Road to H1-B

Ann Tsai
Courant Institute of
Mathematical Sciences
New York University
New York, New York
mt4050@nyu.edu

Ko-Chun, Chiang
Courant Institute of
Mathematical Sciences
New York University
New York, New York
kc4200@nyu.edu

Jiancheng Wang
Courant Institute of
Mathematical Sciences
New York University
New York, New York
jw5865@nyu.edu

***Abstract*— While researching the H1-B acceptance rate, we found that past related studies only focused on the prediction of H-1B results, no work has been done on providing a recipe for international students about what they should learn at school to achieve the highest H1-B acceptance rate. In this paper, we explore what can be learned about H1-B applications from H1-B employer data, Linkedin Job Description, and H1-B salary dataset. We utilize Spark to process our data, statistical methods to verify if the higher salary you have, the higher H1-B acceptance chance you get, and we analyzed the H1-B acceptance and expected salaries in each skills/programming languages. Lastly, we visualize our result using Tableau, which allows readers to understand what skills or programming languages could lead to a higher salary and to a higher possibility to get H1B visas.**

Keywords*— *Machine learning, Statistical analytics, Data analysis, Analytical models, Data visualization, Apache Spark, H-1B working visa

I. Introduction

H-1B is a type of non-immigrant visa that allows US employers to hire foreign workers in specialty occupations that require theoretical or technical expertise in specialized fields such as IT, finance, accounting, engineering, science, medicine, etc. The job must meet some criteria to qualify as a specialty occupation such as minimum entry requirement for the position as bachelor's or higher degree or its equivalent. Additionally, the employer should attest that it will pay the beneficiary a wage which is no less than the wage paid to similarly qualified workers or higher than the prevailing wage for the position in the geographic area.

There are about a million international students in the U.S. per year, and H-1B is the most commonly used working visa after one to three years of OPT (Optional Practical Training). The USCIS (US Citizenship and Immigration Service) grants 85,000 H-1B visas per year which includes 20,000 additional visas for workers with advanced degrees (i.e., MA, PhD) from an accredited U.S. academic institution. Among all the majors, computer science is the most popular major for international students to launch a career in the United States because of the prosperity of the software industry in the United States.

However, most studies [1][2][3] focused on the prediction of H-1B results based on the application data which provided limited information for students. In the competitive computer science schools, students do not have a clear guide which software skills or programming languages would lead them a better chance of getting H1B visas. Not to mention how confused those students are who transferred from other majors. This paper applies empirical studies on how students can maximize their opportunities to get H-1B visas.

This paper applies empirical studies on how students can maximize their opportunities to get H-1B visas. To provide students a clearer guide, this paper will analyze the H-1B acceptance in different aspects including geographic area, fiscal years, and most important, different skill sets/programming languages. Additionally, this paper will also analyze the expected salary of different skill sets/programming languages. Furthermore, an empirical study on the causal relation of salary and H1B acceptance would be conducted. With the paper, international students could have a clearer view of what skills/programming languages are needed to get into companies with a high H-1B acceptance rate and a competitive salary. International computer science students could leverage it for their course selection or career development. International students of other majors who are interested in a career in the software industry in the U.S. could also use the result to evaluate what skills are worth investigating.

II. Motivation

Computer science has become the most popular degree in the United States, especially for international students due to the high volume of positions and salaries in a career of the technology industry. However, students in computer science are confused about what skills or programming languages they should invest their valuable time in. Accordingly, we would like to provide analytics results among different skills and programming languages for computer science students to maximize their career.

The rapid development of the tech industry has spawned a vast number of skill sets and we want to figure out which subset of skills make great contributions to the salary, and indirectly impact the H1-B. For new grads, especially those just switched major to computer science in their last one or two years in academy, they can use this insight to pick up the skills that are most cost-efficient and rule out the skills that are no longer popular in the industry. For experienced developers, they can also cite the result when they want to challenge themselves into getting a higher pay.

III. Related Work

First, the only thing we know about salary and H1-B application is that you meet the application condition requirement as long as your salary reaches the prevailing wage. However, we want to verify if the higher salary you have, the higher H1-B acceptance chance you get. Jena, Olenski, and Blumenthal [4] applied Sensitivity Analysis to analyze the relationship between genders and salaries of physicians. This inspired us to use the same method to verify the relationship between salary and H1-B acceptance rate. Besides, we also utilize linear regression models to get the regression equation.

Second, we want to assess whether the result, which shows the causal relation, we get from linear regression is credible or not. The authors of paper[9] use meta analysis to research the relationship between Oral Health and Brain Injury, try to answer whether the relationship is causal or just casual. We reference this paper and apply meta-analysis on researching the most appropriate to prove the causal relation. Therefore, we look at lots of articles and literature. In the end, we choose hypothesis testing, which seems most appropriate.

