The Slavery-Free Supply Chains Project

William Ieong, Megan Fantes, Shahrez Jan, Sameena Bajwa

INTRODUCTION

For our project, we partnered with Verite, a global, non-profit organization that conducts research and advocacy so that everybody in the world has the opportunity to work under safe, fair, and legal conditions. We will look into certain labor entities that exploit migrant workers during the recruitment process. These labor entities (known as labor brokers or labor agents) are organizations that offer overseas jobs to local workers and help the workers get the appropriate paperwork and transportation to the target work country. But for some of these labor brokers, all of that comes at a price. For the nefarious labor brokers, the exploitation can come in many forms, ranging from charging workers a high broker's fee, to promising migrants a higher wage than what will actually be given to them. We will focus on labor agents from the Philippines, where a large percentage of the labor force uses labor brokers to find work in different countries.

The Philippines Overseas Employment Administration has put together a registry of all of its labor agencies. Currently, the registry has 3667 records, where each record has eight entries, consisting of its contact info and license information. This information is publicly accessible online, however we will not be able to extract the data using APIs. Instead, we will have to use web scraping to collect all of the entries and export them into either a CSV file or Excel file. Because the records appear to follow very similar formats, we believe that scraping is the best way to access the data.

PROBLEM FORMULATION

When we partnered with Verite and began talking about their history and the data they are working with, we identified one core question that we needed to help them answer: How can we identify labor brokers with license violations so Verite can concentrate their efforts in creating slavery-free supply chains? To answer this question, we worked with Justin Shakespeare, our contact at Verite, to identify smaller, more specific questions to meet our goals.

We were first given a set of questions from Verite:

1) Which regions have the highest concentration of labor brokers with license-related violations?

Hypothesis: Labor brokers are concentrated in cities, leading to brokers with higher concentrations also being concentrated in cities.

Explanation: It would be useful for Verite to know where to look in the country for license violations so they can concentrate their efforts in the highest area of effect, instead of spreading themselves thin across the entire country.

2) What labor brokers are associated with specific ownership groups (if there are ownership groups) and which of these have the greatest volume of labor violations?

   Hypothesis: Verite predicts that various groups operate multiple labor agencies. If this is true, maybe one group will be a hotbed for license violations, and multiple of the labor brokers they operate will have license violations.

   Explanation: Similar to how finding regions with high concentrations of license violations will help Verite focus their efforts, finding groups (such as representatives or addresses) that are associated with multiple labor brokers and multiple violations will help them focus their efforts on singular sources of license violations rather than trying to assess every broker, everywhere.

3) What type of labor violations are most common and how do these vary by region?

   Hypothesis: Brokers with the same types of violations congregate together in the same place.

   Explanation: If Verite believes certain violations are worse than others, and those types of violations are common in a specific place, then Verite can focus on brokers in that area to target those violations.

And from there we created additional questions we wanted to explore:

4) How are labor brokers distributed throughout the country? (E.g. Are they focused solely in cities, or are they present in rural areas as well?)

   Hypothesis: Labor brokers are concentrated in cities (given that is where people, i.e. workers, are concentrated).

   Explanation: For a more complete picture of the state of labor brokers in the Philippines, it would be useful to know where they are located in the country.

5) Does having a missing contact field, like website/email/official representative have a correlation with a negative status?

   Hypothesis: There is a significant correlation.

   Explanation: We believe that transparency is one of the critical components for preventing worker abuse. One of the ways that a company can help its transparency is by having a clear line of communication by having a telephone, an email, a website, and/or an official representative. We think that when an agency lacks these things, intentional or not, it leads to more abuse of workers and a heightened likelihood of a negative status.

6) Do agencies that have negative statuses cluster together (geographically)?

   Hypothesis: Labor agents with negatives statuses do cluster together.

   Explanation: We want to explore if there are some geographical/environmental factors that encourage these labor abuses. We think that certain areas, perhaps one where people are of lower income or lack access to electricity for communication, are more likely to have agencies that have negative statuses.

Note: License violations, or negative statuses, are defined as: Delisted, Cancelled, Denied Renewal, Banned Forever, Expired, Revoked, Suspended, Suspended (Document Processing), Cash Bond Withdrawn. These negative statuses are listed in the database of Philippines-based labor brokers under the category "Status."

The trajectory of our project and questions changed as we worked with Verite to identify their needs and our capabilities within the scope of the project. The first change was to look at both databases of broker information and civilian news articles about labor brokers. The intent was to compare reported violations with official license violations. We were to answer such questions as:

- Which regions have the highest concentration of labor brokers with reported violations?
- Is there a correlation between reported violations and license related violations?
  - i.e. Given that someone has a license violation, how likely is it that they also have other reported violations? And vice versa.

However, our contact at Verite quickly noted that our skills and lack of background knowledge on the issue were best suited to identifying patterns in the databases of labor brokers rather than combing through news articles. As such, we will leave the exploration of news articles and reported violations as an intended follow-up.

Similarly, at first we thought we could analyze a database of Taiwan-based labor brokers, but as it is in Mandarin we could not analyze it within the scope of this project. Analyzing this database with the assistance of a Mandarin speaker is left as an intended follow up.

Once we had our major question and the smaller, supplementary questions defined to outline our project, and we had focused our efforts away from news articles and solely on the database, our research progressed smoothly. We found many interesting patterns in the database, which will give new insight to Verite into the labor brokers of the Philippines.

DATA ANALYSIS

Obtaining the data

For our project, we analyzed the data released by the Philippine Overseas Employment Administration on labor brokers.[1] The data, while not downloadable, was uniformly formatted on the webpage. Every agent consisted of eight lines that followed the pattern: labor agent name, address, phone number, email address, website, official representative, status (the target category, this states whether or not the agent has a license violation), and license validity period. Once we scraped the data, we entered it into a text file to be able to access it at any time, then we processed it in Python.

We chose this dataset because it is one of the few comprehensive data sources on labor agencies available. Each of the records was uniformly formatted and seemed regularly updated, meaning the starting dataset was clean and easy to work with. It contained the important and interesting information for identifying and classifying labor agents: license status, location, and contact information (or the lack thereof).

Cleaning the data

We parsed the data stored in the text file with a Python script. The script went line by line and split up the contents of each line by column and value. Lines were split by the colon marker and excess white space was stripped. Once the row was clean and ready, it was appended to a Pandas dataframe.[2] The final dataframe was converted to a csv file, which we used when trying to access the contents of this data set later.

