Global Demographics

Isabel Arvelo

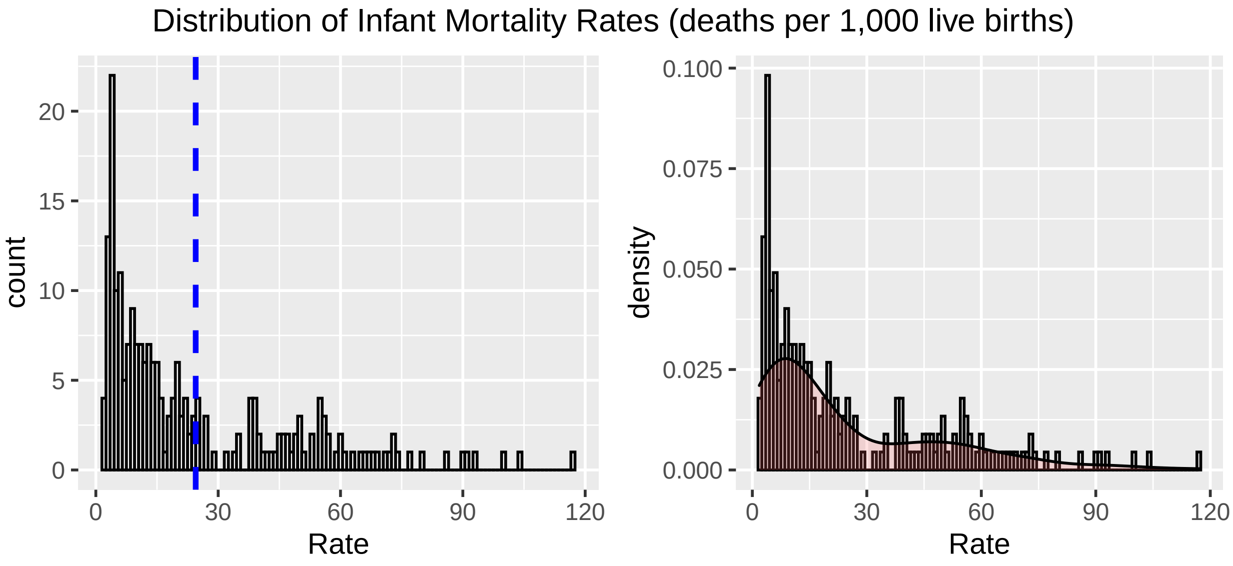
2022-04-29

In this project, I extracted and analyzed global demographic data to explore global trends in population, mortality, infant mortality, GDP, and language. The data for this project came from the 2014 CIA Factbook, a data source that contains basic intelligence for 279 world entities. Although the factbook is released on an annual basis, I chose to use the 2014 version because it is accessible in XML format online, allowing for easy querying and manipulation. I also retreieved geospatial data from Google Developers in order to visually represent the data on a world map. I began with an exploratory data analysis to gain a better understanding of the type of data in the factbook.

Deliverable #1:

1. ISO 3166 codes are part of an internationally recognized codes of letters and/or numbers that are used to refer to countries and their subdivisions in a standardized fashion. Of the 219 countries in the CIA Factbook, 28 don't have an ISO 3166 code.
2. The factbook also uses it own coding system and alots each country a 2-letter abbrevation of its name. 6 of the countries do not an ISO 3166 code or a CIA Factbook 2-letter country abbreviation.

Deliverable #2:



The distribution of infant mortality rates is heavily skewed right, indicating that the vast majority of countries have relatively low infant mortality rates, below 30 deaths per 1,000 live births. Fewer countries experience higher rates of infant mortality The median of the distribution is about 14 deaths per 1,000 live births and about 75% of countries remain below 40 deaths per 1,000 live births. However, the countries with high infant mortality rates, tend to have very high rates. There are several outliers and the maximum of the distribution is more than 8 times the median. The 6 countries with the highest infant mortality experience more than 90 deaths per 1,000 live births.

In order, the countries with the 10 largest infant mortality rates are Afghanistan, Mali, Somalia, Central African Republic, Guinea-Bissau, Chad, Niger, Angola, Burkina Faso and Nigeria. 9/10 of these countries are on the continent of Africa and the other country is Afghanistan, a region that has experienced extreme political turmoil and several wars in the last two decades. From looking at this, it appears that underdeveloped countries, defined by the United Nations to have "widespread chronic poverty and less economic development", tend to have higher infant mortality rates. 5 of the countries with the highest infant mortality are also in the top 10 for the UN's list of least developed countries.

