Analyzing CIA Factbook Data

Malik Williams

July 27, 2019

For this project we will be exploring CIA World Factbook data.

It contains a compilation of statistics over the countries in 2015. We will be looking at data from the facts table. Below is a table detailing each variable in the facts table.

About the Data

The CIA Factbook contains a compilation of statistics over the countries. We will be looking at data from the facts table. Below is a table detailing each variable in the facts table.

Column Name	Description
name	The name of the country
area	The total land and sea area of the country
area_land	The country's land area in square kilometers
area_water	The country's water area in square kilometers
population	The country's population
population_growth	The country's population growth as a percentage
birth_rate	The country's birth rate, or the number of births a year per 1,000 people
death_rate	The country's death rate, or the number of death a year per 1,000 people
migration_rate	The country's migration rate, the ratio between immigrants and emigrants throughout the year

Now let's take a look at the first five rows of the facts table.

```
library(RSQLite)
library(DBI)
library(tidyr)
library(ggplot2)
```

```
conn <- dbConnect(SQLite(), "factbook.db")
tables <- dbListTables(conn)
query <- ("SELECT * FROM facts")
result <- dbSendQuery(conn, query)
first_5_facts <- dbFetch(result, n = 5)
first_5_facts
```

```
##   id code        name    area area_land area_water population
## 1  1   af Afghanistan  652230    652230          0   32564342
## 2  2   al     Albania   28748     27398       1350    3029278
## 3  3   ag     Algeria 2381741   2381741          0   39542166
## 4  4   an     Andorra     468       468          0      85580
## 5  5   ao      Angola 1246700   1246700          0   19625353
##   population_growth birth_rate death_rate migration_rate
## 1              2.32      38.57      13.89           1.51
```

```
## 2              0.30      12.92       6.58           3.30
## 3              1.84      23.67       4.31           0.92
## 4              0.12       8.13       6.96           0.00
## 5              2.78      38.78      11.49           0.46
##                   created_at                 updated_at
## 1 2015-11-01 13:19:49.461734 2015-11-01 13:19:49.461734
## 2 2015-11-01 13:19:54.431082 2015-11-01 13:19:54.431082
## 3 2015-11-01 13:19:59.961286 2015-11-01 13:19:59.961286
## 4 2015-11-01 13:20:03.659945 2015-11-01 13:20:03.659945
## 5 2015-11-01 13:20:08.625072 2015-11-01 13:20:08.625072
```

```
dbClearResult(result)
```

Now let's take a look at the population data by finding the extrema (minimum and maximum) values of the `population` and `population_growth` columns.

```
query <- ("SELECT MIN(population), MAX(population), MIN(population_growth), MAX(population_growth) FROM
result <- dbSendQuery(conn, query)
pop_extrema <- dbFetch(result)
pop_extrema
```

```
##   MIN(population) MAX(population) MIN(population_growth)
## 1               0      7256490011                     0
##   MAX(population_growth)
## 1                   4.02
```

```
dbClearResult(result)
```

That doesn't seem right. It says the minimum population of a country is 0, and the maximum population of a country is greater than 7 billion. We know that the entire world's population is roughly 7.2 billion so this can't be right.

Below, we will find which countries gave those strange values.

```
query <- ("SELECT name, population, population_growth FROM facts WHERE population = 0")
result <- dbSendQuery(conn, query)
country_with_0_pop <- dbFetch(result)
country_with_0_pop
```

```
##         name population population_growth
## 1 Antarctica          0                NA
```

```
dbClearResult(result)
```

```
query <- ("SELECT name, population, population_growth FROM facts WHERE population = 7256490011")
result <- dbSendQuery(conn, query)
country_with_billion_pop = dbFetch(result)
country_with_billion_pop
```

```
##    name population population_growth
## 1 World 7256490011              1.08
```

```
dbClearResult(result)
```

The country with a population of 0 is Anarctica, and the observation with 7.2 billion is the World. This explains why there are such extrema maximum and minimum values. No humans permanently live on Antarctica, so its population of 0 makes sense.

Although these two observations are correct, they are also outliers that will skew our results. Therefore, we will remove the outliers before creating visualizations.