[1] URL: http://www.poea.gov.ph/cgi-bin/aglist.asp?mode=all

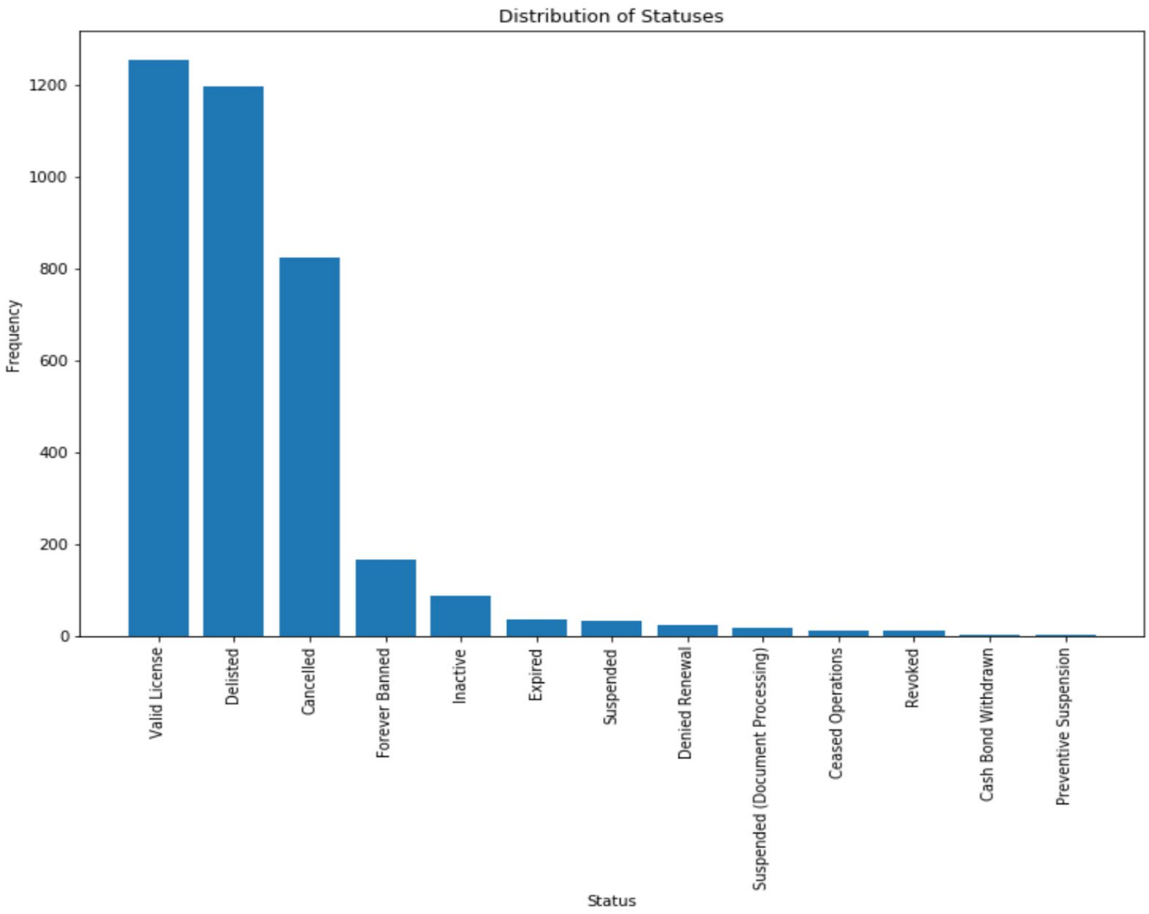
[2] See Appendix A for table showing dataframe structure

Initial exploratory analysis

Frequency of statuses[3]

The first question we wanted to answer was which license violations are the most common (a question posed by Verite). To do so, we ran the Counter() method on the "Status" field of our newly created Pandas dataframe of the Philippines labor agent dataset, and found the following frequencies:

Bar graph of status frequencies:



Observations:

There are 3667 labor agencies. The largest status group is a Valid License (1256 out of 3667), however the majority of the agencies have some form of invalid license. Delisted and Cancelled are the most common, with serious statuses such as Forever Banned and Revoked

[3] See Appendix B for table of statuses and frequencies

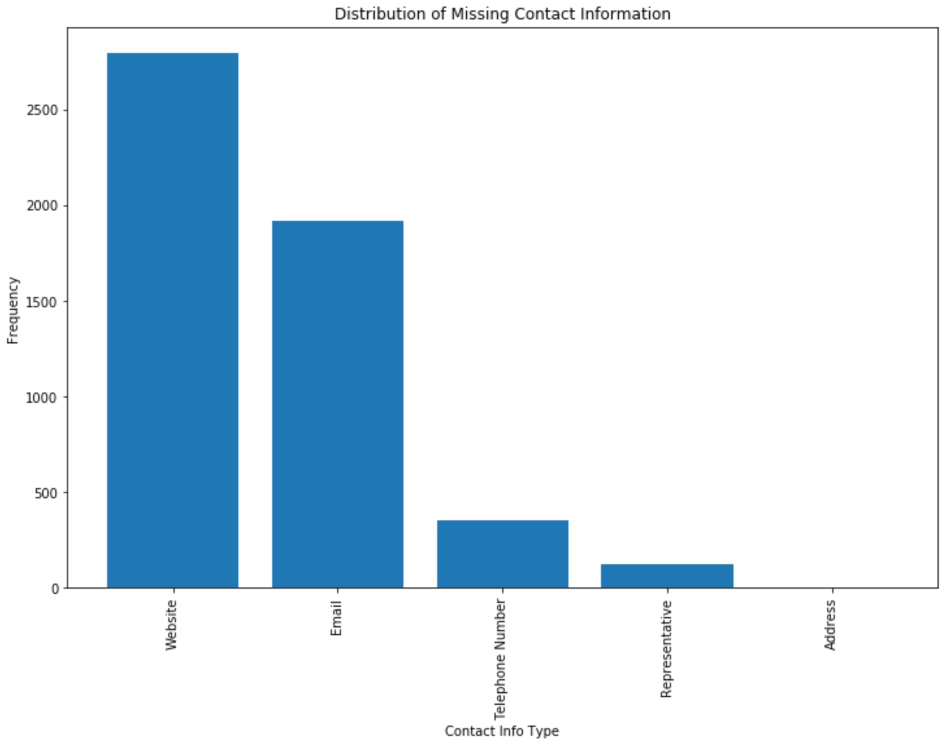
being much more rare. Our contact at Verite instructed us to categorize every status other than "Valid License" as a negative status, meaning the majority of the dataset (2411 out of 3667) is categorized as having a negative status in our later analysis.

Missing contact information[4]

As we cleaned our dataset and entered it into a Pandas dataframe, we found that many labor agents had missing contact information. Our dataset had many ways of indicating whether a communication field was missing: sometimes they used an empty string (‘’), sometimes they used ‘N/A’, etc. As we processed the database for analysis, we checked if a contact field had one of the "missing" values, and changed it to None upon entering it in the dataframe.