Deliverable #3 and #4:

The next step in the project was to locate geolocation data for countries online and prepare the data to be merged with the demographic data from the CIA factbook. I got my data by downloading a csv from Kaggle that has latitude and longitude for every country and state in the U.S and then reading the csv into R. The original source of the data was public data on Google Developers that was released under a Creative Commons 4.0 license.There were no issues with the data set. All of the variables were in the correct form when I downloaded it. I eliminated the columns with USA state data and changed the name of the column with the ISO-3166 code from "country code" to 'iso3166'. Otherwise, all the variables were of the correct type. I then created a dataframe named latlong that contained the latitude and longtitude corresponding to each iso3166 code.

After storing the geolocation data in a data frame, I proceeded to create a single merged data frame with the the country name, the ISO 3166 country code, the 2-letter country abbreviation used in the CIA Factbook, population, mortality, latitude, and longitude. In order to do this, I did a series of full joins to keep all the rows from all of the tables and then eliminated the rows/data I did not want.

I began by joining the countryCodes data table(containing country name, cia, and iso3166) with the latlong table using the ISO 3166 country code as the key to join them. I then performed a full join to merge this table with the population data frame, using the 2-letter CIA codes as the key. At this point, I had a table with all of the information on country name, the ISO 3166 country code, the 2-letter country abbreviation used in the CIA Factbook, population, latitude, and longitude. The last join I performed was a full join of this large dataframe with the mortality data, using CIA country code as the key, which resulted in a data frame with 285 rows. However, since there is different data available for different countries, I had NA values dispersed throughout each variable, resulting from joins in which a country had data for one variable, but not another.

To be specific, this data frame had 4 rows with NA values for the country variable, 3 rows with NA values for the iso3166 variable, 45 rows with NA values for the cia code variable, 61 rows with NA values for the mortality variable, and 40 rows with NA values for the geolocation variables. The first analysis and visualization I wanted to complete required mortality, population, latitude, longitude data for each country so I created a data set that dropped all rows that had missing values for any of these variables.

However, I also wanted to use a k-means clustering algorithm based on geolocation and mortality data, so for this analysis it would not matter if there was population data for a given country. Therefore, I created a second data set that only dropped rows missing values for mortality, latitude, or longitude. However, I discovered that rows missing country data coincided with the rows missing the data for the other variables so this second data set was identical to the first. I procedded forward with a single data set that had all values populated for country name, the ISO 3166 country code, the 2-letter country abbreviation used in the CIA Factbook, population, mortality, latitude, and longitude.

Deliverable #5

The mean mortality rate for all countries with population less than 10 million is 18.863 per 1,000 live births. The mean mortality rate for all countries with population more than 50 million is about 26.051 deaths per 1,000 live births. .

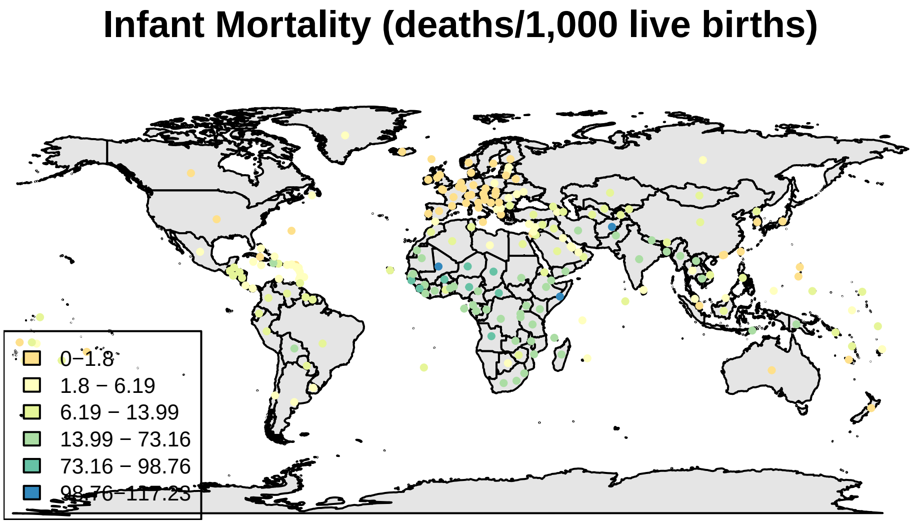
Deliverable #6

Next, I wanted to create a world map showing the infant mortality rate for each country with colored circles, where the color of the circle for each country indicates the approximate mortality rate. In order to do this, I first had to discretize the mortality rates.

