Yuengling DATA 201 Final Project

Olivia Yuengling

2024-12-03

Introduction

The dataset I have selected is the CIA factbook dataset. The data is derived from the CIA 2014 world factbook. “The World Factbook provides basic intelligence on the history, people, government, economy, energy, geography, environment, communications, transportation, military, terrorism, and transnational issues for 265 world entities.” states Andres Renso Caceres Rossi (tidytuesday/data/2024/2024-10-22/readme.md at main · rfordatascience/tidytuesday. (2024). GitHub. https://github.com/rfordatascience/tiddeathytuesday/blob/main/data/2024/2024-10-22/readme.md). The dataset focuses on more population-based information sorted by each country. The dataset would be essential for analyzing basic characteristics of a country including the name, area, internet users, mortality rates, and population. The CIA factbook dataset can be utilized to investigate how the listed variables interact and correlate with one another. The target variable for the prediction algorithms will be the net migration rate, where we will use the other variables in the dataset and observe which ones are the best at predicting the net migration rate. For the classification model, we will attempt to see if we can classify if countries are 3rd world or not. We will understand if the variables in the dataset are proficient at predicting the migration rate and if the classification model to predict if a country is a 3rd world country works proficiently.

Data Cleaning and Processing

```
# loads any necessary libraries
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.3
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```
library(dplyr)
library(ggplot2)
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.3.3
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(ISLR2)
```

```
## Warning: package 'ISLR2' was built under R version 4.3.3
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.3.3
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(gplots)
```

```
## Warning: package 'gplots' was built under R version 4.3.3
```

```
##
## Attaching package: 'gplots'
##
## The following object is masked from 'package:stats':
##
##     lowess
```

```
library(tidyr)
```

```
# loads the dataset

# install.packages("tidytuesdayR")

tuesdata <- tidytuesdayR::tt_load('2024-10-22')
```

```
## ---- Compiling #TidyTuesday Information for 2024-10-22 ----
## --- There is 1 file available ---
##
##
```

```
## ── Downloading files
──────────────────────────────────────────────────────────────
##
##   1 of 1: "cia_factbook.csv"

## OR
tuesdata <- tidytuesdayR::tt_load(2024, week = 43)

## ---- Compiling #TidyTuesday Information for 2024-10-22 ----
## --- There is 1 file available ---
##
##
## ── Downloading files
──────────────────────────────────────────────────────────────
##
##   1 of 1: "cia_factbook.csv"

cia_factbook <- tuesdata$cia_factbook

# displays the rows and columns of the dataset
head(cia_factbook)

## # A tibble: 6 × 11
##   country        area birth_rate death_rate infant_mortality_rate
internet_users
##   <chr>         <dbl>      <dbl>      <dbl>                  <dbl>
<dbl>
## 1 Russia       1.71e7       11.9      13.8                    7.08
40853000
## 2 Canada       9.98e6       10.3       8.31                   4.71
26960000
## 3 United Stat… 9.83e6       13.4       8.15                   6.17
245000000
## 4 China        9.60e6       12.2       7.44                  14.8
389000000
## 5 Brazil       8.51e6       14.7       6.54                  19.2
75982000
## 6 Australia    7.74e6       12.2       7.07                   4.43
15810000