Finding these missing contact fields prompted us to wonder if having missing contact information could imply license violation, leading us to pose the question above (question #5) to search for a correlation between missing contact fields and negative license status. Below is a table summarizing the number of missing contact fields in the dataset:

Bar graph of missing contact information:



[4] See Appendix C for table of frequency of missing contact information

Observation:

We know that there are 3667 Labor Agencies registered in the Philippines. The majority of them have telephone numbers and an official representative. The most common missing contact information is a website, which makes sense given the economic environment of the Philippines; in general it is a more person-based economy, rather than an online economy.

Identifying ownership groups with violations

An ownership group is defined as one entity that owns a group of labor brokers, with violations being defined earlier in the report. For us, it was difficult to identify ownership groups solely based on the information we had. We decided to consider two labor agencies as part of the same ownership group if they shared any type of contact information, like address or official representative. However, this had its limitations, as it is possible for this to alternatively be explained by the natural movement of people from job to job, or the natural movement of agencies from office to office. For example, a representative could have worked for one agency, left their job, and started working for another agency, therefore the representative does not represent an ownership group but rather a single individual working for completely separate agencies.

We found ownership groups by again using the Counter() function on every contact field column in the dataframe to determine how many times each piece of contact information appeared.[5] If an entry of contact information (such as a phone number or an address) appeared more than once, that means more than one labor agency was using that contact information, meaning that contact information could indicate an ownership group. We then entered all of the repeated contact information (i.e. any phone number, address, representative, email, or website that appeared more than once in the dataframe) in a dictionary, then looped through the dataframe to determine how many violations were associated with each entry of repeated contact information. Finally, we sorted the dictionary of repeated contact information and associated violations by number of violations, to see which ownership groups (i.e. which entries of repeated contact information) had the largest number of license violations. This addresses one of the questions posed by Verite (question 2).[6]

Once we had the list of contact information that is associated with ownership groups, we cross referenced the list with the dataframe of labor brokers and found that 276 labor brokers use contact information that is associated with at least one other broker. Of these 276 labor brokers, 176 of them have a license violation. We then ran a hypothesis test on the dataframe to see if

[5] See Appendix D for table of repeated contact information and the number of brokers it appears with

[6] See Appendix E for table of ownership group and violations

listing contact information that is used by another broker is associated with having a license violation, and found that it is ($p < 0.0001$). We then calculated the odds ratio of having a negative status with repeated contact information vs. unique contact information, and found that the odds of having a license violation increase by 76% when a broker lists contact information that another broker uses.

Observations:

Many of the ownership groups are very small, consisting of one or two agencies – yet it is worth noting that there are 3 representatives that are associate with 5 or more labor agencies. However, few ownership groups are associated with multiple violations: we found that 6 official representatives were associated with 3 or more license violations. It is possible that these 6 representatives represent 6 small ownership groups, all of which have 3 or more license violations. Beyond these 6 representatives, there were many phone numbers and addresses that were associated with 2 license violations. (See Appendix E for a table with further detail on ownership groups and their associated number of violations). But if we were looking for a large overarching group controlling a large portion of labor brokers, all of which have license violations, we did not find it.

METHODOLOGIES

Logistic Regression

One question we wanted to answer was whether or not a broker having missing contact information is correlated with that broker having a license violation, or a "negative status" on their work license. We classified each broker as either having a negative status or not, based on whether they had a status of "Valid License" (Negative Status = False) or not (Negative Status = True). As this is a binary outcome, logistic regression is the best tool for analysis.

To start the analysis with logistic regression, we split our data into training and test datasets – 60% of the original dataframe of 3667 labor brokers was allocated for creating the model ("training" data) and 40% of the original dataframe was allocated for testing the model's accuracy once it was created ("test" data).

Using the training data, we started creating the model by entering all 5 fields from the dataset into the model as predictors for Negative Status:

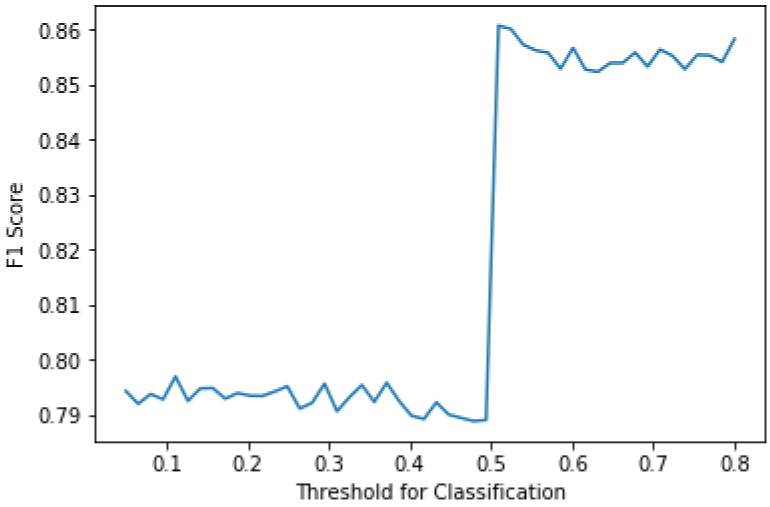
1) Address Missing (boolean)
2) Official Representative Missing (boolean)

3) Phone Number Missing (boolean)
4) Email Address Missing (boolean)
5) Website Missing (boolean)

With all predictors, the model had an R^2 value of 0.3842, but both Address Missing and Website Missing were insignificant predictors for Negative Status ($p = 1$ and $p = 0.965$, respectively). So we removed Address Missing and Website Missing and created a new model. This 3-factor model had an R^2 value of 0.3809, and all predictors were significant ($p < 0.002$ for all three).

Once we had a significant model, we needed to cross-validate it on our test dataset to ensure the model generalized well and was not overfit to the training data. To compare the model's predictions for the Negative Status of the test data to the known target values, we used 2 quantifiers: "precision" and "recall." "Precision" is the fraction of the True values that the model predicted that are correct. "Recall" is the fraction of known True values that the model correctly predicted as True. The 3-factor model above had 94.2% precision on the test data and 78.1% recall. We then wanted to check the precision and recall of the 3-factor model over many iterations of the training and test data, to bootstrap an average precision and recall of the model. Using a for-loop, we randomly split the original data 20 different ways, found the precision and recall of the 3-factor model on each random split, and calculated an average for each metric. The model had an average precision of 95.5% and an average recall of 77.9%.