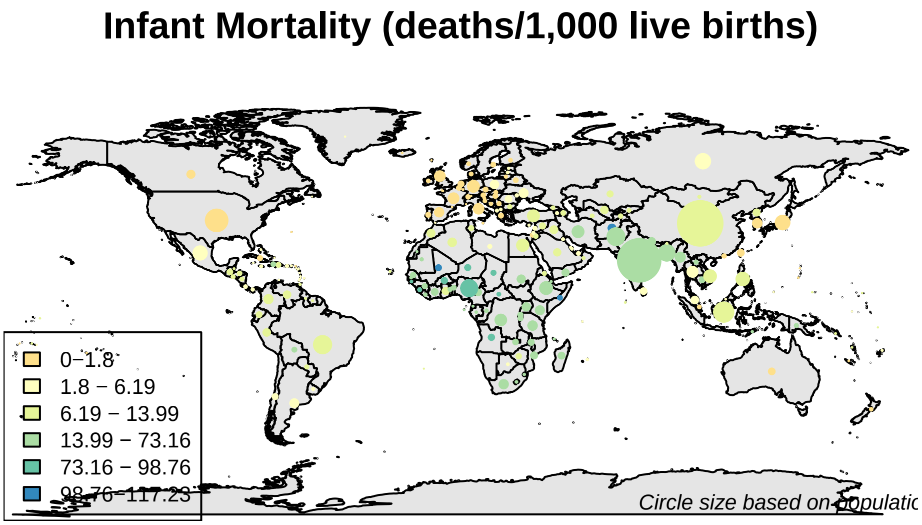
Factors	Freq
25	55
50	55
75	55
95	44
99	8
100	3

I used the quantiles as basis to create a better distribution of the data points across the different levels of the factor. However, I also made the top 5% and top 1% of rates their own factors, because the the countries with such high infant mortality rates merit explicit attention.

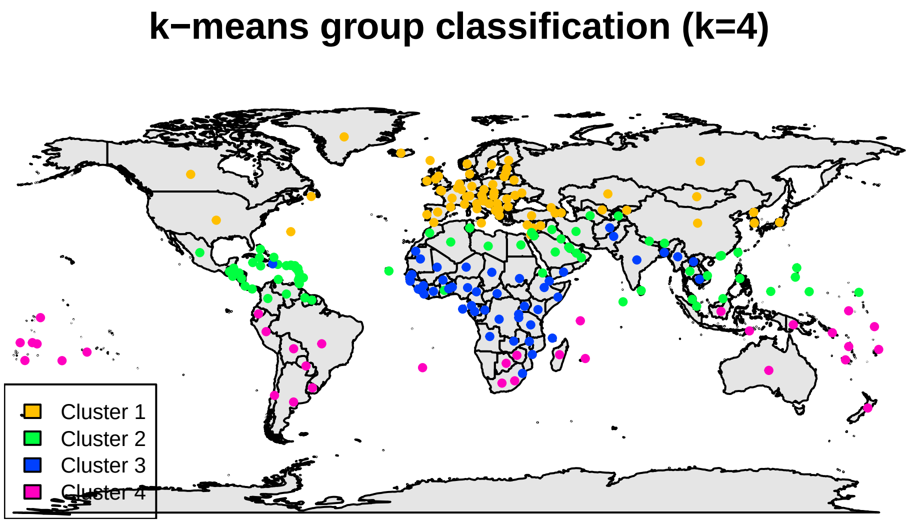
Deliverable #7



Deliverable #8

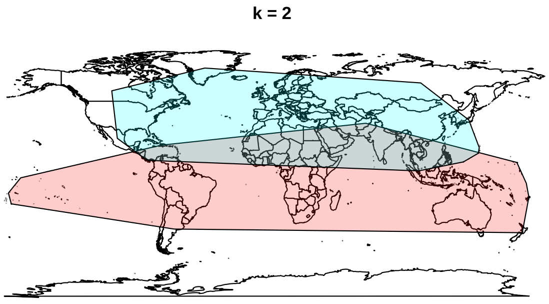

Infant Mortality (deaths/1,000 live births)
0–1.8
1.8 – 6.19
6.19 – 13.99
13.99 – 73.16
73.16 – 98.76
98.76–117.23
Circle size based on populatic