Ke and Qiao [1] applied machine learning on H-1B acceptance and wage prediction. To deal with the problem, the authors leverage two machine learning algorithms for each of the tasks on the H-1B application dataset from the Department of Labor (DOL). The datasets provide the H-1B application information and wage from 2008 to 2017. The features in the dataset are different from 2008 to 2015 and afterwards. The first period includes economic sectors of the employer, the state of the employer, and the citizenship of the applicants. the job level, the pay unit, and the year of the application. The second period has extended features such as the major of the H-1B applicants in college, the education level, the ownership interest of the applicant, prior job experience as the number of months worked, the number of founding years, and the employer's total number of workers. To analyze what contributed to the wage, the authors select the Least Absolute Shrinkage and Selection Operator (LASSO) model to preserve model performance and the interpretability of the model. With 10-fold cross-validation grid-search, the best tuning parameter a is determined as 10^{-6}, and the R^2 of the model on each dataset is 0.54 and 0.68. From analyzing the coefficients, the result shows that applicants working in the healthcare industry or majored in healthcare-related majors may have the highest wages. To analyze what contributes to the acceptance of H-1B, the authors apply Logistic Regression with L1-penalty also for better model interpretability. Also, with 10-fold cross-validation grid-search, the best tuning parameter l is determined as 1, and Area Under Curve (AUC) are 0.674 and 0.676 for each of the dataset. From analyzing the coefficients, the result shows that applicants with higher education levels and majoring in computer science/electronic engineering may have a higher possibility to be approved.

Vegesana [2] uses multiple classification machine learning algorithms for H-1B acceptance prediction. The H-1B dataset is from the U.S. Office of Foreign Labor Certification's iCERT Visa Portal System which contains 3,002,458 of H1-B petition records from 2011 to 2016. The features in the dataset include CASE_STATUS, EMPLOYER_NAME, JOB_TITLE, SOC_CODE, FULL_TIME_POSITION, PREVAILING_WAGE, YEAR, WORKSITE, Lat, Lon. The CASE_STATUS determines the target variable for the task, and the variable contains three categories which are Certified, Denied, Certified-Withdrawn and Withdrawn. The Certified-Withdrawn and Withdrawn categories indicate that the case is not evaluated so the author removed those two. For the training feature, the author selects SOC_NAME, PREVAILING_WAGE and WORKSITE. In the data cleaning process, the author filled the missing field with dummy values. The dataset is split into 70% for training and 30% for testing. For the classification algorithm, four algorithms are used, which are Logistic Regression, Random Forests, K-Nearest Neighbors, and Gaussian Naïve Bayes. Recursive Feature Elimination with Cross Validation (RFECV) is leveraged for parameter turning. With four folds Cross Validation, Random Forests is using max_depth of 5 and estimator of 40. The author evaluates the prediction result with four aspects, accuracy, F1 score, AUC, and performance time. Among the results, the Bagged Random Forests outperforms the other algorithms with 96.8% in accuracy, 0.9 in F1 score, 0.725 in AUC. Logistic Regression has the best running time with 3.6 second, and also has competitive prediction results with 96.7% in accuracy, 0.88 in F1 score, and 0.71 in AUC.

IV. Datasets

A. H1-B employer dataset

The dataset is obtained from the U.S. Citizenship and Immigration Services (USCIS) of the Department of Homeland Security which provides the information on employers petitioning for H-1B workers from 2009 to 2019 [5]. The size of the dataset is 45 MB with around 61 thousands of petitions from 28 thousands employers. Table 1. shows the schema of the dataset.

Table 1. Schama of H-1B employer dataset

Fiscal Year	Int	The fiscal year in which USCIS first recorded an approval or denial (adjudicated date) in the

		electronic systems. A fiscal year covers Oct. 1 of one year to Sept. 30 of the next year.
Employer	String	Petitioner's firm/employer name.
Initial Approvals	Boolean	H-1B petitions with "New employment" or "New concurrent employment" whose first decision is an approval.
Initial Denials	Boolean	H-1B petitions with "New employment" or "New concurrent employment" whose first decision is a denial.
Continuing Approvals	Boolean	H-1B petitions with anything other than "New employment" or "New concurrent employment" whose first decision is an approval.
Continuing Denials	Boolean	H-1B petitions with anything other than "New employment" or "New concurrent employment" whose first decision is a denial.
NAICS	Int	North American Industry Classification System Code: A character string that stands for an industry classification.
State	String	Petitioner's state. This is the State indicated in the mailing address of the employer and is not necessarily the beneficiary(ies) work location.

B. Linkedin Job Description dataset

The job descriptions are scraped from Linkedin.com [6] with keyword searches: roles set to "Software Engineer", "Data Scientist", etc.; locations set to "San Francisco", "New york", etc. The dataset is further normalized and job skills are extracted with hand-written rules.

