Louis Kabelka
4/9/2022
Statistical Computing

Global Demographics Report

This project seeks to analyze global demographics data. We will use XML sheets from the CIA Factbook to get the infant mortality and population data for different countries. Then we will combine this data with online geolocation data. In order to use these data sets together we will cross-reference ISO 3166 codes.
There will be different ways to categorize the data. First, we will categorize the data by diving it into quantiles. Next, we will take a more complex approach by implementing a k-means clustering algorithm. Using these categories and the geo-location data, we will create maps to better visualize global trends.
Finally, we will extend the project by analyzing individual continents using k-means clustering.

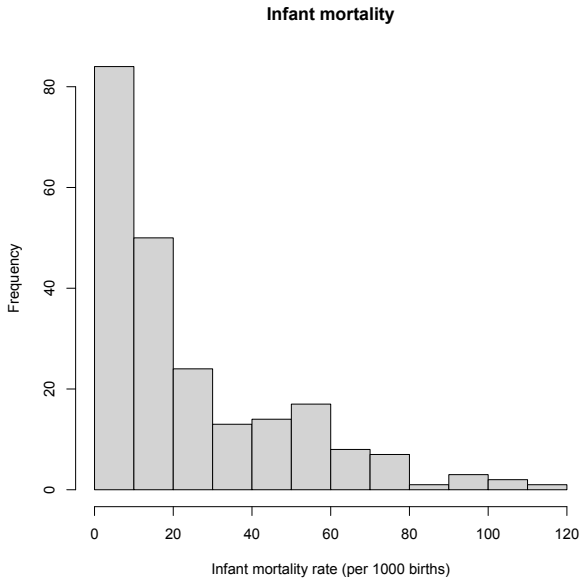
1. CIA Factbook

We have two codes to identify countries. First, the ISO 3166 code. Second, the CIA Factbook 2-letter country abbreviation.

However, not every country in the CIA Factbook has these codes:
a) There are **28 countries** with no ISO 3166 code.
b) There are **6 countries** with neither an ISO 3166 code or a CIA Factbook 2-letter country abbreviation.

2. Infant Mortality

a)



This histogram shows the distribution of the Infant mortality rate of every country in the CIA Factbook. The values indicate the number of infant deaths per 1000 births.

We can see that most of the countries in the CIA Factbook have an infant mortality rate between 0-20. Then there is a considerable amount of countries between 20-80. But beyond that, only very few countries exceed an infant mortality from 80 going up to 120.

b)
The countries with the ten highest infant mortality rates in descending order are the following:

Country	Infant Mortality (per 1000 births)
Afghanistan	117.23
Mali	104.34
Somalia	100.14
Central African Republic	92.86
Guinea-Bissau	90.92
Chad	90.30
Niger	86.27
Angola	79.99
Burkina Faso	76.80
Nigeria	74.09

Every one of these countries is located in Africa. We can tell that the countries with highest infant mortality are very condensed in the African continent. However, we can not tell if there is a particular part of Africa that is representative of the poorest countries: Afghanistan is in South Africa, Mali is in West Africa, Somalia is in East Africa, Central African Republic is in Central Africa.
The lack of other continents means that Africa has some other factors that contribute to it having a particularly high infant mortality. On no other continent do we observe such high infant morality rates. Thus, something or a combination of things must be happening in Africa that are causing high infant mortality. Some possible causations for Africa standing out are disease, war or poverty.

3. Getting Longitude & Latitude Data
We searched " iso3166 latitude longitude csv file" in Google Search. The following GitHub was the first result: https://gist.github.com/tadast/8827699. It contains a .csv file of countries with 3 different code identification and average latitude, longitude data.

The data set had two issues:
1. It contained some repeat entries of essentially the same country with the same iso3166, latitude and longitude data. For example, "Russia" and "Russian Federation" were separate entries.
We isolated the iso3166 code, latitude, longitude data. Then equal rows would actually be detectable. We performed a unique() call to isolate all the rows that were now equivalent.
2. All of the iso3166 codes have a " " character at the beginning. Therefore, it was impossible to merge it with the other data frames, since the Factbook codes did not contain the " " character at the start.
We performed an sapply() on the iso3166 column to get a substring of the 2nd-3rd character. This removed the 1st " " character from every entry.

