CS432/532: Final Project Report

Project Title: H2B Visa Analysis

Team Member(s): Jadhav Mrugank, Vaidya Rujuta

I. Problem

With changes in visa policies, we have tried to identify which factors have a major impact on certification or rejection of visas. This dataset comprises H-2B petition data from the last five years, totalling around 40 thousand entries. Case status, employer name, workplace coordinates, job description, prevailing pay, occupation code, and year filed are some of the columns in the dataset. The raw data given is disorganized and not immediately suited for study. A series of data changes were done to make the data more accessible for speedy examination. The dataset used was taken from GitHub - https://github.com/BuzzFeedNews/H-2-certification-data

This project will evaluate immigrant data to uncover crucial correlations. The major findings of the investigation will be used to know if a person applying for a visa has a good chance of acceptance. Based on the main insights, the final result will be chances of visa certification based on the factors in the dataset.

Analysis tasks performed on the dataset are as follows:
1. Percentage of acceptance of visa depending on employers.

2. Geographical distribution of occupation grouped on the prevailing wage of that occupation.

3. Probability of certification of visa depending on ethnicity of applicant.

II. Software Design and Implementation

A. Software Design and NoSQL-Databse and Tools Used

The database used to import and hold data is MongoDB. The Python package Pymongo enables communication with the MongoDB database. Pymongo is used to connect MongoDB database with Jupyter Notebook.
Jupyter Notebook is used to analyze the imported data. This is done using various python libraries such as pandas, numpy, matplotlib, seaborn, plotly, folium. Jupyter Notebook has allowed us to create and display charts
and formulae to make the analysis more comprehensible.

B. Parts that you have implemented

Implementations that are done for this project are as follows:

1. Importing data to MongoDB and connecting it with Jupyter Notebook. We have used it as MongoDB database can handle large amounts of data extremely efficiently. MongoDB has no restrictions on schema design. Unlike traditional relational databases, where data is stored in clearly defined tables and columns, with each column containing a very specific type of data. MongoDB provides dynamic schemas for unstructured data. This is connected to the local host and the database is added to Mongo compass in the form of a json file. This is converted into a pandas dataframe using the find function. After this, the data is explored and preprocessed.

2. Data preprocessing steps are implemented to explore the dataset and visualize the data in order to understand it and find correlations in it. This is done by finding null values in each, it is understood that the most useful columns are case status (Last_Sig_Event), Prevailing Wage (Prevail_Wage), Employer's State (Emp_state), Job Title (OCC_Title) and Date of Case (Last_Event_Date).

3. It is observed that the prevailing wage is not normalized, it is in various forms like daily, weekly, monthly, and annually. So, this data is normalized by performing normalization operations on them. The date of the case is also not in the correct format making it useless. Two new columns are created extracting month and year from the date column giving is opportunities to calculate and corelate using this.

4. After normalizing data and preparing it for further analysis, we perform exploratory data analysis on the data. The graphs that have been visualized are:
    a. Counts of cases with their case status.
    b. Percentage wise distribution of cases according to year and month.
    c. Ratio of certified and denied cases according to year.
    d. Distribution of cases according to job titles.
    e. Employers that employ the maximum employees of immigrant status.

Few visualizations are shown below:

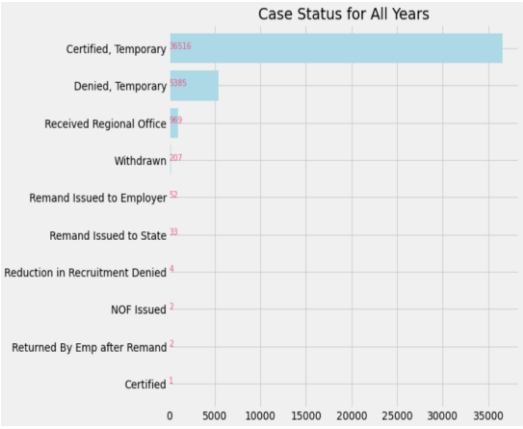

Fig 1 : Case status counts according to their types

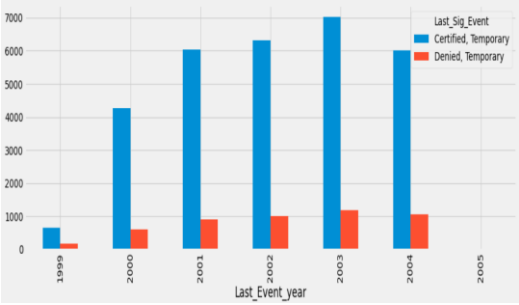

Fig 2: Distribution of certified vs denied according to year

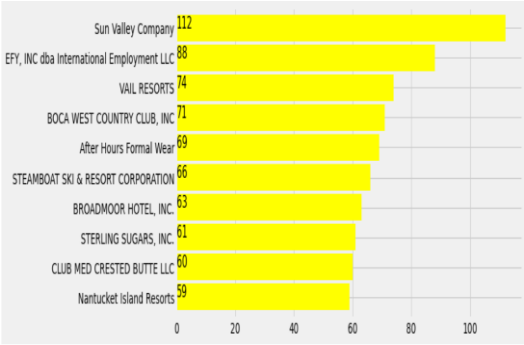

Fig 3: Distribution of people in companies

5. After exploring the data, we are now ready to perform the analysis. We understand that the data being non-numeric, we need to perform mapping and merging of columns to perform analysis as heat maps do not provide sufficient correlation between data to perform prediction level analysis.

III. Project Outcome

We have the following findings from the project:

1. Employers have a significant impact on the acceptance rate of their employees. Depending on the status of the previous visa and employers, we have sorted this data to map them to join the status column with the employer column using the merge functionality of python.

   As displayed in the table below, we see that The Brickman Group, Ltd has maximum acceptance rate of 98%, followed by Harbor View Hotel and Okemo Limited Liability Company with 97% acceptance.

	Emp_Name	Acceptance_rate
17	The Brickman Group, Ltd.	0.980000
21	Harbor View Hotel	0.977273
22	Okemo Limited Liability Company	0.976744
27	WESTIN HOTELS AND RESORTS	0.974359
33	PINEHURST RESORT AND COUNTRY CLUB	0.972222
...	...	...
14664	Mamma DiSalvo's Restorante	NaN
14665	DANNY O's LANDSCAPING	NaN
14666	Mama Rosa's Place, LLC	NaN
14667	DANNY VELAZQUEZ RACING STABLES	NaN
14668	A & M UNDERGROUND IRRIGATION SYSTEMS	NaN

Fig 4: Acceptance rate of people according to employees

2. For the second analysis, we have created a new column that replicated the employer states column as a column named ‘gcode’. This column is then replaced with all the full forms of states in the US and the abbreviations that were invalid or null were checked for and deleted or ignored if they were not in significant columns to be analyzed.

   The new column consists of names of states of employers and using an API key from OpenCage, we have extracted the latitude and longitude of each of the states. According to job title, top occupations are found and are separated into separate dataframes. These dataframes are grouped according to wages. This is then plotted using folium to get the density distribution of occupation according to wages.

   In the map shown below, the occupation is considered as Laborer and we see the outcome that a maximum density of people with the occupation laborer is on the east coast and is lightly distributed across the mid-states with almost no distribution on the west coast.

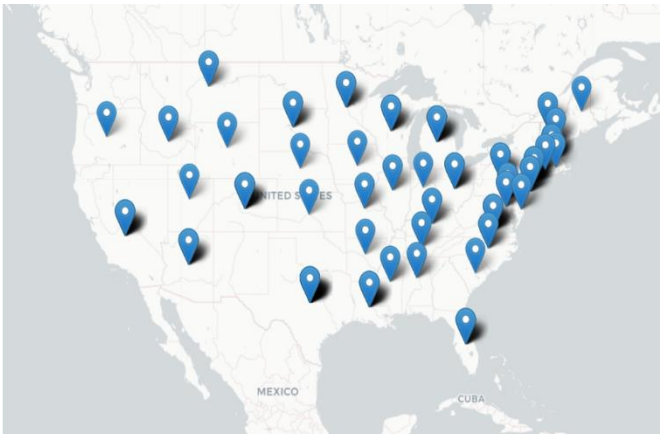

Fig 5: Distribution of laborer according to wages and employer location

This can be done for all the occupations and according to wage and location columns, we can determine the density of the job distribution.