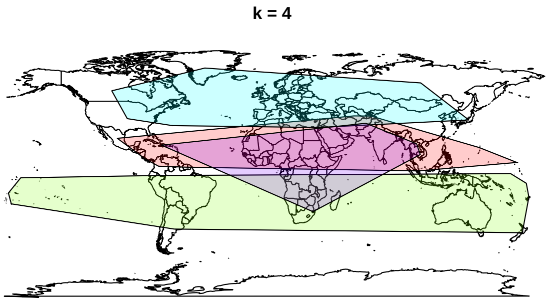
Deliverable #9 Next, I implemented an unsupervised learning algorithm, specifically a k-means clustering algorithm, to classify the countries based on latitude, longitude, and infant mortality. My aim was to partition the observations into k clusters of similar observations with respect to the selected variables.

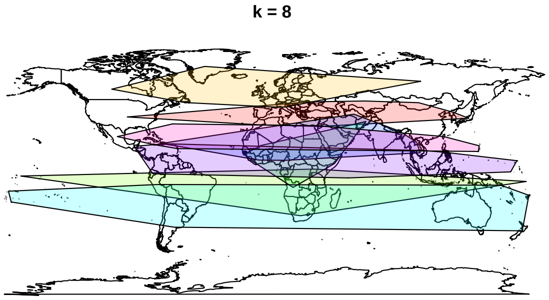


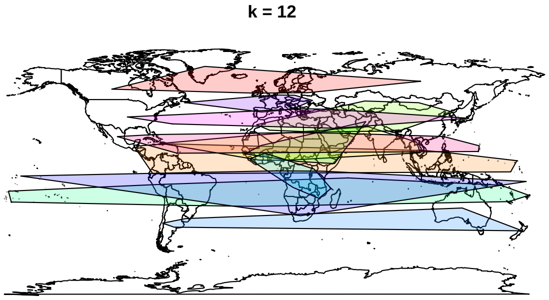
Deliverable #10

The last step was to implement an alternate display of the classification groupings on a world map. In this visualization, a convex hull is drawn around each cluster.


k = 2


k = 4


k = 8


k = 12

Extensions

According to Enyclopedia Brittanica, westernization is “the adoption of the practices and culture of western Europe by societies and countries in other parts of the world, whether through compulsion or influence”. This phenomenon is a consequence of the process of colonialism and can be observed through a linguistic lens. Individuals that speak English are able to travel the world with relative ease because of its widespread adoption across the globe.

While other factors like America’s success in WW2 and it’s status as technological superpower, propogate the popularity of the language, I was curious in further exploring the linguistic legacy of colonialism by examining the dispersion of languages spoken by colonial superpowers like England (English), France (French), Germany (German), Portugal (Portuguese), and Spain (Spanish).