4. Merging with Missing Keys

We decided to not include any rows that had any missing keys. The data frame we started with was the cross-referencing data frame with 279 rows. After the merging, the final data frame had only 222 rows.

However, the data frame only shrunk during the first merging with the infant mortality data frame. The data frame remained the same size while subsequently merging it with the population data frame and the longitude, latitude data frame.

5. Mortality

The mean mortality rate of countries with population less than 10 million is 19.062 deaths per 1000.

The mean mortality rate of countries with population greater than 50 million is 26.05125 deaths per 1000.

6. Mortality Quantiles

Next, we will figure out how to breakdown the infant mortality into different quantiles. This will allow us to categorize countries for visualization. In our world maps, we want to be able to represent a range of mortality values with a particular color.

The following is the typical breakpoints for a quantile distribution: 0, 25%, 50%, 75%, 100%. This is a good way to divide distributions into even categories, where 0%-25% are the lowest values and 75%-100% are highest values.

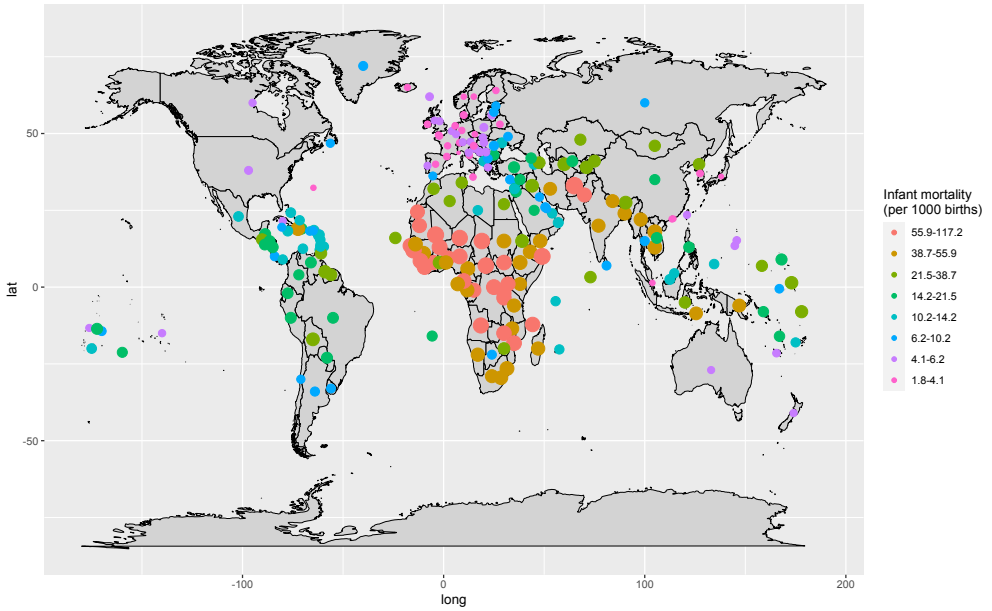
Instead, we chose to divide the distribution into eight percentage quantiles. We added one more break point between all the intervals. This allows for more detailed categories. The percentage break points are the following:

0, 12.5%, 25%, 37.5%, 50%, 62.5%, 75%, 87.5%, 100%.

Using those break points, we end up with the following # of countries per range:

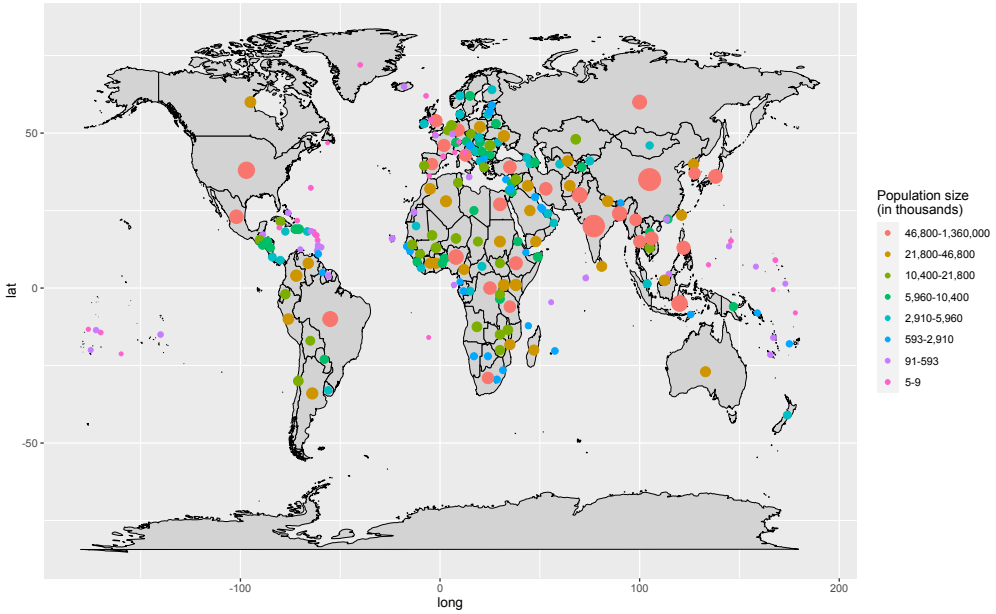
Range of infant mortality	# of countries
(1.81,4.14]	27
(4.14,6.2]	28
(6.2,10.2]	27
(10.2,14.2]	28
(14.2, 21.5]	28
(21.5, 38.7]	27
(38.7, 55.9]	28
(55.9, 117]	28

7. Infant mortality world map



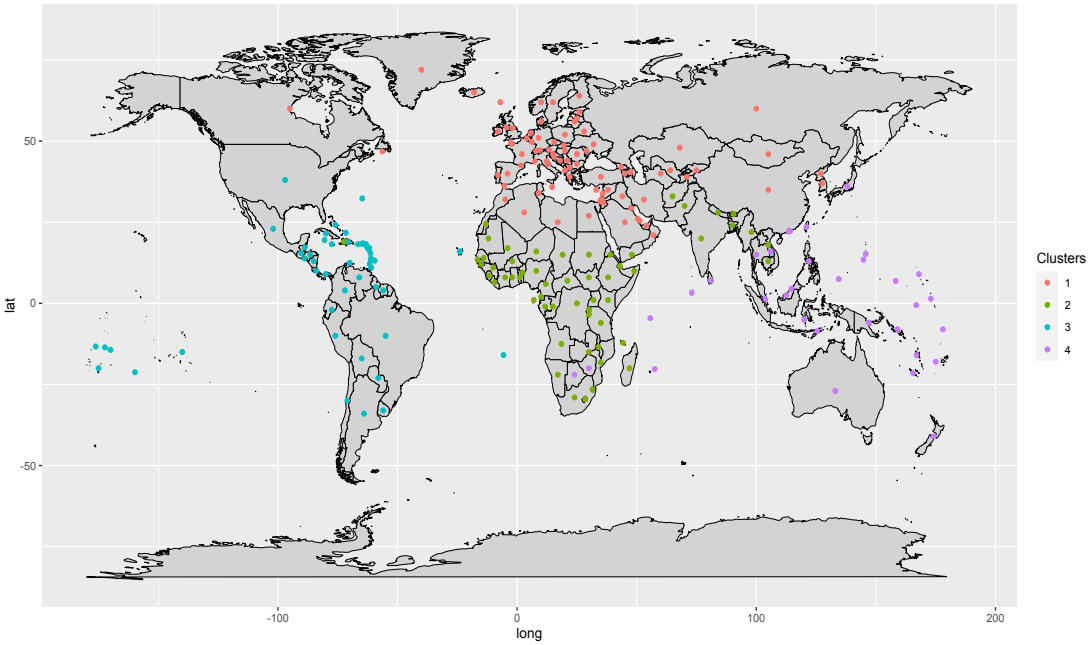
This world map shows infant mortality rates (per 100 births). Like we saw before, the majority of high infant mortality is centralized in Africa. We can see that comparatively there is a cluster of very low infant mortality in Europe and North America.

