US Visa Application Analysis

Jasmine Batra
Dept. of ECE
NYU Tandon
jb7854@nyu.edu

Navya S. Jammalamadaka
Dept. of ECE
NYU Tandon
nsj9072@nyu.edu

Suchitra Kollu
Dept. of ECE
NYU Tandon
sk9857@nyu.edu

Abstract

As part of the project, we intend to analyse the US Visa applications across the globe and present a study about insights received[1]. The next sections of this report explains about the implementation details carried out to analyze the trend in US visa applications throughout the world. The document describes the problem statement, approach, analysis, and proposed scalable architecture. The code and graphical representation of the analysis have been explained in detail for the reader's clarity.

Keywords: SparkML, SparkSQL

1 Problem Statement and Objectives

An unimaginable amount of people applies for US visas every year. The amount of data gathered and information available to process is immense, wherein the processed data can have several use cases.

A permanent labor certification issued by the Department of Labor (DOL) allows an employer to hire a foreign worker to work permanently in the United States. In most instances, before the U.S. employer can submit an immigration petition to the Department of Homeland Security's U.S. Citizenship and Immigration Services (USCIS), the employer must obtain a certified labor certification application from the DOL's Employment and Training Administration (ETA). The DOL must certify to the USCIS that there are not sufficient U.S. workers able, willing, qualified and available to accept the job opportunity in the area of intended employment and that employment of the foreign worker will not adversely affect the wages and working conditions of similarly employed U.S. workers.

Inspired by this scenario, we intend to create an analysis to answer

1. Prediction of visa decisions based on employee/employer/wage
2. Which countries have the highest number of Visa Applications?
3. Which destination states have attracted the majority of visa applicants in the US?
4. Which destination cities have attracted the majority of visa applicants in the US?
5. What do the number of applications per month look like?
6. How have the number of US visa applications changed over the years?
7. What is the change in number of distinct visa case statuses over the years?
8. F1 visa acceptance and rejection analysis

2 Methodology

We started the project by understanding the problem statement, and the available resources. After comprehending our dataset (Section 2.1) we then, resolved the architecture issues (Section 2.2). Then, we proceeded to analyse the data to draw meaningful insights.

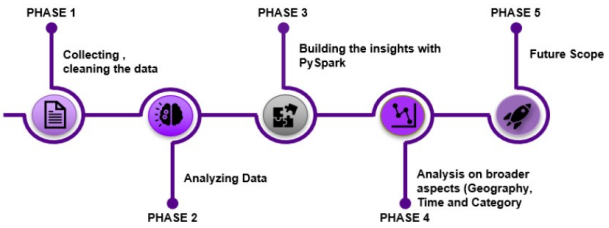

Figure 1: Flow of the project

2.1 Understanding the dataset

The dataset used for this project was collected from Kaggle LINK. Kaggle is an open data portal and it is a publicly available source. The data covers 2012-2017 and includes information on employer, position, wage offered, job posting history, employee education and past visa history, associated lawyers, and final decision. This dataset contains around 374,363 rows and 154 columns.

2.2 Architecture

To perform scalable analytics and machine learning we used PySpark and the libraries it comes with as well as other visualization libraries like plotly, matplotlib and seaborn. The aforementioned dataset can be ingested into PySpark whenever the visualization and prediction needs to be performed. The data which we have used for the overall analysis is reported by the US Department of Labor. Only the years 2012 to 2017 are included in the dataset and real-time data is huge and complex with a lot of missing values and changing formats. The technology architecture shown below can be used in situations where we are analyzing US visa applications across the globe.

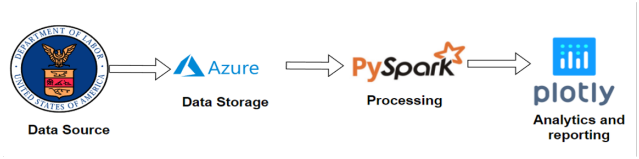

Figure 2: Architecture

2.2.1 Input Data Source

Input data will be the US visa applications as reported by the US Department of Labor. There are two sources of data collection. The Department of Labor (DOL) issues a permanent labor certification that authorizes a business to hire a foreign worker to work permanently in the United States. In most cases, a certified labor certification application from the DOL's Employment and Training Administration is required before a U.S. business can file an immigration petition to the Department of Homeland Security's U.S. Citizenship and Immigration Services (USCIS) (ETA). The Department of Labor must certify to the USCIS that there are insufficient American workers who are able, willing, qualified, and available to accept the job opportunity of intended employment and that the foreign worker's employment will not have an adverse effect on the wages and working conditions of similarly employed Americans.