Table 2. Schema of Linkedin Job Description dataset

employer	String	Name of the employer: "Facebook", "Google", etc.
job_titles	String	Including software engineer, web developer, ML engineer, data scientist, and data analyst
skill_sets	Int	Bitmask where each bit represents if a certain skill is required: Java, Python, C++, etc.
degree	Int	Bitmask where the first three LSB represents if BS/MS/PHD degrees are required in this role

C. H1-B salary dataset

The dataset is obtained from the up-to-date official H1-B data disclosed by the United States Department of Labor [7], which provides information on salary, company, job title, location, … and so on, from 2012 to 2019. The size of the dataset is 36 MB. Table 3. shows the schema of the dataset.

Table 3. Schama of H1-B salary dataset

year	Int	H1-B application submit year. Software engineer / Data analyst: 2012 ~ 2019, Data engineer : 2013 ~ 2019, and Web developer / Data scientist : 2014 ~ 2019.
employer	String	Petitioner's firm/employer name
job_titles	String	Including software engineer, web developer, data engineer, data scientist,and data analyst
salary	Int	annual salary[range : $923 ~ $1350001]
visa_status	String	WITHDRAWN/ CERTIFIED/DENIED.
state	String	Petitioner's state. This is the State indicated in the mailing address of the employer and is not necessarily the beneficiary(ies) work location
month	Int	H1-B application submit month

V. Description of Analytic

H1-B acception

To understand the H1-B acceptance status, we analyzed the H1-B employer dataset using Spark in Scala. First of all, we cleaned the data by removing null, duplicate and typing error rows (6521 rows in 622684 rows). Since we would like to provide information for students who have not had H1-B yet, we would neglect "Continuing Approvals/Denials" which is for the people who switched employers. Among "Initial Approvals/Denials", we used the "groupBy" and "pivot" function to get the H-1B approval rate and counts from 2011 to 2019 among each "State", "Employer", and "NAICS" (North American Industry Classification System Code). For "State" and "Employer" analysis, we concentrated on those jobs in the field of Professional, Scientific, and Technical Services, so we used a "filter" function to select those employers with NAICS: 54.

Programming Language and Software Skills

Each text file was first processed with two preprocessing stages -- tokenization and lower case conversion. After that, we applied rule-based NLP approaches to extract the keywords: programming languages and software skills. To be exact, we wrote a regular expression suite to cover different scenarios and irregular cases. An example of the issues we observed is when extracting language “C”, we need to differentiate it from “C++”, or letter “C” embedded in a word such as “conference”. This was resolved by adding restrictions to the Regex that either “C” shows up in the beginning of the sentence, or there need be one or more stop words/white space before it. The similar restriction applies to the position after it. Nine most popular languages are extracted such as “C”, “C++”, “Java”, and they are not mutually exclusive with regard to a single job qualification. We used a bitmap to store the intermediate result. For example, 0x0001ff indicates that all nine programming languages are mentioned in this job post.

In addition, top-trend software skills are extracted in the same fashion. We first generated candidates in each field of software engineering. For example, [“Django”, “Flask”, “Spring”, “.Net”, “Express.JS”, “Node.JS”, “Ember.JS”, etc.] are spawned as the candidates in the field of “Back-end framework”. We then applied map-reduce to count the occurence of each skill specified in the job description and only selects the top 1/2 popular frameworks in its own language. Other fields of software engineering are “Front-end”, “AWS”, “Container”, “Big Data”, “Database”, and “DL/ML”.

We further used aggregate methods to analyze the popularity of different skills and their impact on salary.

Join

H1-B employer dataset is joined with other two datasets with "company name". Linkedin Job Description dataset and H1-B salary dataset are joined with the primary key of (company name, job_title). Company names are written in different formats in all our three datasets and therefore need to be taken care of. An example for normalization is converting “Amazon Web Service Inc.” to “Amazon”. We did not handle irregular words, since it rarely shows up in our datasets. We also did not deal with subsidiary relationships, for example, “A9” is a subsidiary of “Amazon”, as this cannot be solved without a knowledge graph. The company name is put through a pipeline, built with regular expression patterns, in which five categories of noisy words are filtered out, including stopwords, location, industry identifier, and etc., based on our observation.

Salary and H1-B acceptance with Programming Language and Software Skills

To understand the salary and H-1B acceptance difference among different programming languages and software skills, we join the dataset as the previous paragraph described. Spark dataframe was used to analyse and manipulate data. Spark dataframe functions were heavily leveraged including groupBy, pivot, UDF, explode, and withcolumn, etc.

