Data Visualization & Exploration: Project Report

Nidhi Mundra, 3106511

May 11, 2017

1 Data Information

- **Selected Data:** H1B Visa Petitions
- **Link:** `https://www.kaggle.com/nsharan/h-1b-visa`
- **Description:** The H1B visa is an employment-based, nonimmigrant visa category for temporary workers. For such a visa, an employer must offer a job and apply for your H1B visa petition with the US Immigration Department.

  This dataset contains H-1B Visa Petitions for the period of 2011 till 2016. It contains information like job type, job title, work location, wage, etc and the case status which was provided by the Office of Foreign Labor Certification (OFLC).

  I picked this dataset because I am interested in finding out which locations, employers, job titles and salary range makes H1B petition certification more favourable. I am also interested to analyze the cases of Full time and Part time employees separately. And, for my analysis, I have used data of only the year 2016. This is because of the extremely large size of the data and my analysis did not require the data of all the years.

  Various Case Statuses are:

  - Certified
  - Certified Withdrawn
  - Withdrawn
  - Denied
- **Number of Features in Original Data:** 11
- **Number of Features in Final Data:** 15
- **Size of Original Data:** 106MB
- **Size of Data Used:** 20.3MB
- **Total Number of Records:** 60,261

Feature	Description	Data Type	Present/Inferred
ID	Unique Identifier of a record	Integer	Present
CASE_STATUS	Status associated with the last significant event or decision	String	Present
EMPLOYER_NAME	Name of employer submitting labor condition application	String	Present
SOC_NAME	Occupational name	String	Present
JOB_TITLE	Title of the Job	String	Present
GROUP	Grouped SOC_NAME	String	Inferred
FULL_TIME_POSITION	Boolean value to describe whether the applicant has a full time position or not	String	Present
PREVAILING_WAGE	Prevailing Wage for the job being requested for temporary labor condition	Integer	Present
YEAR	Year of application	Integer	Present
WORKSITE	City information of intended area of employment	String	Present
STATE	State information of intended area of employment	String	Inferred
LON	Longitude of the Worksite	Float	Present
LAT	Latitude of the Worksite	Float	Present
STATE_CODE	State Code	String	Inferred
POSITION_STATUS	Boolean to identify positions favourable for getting H1B	String	Inferred

Table 1: Features of the Final Data

2 Data Analysis

2.1 Issues And Their Fixes

- **Missing LAT and LON:** The LAT and LON values of 2667 records were missing. To fill in the missing values, I chose not to use any traditional method like interpolation, duplication or neighbor voting. Instead, I queried Googles maps API which returned me the location coordinates of the requested city.

  Sample Request: `http://maps.googleapis.com/maps/api/geocode/json?address=AMHERST`

  I parsed the API response and filled the missing values. However, I faced the following issues in this process:

  – Google did not allow me to make too many requests in one second. I limited the number of requests by added a wait timer in my script.

  – 404 Response: For 13 cities, I got a response with 404 status code. To fix this, I filled these values in my data manually.

- **Wrong LAT and LON:** The LAT and LON of 36 cities were wrong and did not lie in the US. For fixing this, I used the same logic of extracting the correct LAT and LON information from the Google API.

- **Missing SOC_TITLE:** SOC_TITLEs of 46 records were missing. For fixing this, I trained a linear model to predict the titles of the remaining data.

  – *Features:* CASE_STATUS, EMPLOYER_NAME, FULL_TIME_POSITION, WORKSITE, STATE, JOB_TITLE, PREVAILING_WAGE

  – *Model:* Linear Regression

In order to check whether the SOC_TITLE were correctly predicted or not, I checked all the records, which had missing SOC_TITLEs, manually.

- *Data Cleaning:* The columns 'CASE_STATUS' and 'SOC_TITLE' had inconsistent data. For e.g.: Some case statuses were 'Certified' and others were 'CERTIFIED', and same title 'Market Research Analyst' was termed as 'Market Research Analysts' at some places. For getting rid of such discordant data, I used regular expressions and replaced the text in the database.

2.2 Inferring New Features

- **States Information:** The states of the location were not given in the data. I used the Google Maps APIs results for filling the state information correctly. Also, along with the state names, I appended the state code in the data.

- **SOC Group:** For grouping the SOC_TITLEs, I performed clustering on the data.

  - *Features:* CASE_STATUS, EMPLOYER_NAME, FULL_TIME_POSITION, WORKSITE, STATE, JOB_TITLE, PREVAILING_WAGE
  - *Model:* K-means Clustering with $K = 10$

  Further, I assigned the cluster label manually by looking at the type of the data in each clusters. In total, there were 7 clusters formed, on which I assigned the following labels:
  Chief, Accountancy, Business, Analytics, Manager, Technology and Others

- **Position Status:** For each Job Position, I intended to classify them as 'Likely' or 'Unlikely', referring to whether they are more probable for getting H1B or not. For this, I added one more feature to the, namely COUNT, which is the number of H1B petitions Certified for that particular Job Position. Further, I performed clustering based classification using K-Means.

  - *Features:* CASE_STATUS, EMPLOYER_NAME, FULL_TIME_POSITION, WORKSITE, STATE, JOB_TITLE, PREVAILING_WAGE, COUNT
  - *Models:*
    - *KMeans Clustering:* First, I clustered the records into two clusters using K-Means Clustering Algorithm. This process divided the data into two groups.
    - *Classification:* Next, I analyzed both the clusters and labeled one of the clusters as 'Likely' and the other one as 'Unlikely'.