2.2.2 Azure Data Source

To store data in the databases, Target Data stores will subscribe to the Azure broker. Any relational database, including PostgreSQL, NoSQL, or a distributed storage system, can be used as the data store. These may be owned by various organizations from various cities that are interested in subscription and data collection.

2.2.3 Apache Spark

Spark is the heart of the processing in this architecture. The in-memory computation of spark enables fast computation for large datasets. In the scalable environment, we can expect thousands and millions of entries coming in every day, and spark is able to handle all the operations from preprocessing to publishing this data. The data filtered by using Spark processing is used for reporting and monitoring.

2.2.4 Data Visualization

In our project, data monitoring and visualization are crucial. To display important information to the user, programs like Plotly and matplotlib can be used. When someone is moving to the United States, this visualization can be helpful in analyzing the distribution of applicants by nation and the frequency of applications by month, which can then be divided into categories based on the type of visa. They can keep an eye on the data and view it, which will provide them with useful information. We deleted it after observing that the column for country of citizenship contained some null entries.

3 Analysis:

3.1 Geography

3.1.1 Distribution of Visa Applications by Country

1. We can see that India tops the list of US visa applications followed by China and South Korea.
2. The Asian subcontinent accounts for the majority of visa requests.
3. The increased rates of visa denial in certain nations may also be the cause.

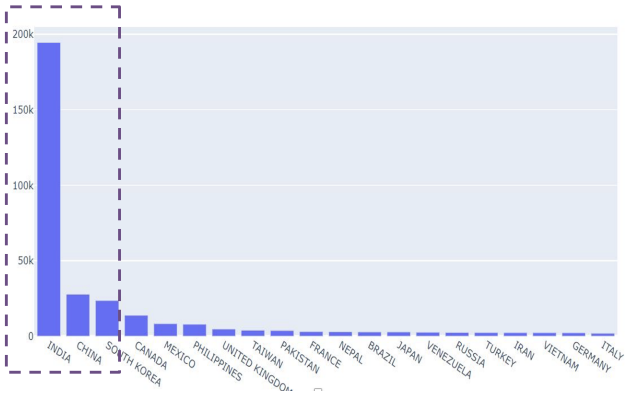

Figure 3: US Visa Application by Country

3.1.2 Destination states that have attracted many visa applicants in the US

1. We can confer from the below plot that the most attracted states for visa applicants are California, Texas, New York, New Jersey.

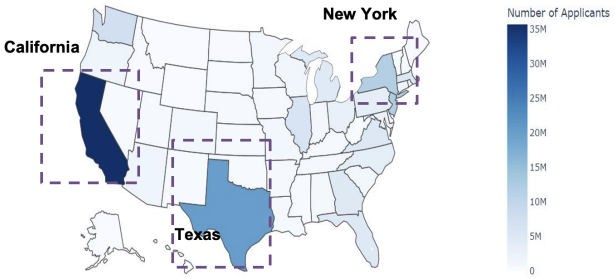

Figure 4: US Visa Application by States

Reasons:

1. Graduates coming out of Californian universities have gone on to found some of the most well-known tech companies.
2. At least 50 countries are represented at NYU, with most international students coming from China, India and South Korea.
3. A quarter of all international students enrolled in universities in Texas come from India.

3.1.3 Destination cities that have attracted many visa applicants in the US

1. According to the below plot, New York City, College Station, and San Jose are the three most popular cities for visa applications.

Figure 5: US Visa Application by City

3.2 Time

3.2.1 The Number of US Visa Applications over the years

1. We can see that from 2012 to 2017, the number of visa requests for the United States surged rapidly.
2. We can see that between 2012 and 2017, the number of visa applications increased by around ten times.
3. The highest increase by percent was between year 2013 to 2014.

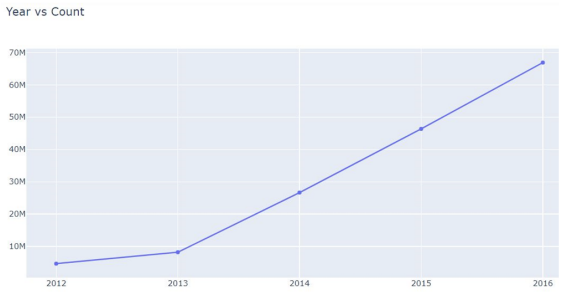

Figure 6: US Visa Application by Year

3.2.2 The Number Applicants per Month

1. The visa applications throughout the year have been almost the same. Following March and January, the month of November saw the largest number of visa applications.

Figure 7: US Visa Application by Month

3.2.3 Distinct visa case statuses change

1. There has been a significant increase in the number of certified visa applications.Between 2013 and 2016, the number of visa applications that were either rejected or withdrawn remained unchanged.