Salary and H1-B acceptance rate

Objective:

To verify if the higher salary you have, the higher H1-B acceptance chance you get, we compare the median of salary per year, state and employer with corresponding H1-B acceptance rate.

Scrapping:

We scrap salary data from the H1-B Visa Salary Database. We only considered job titles related to SOFTWARE ENGINEER, WEB DEVELOPER, DATA ENGINEER, DATA SCIENTIST and DATA ANALYST.

Cleaning:

First, we clean the job titles, we replace SOFTWARE ENGINEER II, JUNIOR SOFTWARE ENGINEER, …… to SOFTWARE ENGINEER, and for other job titles, so on and so forth.

Calculating:

To verify if the higher salary you have, the higher H1-B acceptance chance you get, we calculate the median of salary per year, state and employer. First, we choose to use median rather than average because we want to avoid being affected by outliers. Second, we calculate median from year, state, and employer because year can represent economic situations, which might affect salary and h1-b acceptance rate, besides, same state and same employer are supposed to have similar salary ranges.

Analyzing:

We think that year, state, employer affect median of salary, so we calculate median of salary per year, state, employer, and then joined Median Salary with H1-B Acceptance Rate. We

use mllib to build a linear regression model and get the linear regression equation, which has median of salary as independent variable, and H1-B acceptance rate as dependent variable. If we could find an equation with positive slope, and we can prove this result is significant by hypothesis testing , we can infer that when salary goes up, H1-B acceptance rate also becomes higher. Besides, we also use sensitivity analysis to see when the median of salary changes, how much the level of H1-B acceptance rate will be changed.

VI. Application Design

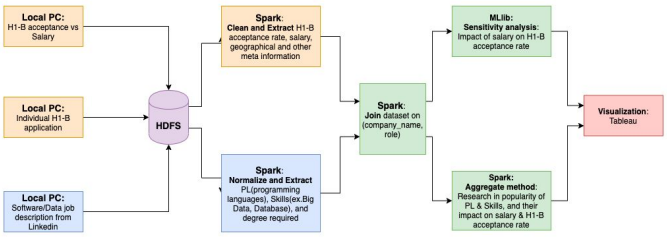

Figure 1 Design diagram

We conducted our experiment both on NYU Dumbo and local PCs. At the first stage, all data is collected locally and then sent to the HDFS. After that, we used Spark to first clean and normalize the data, and then extract the key features. All three datasets are preprocessed separately and joined via a tuple of "company_name" and "role". The joined dataset is then pipelined to build regression models and conduct sensitivity tests using MLlib, and other analytics by applying aggregate methods. The results are eventually sent to Tableau for visualization.

We provide our users with a set of interactive Tableau dashboards to allow them observe the impact of different factors onto the H1-B drawn rate and salary. This includes a set of factors such as skills, programming languages, salary and geographical information. Users could also re-arrange or combine features arbitrarily to see which set of attributes is the most decisive in terms of salary and more importantly, H1-B lottery rate.

VII. Analysis

After we acquired the joined dataset of PL/Skills vs Salary, we further investigated the impact of each factor on the salary using SparkRdd and SparkSQL. The PLs and Skills are not disjoint and therefore, each role could require multiple PLs and Skills.

We created an inverted index that "flatmaps" the salary to the PL/Skill that their roles' job description mentions. As a result, the dataset now looks like "Python: 108000, Python: 109000, Java: 108000, ….". We then computed the median and average for each PL/Skill.

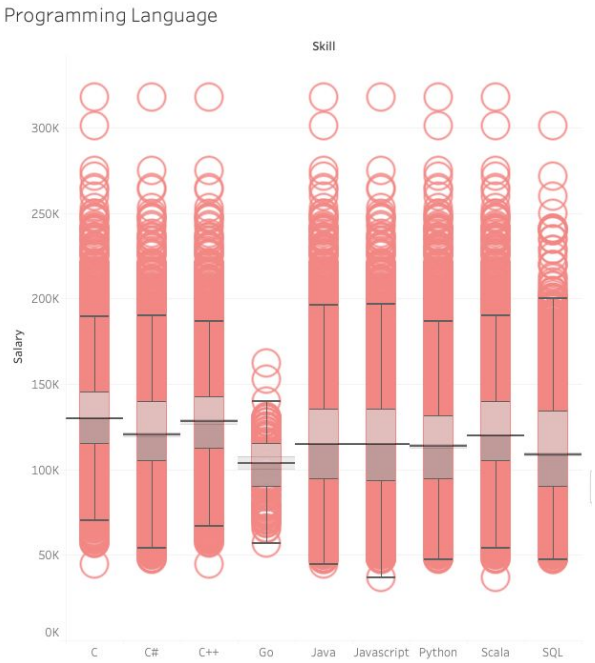

Figure3: Impact of Programming Language on Salary

Figure 3 shows the result of the impact of PL on salary. We can observe that the Top 3 languages that contribute heavily to salary are C, C++ and Scala. This could because C and C++ are required by most big corporations as fundamental programming languages, whereas Scala, mostly used in Big Data platforms like Spark, is essential if the role is involved with big data processing. SQL, on the other hand, is the most popular language, since any role involved with interaction with database(not only database development, but data analytics) requires SQL.