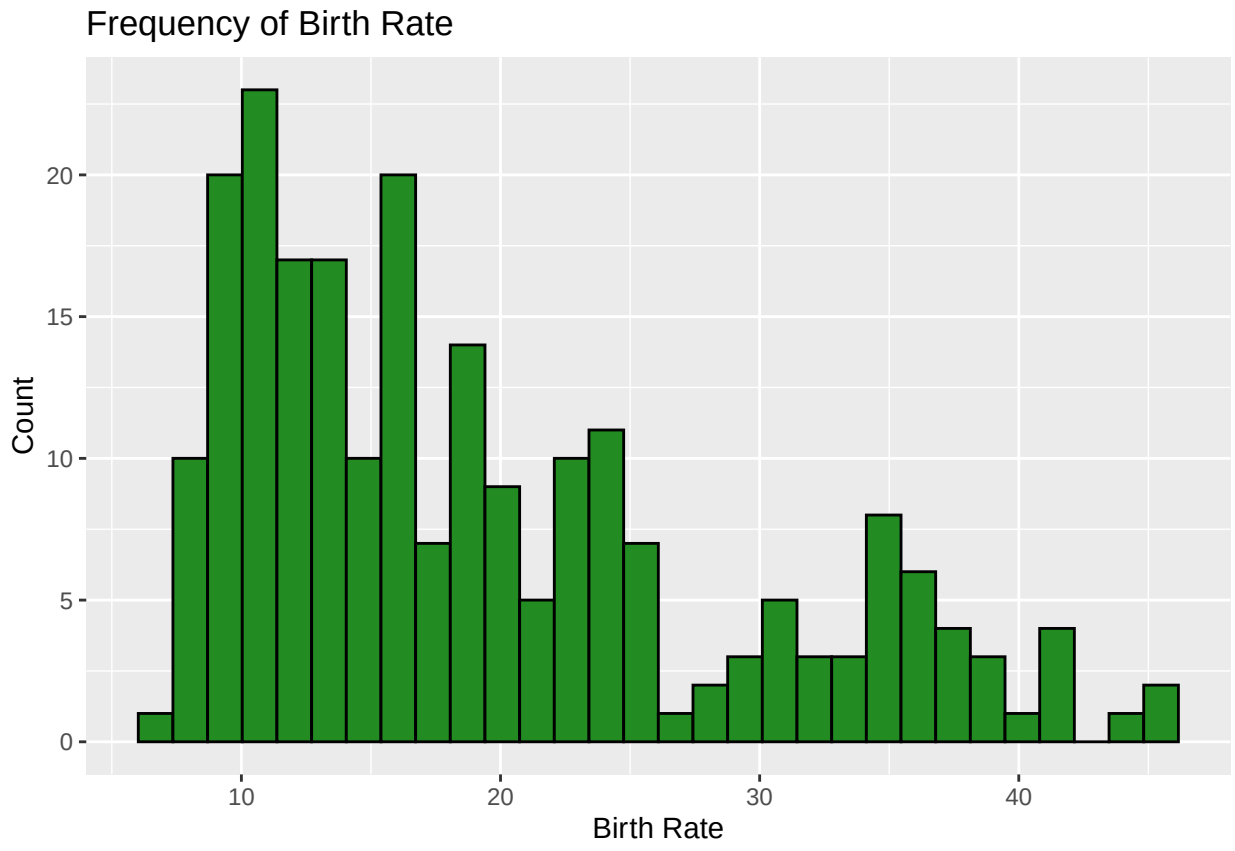
```
query <- "SELECT population, population_growth, birth_rate, death_rate FROM facts WHERE (population !=
no_outliers <- dbGetQuery(conn, query)
```

Creating Visualizations

Now that we have gotten rid of the outliers, let's generate histograms for the `birth_rate`, `death_rate`, `population`, and `population_growth`. This will give us insights on how populations change.