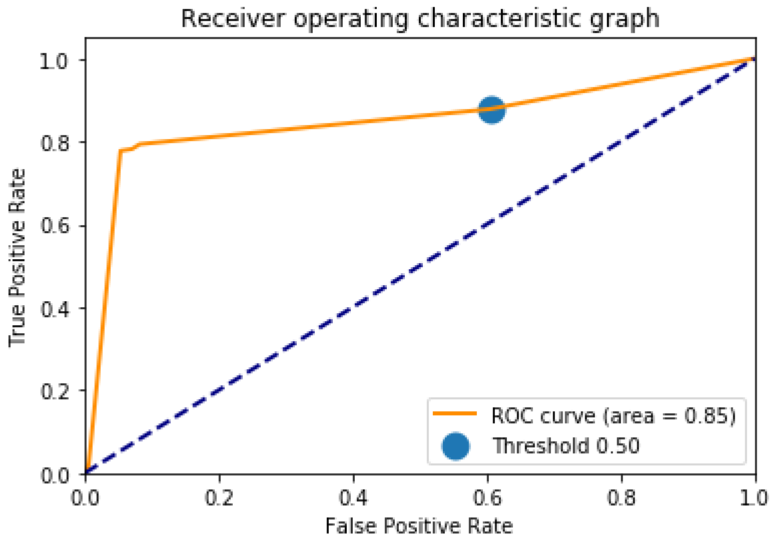
In logistic regression, the model does not output a binary outcome of True or False, it outputs the odds of the outcome being true with the given predictors. We then choose a threshold of odds above which the outcome is classified as True, and below which the outcome is classified as False. Thus far in our analysis, we have used a threshold of 0.5, assuming that True and False were equally likely. We wanted to confirm that this assumption is correct, so we calculated the F1 score of a series of thresholds. An F1 score is a single value that describes how accurate a model is at classifying outcomes with a given threshold, and we calculate the F1 score by finding the harmonic mean of the the precision and recall ($\frac{precison * recall}{precision + recall}$). We visualized the series of F1 scores at different thresholds with a graph:



Graph of F1 Scores for determining threshold of classification

From the graph, we saw that 0.5 was the appropriate threshold to use to classify outcomes as True or False, as it has the largest F1 score of any threshold between 0.05 and 0.8. Thus the precision and recall that we calculated for the model previously were correct.

Finally, once we confirmed a threshold of 0.5, we wanted to evaluate the Receiver Operating Characteristic curve (ROC curve). The ROC curve is a plot of the true positive rate (the fraction of predicted True values that were correctly predicted as True) against the false positive rate (the fraction of predicted True values that were incorrectly predicted as True, i.e. should have been predicted as False), and acts as a another test of model accuracy. With the 3-factor model and a threshold of 0.5, we produced the following ROC curve:



Graph of ROC curve to visualize model accuracy

The farther away the orange line is from the blue line, the more accurate the model is. As we have a large space between the orange and blue lines in our graph, we have further confirmation that the 3-factor model is accurate.

Map Visualizations

We chose to use various map visualizations to visualize if labor agencies with negative statuses cluster together. Heat maps provide a way for us to answer a few of our questions: How are labor brokers distributed throughout the country? Does there exist a correlation between geographic distribution of brokers and areas with higher concentrations of invalid licenses? What type of labor violations are most common and how do these vary by region?

In order to answer these questions, we needed to obtain the longitude and latitude coordinates of each office. The data set obtained from the Philippines Overseas Employment Administration included a column for a labor agents' office address. We wrote a Python script to parse through this column and call the Google Maps API with the addresses listed, in order to obtain the location's latitude and longitude coordinates.

There were some addresses that were either not in a structured enough form or too much shorthand was used for the API to recognize it as a valid location. Since around half of the places ran into this issue, it was not an efficient solution to have one of us manually look up each address and change the wording until the address was in a recognizable form. Our solution, therefore, was to create a helper function which gathered several subsets of the words included in the address to use with the API until a valid result was obtained. For example, Google Maps returned no results for the address “U1,4&5 2F PWU Intl Hse, 1724 Leon Guinto St Cor Malate, Manila.” The helper function then tested the API with the following subsets of the address, starting by eliminating the first word:

- 2F PWU Intl Hse, 1724 Leon Guinto St Cor Malate, Manila
- PWU Intl Hse, 1724 Leon Guinto St Cor Malate, Manila
- Intl Hse, 1724 Leon Guinto St Cor Malate, Manila
- Hse, 1724 Leon Guinto St Cor Malate, Manila
- *1724 Leon Guinto St Cor Malate, Manila*

The coordinates were then used to create a general heat map to observe how labor agencies were distributed across the country.

To find out if there existed a correlation between the geographic distribution of brokers and areas with higher concentrations of invalid licenses, we parsed our data set to only gather the coordinates for agencies with license violations to observe which regions have the highest concentration of violations and if there are any obvious clusters that formed. We had already determined the most common types of license violations during our exploratory analysis, now we wanted to analyze how these types varied by region. Thus, we grabbed the specific type of violation of each labor agency, along with their geographic coordinates.

EXPERIMENTAL RESULTS

Logistic Regression

Using logistic regression as described above, we created a model to predict the odds of a given labor broker having a Negative Status:

$$log(odds) = (3.78)(email\ missing) + (1.15)(representative\ missing) - (0.34)(website\ missing)$$

Where $log(odds)$ is the natural log of the odds, and the odds that a given labor broker has a negative status can be found by raising e to the power of the $log(odds)$.

In logistic regression, we also talk about Odds Ratios, or the comparative increase of the odds when an input factor is in increased. From the above equation, we find that having a missing email has an odds ratio of 44, meaning that a labor broker is 44 times more likely to have a negative status if they do not have an email listed in the database than if they did have an email, when all other variables are held constant. Having a missing official representative has an odds ratio of 3.2, meaning that a labor broker is 3.2 times more likely to have a negative status if they do not have a representative listed in the database than if they did have a representative, when all other variables are held constant. Having a missing website has a very interesting effect on the odds of a labor broker having a negative status: having a missing website has an odds ratio of 0.7, meaning a labor broker is 0.7 times as likely (or 30% less likely) to have a negative status than if they had a website, all other variables held constant. In other words, having a website makes a labor broker more likely to have a negative status, and if they do not have a website, they are more likely to have a valid license.

With the logistic regression model and its proof of accuracy (see the Methodologies section), we have proven that there is a correlation between missing contact information and negative status. In the case of a missing email address or a missing representative, the model fits with our hypothesis that missing contact information relates to a negative status, possibly because it implies a lack of transparency and a desire for the labor broker to hide something by being more anonymous. However, in the case of a missing website, the model goes against our hypothesis by showing that having a missing website is correlated with having a positive status.

Identifying Ownership Groups with Violations

By identifying which labor agencies had violations and looking for common communication information, we were able to identify several owner groups. Of note are agencies whom had the following official representatives:

Ms. Angelina T. Rivera, Capt. Rosendo C. Herrera, Mr Reginaldo A. Oben, Mr Akira S. Kato, and Mr Cesar P. Carandang.