## # ℹ 5 more variables: life_exp_at_birth <dbl>, maternal_mortality_rate
<dbl>,
## #   net_migration_rate <dbl>, population <dbl>, population_growth_rate
<dbl>

str(cia_factbook)

## spc_tbl_ [259 × 11] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ country                : chr [1:259] "Russia" "Canada" "United States"
"China" ...
##  $ area                   : num [1:259] 17098242 9984670 9826675 9596960
8514877 ...
```

```
##  $ birth_rate             : num [1:259] 11.9 10.3 13.4 12.2 14.7 ...
##  $ death_rate             : num [1:259] 13.83 8.31 8.15 7.44 6.54 ...
##  $ infant_mortality_rate  : num [1:259] 7.08 4.71 6.17 14.79 19.21 ...
##  $ internet_users         : num [1:259] 4.09e+07 2.70e+07 2.45e+08
3.89e+08 7.60e+07 ...
##  $ life_exp_at_birth      : num [1:259] 70.2 81.7 79.6 75.2 73.3 ...
##  $ maternal_mortality_rate: num [1:259] 34 12 21 37 56 7 200 77 51 97 ...
##  $ net_migration_rate     : num [1:259] 1.69 5.66 2.45 -0.32 -0.15 5.74 -
0.05 0 0.42 -0.93 ...
##  $ population             : num [1:259] 1.42e+08 3.48e+07 3.19e+08
1.36e+09 2.03e+08 ...
##  $ population_growth_rate : num [1:259] -0.03 0.76 0.77 0.44 0.8 1.09 1.25
0.95 1.17 1.88 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   country = col_character(),
##   ..   area = col_double(),
##   ..   birth_rate = col_double(),
##   ..   death_rate = col_double(),
##   ..   infant_mortality_rate = col_double(),
##   ..   internet_users = col_double(),
##   ..   life_exp_at_birth = col_double(),
##   ..   maternal_mortality_rate = col_double(),
##   ..   net_migration_rate = col_double(),
##   ..   population = col_double(),
##   ..   population_growth_rate = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>

summary(cia_factbook) # prints a 5 number summary of the dataset

##    country               area            birth_rate      death_rate
##  Length:259         Min.   :       0   Min.   : 6.72   Min.   : 1.530
##  Class :character   1st Qu.:     616   1st Qu.:11.84   1st Qu.: 5.930
##  Mode  :character   Median :   51197   Median :16.89   Median : 7.630
##                     Mean   :  530888   Mean   :19.66   Mean   : 7.907
##                     3rd Qu.:  338145   3rd Qu.:24.91   3rd Qu.: 9.450
##                     Max.   :17098242   Max.   :46.12   Max.   :17.490
##                     NA's   :2          NA's   :35      NA's   :34
##  infant_mortality_rate internet_users      life_exp_at_birth
##  Min.   :  1.810       Min.   :      464   Min.   :49.44
##  1st Qu.:  6.185       1st Qu.:    86400   1st Qu.:67.00
##  Median : 13.985       Median :   716400   Median :74.36
##  Mean   : 24.484       Mean   :  8311771   Mean   :71.83
##  3rd Qu.: 38.655       3rd Qu.:  4200000   3rd Qu.:78.29
##  Max.   :117.230       Max.   :389000000   Max.   :89.57
##  NA's   :35            NA's   :46          NA's   :35
##  maternal_mortality_rate net_migration_rate    population
##  Min.   :   2.0          Min.   :-113.5100   Min.   :4.800e+01
##  1st Qu.:  20.0          1st Qu.:  -2.0150   1st Qu.:3.266e+05
```

```
##  Median :  65.5          Median :  -0.0450   Median :5.220e+06
##  Mean   : 178.0          Mean   :  -0.1816   Mean   :3.229e+07
##  3rd Qu.: 240.0          3rd Qu.:   1.2575   3rd Qu.:1.826e+07
##  Max.   :2054.0          Max.   :  83.8200   Max.   :1.356e+09
##  NA's   :75              NA's   :37          NA's   :21
##  population_growth_rate
##  Min.   :-9.730
##  1st Qu.: 0.260
##  Median : 1.020
##  Mean   : 1.101
##  3rd Qu.: 1.920
##  Max.   : 9.370
##  NA's   :26
```

```
dim(cia_factbook) # prints dimensions of dataset
```

```
## [1] 259  11
```