  In total, around 39% records were classified as 'Unlikely' and remaining 61% as 'Likely'.

2.3 Summarizations

- For each state,
  - Total number of visa petitions
  - Total number of visa rejected
  - Total number of visa accepted

- For each employer,
  - Total number of visa petitions
  - Total number of visa rejected
  - Total number of visa accepted
  - Average salary of employees

- For each job position,

  - Total number of visa petitions
  - Total number of visa rejected
  - Total number of visa accepted
  - Average salary of employees with a visa approval

2.4 Key Interests

Questions Answered

- I am interested in finding out the job positions which are more likely to for getting an H1B. I am also interested in digging deeper into the ones which are unfavourable and reduce the chances of getting a work visa.
- Which region in the US files the maximum number of petitions and what is the general distribution of case status in each region? Which region is more / least favourable for getting an H1B?
- I want to find out the effect of wage and employer on visa case status. Are people with higher wages only the ones getting H1B? What should be the minimum wage to be in a have a good probability of getting a visa?
- Are full time employees more favourable to get an H1B?
- Which employers in a state are more favourable for H1B approval?
- Which state files maximum petitions?

Questions Unanswered

- Is the number of H1B granted per year is increasing or decreasing?
- Are the number of applicants increasing or decreasing per year?
- Which employers file the most petitions per year?

Reason: I could not use the data for year-wise analysis because that was diversion from my analysis on the current trend. For the last part, I had to analyze a very number of employers and I could not figure out a good visualization which could describe this well enough.

More Exploration Done

- What percentage of job positions are favourable/unfavourable for a certified H1B petition?
- Which are the top paying employers who are favourable/unfavourable for full time/part time employees, with specific or generalized job position?
- What is the job position hierarchy and what is the trend of H1B petitions at each level?

3 Webpage Link

My webpage consists of two separate links for the viewing visualizations on the full data and the sampled data. I added two separate links to see the interactions quickly as well as see how the graphs look like with the whole data.

URL: `http:\edlab-www.cs.umass.edu/~nmundra`

4 Visualizations and Interactions

4.1 Geospatial Heatmap

This heatmap shows the trend of visa petitions across different states of the US. Different select options provided are:

- Total no. of Petitions: Selecting this option draws a heatmap which shows the total no. of visa petitions filed in all the states of the US.
- Total no. of withdrawn cases: Selecting this option draws a heatmap which shows the total no. of visa petitions which were withdrawn or certified-withdrawn in all the states of the US.
- Total no. of rejected cases: Selecting this option draws a heatmap which shows the total no. of visa petitions which were rejected in all the states of the US.
- Total no. of certified cases: Selecting this option draws a heatmap which shows the total no. of visa petitions which were certified in all the states of the US.

Interactions provided are:

- Selection: Select state(s)
- Probing: Probing on a state displays the statistics of that state

Inference:

4.2 Pie Charts

Position Type

This chart shows the percentage of full time employees and part time employees across different states in the US.
Interactions provided are:

- Selection: Select Position Type

Inference: Nearly equal number of records were of full time and part time employees.

Likely vs Unlikely Job Positions

This chart shows the percentage of job positions which are favourable and unfavourable for certified H1B petitions.
Interactions provided are:

- Selection: Select Position Type

Inference: Lesser number of positions were unfavourable for certified H1B.

Case Status Type

This chart shows the percentage of each case status of H1B petitions.
Interactions provided are:

- Selection: Select Case Status Type

Inference: In the data, maximum records were of certified petitions followed by certified withdrawn, withdrawn and denied respectively.

4.3 Scatter Plot

This chart shows the salary range across different states in the US.
Interaction provided are:

- Selection: Select a data point
- Probing: Displays Salary, Position and the Location

Inference: Full time employees get much more salary as compared to part time employees across all the states in the US.

4.4 Row Chart

This chart displays the count of H1B petitions across different states of the US.
Interactions provided are:

- Selection: Select one or multiple state(s)
- Probing: Displays the count of petitions in the state

Inference: California files the maximum number of petitions across US.

4.5 Bubble Chart

This chart displays the count of H1B petitions filed and average salaries of various Job Positions.
Interactions provided are:

- Selection: Select a Job Position
- Probing: Displays Job Title and the count of H1B petitions

Inference: Maximum petitions were filed for Business Analysts in 2016.

4.6 Sunburst With Line Plot

This chart displays the hierarchy of various Job Positions in 2016. The line plot for the corresponding hierarchical level shows the trend of various case statuses of the selected hierarchy. The line plot changes as one goes deeper into the hierarchy.
Interactions provided are:

- Mouse hover: On mouse hover of any level of the hierarchical map, the line plot changes and displays the trend of corresponding job title.

Inference: Maximum applications were for managers and most H1B petitions were certified for Management Analyst, under Analytics followed by Technology.

4.7 Data Table

This table list the top paying employers with their respective average salaries.

5 Technologies Used

- Mongodb v3.4.3: To store all the data
- Python v2.7.12: For data preprocessing and analysis
- D3.js v4.7.4
- DC.js v2.1.4
- Javascript with JQuery v3.1.1