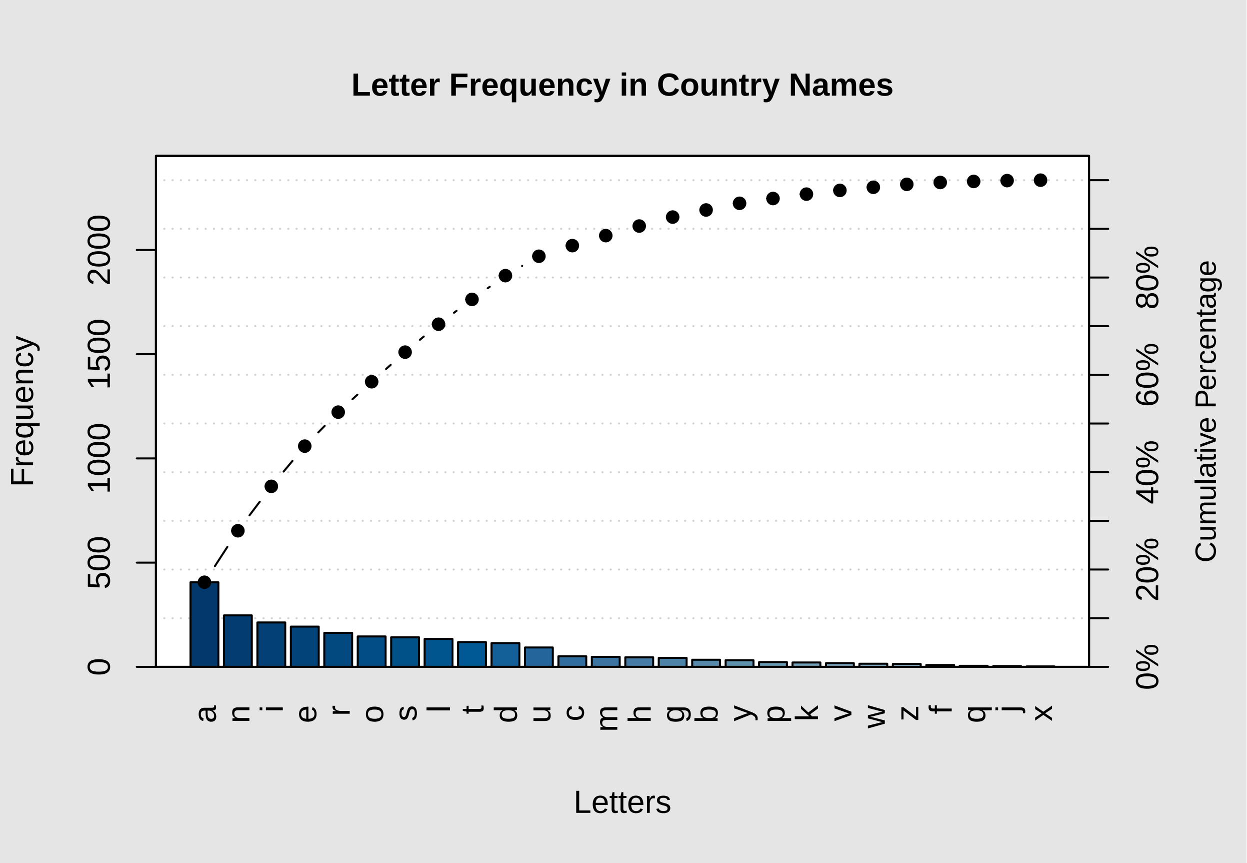
I extracted data from the CIA World Factbook, and found the number of countries that have each of these languages as an official language of the country.

```
##     Language Frequency
## 1    English        57
## 2     French        34
## 3    Spanish        17
## 4 Portuguese         7
## 5     German         5
```

The table above illustrates that the top 3 “official languages” spoken in countries across the world are english, french, and spanish; languages spoken by three of the most prominent imperial powers.

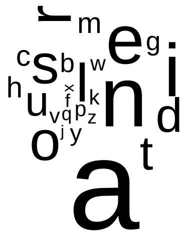
Of the 57 countries that have English as an official language, it is only the most commonly spoken language in 31 (54.39%) of those.

Another aspect of the data I was curious about was which letters are most common in country names. All 26 letters in the English alphabet appear at least once in the English spelling of the country names. However, some are in several hundred country names, while others only appear in a handful.