Creating the Classification/Binary Variable

A stated in the beginning of the project, we want to observe if we can classify whether a country is third world or not. Unfortunately, the dataset does not come with a variable for this so we will have to code it ourselves.

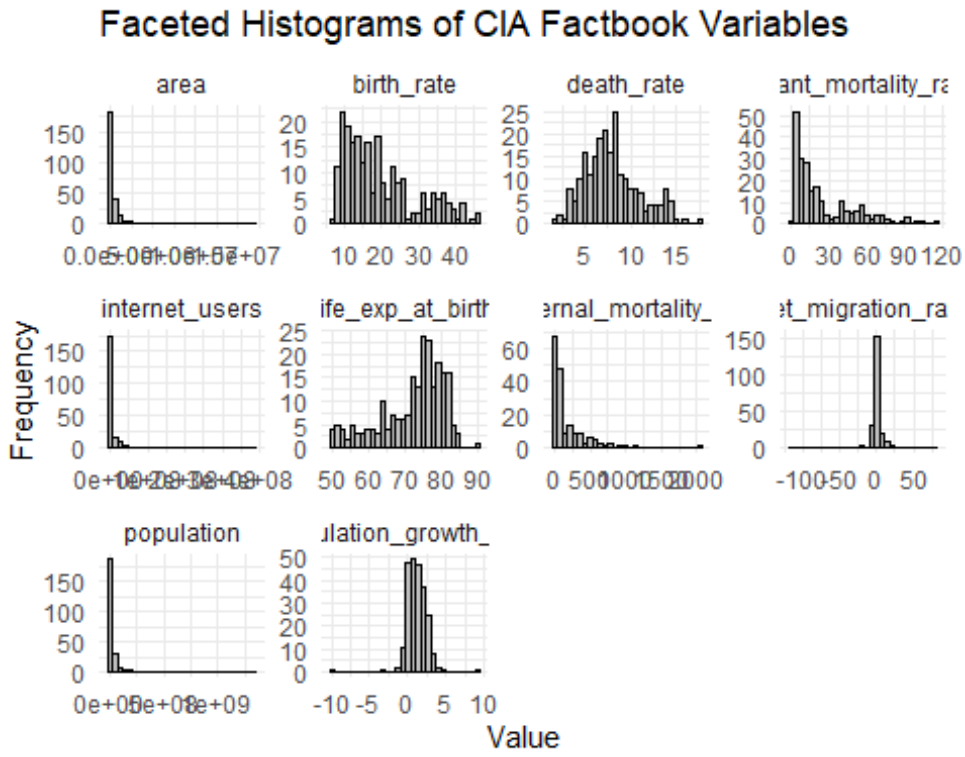
We will create a vector list of all of the countries that have been classified as underdeveloped in 2014 according the the United Nations (The Least Developed Countries Report 2014 | Department of Economic and Social Affairs. (2014). Un.org. https://sdgs.un.org/publications/least-developed-countries-report-2014-17949). After that we will use the mutate function from the r package dplyr to create our target classification variable “third_world”.

```
undeveloped_countries <- c(
  "Afghanistan", "Angola", "Bangladesh", "Benin", "Bhutan", "Burkina Faso",
"Burundi", "Cambodia",
  "Central African Republic", "Chad", "Comoros", "Democratic Republic of the
Congo", "Djibouti",
  "Equatorial Guinea", "Eritrea", "Ethiopia", "The Gambia", "Guinea",
"Guinea-Bissau", "Haiti",
  "Kiribati", "Lao People's Democratic Republic", "Lesotho", "Liberia",
"Madagascar", "Malawi",
  "Mali", "Mauritania", "Mozambique", "Myanmar", "Nepal", "Niger", "Rwanda",
"Sao Tome and Principe",
  "Senegal", "Sierra Leone", "Solomon Islands", "Somalia", "South Sudan",
"Sudan", "Timor-Leste",
  "Togo", "Tuvalu", "Uganda", "United Republic of Tanzania", "Vanuatu",
"Yemen", "Zambia"
) # creates a vector of a list of underdeveloped countries, derived from the
UN

cia_factbook <- cia_factbook %>% # loads mutate command into dataset
```

```
  mutate(third_world = ifelse(country %in% undeveloped_countries, # creates a
new binary variable of 3rd world status
                                      1,  # undeveloped country
                                      0)) # developed country
```

General EDA with Dataset

```
# converts the data to long format for the facet
cia_factbook_long <- cia_factbook %>%
  pivot_longer(cols = c(area, birth_rate, death_rate, net_migration_rate,
                        infant_mortality_rate, internet_users,
life_exp_at_birth,
                        maternal_mortality_rate, population_growth_rate,
population),
               names_to = "variable", values_to = "value")

# creates faceted histogram plot
ggplot(cia_factbook_long, aes(x = value)) +
  geom_histogram(bins = 30, fill = "grey", color = "black") +
  facet_wrap(~variable, scales = "free") + # facet by variable, allow
different scales
  theme_minimal() +
  labs(title = "Faceted Histograms of CIA Factbook Variables",
       x = "Value",
       y = "Frequency")

## Warning: Removed 346 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



Handling Dataset NA's

```
# removes rows with more than 4 NA's
row_na_count <- rowSums(is.na(cia_factbook)) # counts the number of NA's in
each row
limit <- 4 # sets the threshold
cia_factbook <- cia_factbook[row_na_count <= limit, ] # removes rows with
more than 4 NA's
```

For the remaining NA's in the data we will use mean or median imputation depending if the data is skewed or normally distributed based on the shape from the histograms.The histograms with a normal distribution are net migration rate and population growth rate, so we will use mean imputation because the mean is not as influenced by outliers compared to the other values in the dataset.