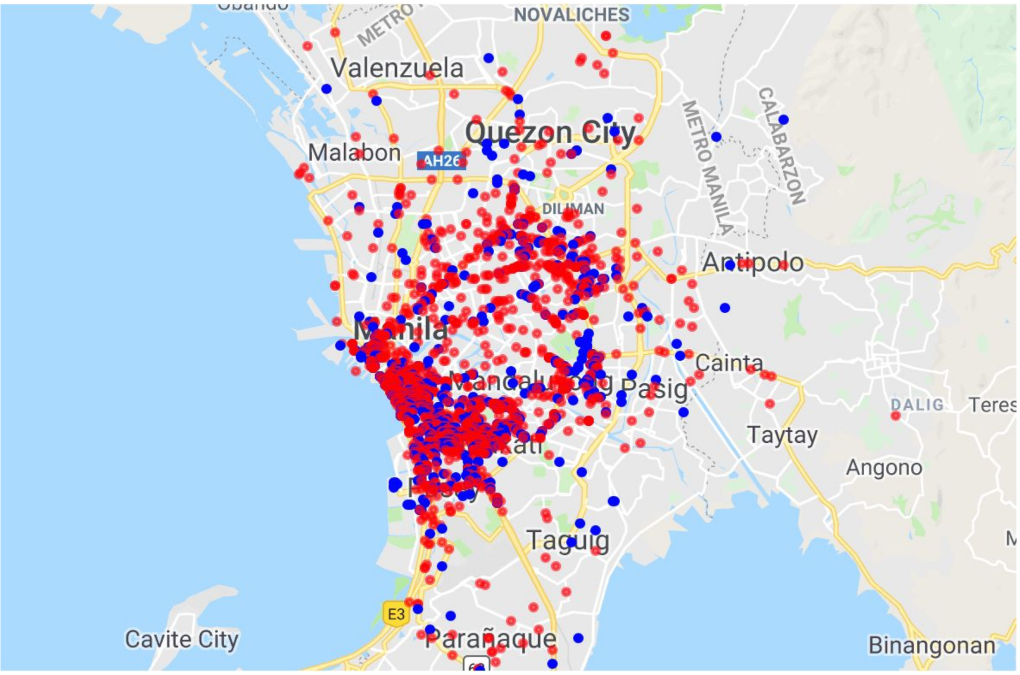
All of the official representatives above were associated with three or more labor agencies that had a violation.

As we mentioned in the data analysis section, our method of identifying ownership groups has its limitations. It is possible that the entity that owns these companies is not public facing, meaning that they do not publicly list any affiliation with this labor agencies. This would make sense, as many of these labor agencies have violations, meaning the people who own the labor agencies would want a degree of separation to avoid accountability. It is also possible that some of these companies simply shut down, and representatives, whom are already in the field, start to work for another labor agency that also eventually runs afoul.

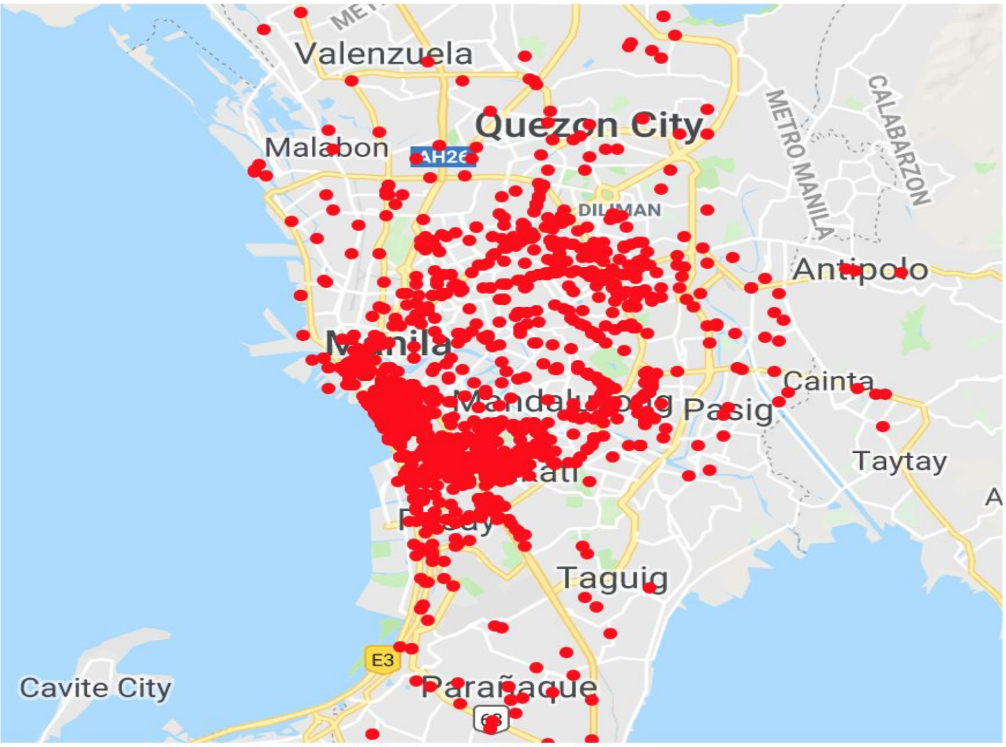
After we had identified contact information associated with multiple brokers and multiple violations, we were able to make a list of ownership groups, the labor brokers in each group, and the number of violations associated with each group. This file identifies 128 ownership groups with 2-7 labor brokers in each, each with 0-5 violations. This file can be used by Verité to investigate the ownership groups with a large number of violations and perhaps even give them the proof they need to shut down the nefarious ownership groups and take one step closer to a slavery-free supply chain.

Map Visualizations

To visually understand the distribution of labor brokers throughout the Philippines, as well as the distribution of brokers with license violations, we plotted the latitude and longitude values we found (as described in the Methodologies section) on a map using the gmaps Python package:



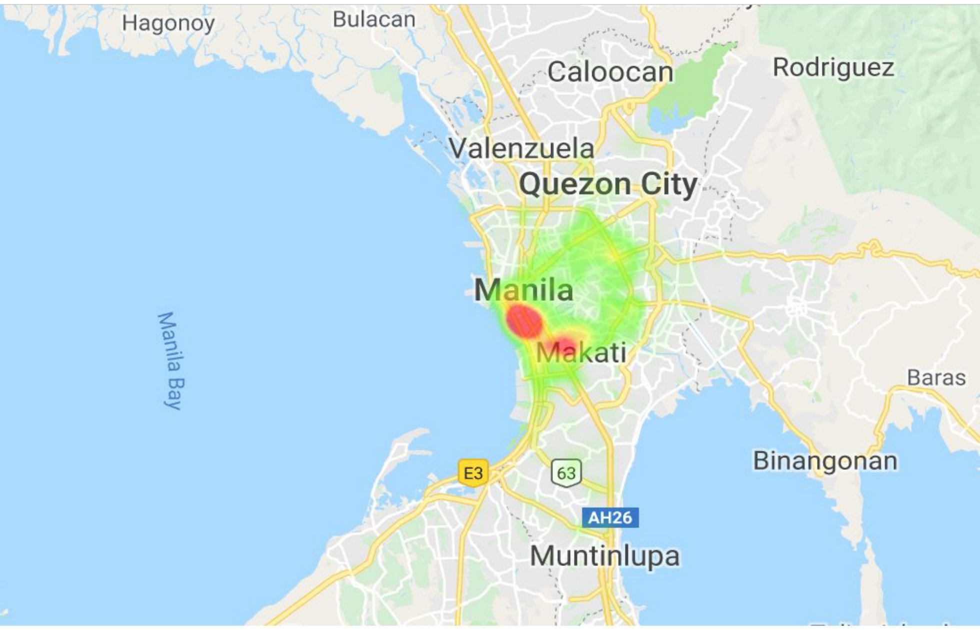
A map visualization showing labor brokers with valid licenses (blue) and negative statuses (red)



A map visualizing the distribution of only labor brokers with license violations

On the map visualizations we see that labor brokers are almost entirely located around Manila, the largest city in the Philippines, which confirms our hypothesis that labor brokers tend to congregate in cities with a high population density.