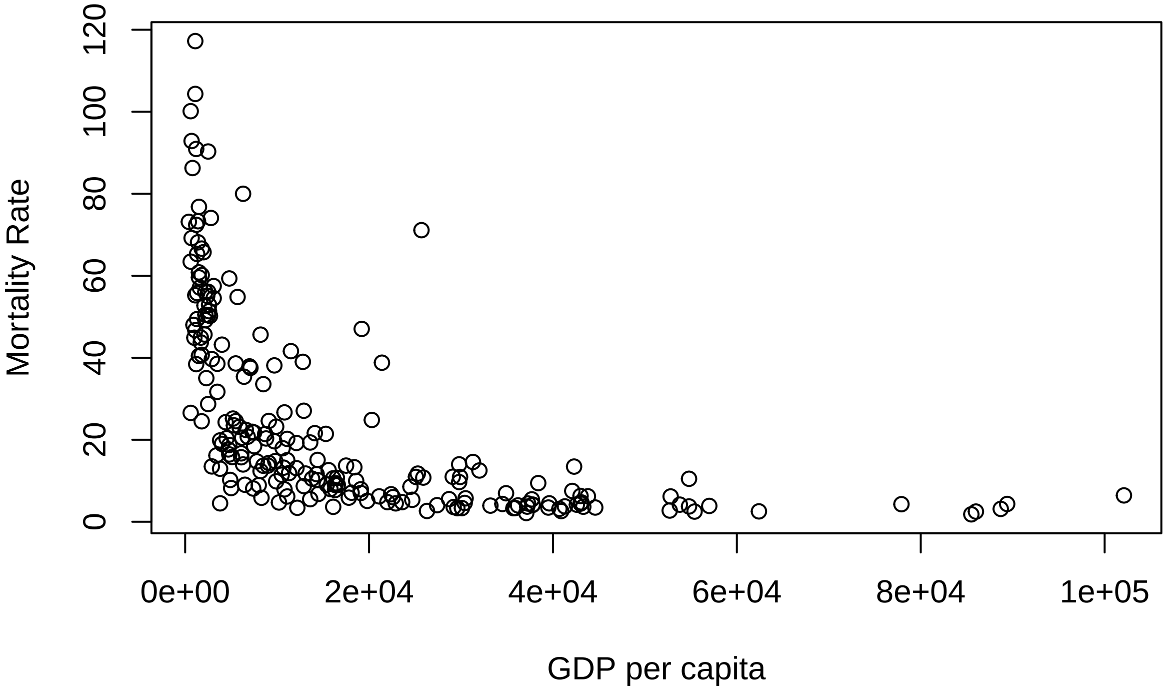


```
##
## Pareto chart analysis for letters
##        Frequency    Cum.Freq.   Percentage Cum.Percent.
##   a  406.0000000  406.0000000   17.3875803   17.3875803
##   n  247.0000000  653.0000000   10.5781585   27.9657388
##   i  213.0000000  866.0000000    9.1220557   37.0877944
##   e  193.0000000 1059.0000000    8.2655246   45.3533191
##   r  163.0000000 1222.0000000    6.9807281   52.3340471
##   o  146.0000000 1368.0000000    6.2526767   58.5867238
##   s  142.0000000 1510.0000000    6.0813704   64.6680942
##   l  134.0000000 1644.0000000    5.7387580   70.4068522
##   t  119.0000000 1763.0000000    5.0963597   75.5032120
##   d  114.0000000 1877.0000000    4.8822270   80.3854390
##   u   93.0000000 1970.0000000    3.9828694   84.3683084
##   c   51.0000000 2021.0000000    2.1841542   86.5524625
##   m   48.0000000 2069.0000000    2.0556745   88.6081370
##   h   46.0000000 2115.0000000    1.9700214   90.5781585
##   g   43.0000000 2158.0000000    1.8415418   92.4197002
##   b   34.0000000 2192.0000000    1.4561028   93.8758030
##   y   32.0000000 2224.0000000    1.3704497   95.2462527
##   p   23.0000000 2247.0000000    0.9850107   96.2312634
##   k   21.0000000 2268.0000000    0.8993576   97.1306210
##   v   18.0000000 2286.0000000    0.7708779   97.9014989
##   w   15.0000000 2301.0000000    0.6423983   98.5438972
##   z   14.0000000 2315.0000000    0.5995717   99.1434690
##   f    9.0000000 2324.0000000    0.3854390   99.5289079
##   q    5.0000000 2329.0000000    0.2141328   99.7430407
##   j    4.0000000 2333.0000000    0.1713062   99.9143469
##   x    2.0000000 2335.0000000    0.0856531  100.0000000
```

In order to create a visual representation of which letters are most common, I created a word cloud in which the frequency of the character determines its relative size in the cloud.