8. Population size world map



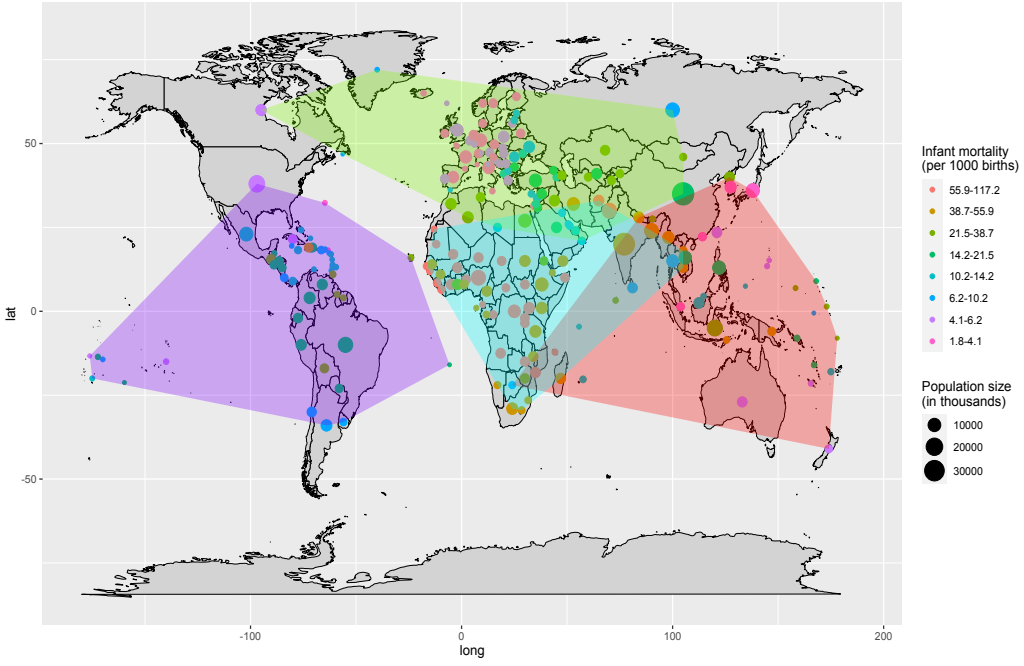
This world map shows the population size (in thousands). The countries with the highest population tend to be in Asia. India and China are obvious outliers. We can also see that although Africa has a high infant mortality rate, they have a comparatively lower population size. The lowest populations tend to be islands in the Pacific Ocean.

9. K-Means Clustering Algorithm



This world map has 4 different clusters based on location & mortality with each cluster being represented by a different circle color.

10. Regional Map



This world map has infant mortality as the color of each circle, population size as the size of each circle and convex hulls to represents the clusters based on location & mortality.

The world maps in 9. & 10. represent 4 different clusters of countries. Each cluster is based on the countries location and its infant mortality rate. So, each cluster will be countries that are near each other with similar infant mortality rates.

The algorithm that was used to determine the clusters is K-Means Clustering. This algorithm is an unsupervised learning algorithm. We use different features to cluster points together. In this case, the features are longitude, latitude, infant mortality rate and the points are countries It continuously finds the center of the four groupings
(ie. the mean of the different features). Then for each country, it sorts it into the cluster based on which each country is the closest to it in terms of Euclidean Distance for all the featurers. When the clusters are no longer changing, the algorithm stops. Since, the first centers are determined randomly, we will get a slightly different result every time the algorithm is run.

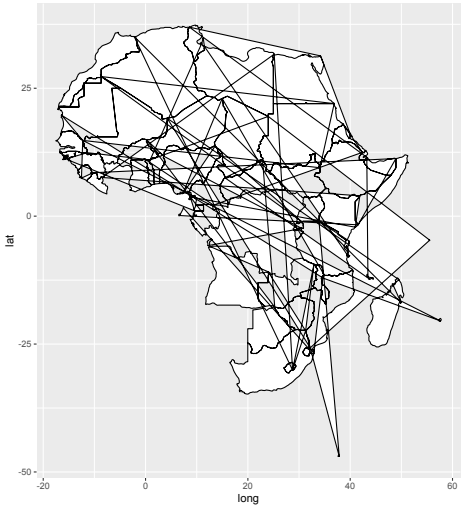
[The R-Implementation of the K-Means Clustering Algorithm is included in the Appendix.]

Extension

Near the start of this project, we saw that most of the countries with high infant mortality were in Africa. But it was unclear just from the data which parts of Africa have the highest mortality. From our maps, we can only tell how the infant mortality of Africa is relative to the rest of the world. My extension seeks to better improve our understanding of intracontinental infant mortality rates. We will create continent maps that cluster infant mortality rates based within the continent.

