Section 1

Question 1. Which companies applied for : recoverable: many different cities with the name NYC (use state to select relevant rows as well)
where the job opening was located in NYC? Please describe any issues you may encounter summarizing the data by employer name.

Answer: To answer the above question one has to (1) identify all the rows that describe cases where the job location is in NYC, (2) group those rows on the column(s) that describe the name of the employer (i.e.'lca_case_employer_name'), (3) sort the grouped rows in descending order and (4) possibly retrieve the N (e.g. 10) employers the with the largest number of H1B visas.

After inspecting the schema of dataset and studying the description of its semantics it seems to me that there are two columns that describe the location (city) of the job posting, i.e. 'lca_case_workloc1_city' and 'lca_case_workloc2_city', so to get all the relative case we need to union the case where NYC is either the first or the second location.

The issues we we may encounter summarizing the data by employer name include:

- It may be the case that there are more than one different cities with the name NYC. To recover from that I will include the state when selecting the rows
- There may missing values or wrong values in the column that we need to filter rows with (e.g.there exists the value '12TH FLOOR' as city name)
- Word case (capitals, small letters) may vary
- There may be additional characters/info along with the city name. Examples that are found in the job city column include 'NEW YORK, NEW YORK', 'NEW YORK, NY', 'NEW YORK, NEW YORK 10003', 'NEW YORK,')
- The name of the city may be misspelled. Specific example 'NEW YORKI', 'NEW YOK', 'NEW YROK', 'NEW YORK CIY'
- Different names may be used for NYC e.g. 'NEW YORK', 'NEW YORK CITY',
- 'MANHATTAN', Brooklyn?

Similar issues may exist for the employer name that we need to group by with.

The tries to from am many of the above issues and returns the top 10 companies with the largest number of VISA applications.

2. Calculate the mean and standard deviation of wages proposed for workers located in New York City and Mountain View. Are the average wages in these two locations statistically different? What factors could explain the results?

Answer: To answer the this question we need to (1) identify all the rows that describe cases where the job location is in NYC, (2) identify the column that holds the proposed wages for workers (to my understanding this is 'lca_case_wage_rate_from'), (3) convert all the wages to the same unit (I chose to convert them to yearly salaries), (3) computer the mean and standard

deviations of the converted wages (4) repeat (1), (2), (3) for the rows that describe cases where the job location is in Mountain View and (5) statistically compare the two results.

Running the attached python code it returns:
“
Mean and standard deviation of NYC wages: 176921.81136 3770387.7896
Mean and standard deviation of Mountain View wages: 375891.065427 7534964.53268
“

To statistically compare the two results we find the 95% confidence intervals of the two population which is: (-4639.57230417, 402578.080439). This interval contains the zero, so the difference is not statistically significant.

Reasons for this difference may include the higher cost of living in Bay Area, the existence of more startups in Bay Area that drives job supply higher and as a result increases salaries offered, and the fact that more engineering jobs are offered in Bay Area (which in NYC is may be a mix of other job types), which in average pay a higher salary.

3. For NYC, what is the relationship between the total number of H1B visas requested by an employer and the average wages proposed? Visually represent this relationship if appropriate. Is the relationship statistically significant? What might explain this relationship?

Answer: To answer the above question one has to (1) identify all the rows that describe cases where the job location is in NYC, (2) group those rows on the column(s) that describe the name of the employer (i.e.’lca_case_employer_name’) and count the number of rows per group, (3) group those rows on the column(s) that describe the name of the employer again and and take the average of the (transformed to yearly) salary per group.

The scatterplot of the results from (2) and (3), which is shown in Figure 1, hints that there may be an exponential relationship between the number of H1B visas requested and the average proposed salary by employer. However, if we compute the Pearson correlation coefficient between the number of H1B visas requested and the logarithm of average proposed salary by employer (it is computed in the provided code and is equal to 0.17) reveal that this relationship is not very strong.

Nonetheless, some observations that can be deduced from the data are (1) most companies file a small number of H1B VISA requests and (2) the ones with the large number of H1B VISA requests propose a very small average salary in comparison with the rest of the cases.

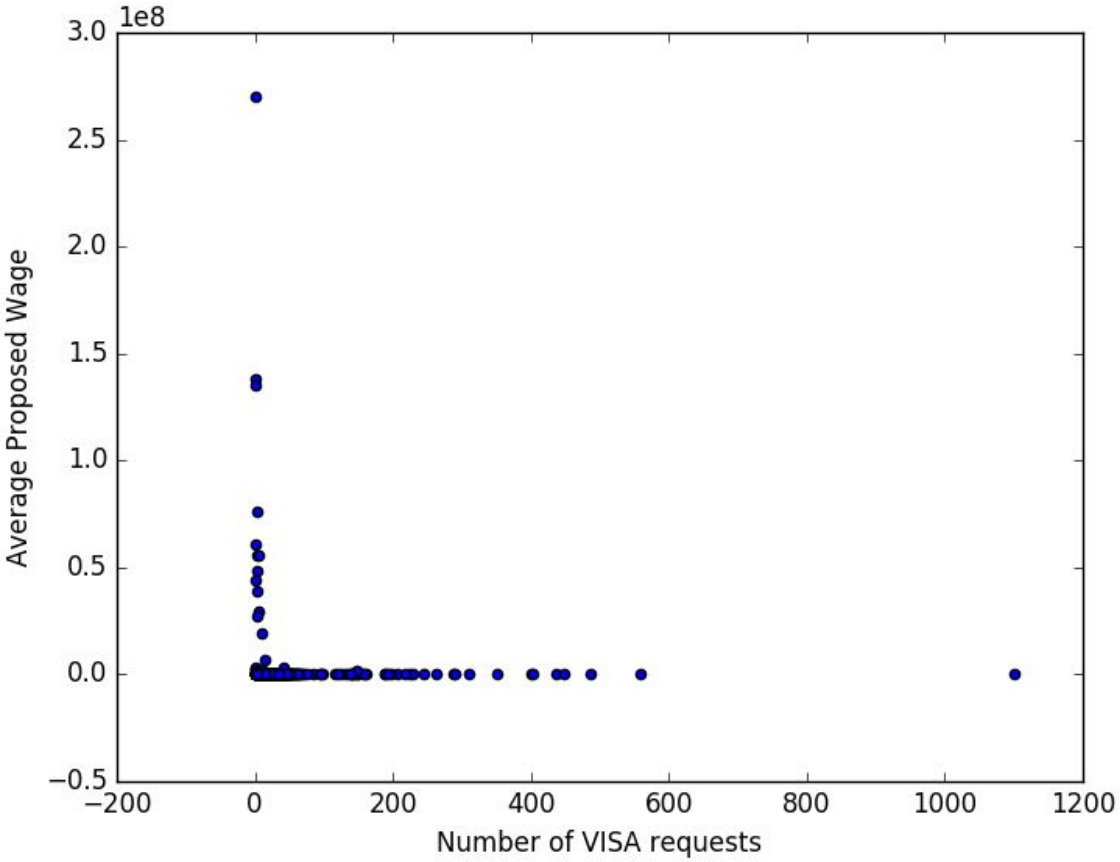
1e8
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
Average Proposed Wage
−200
0
200
400
600
800
1000
1200
Number of VISA requests

Figure 1.