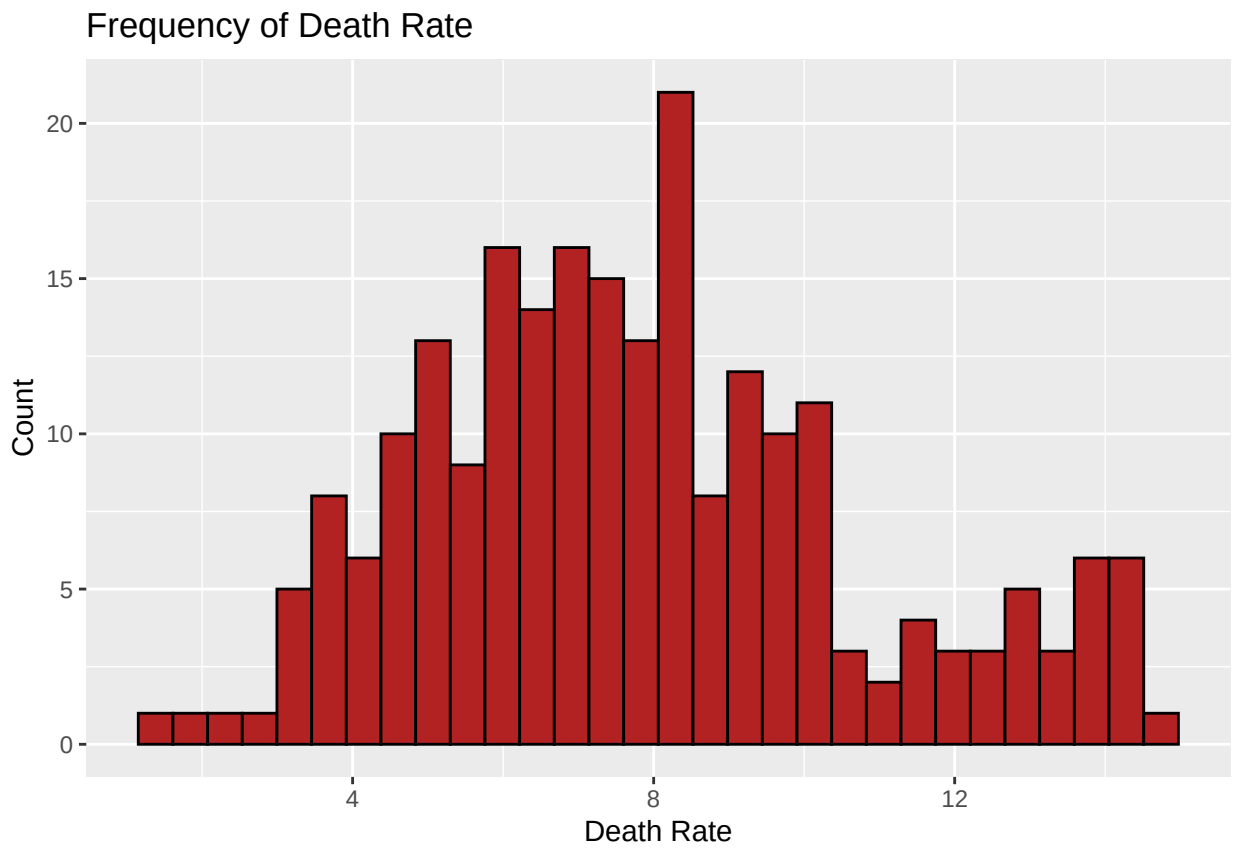
```
# Birth Rate
ggplot(data = no_outliers) +
  aes(x = birth_rate) +
  geom_histogram(bins = 30, color = "black", fill = "forestgreen") +
  labs(x = "Birth Rate", y = "Count", title = "Frequency of Birth Rate")
```

```
## Warning: Removed 13 rows containing non-finite values (stat_bin).
```