We further observed that the hot skills required by software related roles can be partitioned into seven categories: Cloud Computing, Big Data, DevOps, Database, Back-end, Front-end, and Deep Learning. Therefore, we first manually crafted the candidates of each category and selected the top 1 or 2 popular skill. For example, MXNet is outperformed by Tensorflow and Pytorch in terms of the count of roles mention it, and therefore is excluded from "Deep Learning".

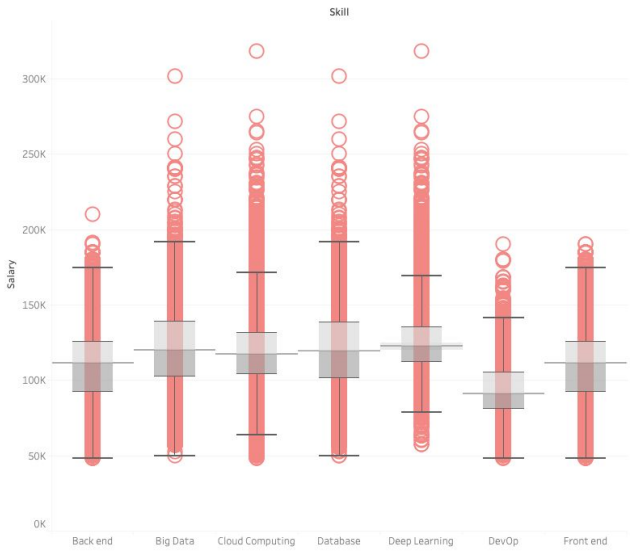

Figure4. Impact of Software Skills on Salary

Figure 4 shows the result of the impact of Software Skills on salary. We can tell that the Top 3 skills heavily contribute to salary are Deep Learning, Big Data and Database. In fact, database is also the most demanded skill in terms of the popularity count. Therefore, learning database would be the most cost-efficient for anyone just switched to software engineering or want to get a salary pump.

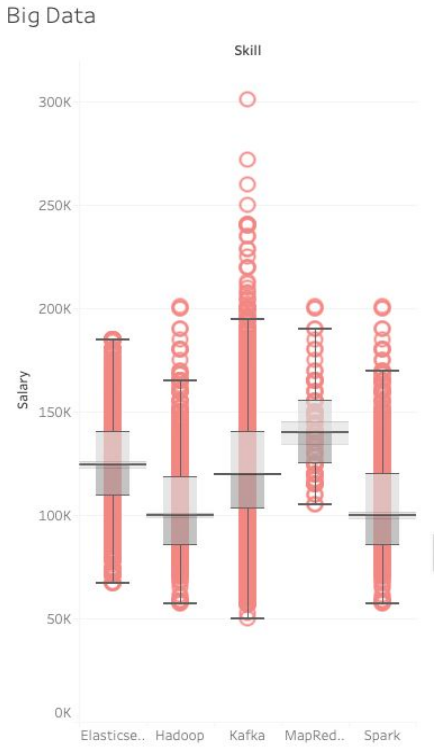

Figure5. Impact of Big Data Skills on Salary

Figure 5 shows the impact of different Big Data skills on salary. Kafka has gained a large portion in the marketplace and candidates who master Kafka tend to have high salary. On the other hand, Spark is still the most dominating language in the Big Data field, though the average salary is not the top.

H1B acceptance analysis

Analytics on the H1B acceptance versus different states, fiscal years, and industries were conducted using SparkSQL and SparkRDD within the H1B acceptance dataset.

Figure 6 shows the H1b acceptance rate box plot from 2012 to 2019 and the H1B acceptance rate for NAICS 54. From the upper box plot, we can see that the acceptance rate is trending down. The median of the acceptance rate dropped from 100% to 94% which is still pretty high. However, the lower whisker (25%) dropped from 91% to 52% which is extremely severare. The second part of the graph presents the H1B acceptance rate trend for top five industries. The NAICS 54 (Professional, Scientific, and Technical Services) is highlighted in green as it is the industry for most technology companies. The green line shows that the acceptance rate of the tech industry dropped from 96% to 79% while the overall number of applicants remains similar between 2012 and 2019. However, though the drop is not subtle, the tech industry still holds the highest approvals among all industries. The phenomenon shows that it is harder for applicants to get a H1B approval now, but the technology industry still the most possible industry to get an H1B visa.

