ENIGMA DATA CHALLENGE

Benjamin Pujol

SECTION 1: Warm-up

1. The company that applied for the most visas in NY is MPHASIS Corporation with 946 applications. One problem encountered was that New York is referred as "New York", "NY" or "NYC" in the database
2. For New York:
   Mean salary of $96,504
   Standard deviation of $133,850

   For Mountain View:
   Mean salary of $118,795
   Standard deviation of $34,370

   Average wage is higher in Mountain View but the standard deviation is lower than New York. This can be explained by the fact that Mountain View gathers the same kind of high-paying job in technology companies such as Google whereas jobs in New York are much more varied.

3.

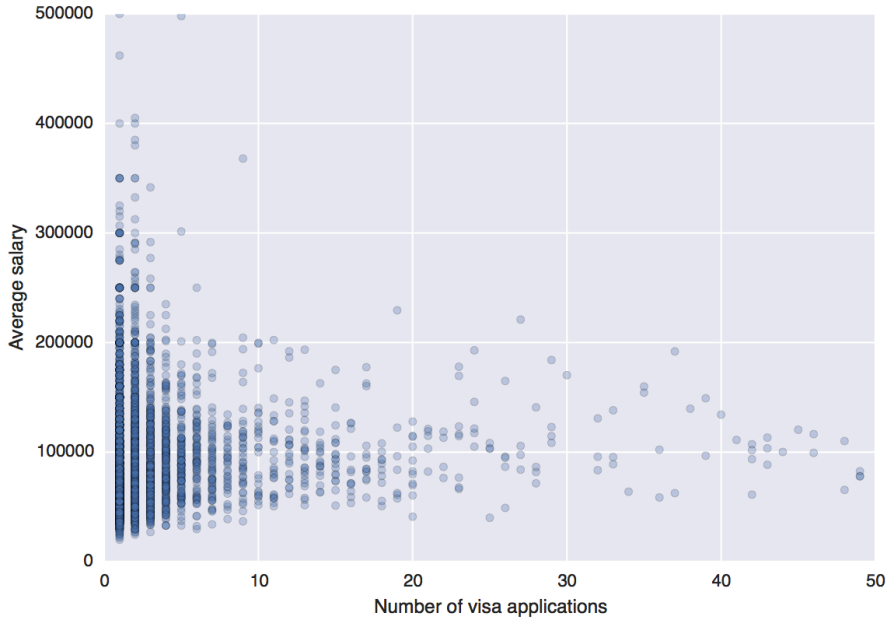

Figure 1: Relationship between number of applications and average salary in New York

What we learn from this distribution is that when the number of visas applied for increase for a company the wages tend to be much more gathered around a mean value of 100,000$. This can mean that company that hire a lot of foreign workers hire them for the same kind of position which are "mid-rank" positions.
Between 0 and 10 applications we can see that the standard deviation for wages is much higher.

SECTION 2: Brainstorming

The depth of knowledge contained in this dataset is impressive. It basically gives us access to the wage of every H1B worker for a specific company and a specific position. From this information we can get a lot of insights on the US job markets and the immigration of foreign workers

1. For a specific job position in a specific city we can build the statistical distribution of wage. What is the salary for this job? Do we need to pay a foreign worker more than a US candidate? Is it financially interesting for our company? The recruiter can then adjust the salary based on the profile of the candidate: is he experienced or not?

Implementing this is rather straightforward:
- Select the job position considered using pandas Dataframes
- Plot the distribution of salaries for the job position

2. On a geographical point of view, this dataset can give insights on how the US job market (for foreign workers) is organized. By looking at the type of job that dominates in a particular region of the country or of a city, we can build a map of the US based on jobs!