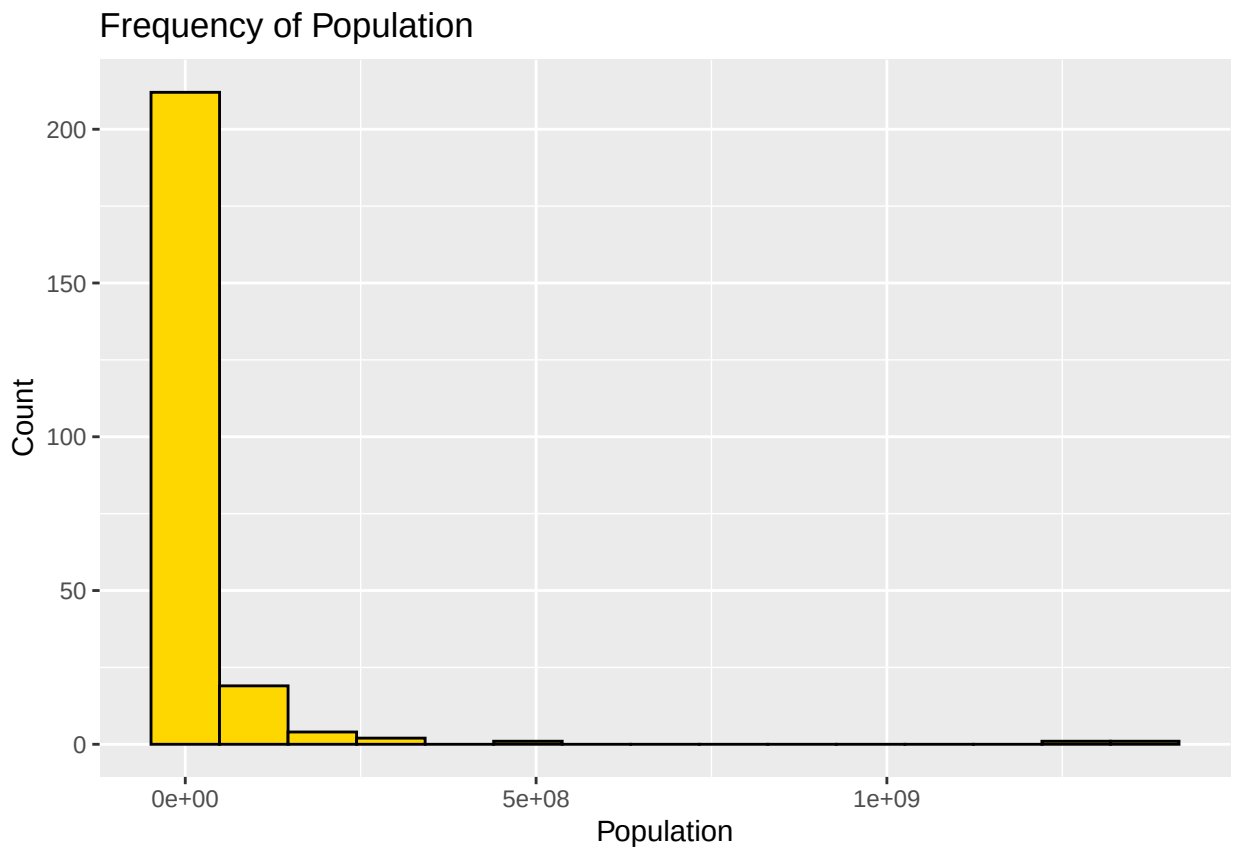



```
# Death Rate
ggplot(data = no_outliers) +
  aes(x = death_rate) +
  geom_histogram(bins = 30, color = "black", fill = "firebrick") +
  labs(x = "Death Rate", y = "Count", title = "Frequency of Death Rate")
```

```
## Warning: Removed 13 rows containing non-finite values (stat_bin).
```