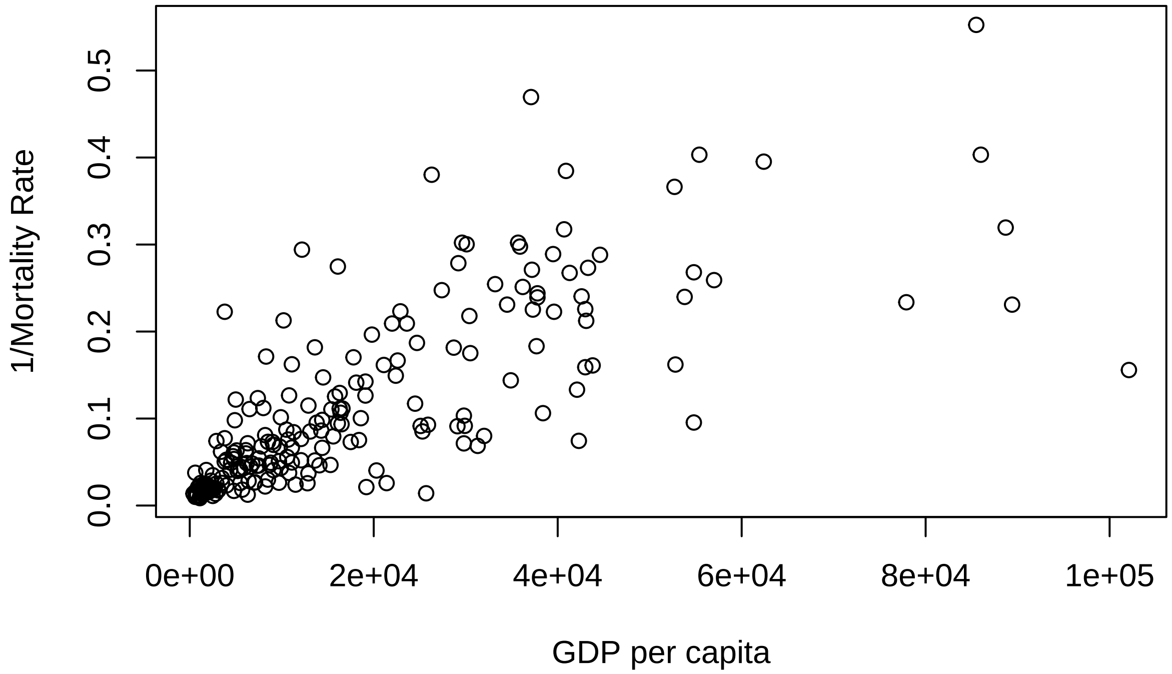
Given that the countries with highest infant mortality rates tended to be some of the poorest countries in the world, I also wanted to look further into the relationship between GDP per capita and mortality rate.



The relationship between GDP per Capita and mortality rate appears to be negatively exponential.

I decided to transform mortality to $\frac{1}{mortality}$ and created a linear model between GDP per capita and $\frac{1}{mortality}$.

$\widehat{\frac{1}{mortality}} = .03397 + .000004222(GDPpercapita)$



Each one 1 dollar increase in GDP per capita is associated with an expected average increase of 4.222e-06 dollars in $\frac{1}{mortality}$. Since the p-value (< 2e-16) is less than .05, we can reject the null hypothesis to conclude that GDP per capita is a statistically significant predictor variable for $\frac{1}{mortality}$. Changes in the GDP per capita are associated with changes in $\frac{1}{mortality}$ at the population level.

Appendix

```r
kmeans <- function(k = 3, DM) {
  changing = 1
  c = matrix()
  n = nrow(DM) #5
  p = ncol(DM) #5

  #standardizing features
  for (i in 1:p) {
    DM[ ,p] <- as.numeric(scale(DM[ ,p]))
  }

  rows <- sample(1:n, k)

  #initialize centroids
  centroids <- DM[rows, ]
  newcentroids <- DM[rows, ]

  #create a length-n vector data structure to hold groupings
  X <- vector(mode="numeric", length=n)

  while (changing) {

    #reclassify all points
    for (i in 1:n) {
      results = c()
      for (j in 1:k) {
        results[j] <- sum ((DM[i,] - centroids[j, ] )^2)
      }
      X[i] <- which.min(results)
    }

    #recalculate all centroids
    for (i in 1:k) {
      if (length(which(X == i)) == 1 ) {
        newcentroids[i, ] <- DM[which(X == i), ]
      }else {
        newcentroids[i, ] <- colMeans(DM[which(X == i), ])
      }
    }

    if (identical(newcentroids, centroids ) ){
      changing = 0
    } else {
      centroids <- newcentroids
    }

  }

  return(X)
}
```

```
regionalMap <- function(k) {

  #Create a map displaying population-sized mortality-colored circles for each country.
  map("world", fill = TRUE, col = "white")
  title(paste("k =", k, ""))


  #Perform k-means classification of the countries using the standardized latitude, longitude, and infa
  countryData$clust <- kmeans(k, DM)

  colors = rainbow(k, alpha = .2)

  for (i in 1:k) {
    sub <- subset(countryData, clust == i)
    x <- sub$longitude
    y <- sub$latitude
    ind <- chull(x, y)
    hull_x <- x[ind]
    hull_y <- y[ind]
    polygon(hull_x, hull_y, col = colors[i])
  }
}
```