The first step is to analyse geographically the datasets:

- First let's have a global view: which states have the most applications for H1-B visas (every job position considered)?
- Split job positions in clusters of related jobs (Engineering, Business, Education, Healthcare, Agriculture, ...)
- Analyse the distribution of these clusters over the territory (we can even zoom at the scale of a city given that we have access to the post codes of each company)
- Visualize the data through plots and maps

3. We can also study what features impact the Certification or the Withdrawal of a visa. This is interesting for a company considering the time and financial resources needed to get a visa approved. Why apply for a visa for architects if they never get approved?

- The idea is to study the correlation between the different features and the "Status" feature
- For numerical variable, we can compute Pearson correlation coefficient
- For non numerical variables we can use visualization tools (company location, job position)

There are 6 possible statuses for a H1B visa:
- Certified: Employer filed the LCA, which was approved by DOL
- Certified Withdrawn: LCA was approved but later withdrawn by employee
- Withdrawn: LCA was withdrawn by employer before approval
- Denied: LCA was denied by DOL
- Rejected or invalidated

The repartition of the data set is the following:

Certified	455144
Certified-withdrawn	36350
Withdrawn	16069
Denied	11938
Rejected	2
Invalidated	1

This is a 2.5% chance of a visa to be rejected.

SECTION 3: Exploration

I decided to pursue on my second idea from the brainstorming session: mapping the US job market using H1B visas data. The first objective is to get a sense of which category of jobs dominates in a particular state of the country: is Texas a tech or a business state?

To do so, I tried to classified jobs in different clusters: Tech, Business, Education & Research, Other
This is a very basic clustering but it can be easily improved with additional time to get a really precise understanding

Classification:

- The goal is to analyse each job position title and be able to classify it as 'Tech', 'Business', 'Education' or 'Other' job
- To do so we have to do some Natural Language Processing (NLP):

  Example:

- "Director of Software Development" will be converted into the following list of key words : [director, software, develop] through NLP. I put the words in lower case, remove punctuation, remove stop words ('of'), and stemmed the words (development → develop).

- This is list of keywords is then mapped to the corresponding cluster for each key words. Tech:0, Business:1, Education:2, Other:3.

- For the clustering we take only the 50 most common key words, which cover most of the job positions. So some key words from the Dataframe may not be attributed to a cluster. For our example the list of keywords is mapped to [(None), 0, 0]
  We then take the minimum of this list and it corresponds to the cluster. Here it returns 1. So "Director of Software Development" is a Tech job

  Why the minimum? Because 0 corresponds to Tech jobs and a job position that contains a tech keyword and a business keyword can be considered as a tech job and not a business one.

  For example: "Software engineering consultant" which gives the key words [software, engin, consult] mapping to the list [0, 0, 1] is a tech job and not a business one.

- When the clustering is done we drop the few job positions that couldn't be mapped

- We can then visualize the repartition of different job clusters between different states as below (I run the program of 200,000 positions). The best visualization would be to project the results on a map