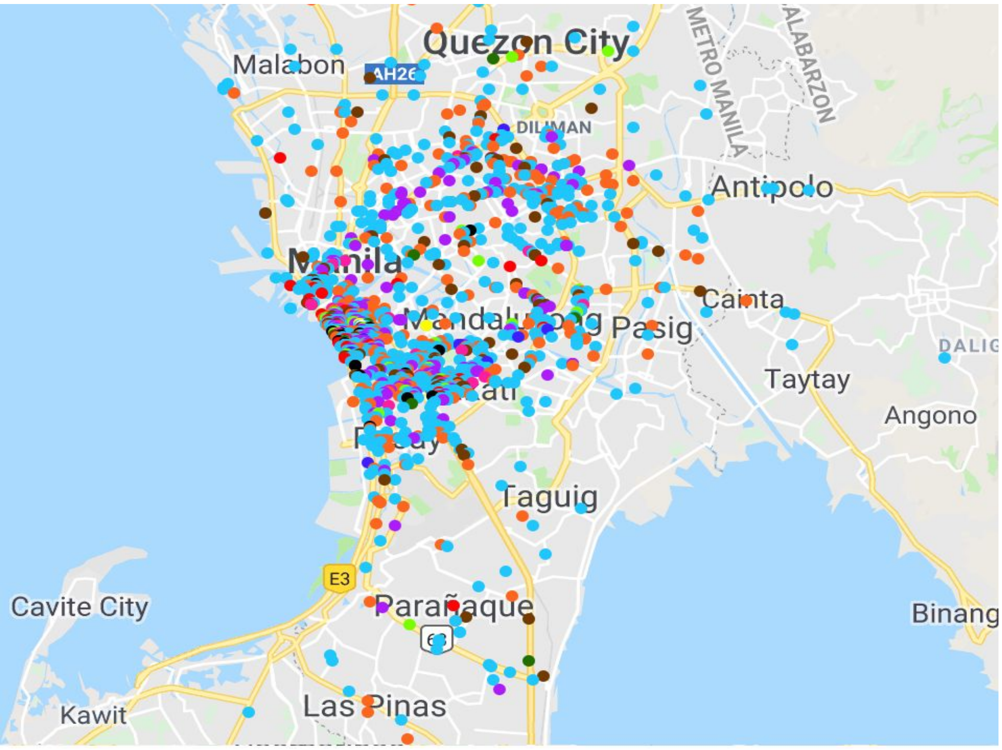
Once we visualized the basic distribution of labor brokers, we wanted to see the relative distribution of labor brokers using a heatmap, to see where brokers were mostly densely concentrated:



A map visualizing the distribution of only labor brokers with license violations, with light green being low density and orange/red being high density

As we can see from the above visualization, most labor brokers tend to cluster around the city of Makati and the district Malate, this is possibly due to the proximity to the port of Manila and the diplomatic area.

Once we had the heat map, we decided to plot the latitude and longitude of the brokers by type of license violation:



A map visualizing the distribution of labor brokers with different license violations, see Appendix F for legend

The resulting visualization showed us how labor brokers with the different types of license violations tend to cluster together. The most common license violations are Delisted, Cancelled and Forever Banned, and they are predominantly concentrated around the port of Manila and diplomatic area in the district of Malate.[7]

CONCLUSIONS

The goal of this project was to find answers to our original six questions posed in the beginning and to use these findings to answer our main question: How can we identify labor brokers with license violations so Verite can concentrate their efforts in creating slavery-free supply chains? Our first task we set out to accomplish was to determine what type of labor violations were the most common among all of the labor agencies? During our initial exploratory analysis, we found that more than half of the agencies listed had some form of an invalid license. This alone showed us the magnitude of the problem at hand.

[7] See Appendix F for an index of license violation, frequency, and corresponding color

We then set out to answer which areas have the highest concentration of labor brokers with license-related violations. We hypothesized that most labor brokers congregate around cities with large populations, and with a higher concentration of labor brokers we can find more labor brokers with license-related violations. Our results concluded that most labor brokers are located at the district of Malate and the city of Makati. These locations are extremely close to the port of Manila. This information would be extremely useful for Verite as they can concentrate their effort in these areas.

The three most common forms of license violations where Delisted, Cancelled and Forever Banned. According to our map visualization, most of these license-violating brokers seem to most commonly appear around the port of Manila and diplomatic area in the district of Malate, this may be due to a formation of a syndicate amongst the violating brokers. This result will allow Verite to go after the collusion between violating brokers in order to have a big impact in reducing labour violations.

Because of how common license violations seemed to be, we wanted to help Verite in focusing their efforts on the main sources of these violations. We sought out to find if there were any ownership groups that operated multiple labor agencies. We hypothesized that there may be individual groups responsible for running multiple agencies, many of which with license violations. From our results, we found that there were a handful of agencies with infractions that were associated with the same official representative or address. When deciding where to concentrate their efforts in cracking down on labor offenses, Verite can look straight to the ownership groups.

The outcome of the logistic regression we performed gave us significant insight into patterns in behavior of common offenders. As stated in our results, agencies that don't have email addresses and official representatives listed are 44 times and 3.2 times, respectively, more likely to have a license violation. These numbers showed us that labor brokers that are participating in worker abuse are actively trying to obscure their lines of communication. The less contact information they put out, the less they feel that they are going to be held accountable for their actions.

Our map visualizations show that license violating brokers tend to cluster together. They tend to cluster around the port of Manila, this may be because the port is the premier gateway into and out of the country. This clustering can also be a sign that license violating brokers have formed a syndicate in order to abuse and extort the migrant workers. This information can be used by Verite to work with the Philippine government to break up collusion between license violating brokers and ensure migrant worker rights.

Over the course of this project, we have documented different types of license violations by labor brokers and have created models to predict whether a given broker has a negative status. We have answered all the questions posed by Verite and our own exploratory questions, which resulted in a numerical model to predict if a labor agent will have a license violation based on their contact information, as well as a map visualization showing the clustering of labor violations by classification. The clustering on the maps we created shows possible signs of a of labor broker syndicates exploiting migrant workers. We hope that Verite can use our analysis to work towards ensuring the rights of migrant workers.