1. We will get data to categorize countries into different continents.
2. We will creates a continentMap() function that returns a map of a continent with k-means clustering chulls. The dots of countries will have both the size and color based on infant mortality.
3. We will analyze how to the maps show trends within Africa, Asia and Europe.

1. Get data.



The plot on the left shows a failed attempt to plot maps of the using the countries from our original data frame. Since, we don't have enough countries, there are a bunch of missing spots.

So, instead we downloaded a list of countries from an R tutorial on mapping continents: https://warin.ca/posts/rcourse-datavisualizationwithr-maps/

We won't merge the list of countries with our data frame this time. Instead, we will subset our data frame based on the list of countries.

2. continentMap()

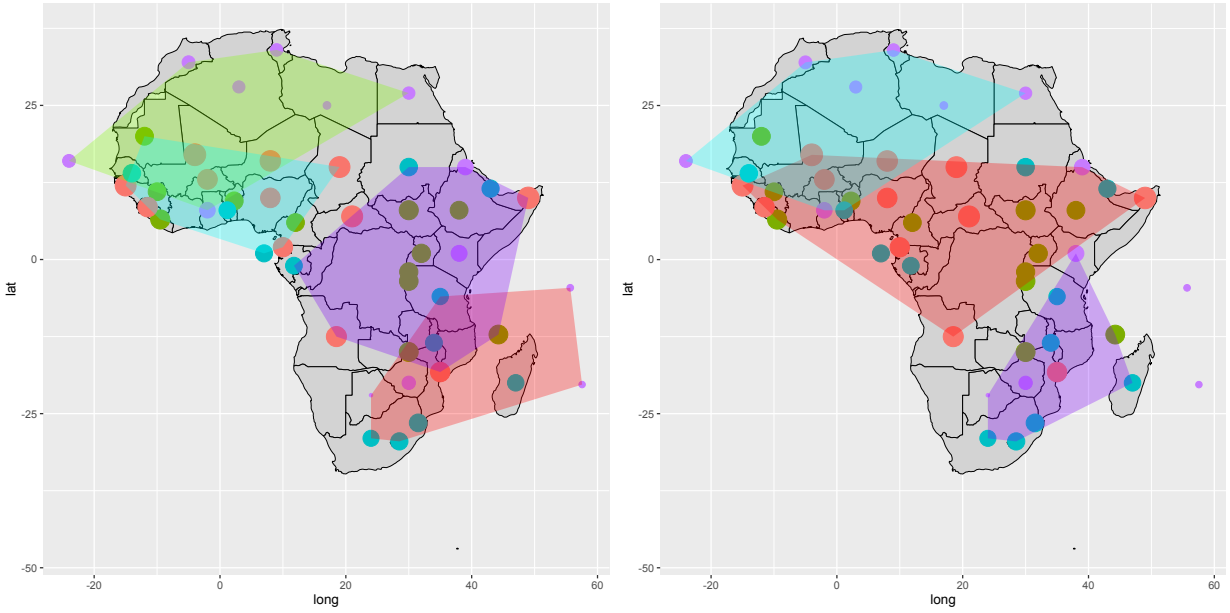
We made a general function that takes a list of countries to create the map and the k-value for k-means clustering, like our previous function. As mentioned, the list of countries will subset our original data frame.

This time we will use different breakpoints for determining colors. Since, we're only plotting a single continent, there will be far less countries to display. Therefore, having many colors will only make our map confusion We decided to scale down the more traditional break points: 0%, 25%, 50%, 75%, 100%. This gives us 4 quantiles.

3. Analysis

When analyzing these countries, it's important to note that all the infant mortality rates we can see are relative to the continent. A high infant mortality rate in Europe is going to be a lot lower than a high infant mortality rate in Africa.