2. The number of certified but expired visa applications has likewise slowly but steadily increased over the years.

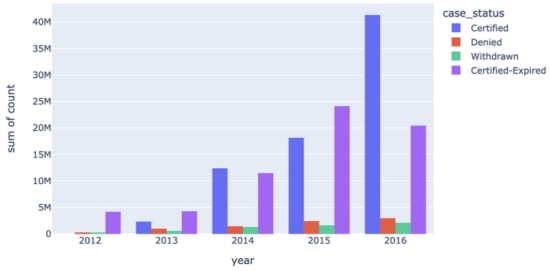

Figure 8: Distinct visa case statuses change.

3.3 Category

3.3.1 The Distribution of the different visa case status

1. The percentage of visas certified is 48%. About 40% of the visas were certified but were already expired. It would be convenient to state that around 88% of all visas were certified.
2. Only approximately 5% of those who applied withdrew their application, and a very small percentage of visa requests were turned down.

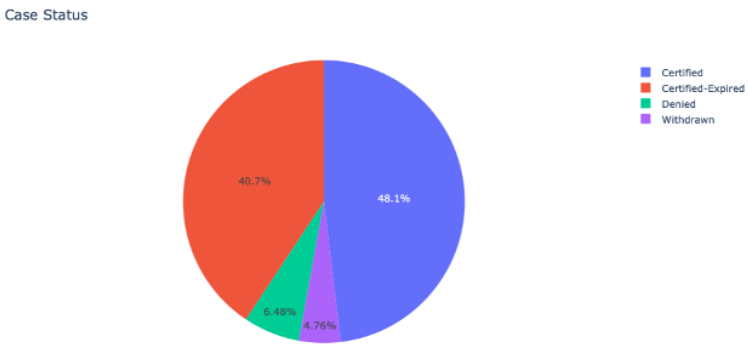

Figure 9: US Visa Application by different Visa Case

3.3.2 The distribution of the different visa types

1. As may be seen, the vast majority of visa requests are of the H-1B variety. This indicates that there are a lot of employment options in the US.
2. The F-1 visa, which receives about 5% of applicants, is another lucrative area. This explains the lax regulations governing educational institutions in the US.

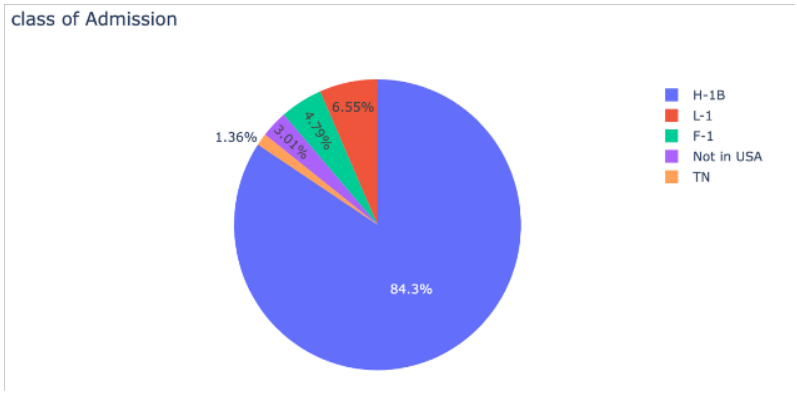

Figure 10: US Visa Application by class of Admission

3.3.3 The distribution of acceptance for F1 visas for different educational backgrounds

1. From a broad perspective, it appears that most candidates have a history in medicine.
2. This is due to strong demand among applicants brought on by the US health sector's expanding developments.

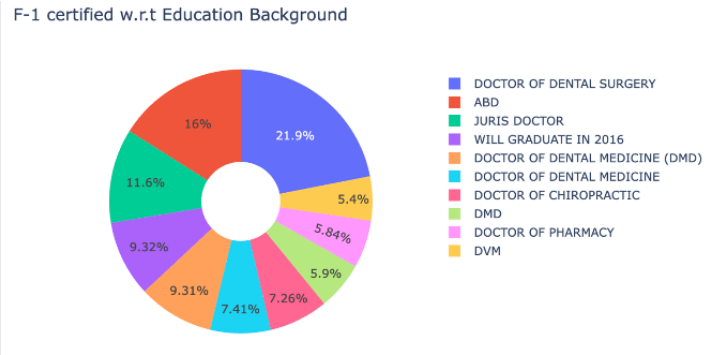

Figure 11: Visa Application acceptance- Study

3.3.4 The distribution of rejection for F1 visas for different educational backgrounds

1. Again and, the F1 visa category has the greatest rejection rates for educational backgrounds in the medical field.This is because there are so many applications from students in the medical field.