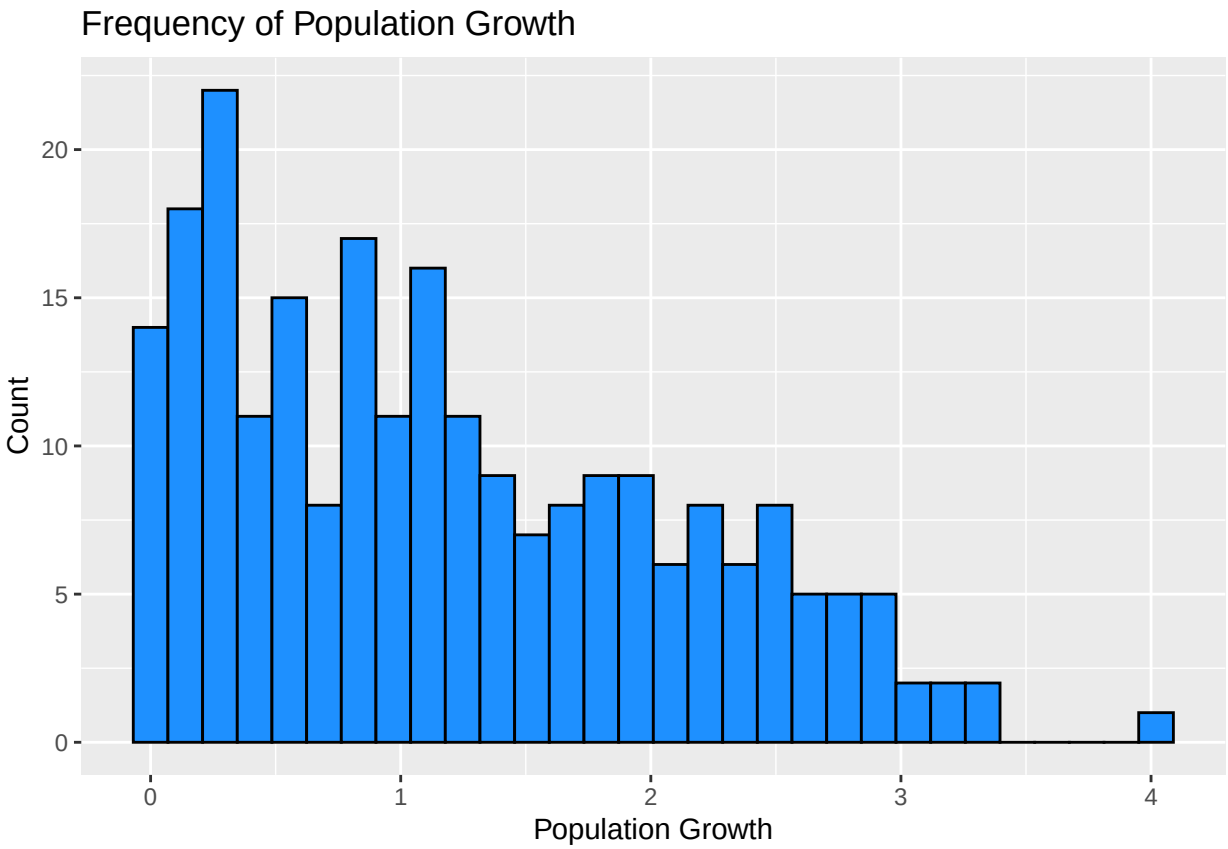



```
# Population
ggplot(data = no_outliers) +
  aes(x = population) +
  geom_histogram(bins = 15, color = "black", fill = "gold") +
  labs(x = "Population", y = "Count", title = "Frequency of Population")
```




```
# Population Growth
ggplot(data = no_outliers) +
  aes(x = population_growth) +
  geom_histogram(bins = 30, color = "black", fill = "dodgerblue") +
  labs(x = "Population Growth", y = "Count", title = "Frequency of Population Growth")
```

```
## Warning: Removed 5 rows containing non-finite values (stat_bin).
```



The birth rate, population, and population growth histograms are all right skewed. This means that their median values are greater than the means. The death rate is close to being normally distributed.