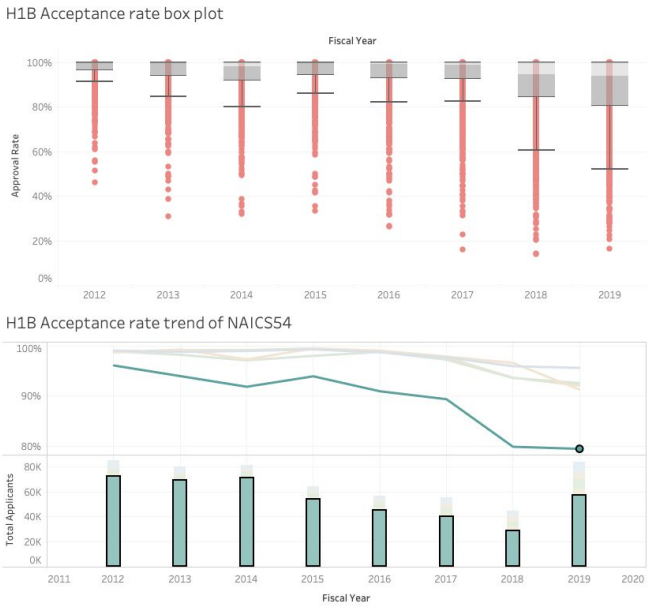

Figure6. H1B acceptance rate through years

Figure 7 demonstrates the H1B approvals status of the top six states, CA, NJ, TX, NY, WA, and MA. The upper part is the map of the overall approvals from 2012 to 2019. California holds the most of the approvals without surprise, but NJ being the second highest is out of expectation. The lower part of the graph presents the trend line of these states. We can see that there is a great increase from 2018 to 2019 especially for California.

To understand which skill or programming language could bring a better chance of getting H1B, we joined the H1B

dataset with LinkedIn dataset on normalized company name to get information in the job description.

Figure 8 shows the H1B acceptance for each programming language per year. We can find the top three programming languages are SQL, Python, and Java. SQL and python being the top two might be contributed by the prosperity of big data, data science and AI related jobs.

Figure 9 presents the H1B acceptance of each software skill group. Cloud computing is the top software skill to get a H1B with about 5000 more H1B counts per year than the second skill group, Frontend. Figure 10 breakdowns each group into detailed software skills. We can see that .NET and Node.js are the top two skills that could lead to a H1B visa in the Backend; for Frontend, it is Angular.JS with around 9000 counts more than the second skill, React. The H1B acceptance count for each skill in the Database is similar that the top one, MySQL, only has 500 more counts than the second one. For Big Data skills, Spark is the top one with over 2000 counts than the second groups, Hadoop and Kafka. For Cloud Computing, AWS and Azure have around 8000 more counts per year than the third skill, GCP. For DevOps, it contributes least in H1B visa from 2012 to 2019, and the most popular skill is Jerkins. For Deep Learning, it is surprising that it is the second last among all skill groups considering the popularity of AI these days. That might be due to AI being only popular in recent years. Among Tensowflows and Pytorch, Tensowflows contributes 2000 more than Pytorch.