```
cia_factbook$net_migration_rate[is.na(cia_factbook$net_migration_rate)] <-
  mean(cia_factbook$net_migration_rate, na.rm = TRUE)

cia_factbook$population_growth_rate[is.na(cia_factbook$population_growth_rate
)] <- mean(cia_factbook$net_migration_rate, na.rm = TRUE)

summary(cia_factbook) # prints summary to confirm there are no NA's in
normally distributed variables
```

```
##    country                area            birth_rate      death_rate
##  Length:224         Min.   :       2   Min.   : 6.72   Min.   : 1.530
```

```
##  Class :character   1st Qu.:    5836   1st Qu.:11.84   1st Qu.: 5.930
##  Mode  :character   Median :   87971   Median :16.89   Median : 7.540
##                     Mean   :  608449   Mean   :19.66   Mean   : 7.907
##                     3rd Qu.:  448124   3rd Qu.:24.91   3rd Qu.: 9.457
##                     Max.   :17098242   Max.   :46.12   Max.   :17.490
##
##  infant_mortality_rate internet_users      life_exp_at_birth
##  Min.   :  1.81        Min.   :      900   Min.   :49.44
##  1st Qu.:  6.20        1st Qu.:    95000   1st Qu.:66.90
##  Median : 14.00        Median :   746000   Median :74.29
##  Mean   : 24.57        Mean   :  8470823   Mean   :71.76
##  3rd Qu.: 38.70        3rd Qu.:  4393000   3rd Qu.:78.28
##  Max.   :117.23        Max.   :389000000   Max.   :89.57
##  NA's   :1             NA's   :15          NA's   :2
##  maternal_mortality_rate net_migration_rate    population
##  Min.   :   2.0          Min.   :-113.5100   Min.   :5.215e+03
##  1st Qu.:  20.0          1st Qu.:  -2.0050   1st Qu.:5.843e+05
##  Median :  65.5          Median :  -0.0700   Median :5.617e+06
##  Mean   : 178.0          Mean   :  -0.1881   Mean   :3.202e+07
##  3rd Qu.: 240.0          3rd Qu.:   1.2200   3rd Qu.:2.176e+07
##  Max.   :2054.0          Max.   :  83.8200   Max.   :1.356e+09
##  NA's   :40
##  population_growth_rate  third_world
##  Min.   :-9.730         Min.   :0.000
##  1st Qu.: 0.330         1st Qu.:0.000
##  Median : 1.075         Median :0.000
##  Mean   : 1.140         Mean   :0.192
##  3rd Qu.: 1.923         3rd Qu.:0.000
##  Max.   : 9.370         Max.   :1.000
##
```

Looking at the summaries now, we can now see that there are no NA's for the normally distributed variables in the dataset. Now, let's do the remainder of the numerical variables but with their respective median value.

```r
# median imputation

cia_factbook$net_migration_rate[is.na(cia_factbook$infant_mortality_rate)] <-
  median(cia_factbook$infant_mortality_rate, na.rm = TRUE)

cia_factbook$infant_mortality_rate[is.na(cia_factbook$infant_mortality_rate)]
<-
  median(cia_factbook$infant_mortality_rate, na.rm = TRUE)

cia_factbook$internet_users[is.na(cia_factbook$internet_users)] <-
  median(cia_factbook$internet_users, na.rm = TRUE)

cia_factbook$life_exp_at_birth[is.na(cia_factbook$life_exp_at_birth)] <-
  median(cia_factbook$life_exp_at_birth, na.rm = TRUE)
```

```r
cia_factbook$maternal_mortality_rate[is.na(cia_factbook$maternal_mortality_ra
te)] <-
  median(cia_factbook$maternal_mortality_rate, na.rm = TRUE)

cia_factbook_no_country <- cia_factbook %>%
  select(-country)