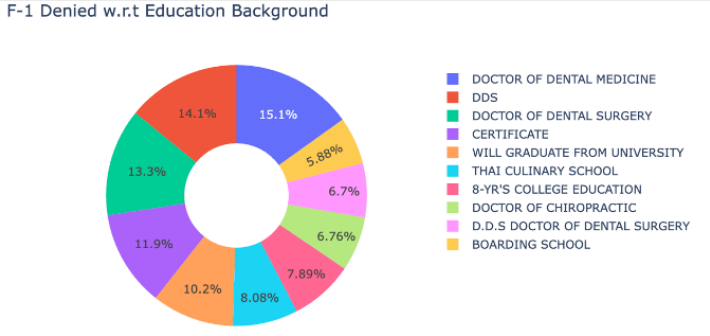

Figure 12: Visa Application Rejection-Study

3.3.5 The Distribution of Acceptance for F-1 visas for different countries

1. India leads in the acceptance of F-1 visas. It is followed by South Korea and China.The high volume of applications from India, where a sizable number of students are looking for educational opportunities in the US, can be attributed to this.

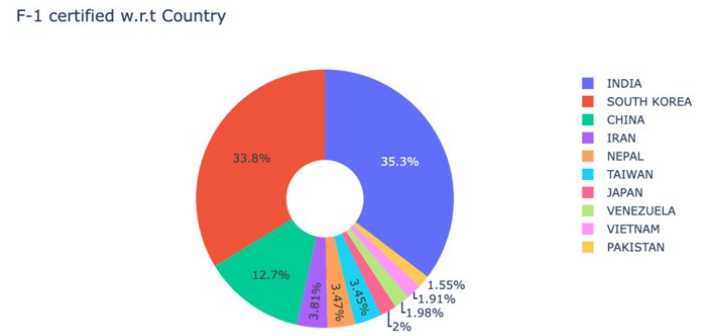

Figure 13: Visa Application acceptance- Country

3.3.6 The distribution of rejection for F-1 visas for different countries

1. We can see that contradictory to acceptance, South Korea leads the rejection rate followed by India and China.

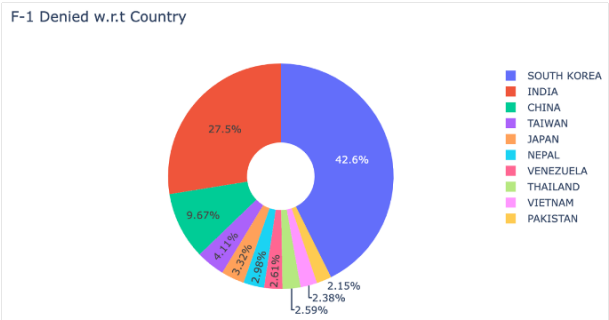

Figure 14: Visa Application Rejection-Country

3.3.7 Distribution of H-1B visa ordered by Employer State

1. In comparison to other states, we can see that Texas and California have exceptionally high acceptance rates. This reveals the inequality in opportunities among states. The rejection rate is rather stable throughout all states when compared to acceptance.

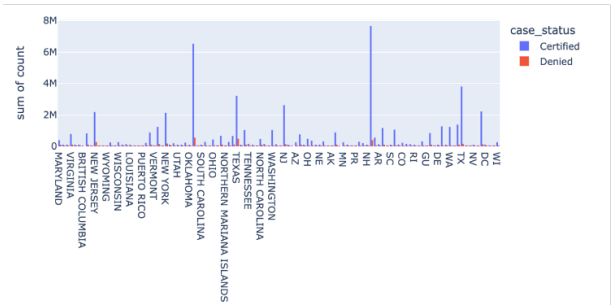

Figure 15: US Visa Application by Employer State

3.3.8 Distribution of H-1B visa ordered by Employer

1. We can see that Golden Valley Health Centers has the highest number of certified visa applications.

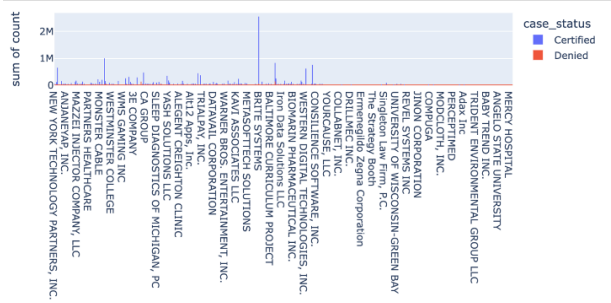

Figure 16: US Visa Application by Employer

4 Future Scope

1. Study COVID-19 data to analyze the impact of the global pandemic on US Immigration statistics.
2. Implement a Machine Learning model to classify Visa decisions based on given input attributes.

5 References

1. Link to the Dataset
2. https://www.dol.gov/agencies/eta/foreign-labor/programs/permanent
3. https://idl.cs.washington.edu/papers/immens/
4. Guide to Spark-ML
5. Spark- SQL
6. Library-Plotly