**How similar are Indonesian Embassies based on their location**

boy.setiawan

April 4th, 2021

1. **Introduction**
   1. **Problem Description**

**An Embassy** is a representative of a country in other countries, their existence helps to indicate a relation among countries and serve as a way to communicate or strengthen the ties. Their location follows a strict and complex requirements both from the country it comes from and the country it resides, usually in a special diplomatic compound or district. Despite all of the careful planning and requirements, the decision to establish an embassy could have come from other necessities such as certain neighborhood/district/area, near to and close from certain amneties, places that could support the embassy mission etc.

The knowledge of how a certain embassies are similar or different could **help give a bigger view to Indonesian Foreign Affairs Officials** to understand the general environment their embassies are located. Should a certain embassies need to be treated differently, do embassies with certain criteria experience the same or different stress level of working for their staffs, do certain embassies experience certain disturbance etc.

The task to find and group embassies into their own similar group could be a daunting task. From forming the basic data, to gathering and analyzing the data must be done without bias and subjectivity. The end result must be able to give an insight of what the embassies really are so they can be interpreted and understand as a whole new information or to strengthen the already known facts.

* 1. **Data Description**

In order to accomplish the analysis, we will need data on Indonesian Embassies abroad and their latitude and longitude so we can combine it with data about venues surrounding the embassies which we intend to explore from FourSquare API to give a better understanding about the embassy's neighborhood, venues and places.

* 1. **Solution**

Our solution is to cluster the embassies and see the differences and similarities between them based on the data gathered about venues surrounding the embassies.

* 1. **Methodology**

1. **Data Gathering**
2. List of Indonesian Embassies Abroad

Apparently to get data about the list of Indonesian Embassies abroad is going to be a bit difficult, this is because the Ministry of Foreign Affairs website doesn't display the data to be easily analyzable. But with the help of google and wikipedia we find a page that we could extract the data that we need as a starting point.

<https://id.wikipedia.org/wiki/Kedutaan_besar_Republik_Indonesia>

Table

Description automatically generated

After scraping the page with BeautifulSoup we manage to import the data in a Data Frame as so. There are 98 Indonesian Embassies according to our data. As we can see the data needs to be cleaned since the scrapping process leave us with some characters we need to eliminate.

Table

Description automatically generated with low confidence

After we clean the Data Frame we need to change some values and drop some rows from the Data Frame. This is because some of the data used pronunciation of capitals in Indonesian language, others not so common name and data that does not interest us such as Indonesian representatives for the United Nations.

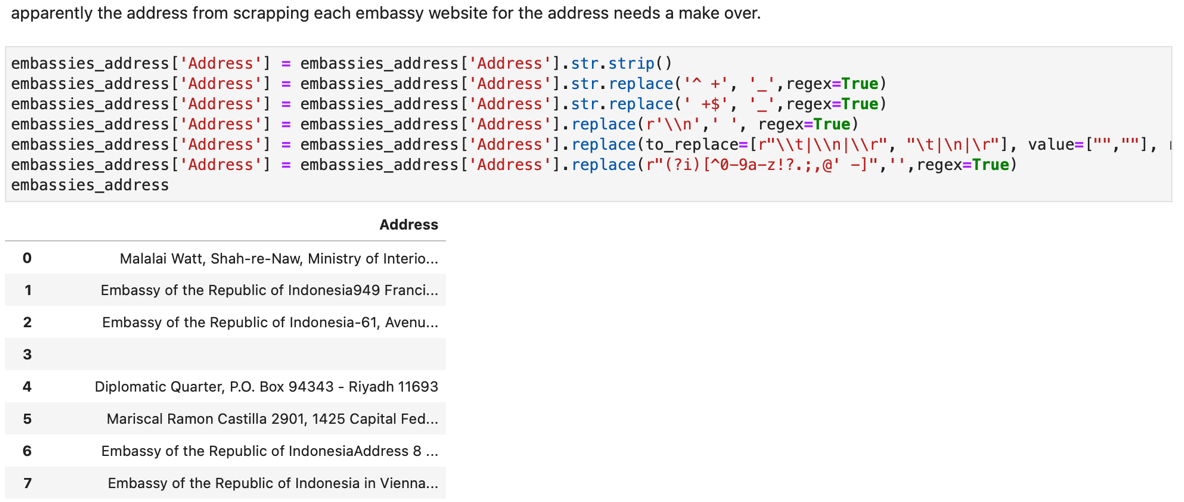
The final tally is 93 Indonesian Embassies abroad.