summary(cia_factbook) # prints summary to confirm there are no NA's
```

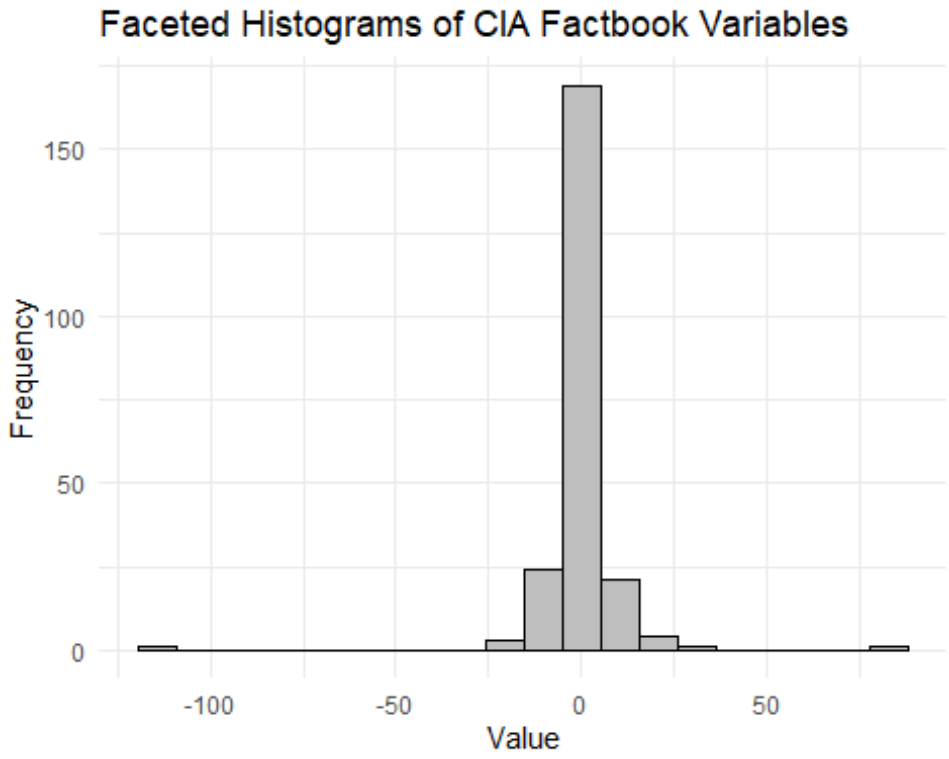
```
##    country               area            birth_rate      death_rate
##  Length:224         Min.   :       2   Min.   : 6.72   Min.   : 1.530
##  Class :character   1st Qu.:    5836   1st Qu.:11.84   1st Qu.: 5.930
##  Mode  :character   Median :   87971   Median :16.89   Median : 7.540
##                     Mean   :  608449   Mean   :19.66   Mean   : 7.907
##                     3rd Qu.:  448124   3rd Qu.:24.91   3rd Qu.: 9.457
##                     Max.   :17098242   Max.   :46.12   Max.   :17.490
##  infant_mortality_rate internet_users      life_exp_at_birth
##  Min.   :  1.810       Min.   :      900   Min.   :49.44
##  1st Qu.:  6.205       1st Qu.:   113150   1st Qu.:67.00
##  Median : 14.000       Median :   746000   Median :74.29
##  Mean   : 24.528       Mean   :  7953536   Mean   :71.78
##  3rd Qu.: 38.655       3rd Qu.:  4012750   3rd Qu.:78.25
##  Max.   :117.230       Max.   :389000000   Max.   :89.57
##  maternal_mortality_rate net_migration_rate    population
##  Min.   :   2.00         Min.   :-113.5100   Min.   :5.215e+03
##  1st Qu.:  26.75         1st Qu.:  -2.0050   1st Qu.:5.843e+05
##  Median :  65.50         Median :  -0.0550   Median :5.617e+06
##  Mean   : 157.89         Mean   :  -0.1248   Mean   :3.202e+07
##  3rd Qu.: 200.00         3rd Qu.:   1.2500   3rd Qu.:2.176e+07
##  Max.   :2054.00         Max.   :  83.8200   Max.   :1.356e+09
##  population_growth_rate  third_world
##  Min.   :-9.730         Min.   :0.000
##  1st Qu.: 0.330         1st Qu.:0.000
##  Median : 1.075         Median :0.000
##  Mean   : 1.140         Mean   :0.192
##  3rd Qu.: 1.923         3rd Qu.:0.000
##  Max.   : 9.370         Max.   :1.000
```

Now, there are no NA's in the dataset. Let's now analyze our target variable, net migration, in more detail.

EDA with Regression Target Variable