WHAT VERITÉ CAN TAKE AWAY FROM OUR PROJECT

From the map visualizations we have created, Verité can see that almost all labor brokers in the Philippines are located in Manila, with most of them around the port of Manila. There are also labor brokers scattered around the edges of the city of Manila, and a few in other rural areas of the country, but the largest portion of labor brokers by far is around the port of Manila. Verité has more specialized knowledge about the Philippine economy, meaning they likely have more informed hypotheses about why there are clusters at the port, but we hypothesize that the clustering is related to the cruise ship industry in the Philippines and the fact that labor brokers may often recruit employees to these ships. Alternatively, the port is a primary way to get into and out of the country, so being near the port may make it easier to facilitate transporting workers into and out of the country.

From the analysis of types of license violations, Verité can see that Valid License is the largest group of license statuses, which is comforting to know, but it makes up less than half of the total number of labor brokers, and the majority of labor brokers have some type of license violation. The most common license violations are Delisted, Cancelled, and Forever Banned. These common license violations do not tend to appear in isolated clusters, but because most labor agents are clusters around the port of Manila, all types of license violations tend to cluster there as well.

From the logistic regression analysis, Verité can see how significantly correlated missing contact information is with a license violation (or “negative status”). We found that not having an email or representative listed as part of their contact information is a strong indicator that a labor broker has a license violation. If a labor broker does not have an email listed, they are over 40 times more likely to have a license violation, and if they do not have a representative listed, they are over 3 times more likely to have a license violation. On the other hand, we found that *having* a website listed is significantly correlated with having a violation – a labor broker is 30% more likely to have a license violation is they *do* have a website listed. Verité could apply their knowledge of the Philippine work environment to more solidly understand why this might be the case.

From the analysis of groups with shared contact information, Verité can see that there are a considerable amount of labor brokers that share contact information (address, phone number, representative, etc.) with another broker – 276 labor brokers share 198 entries of contact information.[8] However, there are 3667 labor brokers in the database in total, meaning the vast majority of labor brokers operate on their own. Additionally, not all labor brokers with repeated contact information have a license violation – only 176 brokers among the 276 with repeated information have a license violation. Having repeated contact information, i.e. possibly being part of an ownership group, is significantly correlated with having a license violation: the odds of having a license violation increase by 76% when a labor broker lists contact information that is used by another broker.[9]

In our analysis, we first looked for any contact information that is shared by multiple brokers. That list can be seen in Appendix D. Once we found this list, we then looked at the number of violations with which each of these entries of contact information were associated. That list can be seen in Appendix E. The list of overlapping contact information with the number of times each entry appears in the database (in Appendix D) is interesting for seeing potential "ownership groups," or single organizations that operate multiple labor brokers. The list of overlapping contact information with multiple associated violations (in Appendix E) is a preliminary tool for detecting potentially nefarious ownership groups. The fact that the list of overlapping contact information with associated violations is shorter than the raw list of overlapping contact information shows that there may exist some ownership groups that are not bad.

INTENDED FOLLOW-UP

Given the results we have obtained from our exploratory data analysis, logistic regression models, and various heat maps, we believe that we have obtained meaningful answers to not only our major problem, but to also all of our supplementary questions. Verite will be able to use our findings as evidence when identifying labor brokers in the Philippines that may have license violations. Their classifications will be based on patterns that we have established through this project: region, ownership groups, and availability of certain contact information.

[8] See list in Appendix D

[9] Justine Shakespeare was sent CSV files containing the contact information that was repeated across multiple brokers and the number of violations associated with each repeated contact field, as well as a file containing a list of ownership groups and the brokers that belong to them.

Through our discussions with Verite, we have recognized areas for further research. The first, as briefly mentioned earlier in this report, is the presence of reported violations and how those might correlate with license-related violations. This raises the question that if a labor agency has been reported for questionable or illegal practices, what is the status of their license? Answering this question involves searching the web and scraping civilian news articles about labor brokers, creating a script to read the article and detect any offense reported, and then connecting the labor broker mentioned to one listed in the data set we obtained.

A second area for additional exploration is the database of Taiwan-based labor brokers. The majority of migrant workers from the Philippines are sent to Taiwan to work. Therefore, there exists two labor brokers during this process, one from the sending side (the Philippines) and one from the receiving side (Taiwan). The problem we found with the Taiwan database is that it was in Mandarin, which posed as a major obstacle for us when it came to scraping and understanding their data. Doing a full analysis of Taiwan-based labor brokers would require the assistance of a Mandarin speaker to help us understand our findings.

Furthermore, a couple of new questions arose from our initial analysis. We classified several license statuses as "negative" in the Problem Formulation section of this report. Given that our questions and conclusions are all based on a labor agencies' license status, we would like to get a better sense of what statuses are truly negative and which may not be as bad. For example, the status of "Forever Banned" clearly indicates the labor broker has engaged in problematic behavior, but a status of "Inactive" could be for reasons that are not necessarily as bad.

Finally, one of the most interesting discoveries we came across was that when a labor broker has a website, they are more likely to have a negative license status. It would be interesting to look into internet use in the Philippines to understand the reason behind this finding. Could a website imply that a labor broker is trying to hide something by looking more advanced? We could explore this question more by scraping and analyzing sources that profile internet users in the Philippines.

APPENDIX A
Excerpt of dataframe containing cleaned labor agency information and indicator variables

Index	Address	Address Missing	Email Address	Email Missing	License Validity	Name	...
0	R 20 G/F & 33B 2/F...	False	brilliant_minds1015@yahoo.com	False	9/11/2017 to 9/11/2019	1015 BRILLIANT MINDS INC...	...
1	R602 DOÃƒâ€˜ A F...	False	CARGOFLEET@MARC-SHIPS.COM	False	11/4/2011 to 11/3/2015	1022 MARITIME SERVICES...	...
2	2F & 3F, 1523-1527...	False	tenthstorypai@yahoo.com	False	4/2/2016 to 4/1/2020	10TH STORY PLACEMENT AGENCY ...	...
3	2F&3FCORA ROSE ...	False	None	True	3/28/2001 to 3/28/2003	168 PLACEMENT CORPORATION Private...	...

