CIA Factbook Data

Mari Osmundson

November 2019

I downloaded the factbook.db from GitHub to use to run some analyses of the “facts” table contained within the database file. <https://github.com/factbook/factbook.sql/blob/master/README.md>. The original source is <https://www.cia.gov/library/publications/the-world-factbook/>.

## I ran querys using SQLite for population, population\_growth, birth\_rate, and death\_rate, and plotted the data in histograms.

I opened connection with the database and created a dataframe.

conn <- dbConnect(SQLite(), "./factbook.db")  
query <- "SELECT \* FROM facts LIMIT 5"  
result <- dbGetQuery(conn, query)  
view(result)

Ran a query to select the min and max values of the population and population\_growth columns. The result shows a country with a population of zero, and another country with a population over 7 billion people.

query2 <- "SELECT MIN(population), MIN(population\_growth), MAX(population), MAX(population\_growth) FROM facts"  
result2 <- dbGetQuery(conn, query2)  
view(result2)

I examined the results of query2 more closely, finding that the population of zero belongs to Antarctica. The population of 7 billion plus is the total global population of Earth.

query3 <- "SELECT name, population FROM facts  
WHERE population = 0 OR population = 7256490011"  
result3 <- dbGetQuery(conn, query3)  
view(result3)

Excluded the outliers from query 2-3.

query4 <- "SELECT population, population\_growth, birth\_rate, death\_rate FROM facts  
WHERE (population != MAX(population) AND population != MIN(population))"  
result4 <- dbGetQuery(conn, query4)  
view(result4)

I decided to separate three of the four variables (population\_growth, birth\_rate, death\_rate) into one group, since their values were significantly smaller than population values. Putting them on the same graph as population skewed the results so that the majority of variables were at zero.

rates\_query <- "SELECT population\_growth, birth\_rate, death\_rate FROM facts WHERE population != 0 AND population != 7256490011"  
rates\_result <- dbGetQuery(conn, rates\_query)  
view(rates\_result)

Removed scientific notation from plots.

options(scipen=999)

I gathered population\_growth, birth\_rate, and death\_rate into a tidy dataframe.

result5 = rates\_result %>%  
 gather(key = "variable", value = "val")

Plotted result5.

pop\_histo <- ggplot(data = result5) +  
 aes(x = val) +  
 geom\_histogram(binwidth = 1) +  
 facet\_grid(~ variable)  
print(pop\_histo)

Ran a query to select population.

pop\_query <- "SELECT population FROM facts  
WHERE population != 0 AND population <= 1000000000"  
pop\_result <- dbGetQuery(conn, pop\_query)

Plotted population as a histogram. The majority of values were between 0 and 100,000,000.

pop\_plot <- ggplot(data = pop\_result) +   
 aes(x = population) +  
 geom\_histogram(bin=15)  
print(pop\_plot)

I replotted population, this time restricting x-axis values to get a better picture of results. The majority of countries have populations under 50 million.

pop\_plot2 <- ggplot(data = pop\_result) +   
 aes(x = population) +  
 geom\_histogram(bins=20) +   
 xlim(0, 50000000)  
print(pop\_plot2)