```r
ggplot(cia_factbook, aes(x = net_migration_rate)) +
  geom_histogram(bins = 20, fill = "grey", color = "black") + # sets bins and
color
  theme_minimal() + # sets theme
  labs(title = "Faceted Histograms of CIA Factbook Variables", # sets label
```

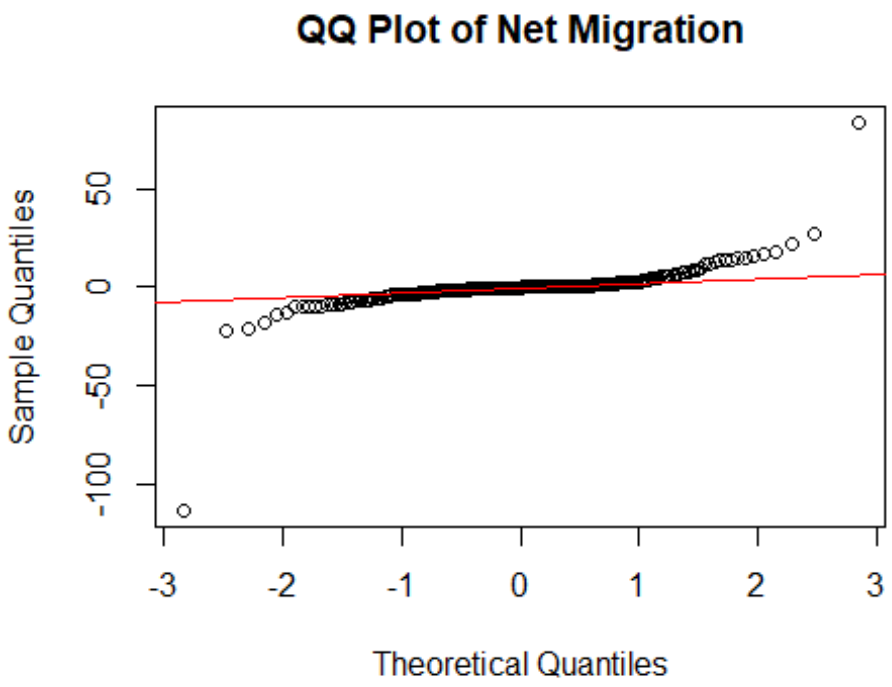
```
       x = "Value",
       y = "Frequency")
```



As we can see there are two distinct outliers on the outside of the distribution. One of these outliers goes beyond the -100 threshold, and the variable is measured as a percentage. This is from Syria and the value is -113. This is obviously impossible because this would mean 113 percent of the country's population is leaving the country, and there can only be 100 percent of the country leaving. This is an error that may be changed depending on context. The other outlier, Lebanon, has a migration rate of 83.3 percent. We should take context into account before removing the outlier.

Let's do a brief history lesson. In 2011, there has been a conflict in Syria (Syrian Civil War) and a large number of citizens have been fleeing to Lebanon (The number of Syrian refugees in Lebanon passes the 1 million mark | UNHCR US. (2014). UNHCR US. https://www.unhcr.org/us/news/stories/number-syrian-refugees-lebanon-passes-1-million-mark). The year the data has been taken was 2014, which is only three years after the conflict has begun. Therefore, these outliers have significant insights within the dataset so we shall keep them.

```
qqnorm(cia_factbook$net_migration_rate, main = "QQ Plot of Net Migration") # creates qqlpot
qqline(cia_factbook$net_migration_rate, col = "red") # creates regression line
```



The histogram and QQ plot of net migration rates clearly show two major outliers: Syria (-113) and Lebanon (83.3). These values tie back to the Syrian Civil War, which started in 2011 and caused massive emigration from Syria while Lebanon experienced a huge influx of refugees. The -113 for Syria is technically impossible as a percentage and probably a reporting error, but both outliers are critical for understanding the dataset because they reflect the impact of this crisis. The QQ plot makes it even clearer that the data isn't normally distributed thanks to these extreme values. Since these outliers provide meaningful context, I'm still keeping them in the analysis.

Now that we've explored our target variable, net migration, let's explore the target variable (third_world) for classification.

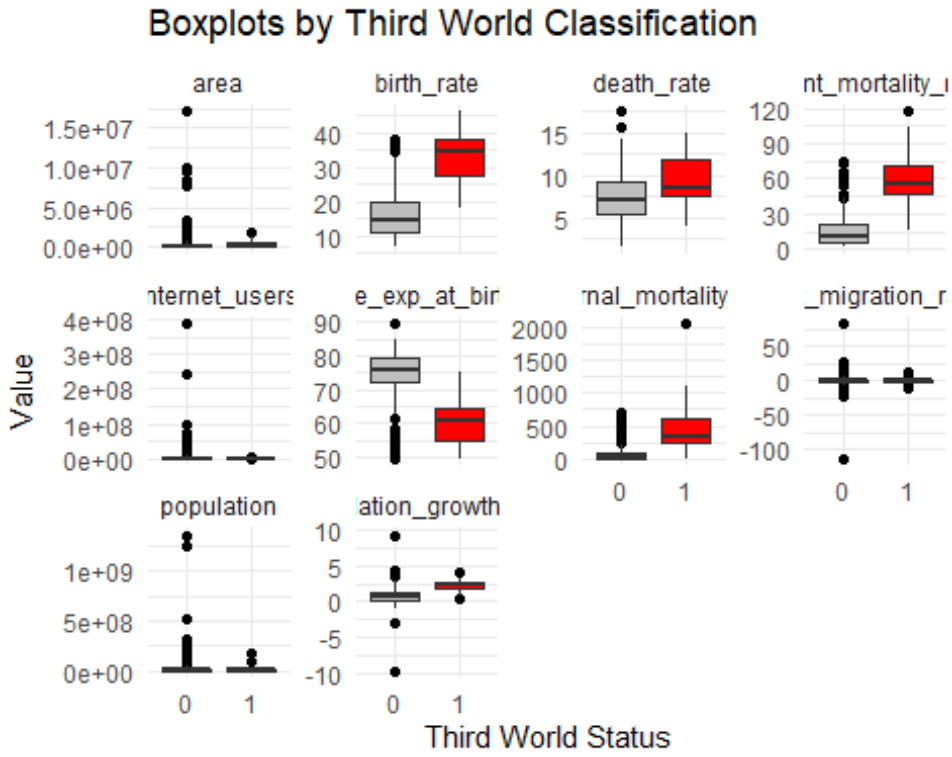
EDA with Classification target variable

Just a refresher– the target for classification is whether a country is a third world country or not, so let's explore using a bar plot to visualize the distribution of the value for each country.