1. The Embassies Address

It would be nice if everything was served on a plate but in this task, we need to have some wits up our sleeves. What we need are longitudes and latitudes of the embassies, as with the problem above it is not readily available to us. But there is hope since the address is available for us to extract from the Ministry of Foreign Affairs website for they have a common URL that we can explore.

<https://kemlu.go.id/CAPITAL>

with this information we need to iterate this URL with our Data Frame and do another scrapping to get the all the addresses.



Table

Description automatically generatedAfter looking at the result we need to take a decision to save the data frame and edit the address manually since the format is not unison. Not a very satisfying task but something we need to take since creating regex rules does not look like a viable solution for this problem. After that we integrate it with our Data Frame and voila a list of Indonesian Embassies abroad with their address.

1. The Embassies Geo Location

Graphical user interface, text, application, email

Description automatically generated

As our initial data requirement is to have a list of embassies with their geolocation, we need to translate the address into latitude and longitude coordinate. To do that we use a common library in python called geocoder and loop through the address.

And then format the result in a Data Frame and combine with our embassies Data Frame.

Table

Description automatically generated

Unfortunately, there are some embassies without geolocation information.

Table

Description automatically generated

In this case, 40 embassies are still without geolocation information.

Graphical user interface, text, application, email

Description automatically generatedTo overcome this obstacle and after googling about free forward geocoding, we stumble open positionstack.com which offer a diverse and rich API in their service. After following the instruction and documentation, we translate the remaining address to get their latitude and longitude.

And do the process all over again and see if there is any embassy without geolocation information, which there are and decide to do it manually since the number is quite small.

Table

Description automatically generated

Let’s see if we get our geolocation data right with displaying a sample.

Map

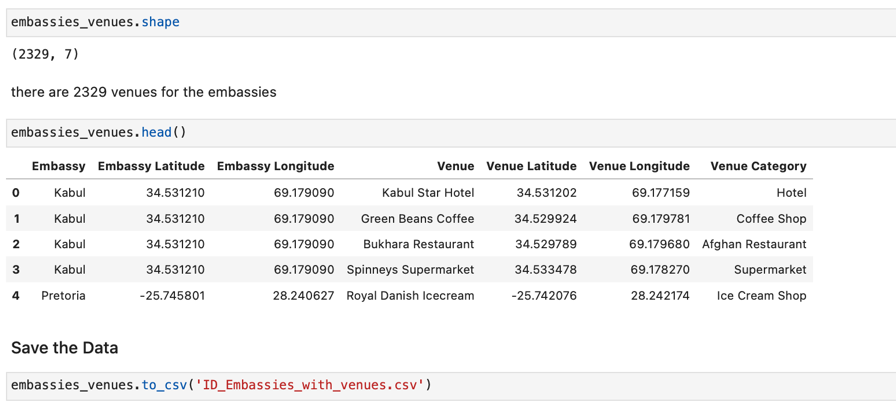
Description automatically generated

1. The Embassies Venues

Graphical user interface, text, application

Description automatically generated

Now let’s get venues data surrounding the embassies from FourSquare.



Which resulted in an astonishing 2.329 venues which we can work on.

