**Analyzing H-1B Visa Petitions**

**Yash Bavishi, Xingchen Zhou, Jennifer Liang**

**Data Dictionary**

The following fields in the dataset will be used in our analysis:

1. **CASE\_STATUS**: Status associated with the last significant event or decision. Valid values include “Certified,” “Certified-Withdrawn,” Denied,” and “Withdrawn”.
2. **EMPLOYER\_NAME**: Name of employer submitting labor condition application.
3. **SOC\_NAME**: Occupational name associated with the SOC\_CODE. SOC\_CODE is the occupational code associated with the job being requested for temporary labor condition, as classified by the Standard Occupational Classification (SOC) System.
4. **PREVAILING\_WAGE**: Prevailing Wage for the job being requested for temporary labor condition. The wage is listed at annual scale in USD. The prevailing wage for a job position is defined as the average wage paid to similarly employed workers in the requested occupation in the area of intended employment. The prevailing wage is based on the employer’s minimum requirements for the position.
5. **YEAR**: Year in which the H-1B visa petition was filed
6. **WORKSITE**: City and State information of the foreign worker's intended area of employment

Total number of entries in the dataset: 3002458

Total number of unique EMPLOYER\_NAMES: 236014

Total number of unique WORKSITE: 18622

Total number of unique SOC\_NAME: 2133

CASE\_STATUS counts:

1. ‘CERTIFIED’: 2615623
2. ‘CERTIFIED-WITHDRAWN’: 202659
3. ‘WITHDRAWN’: 89799
4. ‘DENIED’: 94346

**Preprocessing the Dataset**

Since the dataset is well structured, it requires minimal preprocessing. However, we did run into a few problems and the following are how we fixed the data.

1. Filtered data with invalid value or empty data field
2. Fixed typos in employer name and worksite
3. When creating feature vectors, got rid of special characters and convert all uppercase letters to lowercase letters

**Clustering**

The goal here is to identify various industries that hire international employees. We will cluster the dataset on national level.

*Data processing for Clustering:*

Steps:

1. Identify unique pairs of Employers and their corresponding location from ‘EMPLOYER\_NAME’ and ‘WORKSITE’ respectively.
2. For each unique pair identified we extract all the soc\_names associated with it. This data is then vectorised.
3. A feature vector is then created for each unique pair in the following format:

[ ‘WORKSITE’, \*Vector of SOC\_NAME\* ]

These vectors are fed to clustering algorithms.

We will run K-Means++, Hierarchical, and GMM methods and compare the results.

Next step would be to repeat the process but after the dividing the data based on the year. By doing this we expect to see changing trends over the past 5 years.

**Statistical Questions**

**Compare occupation and case status:** For this statistical analysis, we will determine the top 25 most frequently occurring job titles in the data set and for each job title, determine the percentage of applications that were “Certified” or “Certified-Withdrawn” out of the total number of applications for the occupational name. This information will be displayed in the form of a table.

**Top employers with most applicants:** We will determine the top 25 most frequently occurring employers in the data set and for each employer, determine the percentage of applications that were “Certified” or “Certified-Withdrawn” out of the total number of applications for the employer. The results will be provided in the form of a table and a bar graph.

**Employers paying most wages to applicants:** We will determine the top 25 employers that employ workers with the highest prevailing wages. In order to determine these employers, we will average the prevailing wages of the workers of each employer, sort the employers, and retrieve the 25 employers with the highest average. We will display this information in the form of a bar graph.

**Number of applications a year and percentage certified per year:** We will plot the number of applications per year in a line graph. We will determine the percentage of applications that were “Certified” or “Certified-Withdrawn” out of the total number of applications per year for each of the years and display this information in the form of a table and a bar graph.

**Correlation between certification and wage:** We will determine if there is a correlation between whether or not an application is “Certified”/”Certified-Withdrawn” and the wage of the applicant by first creating tuples for each of the applicants in the form of (Certified, Wage) with the Certified value equalling 1 if the application is certified and 0 if it is not. With this data, we will determine the correlation coefficient and p-value to determine the significance of the correlation. We will create a histogram of the wages of certified applicants to see the distribution of the wages.

|  |  |  |
| --- | --- | --- |
| **Deadline** | **Subject** | **Assigned to** |
| 11/27 | Clustering | Xingchen Zhou, Jennifer Liang |
| 11/27 | Classification | Yash Bavishi |
| 12/4 | Statistical Questions and Visualizations | Xingchen Zhou, Jennifer Liang, Yash Bavishi |
| 12/12 | Poster | Xingchen Zhou, Jennifer Liang, Yash Bavishi |