```
ggplot(cia_factbook_long, aes(x = factor(third_world), y = value, fill =
factor(third_world))) +
  geom_boxplot(outlier.color = "black", outlier.shape = 16) +
  facet_wrap(~variable, scales = "free_y") +  # Facet by variable
  theme_minimal() +
  labs(
```

```
    title = "Boxplots by Third World Classification",
    x = "Third World Status",
    y = "Value"
  ) +
  scale_fill_manual(values = c("grey", "red"), labels = c("Developed", "Third
World")) +
  theme(legend.position = "none")
```

```
## Warning: Removed 346 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



The boxplot shows big differences between developed and Third World countries. Third World countries have higher birth and death rates, lower life expectancy, and more infant and maternal deaths. They also have faster population growth, while developed countries grow slowly or not at all. Developed countries have way more internet users, showing a big gap in access to technology. Migration rates are more mixed in developed countries, while Third World countries stay closer to zero. Overall, the chart shows the big gaps in health, resources, and living conditions between the two groups.

Let's also look at a frequency table of the target variable:

```
table(cia_factbook$third_world)
```

```
##
##   0   1
## 181  43
```

There are 181 countries that are not third world, and 43 countries classify as third world. That is about 19 percent of the observations in the dataset.

Regression Modeling

Let's start to create our regression model for net migration rate. We will use ggpairs and a summary of our full model to understand if there is a correlation between net migration and an variables in the dataset.

Correlation Matrix

```
cormat <- round(cor(cia_factbook_no_country),2)
head(cormat)
```

```
##                          area birth_rate death_rate infant_mortality_rate
## area                     1.00      -0.04       0.09                 -0.01
## birth_rate              -0.04       1.00       0.15                  0.87
## death_rate               0.09       0.15       1.00                  0.36
## infant_mortality_rate -0.01       0.87       0.36                  1.00
## internet_users          0.57      -0.15       0.01                 -0.12
## life_exp_at_birth      -0.01      -0.83      -0.48                 -0.88
##                         internet_users life_exp_at_birth
maternal_mortality_rate
## area                              0.57             -0.01                 -
0.01
## birth_rate                       -0.15             -0.83
0.73
## death_rate                        0.01             -0.48
0.35
## infant_mortality_rate            -0.12             -0.88
0.78
## internet_users                    1.00              0.13                 -
0.11
## life_exp_at_birth                 0.13              1.00                 -
0.67
##                       net_migration_rate population population_growth_rate
## area                                0.03       0.46                 -0.02
## birth_rate                         -0.13      -0.02                  0.56
## death_rate                         -0.07       0.00                 -0.17
## infant_mortality_rate              -0.10       0.04                  0.45
## internet_users                      0.01       0.76                 -0.09
## life_exp_at_birth                   0.15      -0.02                 -0.35
##                       third_world
## area                        -0.06
## birth_rate                   0.70
## death_rate                   0.25
## infant_mortality_rate        0.72
## internet_users              -0.11
## life_exp_at_birth           -0.61
```

```r
library(reshape2)
```