1. **Data Exploration**

Graphical user interface, text, application, email

Description automatically generated

To get a better understanding of the data, we need to explore our data. Let’s see what kind of data we have exactly. We can see that each embassy has venues and their categories. The one we will be using is the value in the Venue Category, because this value will be available across embassies. Let’s see how many categories there are in our Data Frame.

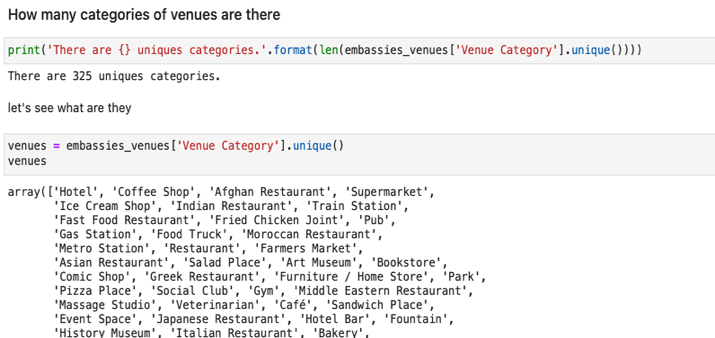
Table

Description automatically generatedWe can see that that the biggest venues are in Manila, Mexico City and Washington. And the least venues are in Ottawa, Addis Ababa and Abu Dhabi.

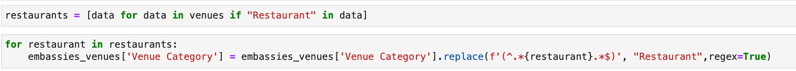
Table

Description automatically generated

Let’s see how many categories we have in the Data Frame.



There are 325 categories and if we see some are generally the same. A Restaurant might have specialty but basically it is a restaurant, so we have to gather identical categories into a more general category so we can have a better model as a result.

We do this repeatedly for the categories we want until we are satisfied that there no more ambiguous data in the Data Frame.

In this case from 325 we managed to shrink it to 143 categories. For the last exploration let’s find out if there is any embassy with no venues at all.

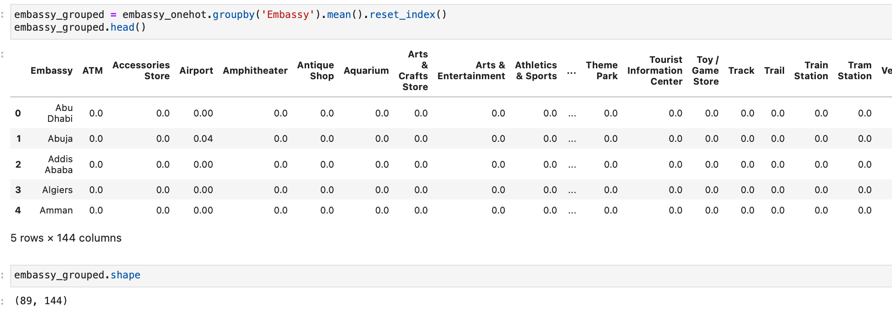
Graphical user interface, text

Description automatically generated

Apparently, there are 4 embassies with no venues data. So, we will disregard these embassies in the final result.

The next step is to make the categories into their own column in the Data Frame.

Which resulted in 2.329 rows and 144 columns. In the current state our Data Frame can be used to make a cluster.

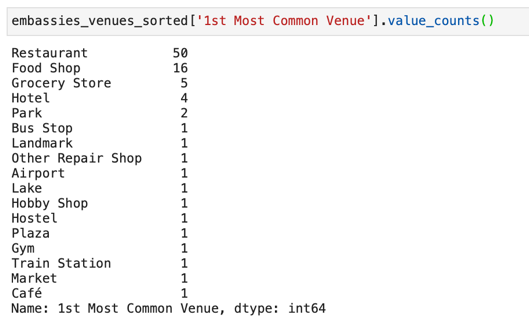
If we looked back, our first data consist of 93 embassies and after exploring FourSquare we end up with 4 embassies with no venues data. So, if our calculation is correct, we need to sum up 2.329 rows into 89 rows. We do that by grouping the Data Frame by the embassy column and count the mean of each column in the group.

