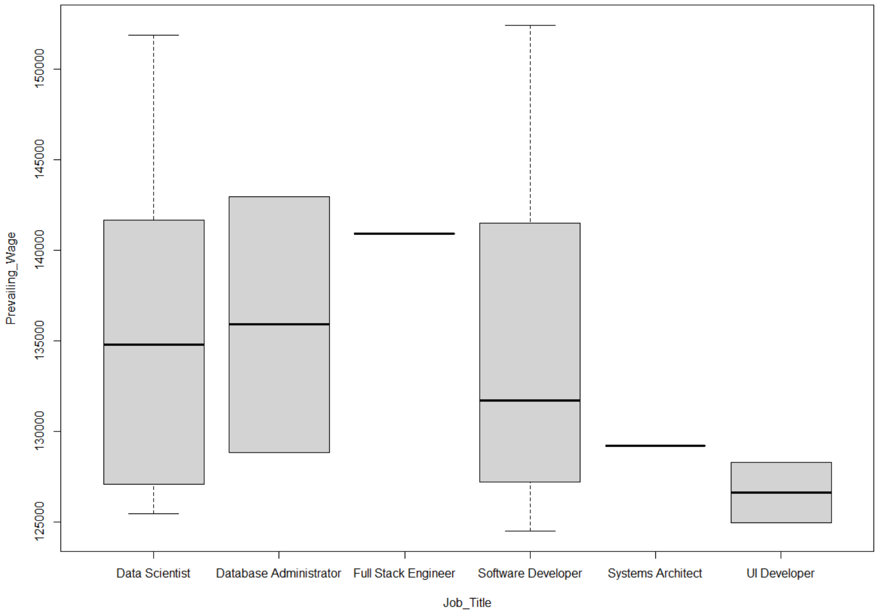


Figure 5: Title: Wage Variance Among the US Visa Applicants in Cluster 4

2.5 Linear regression

Wages are extremely important information to both employees and companies. Predicting wage could help people know where and what demand for jobs will be in the future. And companies could use a similar model to predict what wages they need to offer to remain a competitive employer attracting valuable employees.

Linear modeling simulates how each of the predicting variables effects a single response variable. In this case job titles, Visa class, and year were chosen as predictors while the response variable was Wage. This model predicted $100,000 as the average salary for A UX Designer (the intercept of the model). Full Stack Developer, at a p-value of .05, was linked with the largest positive effect on UX Designer wage. On average Software Developers positions had a positive increase of $26,000 over UX Designer wage. The position most negatively associated to UX Designer wage was, Database Administrator, with an average estimate of $-14,000. However, this correlation was weaker with a p-value of only .072. Considering this data is such a small subset of the technology industry, this data is only a very rough estimate.

The limited Visa applicant data available is a perfect example of the data restricted data. H-1B1 Visa applicants from Chile were associated with the highest average wages at an average increase of $41,000 over the H-1B applicants, with E-3 Australian applicants closely behind being associated with an average increase of $39,000. The categories had p-values of .07 and .05 respectively. However, this is not a representation of every country's applicant, as the overwhelmingly majority of datapoints are H-1B applicants which cover a huge number of countries.

2.6 Modeling Discussion.

It is also pertinent to remember that the wages are Visa applicants being hired to move from overseas. According to nnroad.com many Tech companies hire people from India as the wages offered to people from India are lower relative to that of domestic job searchers (NNRoad). The way these companies value each employee may vary. Although, this data still provides people searching for a job the technology industry a better understanding of the wages for each job and location.

3. Classification

Classification models the effect of a set of predicting variables on a binary response variable. Same as linear regression model the predicting variables are job titles, Visa class, and year. The response variable was wage however, it was spilt in to two categories higher than or equal to $150,000 (referred to as class 1) or lower than to $150,000 (referred to as class 0). The classification analysis will focus on the job title as there are too many job locations with a single wage associated. This means that most of the locations are associated with a single wage and therefore hold little significance.

Data Scientist had a one of the smallest associations with class 1. However, this did not fit with what had been predicted in other models. Both the linear regression and clustering models placed Data Scientist as one of the highest paying jobs in the data set. This was re-affirmed by comparing the mean of the Data Scientist position wage $100,000, to the mean of the most positively associated job title to class 1, Cloud Engineer. With a mean of $96,000 Cloud Engineer has a lesser mean than that of Data Scientist yet, had a higher association with class 1. The culprit of this strange behavior is the number of entries to each of the job positions vary greatly. Of the 117 Data Scientist entries there were 113 that were not associated with class 1, as for Cloud engineer, only 13 of the 13 entries were not associated with class 1. This means that the model predicted that the Data Scientist more often than the Cloud Engineer associated with class 0. This data should not, be used, as it is misleading.

Although the job title data remains unused the year data has a relatively even number of fields as was demonstrated in figure 1. This holds true in this truncated version of the data as well. Year data in this smaller subset of data is only available from the years 2021 and 2022 with 337 and 276 points of data respectively. The coefficients from these data points are accurate enough to be interpreted. Both years were negatively associated with class 1 (wages above $150,000). The year 2021 was more so negatively affected than that of 2022. There are missing years in this analysis that is because of the necessary filtering of certain job titles earlier in the modeling process.

3.1 Classification Discussion

Because the Visa application process has no standardized way to list job titles there is no simple way to capture every job that was selected for. An example of this would be: a Data Scientist position title could have been listed as “Senior Data Scientist” and the selection would not have captured this as an entry. There certainly are many job titles that were not factored into the data for this reason. This is one reason for the strange model behavior observed.

3.2 Conclusion

There are many take aways from this dataset, however the results discussed in this report has focused on the a few key points that a worth reiterating. The H-1B, H-1B1, E-3 Vias Petitions 2017-2022 data set contains Visa applications from an assortment of countries. The data set contains mostly H-1B Visas with the majority of Visas petition salaries between $60-000 and $123,000 per year. 10 job titles were chosen to model with, these were picked from Indeed.com’s list of 20 In-Demand Information Technology Jobs that Pay Well (Indeed Editorial Team, 2022). Of the three models interpreted the jobs Software Developer and Data Scientist were on average the two highest paying positions. And finally, despite the majority of applications submitted in 2021 it was also the year with the greatest negative association with jobs of a wage above $150,000. Individuals looking to use this report for insight on job wage data in the US should understand that the data contains Visa applicants from many different countries and backgrounds. The data here may not be representative of the wages that a company may pay a US citizen.

As for people applying for a US work Visa. The Visa application process is very stressful. The more information that can be revealed on the jobs and qualifications that potential applicants should prepare themselves with, the less difficult the process will be. The analysis of the H-1B, H-1B1, E-3 Vias Petitions 2017-2022 data set aimed to expose some of the mystery in the Visa application process.

Reference

Indeed Editorial Team. (2022, April 14). *20 In-Demand Information Technology Jobs That Pay Well*. Indeed Career Guide. https://www.indeed.com/career-advice/finding-a-job/it-jobs-list

Why Tech Companies are Hiring Employees from India | NNRoad. (2020, May 11). Nnroad.com. https://nnroad.com/blog/why-tech-companies-are-hiring-employees-from-india/