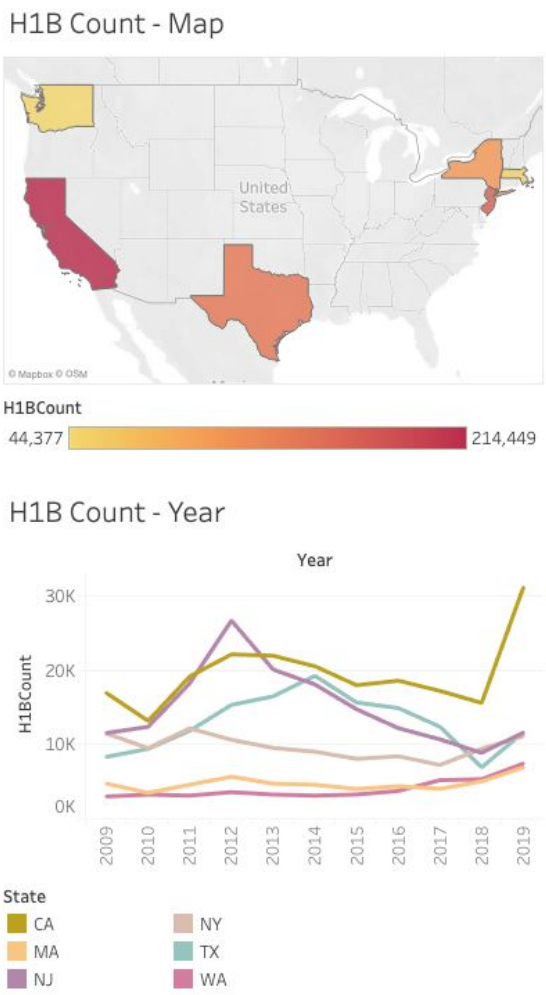

Figure. 7 H1B acceptance counts vs State

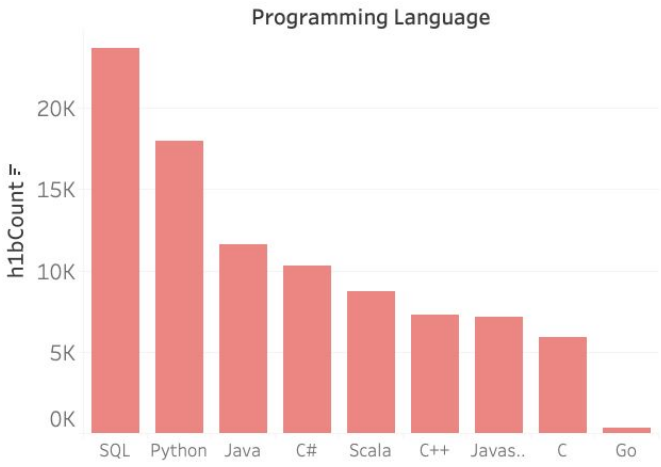

Figure.8 H1B acceptance counts per year vs Programming languages

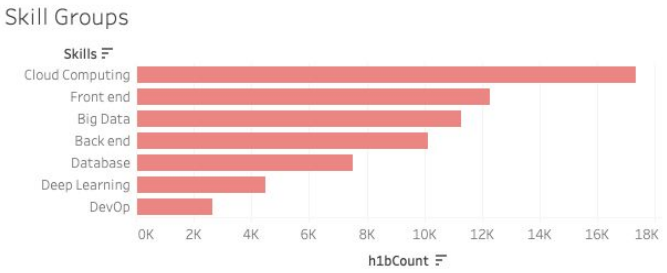

Figure.9 H1B acceptance counts per year vs Software Skill Groups

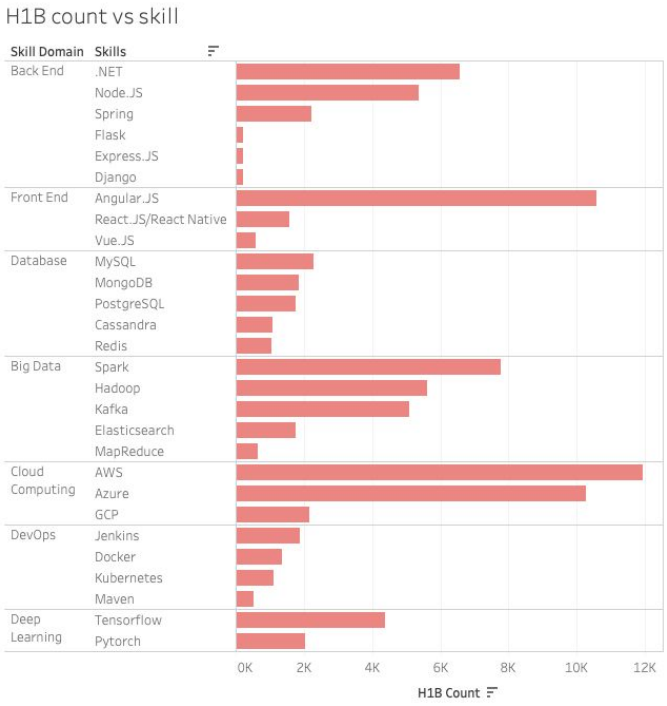

Figure.10 H1B acceptance counts per year vs Software Skills

H1B Salary analysis

Analytics on the H1B median salary versus different states, years, and job titles were conducted using Tableau and within the H1B Salary dataset.

Figure 11 demonstrates that DATA ENGINEER and DATA SCIENTIST have higher maximum and average of Medium Salary, while DATA ANALYST and WEB DEVELOPER have lower maximum and average.

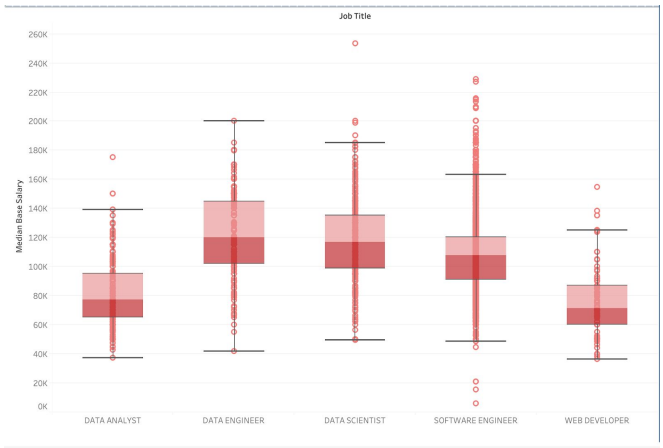

Figure. 11 H1B median salary v.s. job title

Salary and H1-B acceptance rate analysis

Analytics on the relationship between H1B median salary and H1-B acceptance rate versus different states, years, and employers were conducted using sparkSQL and mllib linear regression. In the end, we utilize hypothesis testing to test the credibility of the result.