Graphical user interface, text, application, email

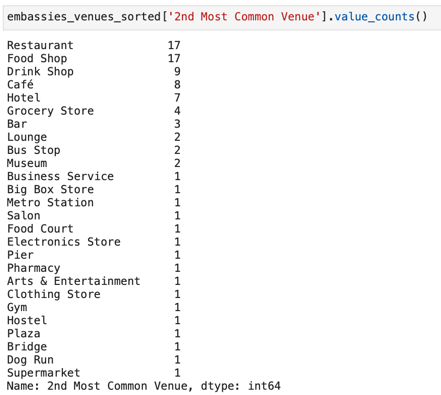
Description automatically generated

Now let’s find out what are the most common venues in each embassy.



We can see from the Data Frame that Restaurant comes up a lot on the 1st Most Common Venue. Let’s count how many of them.

We can see that 50 embassies have Restaurant as their 1st most common venue and Food Shop in second.

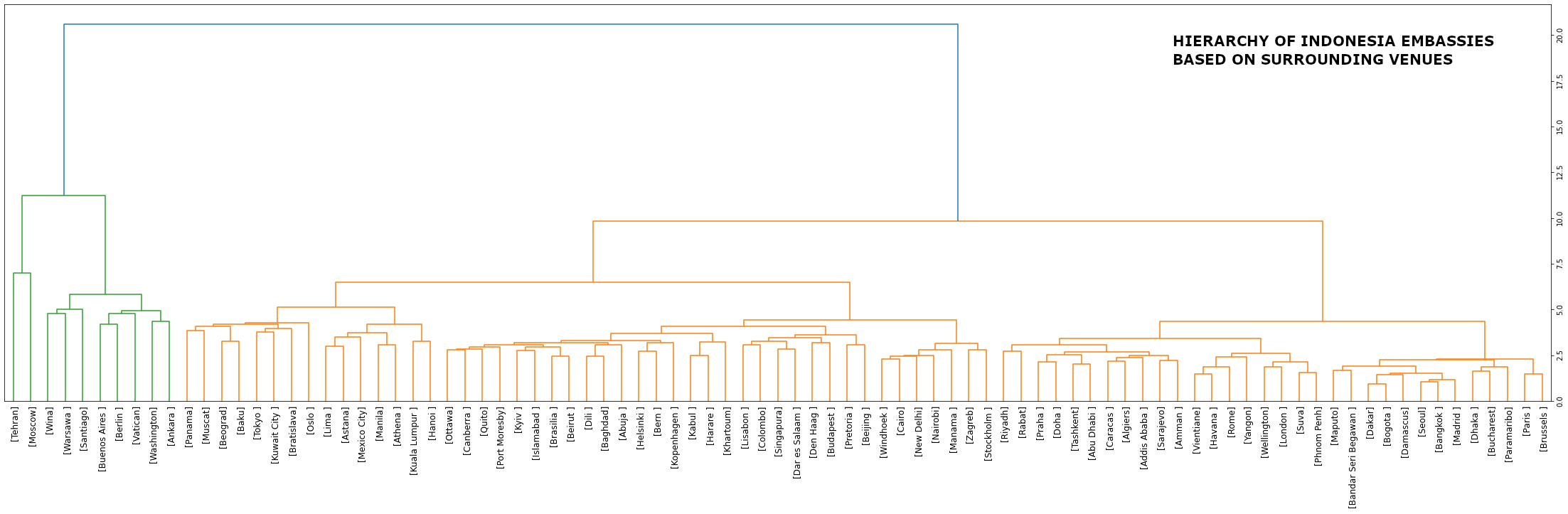
What’s the most 2nd Most Common Venue?. Apparently is the same as before with Restaurant and Food Shop as the top two.