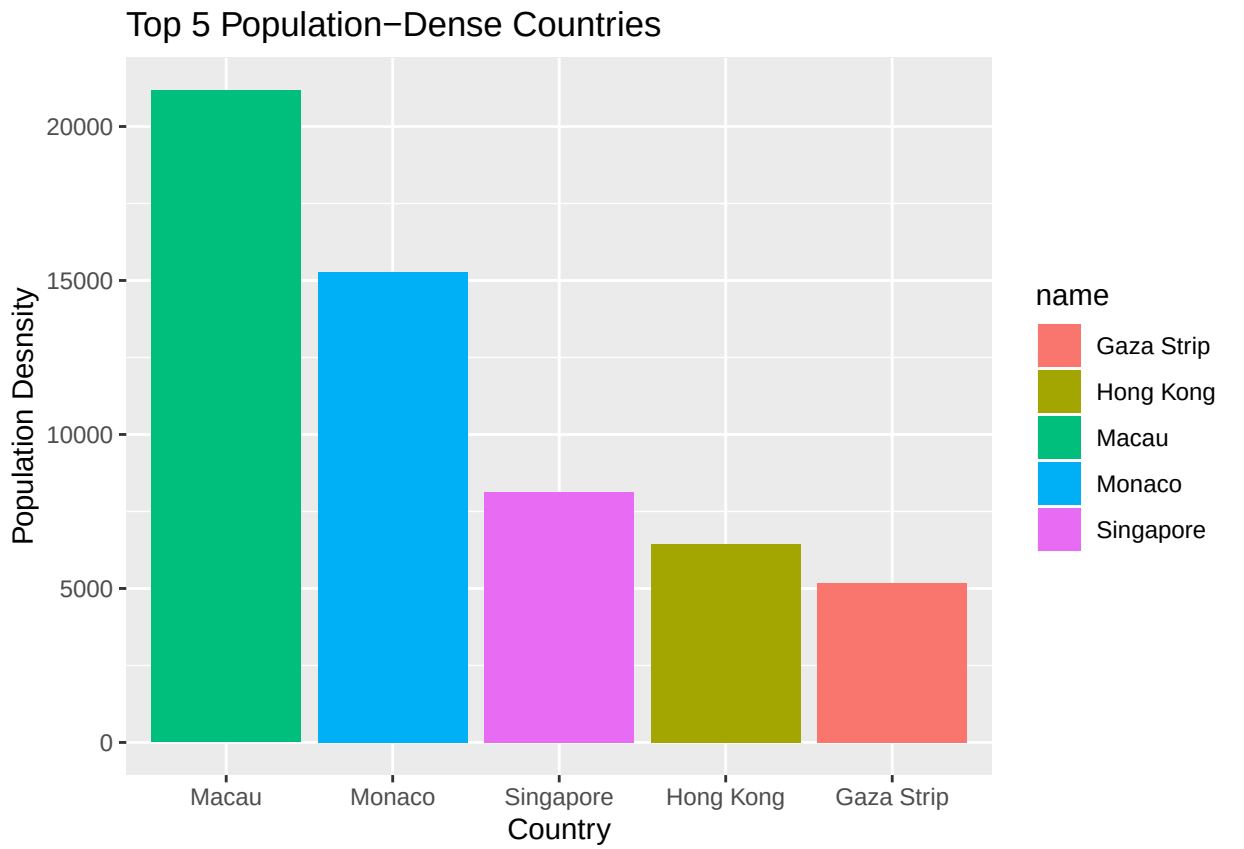
Finding the Countries With the Highest Population Density

Now that we have gotten an idea of the distributions of the population-related variables, we want to find the most population dense countries. The population density is a measurement of population per unit area or unit volume. In other words, it is the `population` divided by the `area`.

```
query <- "SELECT name, population, population_growth, birth_rate, death_rate, (CAST(population as float)
pop_density <- dbGetQuery(conn, query)
top <- head(pop_density, 5)
top
```

```
##         name population population_growth birth_rate death_rate
## 1      Macau     592731              0.80       8.88       4.22
## 2     Monaco      30535              0.12       6.65       9.24
## 3  Singapore    5674472              1.89       8.27       3.43
## 4  Hong Kong    7141106              0.38       9.23       7.07
## 5 Gaza Strip    1869055              2.81      31.11       3.04
##   population_density
## 1          21168.964
## 2          15267.500
## 3           8141.280
## 4           6445.042
## 5           5191.819
```

```
ggplot(data = top) +
  aes(x = reorder(name, -population_density), y = population_density, fill = name) +
  geom_bar(stat = "identity") +
  labs(x = "Country", y = "Population Desnsity", title = "Top 5 Population-Dense Countries")
```