Before analytics

We only know that you meet the application condition requirement as long as your salary reaches the prevailing wage.

There seems to be no relationship between Median Salary and H1-B acceptance rate from the naked eyes, because all the values of H1-B acceptance rate looks extremely high.

After analytics

Salary indeed can enhance H1-B acceptance rate.

a. Linear regression:

We treat all the records with rate = 0 as anomalys, and ignore them. Then we set Median Salary as features, and H1B Acceptance Rate as labels. For training, we calculate the linear regression equation between Median Salary and H1B Acceptance Rate year by year from 2012 to 2019.

b. Hypothesis Testing:

First, we set $\alpha = 0.05$, and compare the p-value of each year with α.

Second, The result shows that only 2012 and 2013 are insignificant. However, after data joining and removing all the anomalys, we only have less than 100 records among these two years, so we think the result of these two years can be ignored

In the end, the results of 2014 to 2019 are all significant, so we could infer that Salary indeed can enhance H1-B acceptance rate.

VIII. Conclusion

The paper successfully analyzed the three datasets in three different angles and provided a comprehensive guide for international students in the United States to get H1B in the software industry. First, the skills/PL information was joined with H1B acceptance data, and we derived the top skills/PL that could lead to a H1B visa in the future. Secondly, we joined the salary with H1B acceptance data and successfully proved that higher salary does lead to a higher chance of getting a H1B visa. Last but not least, since higher salary is positive to getting a H1B, we found out which skills/PL could lead to a higher salary by joining the skills/PL information with salaries. Accordingly, international Computer Science students could leverage our analytics result to optimize their learning strategy/course plan and maximize their chance to get a H1B visa and higher salary in the future.

IX. Future Work

1) . *Company name normalization.*

Company_name plays a crucial role in determining the amount of data we can successfully join, and so far we rely on lots of hand-crafted rules to normalize the noisy text. In the future, we would scrape more job description data and run Word2Vec with cosine similarities to improve the normalization. We will also look into the probability of using Edit distance on photonic expression when it comes to traditional NLP methods.

2) . *Web application.*

We want to give our users the ability to picture their H1B acceptance rate with their existing jobs. To do so, we would like to create an interactive website backed by a simple SQL database along with the regression model.

The website provides two different views for users to visualize what they searched for. In the first view, users can query the prediction results by selecting fields from the drop bar. This includes information such as employer, job title and geolocation as shown in Figure 11. As soon as users hit the search button, the confidence level of their H1B applications getting accepted will be prompted, with 0 being no chance and 1 being must accepted. This is done by running our pre-trained regression model on the back-end.

In the second view, the webpage provides users with the functionality to arbitrarily add more skills to their existing skill set and see if it can increase the confidence level. It is well-known that employees from big tech companies like Google and Facebook tend to be accepted with higher chances, and in fact, lots of the skill set these tech giants ask from candidates are in common. Having this view could potentially help users decide which skill set is more cost-effective to learn in the short time.

Figure 11. UI to query H1-B lottery rate

Acknowledgment

Thanks to Eberly College of Science - Penn State for sharing the course content of Probability Theory and Mathematical Statistics, which allows us to reference when measuring our linear regression model.

References

[1] KE, Barry; QIAO, Angela. Who Gets the Job and How are They Paid? Machine Learning Application on H-1B Case Data. *arXiv* preprint arXiv:1904.10580, 2019.

[2] Vegesana, Sharmila. "Predictive analytics for classification of immigration visa applications: a discriminative machine learning approach." (2018).

[3] GUNEL, Beliz; MUTLU, Onur Cezmi. Predicting the Outcome of H-1B Visa Applications.

[4] VEGESANA, Sharmila. Predictive analytics for classification of immigration visa applications: a discriminative machine learning approach. 2018.

[5] Sex Differences in Physician Salary in US Public Medical Schools, Anupam B. Jena, MD, PhD; Andrew R. Olenski, BS; Daniel M. Blumenthal, MD, MBA, September 2016

[6] H1-B employer data

[7] LinkedIn Job description data

[8] H1B Salary Database

[9] Oral Health and Brain Injury: Causal or Casual Relation? Pillai R.S., Iyer K., Spin-Neto R., Kothari S.F., Nielsen J.F., Kothari M.