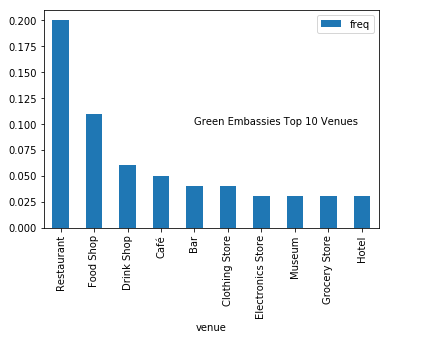


What’s the most 3rd Most Common Venue?. Looks like Hotel comes up with 13 embassies have it as their 3rd Most Common Venue.

1. **Clustering**

Before we begin clustering, we must answer the question “do we know how many clusters the data have?”. Based on the problem domain and your knowledge of the data, usually we can answer the question. But if we are clueless about the cluster in the data, there are tools that can help us. First let’s see if our data have any hierarchy in them and built our cluster in a hierarchy.



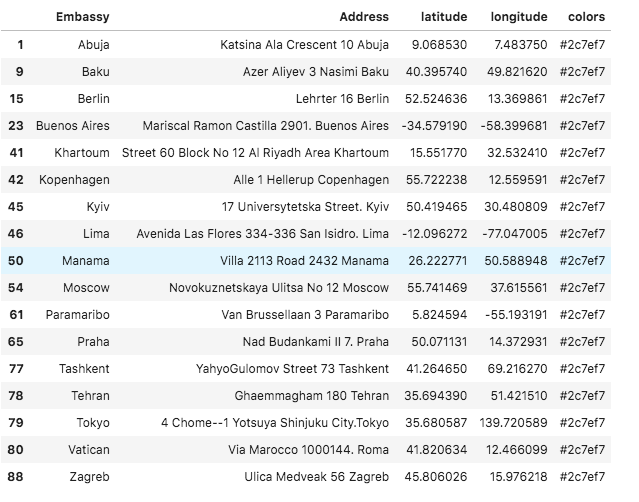


We can see that the embassy in Wina and Warsawa is very close in venues with the embassy in Santiago. Combine they are very close with the embassies in Buenos Aires, Berlin, Vatican, Washington and Ankara. What’s interesting in the hierarchy, the embassies in the orange hierarchy will eventually resemble the embassies in the green hierarchy. Let’s combine the embassies in the green hierarchy and see what their top venues are to get a better understanding of the hierarchy. We can see that most of them are something that we are expected from the data and nothing is out of the ordinary.

Now let’s cluster our embassies with a question like ‘an embassy is in a group all together if there are at least 7 embassies with the same value, if not then they are an outlier?’. We can do that with the help of DBSCAN which clusters based on the data and the parameters we wanted and also produce a set of outlier data that doesn’t belong to any of the generated clusters. We can see the result below.



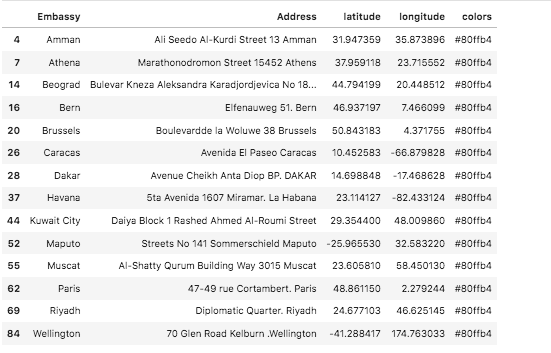
The embassies in the list are outliers, they are the embassies whose values are very far apart from the rest of the data and doesn’t have at least 7 embassies with similar data.



These are the first cluster, as we can see some of the embassies from the green hierarchy are here like Moscow, Tehran, Berlin and Buenos Aires belong in this cluster.