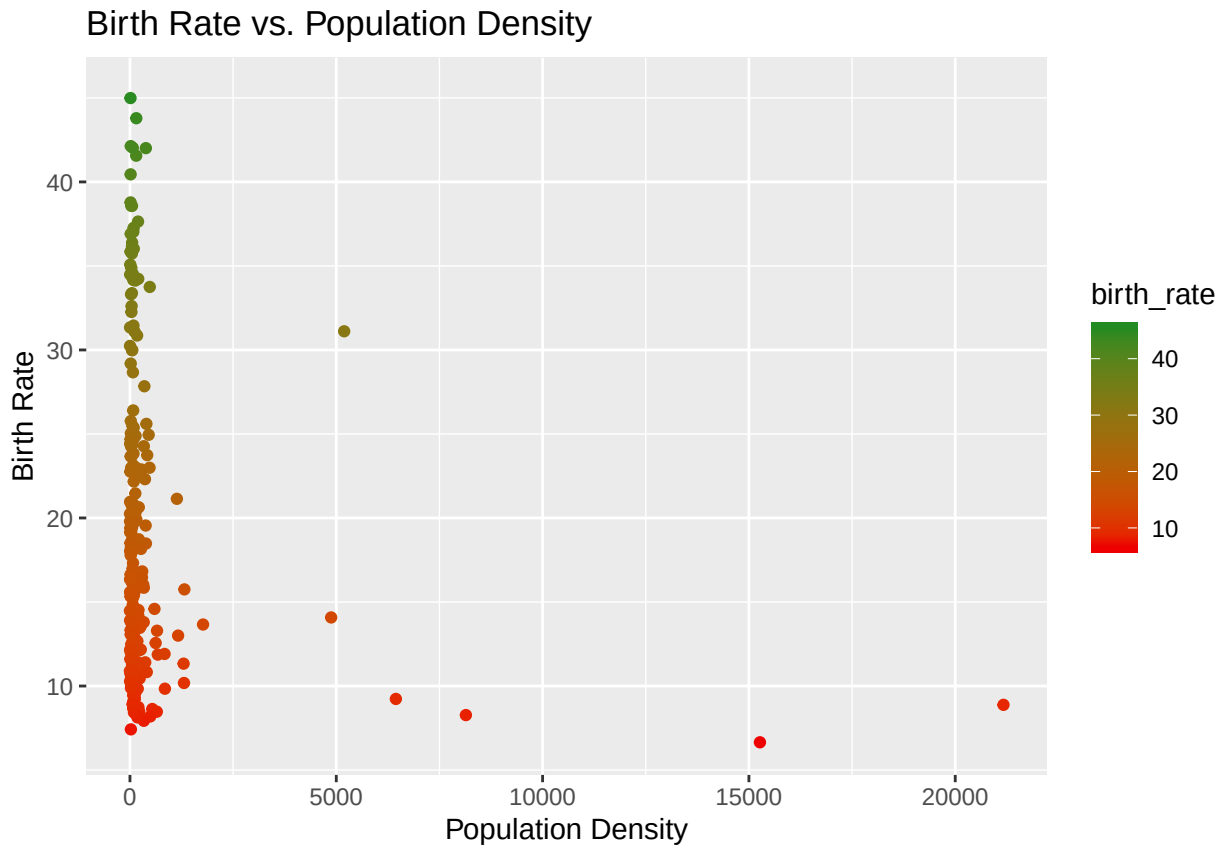
From this data, we see that the most population-dense countries are Macau, Monaco, Singapore, Hong Kong, and Gaza Strip. If we compare our findings with what is listed on Wikipedia, we see that the top 4 countries match, but on Wikipedia, the 5th country is Gibraltar instead of the Gaza Strip.

The Relation Between Population Density and Birth Rate/Death Rate

Now that we have explored population density a bit, let's see if it has any discernable correlations with birth rate or death rate.