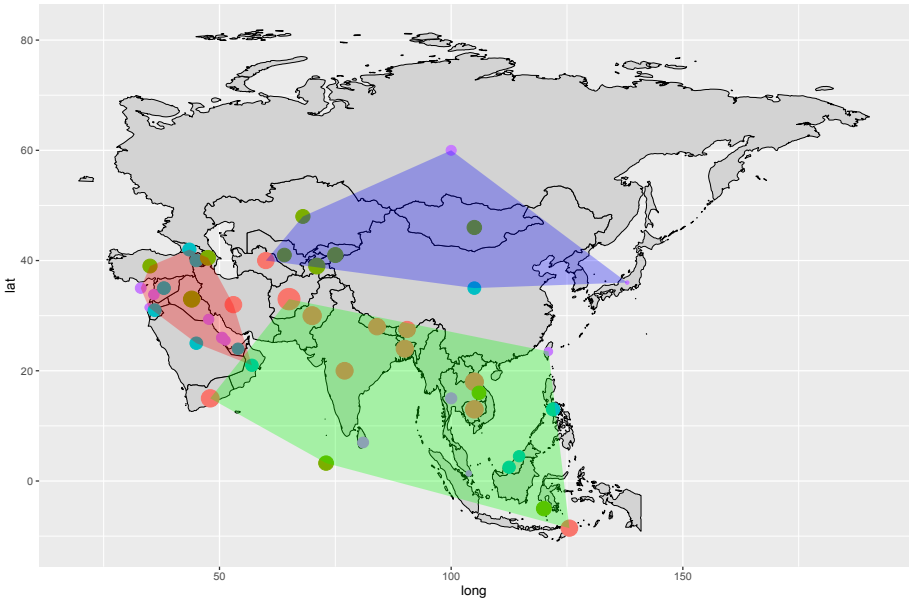
a. Africa



The above shows two different results from the continent map of Africa with k=3 and k=4. From the first map, we can see that the infant mortality rate is especially high at the north-west coast of Africa. From the second map, see that the highest infant mortality rates also could be seen as extending more throughout central Africa.

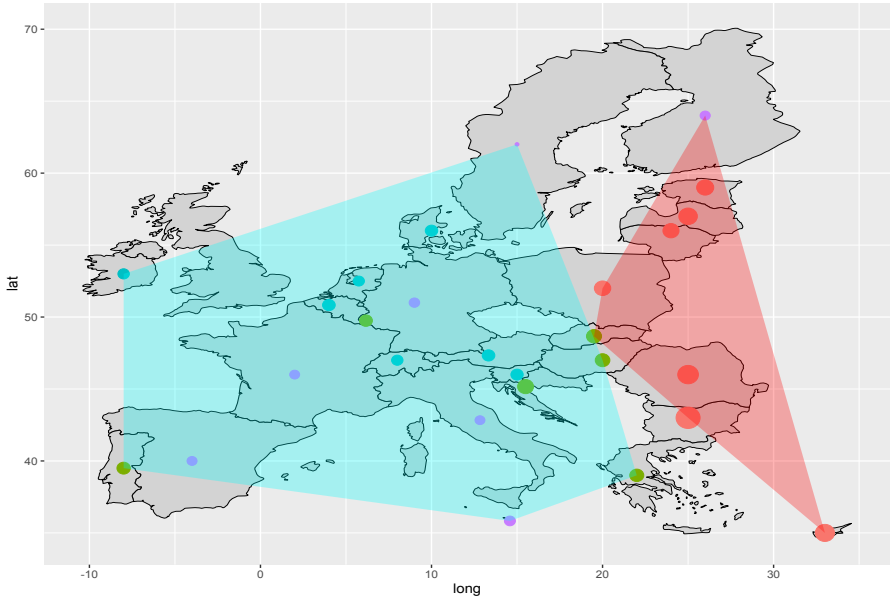
The two maps share in common that they have a clustering in the north with the lowest infant mortality rates. This is particularity interesting, since it is right above the clustering with the highest infant morality rates.

b. Asia



The continent map for Asia shows the clustering with the highest infant mortality rate as a big clustering in South Asia. There is a size contrast between that clustering and the one to the left of it. There is also much more variation of infant mortality rate in the clustering to the left, despite it covering less area. This could be because there are more first-world countries as we get closer to Europe.

c. Europe



The continent map for Europe shows a difference between Eastern Europe and Western Europe, represented in these two clustering. The analysis with k=2 worked the best to represent this difference. We can see that there is Western Europe tends to have very low infant mortality rates, Eastern Europe tends to have very high infant mortality rates.