Negative Status	Official Representative	Representative Missing	Status	Tel No/s	Telephone Missing	Website	Website Missing
False	ANITA A COBER	False	Valid License	(02) 2522338 / 2515100	False	None	True
True	None	True	Ceased Operations	5362997 5360831	False	None	True
False	MR ALFONSO UY NG	False	Valid License	3531581	False	www.10thstory.com	False
True	MA LISA G LOPEZ	False	Delisted	4336267	False	None	True

APPENDIX B
Table of License Statuses and Frequencies

Status	**Frequency**
Cancelled	823
Cash Bond Withdrawn	3
Ceased Operations	11
Delisted	1197
Denied Renewal	25
Expired	35
Forever Banned	167
Inactive	86
Preventive Suspension	1
Revoked	11
Suspended	32
Suspended (Document Processing)	19
Valid License	1256

APPENDIX C
Table of Contact Information Fields and the Number of Brokers Missing Each Field

Contact Field	**Number of Brokers Missing that Contact Field**
Address	4
Phone Number	352
Official Representative	121
Email	1916
Website	2792

APPENDIX D
Contact Information that appears for multiple labor brokers, and the number of labor brokers it appear with

	Contact Info	Frequency
0	GREGORIO F. ORTEGA	6
1	GREGORIO F ORTEGA	5
2	MS ANGELINA T RIVERA	5
3	CAPT ROSENDO C HERRERA	5
4	5211566 / 5211567	4
5	MR REGINALDO A OBEN	4
6	MR CESAR P CARANDANG	3
7	NARCISSUS L. DURAN	3
8	MR AKIRA S KATO	3
9	ARLEEN V ASUNCION	3
10	RCM BLDG 1418 SAN MARCELINO ST ERMITA, MANILA	3
11	CHRISTOPHER DINO DUMATOL	3
12	JOSEPHINE J FRANCISCO	3
13	www.tsmphil.com.ph	3
14	MR BONIFACIO F GOMEZ	3
15	MS RIZALINA LAMZON	2
16	MR ERICSON M MARQUEZ	2
17	MR ERWIN L CHIONGBIAN	2
18	MR EMMANUEL L REGIO	2
19	MS THELMA A MANUEL	2
20	NO ACKNOWLEDGED REP.	2
21	MR GREGORIO F ORTEGA	2
22	MR OSCAR M LOPEZ	2
23	CAPT ALEJO M VINLUAN	2

24	MR ROGELIO S MANALO	2
25	MR LEVI S DE MESA	2
26	MS TARCIANA SANTOS NG	2
27	CAPT EDGARDO M GUALBERTO	2
28	ALEXANDER M NUNEZ JR	2
29	CRISTY LYN P MASONSONG	2
...	...	...
168	5277991 TO 97	2
169	RM 202 2F KIMVI BLDG 1191 MA OROSA ST ERMITA, ...	2
170	MR ROLANDO P MALIG	2
171	CAPT REYNOLD L TORRES	2
172	MA PAULA L SAN AGUSTIN	2
173	NIERESA A CABAHUG	2
174	5214533	2
175	503341	2
176	4905636	2
177	5210836 / 5211075	2
178	8981111	2
179	516-5640 / 0917 823 7153	2
180	(032) 5208855/5207278/2687274/5207663/09188160000	2
181	5270772 / 5270779	2
182	593095 / 504940	2
183	505805 / 595136	2
184	7437391 / 7311175	2
185	8671476	2
186	5218162	2
187	5222466 / 581643	2

188	5279980 / 3608800	2
189	7315984	2
190	500671 / 500675	2
191	5518870/5518885/5518889	2
192	587766 / 594377	2
193	7864226 / 7864155 / 09177265564/09258337933	2
194	5238646-50	2
195	8164875 / 8171312	2
196	813-0696/ 759-2174/ 885-7740	2
197	egp@wallem.com.ph	2

APPENDIX E
Ownership groups and their license violations

	Ownership Group	**Number of Violations**
0	MS ANGELINA T RIVERA	5
1	CAPT ROSENDO C HERRERA	4
2	MR REGINALDO A OBEN	3
3	MR AKIRA S KATO	3
4	MR CESAR P CARANDANG	3
5	MR BONIFACIO F GOMEZ	3
6	RM 202 2F KIMVI BLDG 1191 MA OROSA ST ERMITA, ...	2
7	ims_3r@yahoo.com	2
8	MR ROLANDO P MALIG	2
9	886870 / 8163199	2
10	NIERESA A CABAHUG	2
11	587955 / 586087	2
12	MR ANTONIO P MADRIGAL	2
13	MR VIRGILIO L LEYEZA	2

14	MR TRANQUILINO VENTURA JR	2
15	S501 5F 1377 A MABINI COR STA MONICA ST ERMITA...	2
16	aumsimla@pldtdsl.net	2
17	5223854/525-1205	2
18	www.aumsi.com	2
19	4003915	2
20	4905636	2
21	FERROS BLDG 176 SALCEDO ST LEGASPI VILL MAKATI	2
22	MR CARLOS C SALINAS	2
23	8981111	2
24	NARCISSUS L. DURAN	2
25	LYDIA S. GURANGO	2
26	MR VICTORINO A BASCO	2
27	5270772 / 5270779	2
28	SABAS ALMEDA BLDG 505 A.FLORES ERMITA MANILA	2
29	MR VIRGILIO G DEL ROSARIO	2
...	...	...
135	MR LEVI S DE MESA	1
136	MR EMMANUEL L REGIO	1
137	ANGEL S RACOMA	1
138	915 PRES.QUIRINO AVE COR LEON GUINTO ST MALATE...	1
139	leonismanning@leonisnav.com.ph	1
140	5238646-50	1
141	8164875 / 8171312	1
142	CAPT EDGARDO M GUALBERTO	1
143	CRISTY LYN P MASONSONG	1
144	RECRAA BLDG VITALEZ COMPOUND SUCAT PARA#AQUE M.M.	1
145	MR ERWIN L CHIONGBIAN	1
146	brothersinlaw@yahoo.com	1
147	TRINA T DIZON	1

148	RUEL A. BENISANO	1
149	GREGORIO F. ORTEGA	1
150	CECILIO A C VILLANUEVA	1
151	UPL BLDG STA CLARA ST INTRAMUROS, MANILA	1
152	mailadmin@uplines.net	1
153	5277491	1
154	www.uplines.net	1
155	JOSE MARI MORAZA	1
156	ELMER A PULUMBARIT	1
157	3RD FLOOR, CASA MARITIMA, GENERAL LUNA STREET ...	1
158	vintex@vintex.com.ph	1
159	5218822 TO 24/0917-8173701	1
160	www.vintex.com.ph	1
161	WALLEM PHILS BLDG LEGASPI COR BEATERIO INTRAMU...	1
162	egp@wallem.com.ph	1
163	5277991 TO 97	1
164	www.walcrew.com.ph	1

APPENDIX F
License violation, frequency, and correspond color on map visualization

Ceased Operations: 11 - Blue
Delisted: 1179 - Sky
Cancelled 807 - Orange
Denied Renewal: 24 - Pink
Forever Banned: 167 - Purple
Inactive: 85 - Brown
Revoked: 11 - Dark Green
Suspended: 31 - Green
Expired: 34 - Red
Cash Bond Withdrawn: 3 - Yellow
Suspended Document Processing: 19 - Black
Preventive Suspension: 1 - White

APPENDIX G
Github Repository

You can view all the code and their outputs here:
https://github.com/Shahrez19/CS506-Data-Mongers