Extensions Code:

```
languageNodes <- getNodeSet(root, '//field[@ref="f2098"]')
languages <- lapply(languageNodes, function(x) xmlValue(x[[1]]))

e <- sum(lapply(languages, function(x) grepl("English (official", x, fixed = TRUE)) == TRUE)
f <- sum(lapply(languages, function(x) grepl("French (official", x, fixed = TRUE)) == TRUE)
s <- sum(lapply(languages, function(x) grepl("Spanish (official", x, fixed = TRUE)) == TRUE)
p <- sum(lapply(languages, function(x) grepl("Portuguese (official", x, fixed = TRUE)) == TRUE)
g <- sum(lapply(languages, function(x) grepl("German (official", x, fixed = TRUE)) == TRUE)

data <- data.frame(Language  = c("English" ,  "French", "Spanish", "Portuguese", "German"),
                   Frequency = c(e,f,s,p,g))
```

```
ml <- c("Setswana", "24 major African language groups", "Tigrinya", "English", "Asante", "English", "Se
"English", "English", "English", "English", "Cantonese","Tok Pisin","Filipino","Mandarin","English",
"Maltese","English","English","Punjabi")

length(ml)
```

```
## [1] 57
```

```
sum(ml == "English")
```

```
## [1] 31
```

```
countryletters <- str_flatten(countryCodes$country)
countryletters  <- gsub("[^a-z]", "", countryletters)
```

```r
character_array <- unlist(strsplit(countryletters, ""))

letters <- sort(table(character_array), decreasing = TRUE)
```

```r
#pareto.chart(letters,  xlab = "Letters",  ylab="Frequency",  cumperc = seq(0, 100, by = 10),  ylab2 =



chars <- as.data.frame(letters)
set.seed(100)
#wordcloud(words = chars$character_array, freq = chars$Freq, random.order=TRUE, min.freq = 0)
```

```r
gdpNodes <- getNodeSet(root, '//field[@name="GDP - per capita (PPP)"]/rank')
gdp_rate<- sapply(gdpNodes, function(x) xmlGetAttr(x, "number"))
gdp_country<- sapply(gdpNodes, function(x) xmlGetAttr(x, "country"))

gdp_df <- data.frame(gdpPerCap = as.numeric(gdp_rate), country = gdp_country)

gdp_join_1 <- gdp_df  %>% full_join(mortality_data,
  by = c("country" = "c_codes")
)

gdp_join_1 <- gdp_join_1  %>% drop_na( rates, gdpPerCap)
```

```r
lm1 <- lm((1/gdp_join_1$rates) ~ gdp_join_1$gdpPerCap)
summary(lm1)
```

```
##
## Call:
## lm(formula = (1/gdp_join_1$rates) ~ gdp_join_1$gdpPerCap)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30928 -0.02796 -0.01557  0.02617  0.27887
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.397e-02  6.153e-03   5.521 9.45e-08 ***
## gdp_join_1$gdpPerCap 4.222e-06  2.395e-07  17.629  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06705 on 220 degrees of freedom
## Multiple R-squared:  0.5855, Adjusted R-squared:  0.5836
## F-statistic: 310.8 on 1 and 220 DF,  p-value: < 2.2e-16
```

References

"Least Developed Countries (Ldcs) ." United Nations, United Nations, https://www.un.org/development/desa/dpad/least-developed-country-category.html.

"Westernization." Encyclopedia Britannica, Encyclopedia Britannica, Inc., https://www.britannica.com/topic/Westernization.