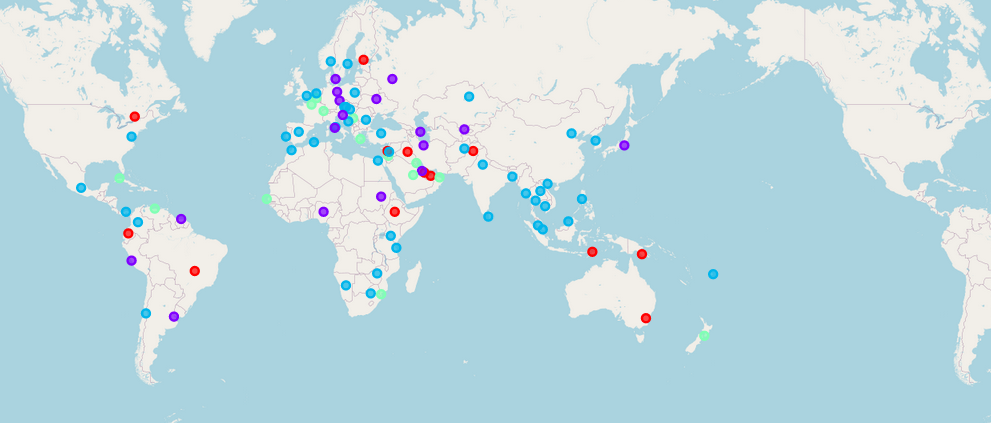


And this is the second cluster of our embassies where the rest of the green hierarchy belongs such as Ankara, Washington, Wina, Warsawa, Vatican and Santiago.

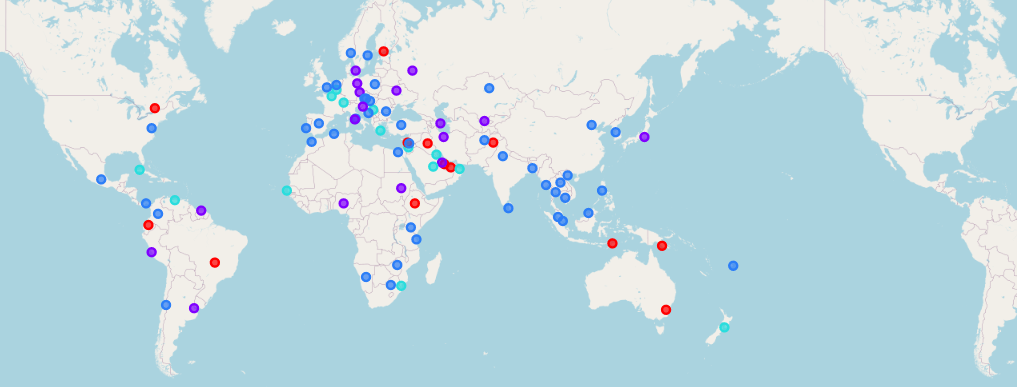


This is a cluster of other embassies that have similarities between them and form their own hierarchy.

Let’s visualize our clusters and the outliers together.



What if we have a certain number of clusters in mind, let’s say we want to cluster our data into 4 clusters based on a certain knowledge previously known to us. We can use KMeans and see what the results are and visualize our new clusters. As we can see in general the result is quite consistent with the previous cluster.



* 1. **Discussion**

The result of our cluster shows there are some similarities between the Indonesian Embassies. We based the similarities on venues around the embassies, although we could also based it on any other data according to the problem we are trying to solve. The conclusion of the result of an unlabeled data cluster like we have, usually ends out in the domain knowledge of our intended research in this case the staff on Indonesian Foreign Ministry. But the general audience also can enjoy the result with knowing that their embassies in Australia, Timor Leste and Papua Nugini are almost similar based on the venues around them compare with the embassies in south east asia. Majorities of the embassies in the south and east asia are alike in both clusters.

Another improvements we need to consider is to get get data about public venues such as embassy, government building, public services which usually common around an embassy. Unfortunately we couldn’t get the data without resorting to paid services, which is beyond our reach at the moment.