3. We have a column for ethnicity, this is merged with the case status column. Using one hot encoding, we have encoded the five categories of ethnicities. This encoding is used to calculate the counts and distribution of ethnicities.

   From this, we understand that the ethnicity – 'Caucasian' is the one with the maximum counts. Later we visualize the ratios of various ethnicities with the case status. This can be seen from the table shown below:

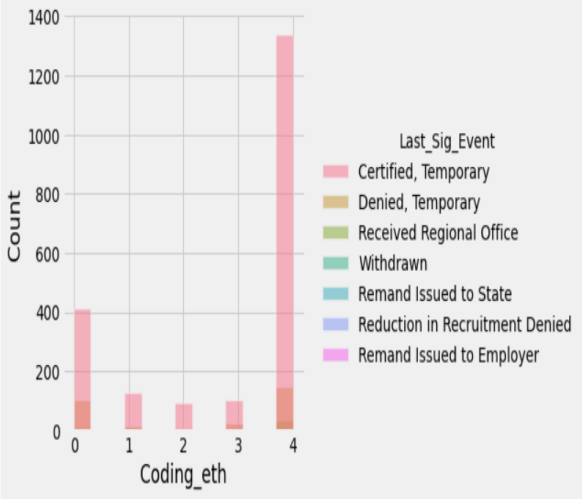

Fig 6 : Ratio of case status according to ethnicities.

After understanding this data, we have calculated the probability of getting a certified visa according to ethnicity.

This is done by counting the value of certified number and total number of applicants of every ethnicity separately and a new table is created. The probability of this is calculated and displayed in a new table (shown below).

	Ethnicity	Probablity of Certified
0	Hispanic	0.7729831144465291
1	Asian	0.8269230769230769
2	African	0.8888888888888888
3	Italian	0.7878787878787878
4	Caucasian	0.8767213114754099

Fig 7 : Probability distribution according to ethnicities

As we can see that, the highest probability of certification if of people with the ethnicity of 'African.'

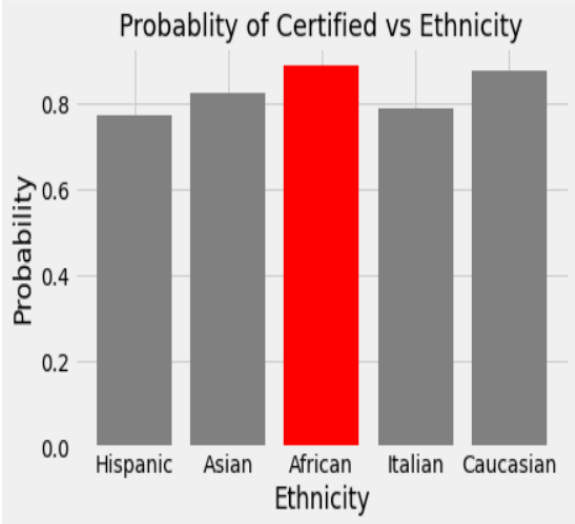

Fig 8 : Highest probability of certified visa according to ethnicity.

We have also visualized this using a table that clearly shows this output.

Results of the project:
1. The Brickman Group has the highest probability of acceptance with an acceptance rate of 98%.

2. The people with jobs of category laborer have a great density of jobs on the east coast.

3. The highest number of applicants is contrasting with the highest probability of acceptance.

References

[1] https://github.com/BuzzFeedNews/H-2-certification-data
[2] https://www.mongodb.com/home
[3] https://pandas.pydata.org/
[4] https://www.kaggle.com/
[5] https://matplotlib.org/
[6] https://medium.com/@pragya_paudyal/connecting-mongodb-to-jupyter-notebook-e3f636a85830
[7] https://python-visualization.github.io/folium/
[8] https://opencagedata.com/
[9] https://stackoverflow.com/