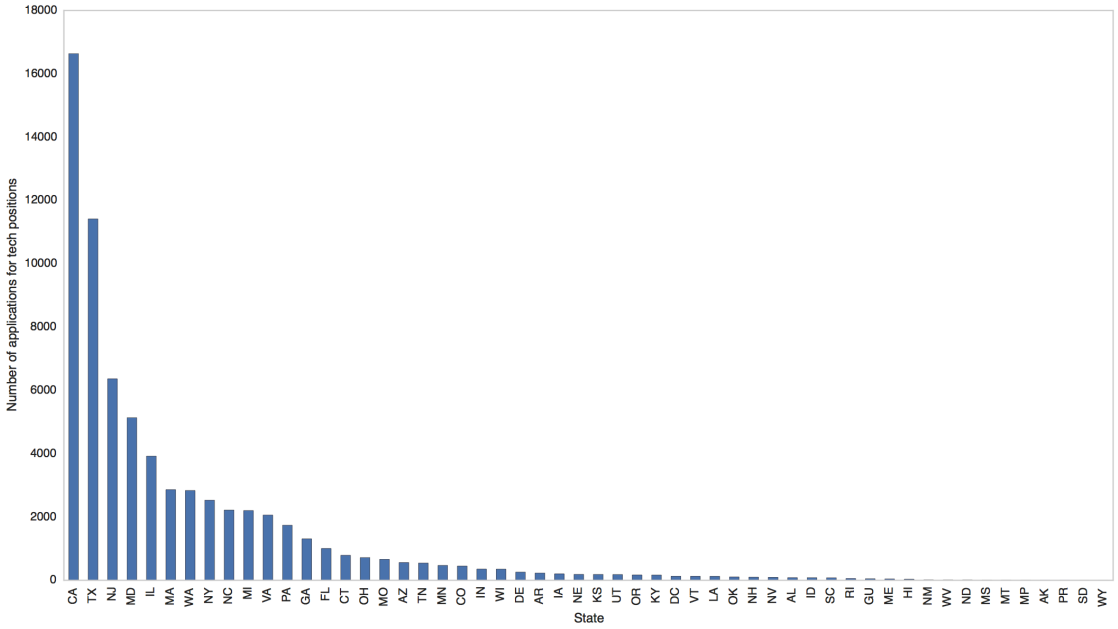

Figure 2: Number of applications for H1B visas for Tech positions per state

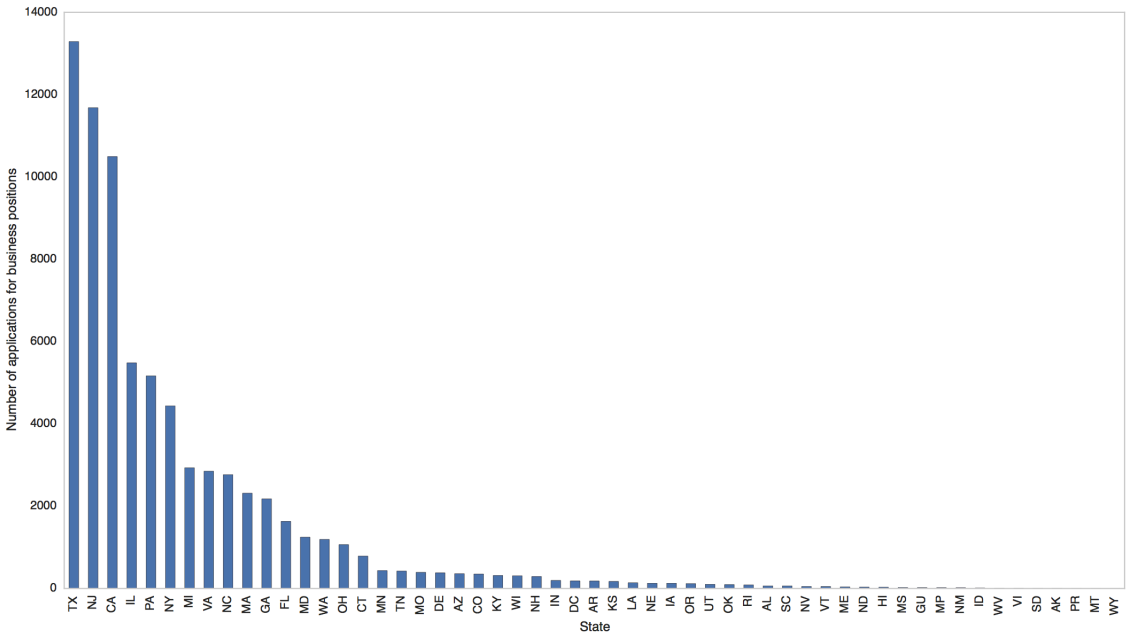

Figure 3: Number of applications for H1B visas for Business positions per state

Conclusion:

From this classification we can get lot of other insights on the US job market. Given the precise information we have about the employers we can zoom even more than states and map the distribution of jobs in a particular city. Concerning visualization, we can do a "scatter plot" of jobs on the map of the US (which would look like a satellite image of the country at night) and choose which kind of jobs we want to visualize. As well as that, the clustering can be much more precise than only 4 categories.