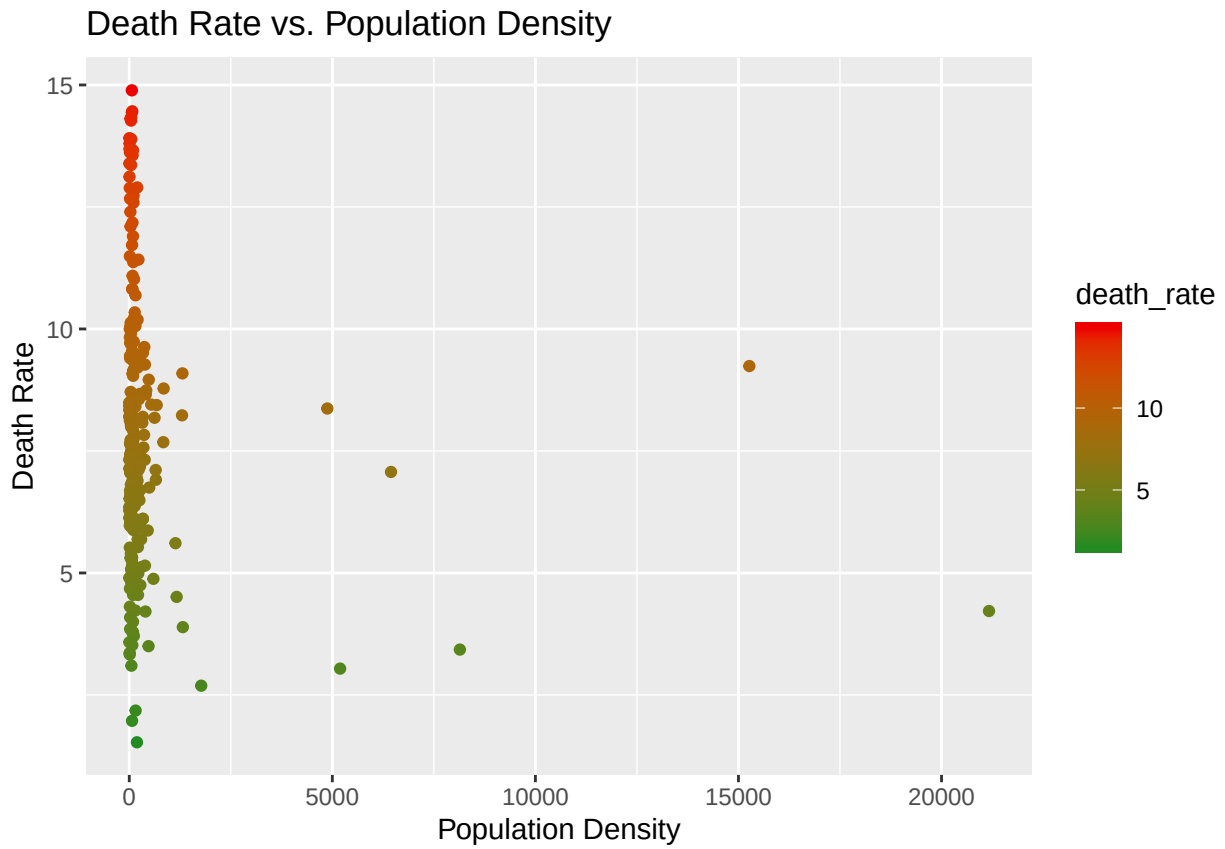
```
ggplot(data = pop_density) +
  aes(x = population_density, y = birth_rate, color = birth_rate) +
  geom_point() +
  scale_color_gradient(low = "red2", high = "forestgreen") +
  labs(x = "Population Density", y = "Birth Rate", title = "Birth Rate vs. Population Density")
```

```
## Warning: Removed 36 rows containing missing values (geom_point).
```




```
ggplot(data = pop_density) +
  aes(x = population_density, y = death_rate, color = death_rate) +
  geom_point() +
  scale_color_gradient(low = "forestgreen", high = "red2") +
  labs(x = "Population Density", y = "Death Rate", title = "Death Rate vs. Population Density")
```

```
## Warning: Removed 36 rows containing missing values (geom_point).
```




```
cor(pop_density$birth_rate, pop_density$population_density, use = "pairwise.complete.obs")
```

```
## [1] -0.1540137
```

```
cor(pop_density$death_rate, pop_density$population_density, use = "pairwise.complete.obs")
```

```
## [1] -0.1149941
```

From these scatterplots and correlation coefficients, there appears to be a slight negative correlation between Birth Rate and Population Density. This means that as population density increases, birth rates start to decrease. This makes sense because in more population dense, people are less likely to want have children. There is an even smaller negative correlation between death rate and population density.

Conclusion

Looking at the CIA Factbook Data seems to indicate that having a large population density correlates with having a smaller birth rate. Death rate seems to be unaffected by population density. This is probably because medicine has advanced to accomodate large populations.