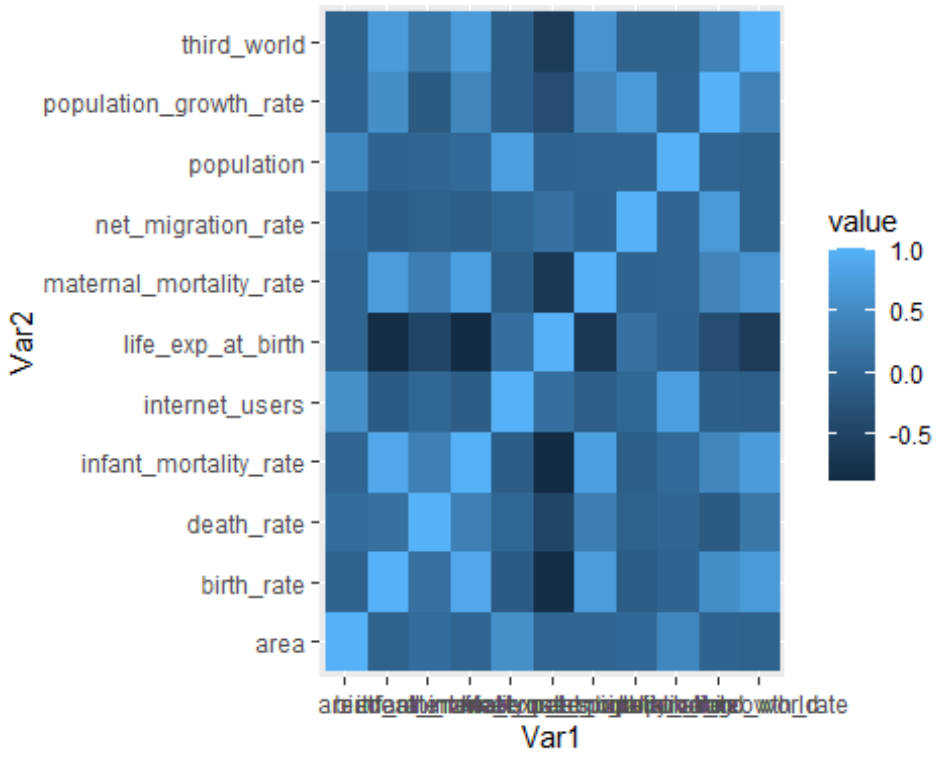
```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```r
melted_cormat <- melt(cormat)
head(melted_cormat)
```

```
##                    Var1 Var2 value
## 1                  area area  1.00
## 2            birth_rate area -0.04
## 3            death_rate area  0.09
## 4 infant_mortality_rate area -0.01
## 5        internet_users area  0.57
## 6     life_exp_at_birth area -0.01
```

```r
library(ggplot2)
ggplot(data = melted_cormat, aes(x=Var1, y=Var2, fill=value)) +
  geom_raster()
```



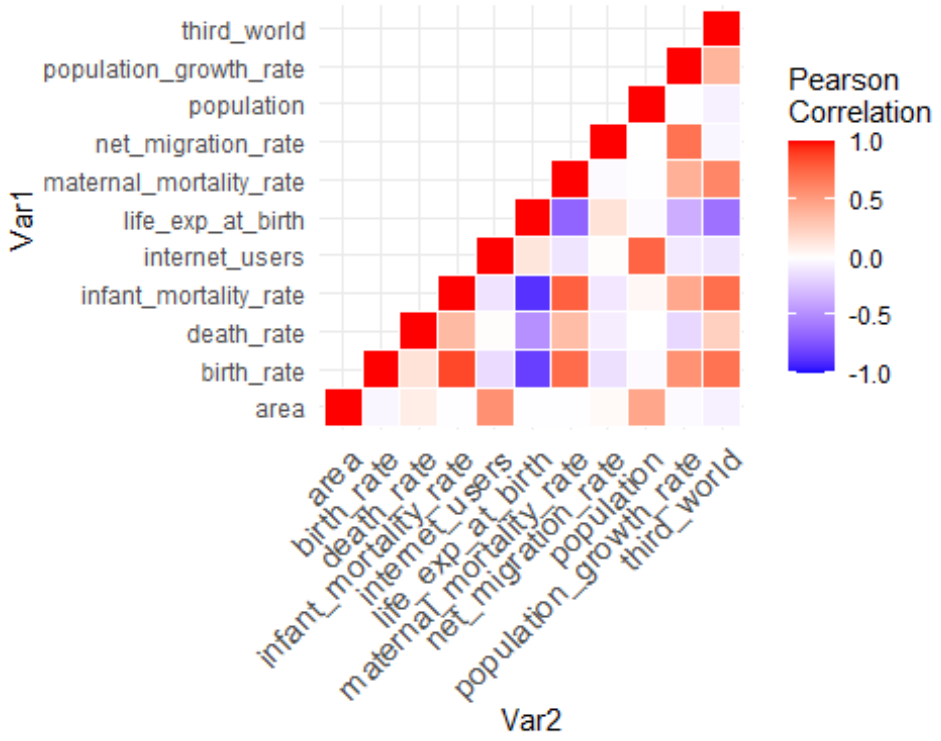

```r
# Get lower triangle of the correlation matrix
  get_lower_tri<-function(cormat){
    cormat[upper.tri(cormat)] <- NA
    return(cormat)
```

```
  }
  # Get upper triangle of the correlation matrix
  get_upper_tri <- function(cormat){
    cormat[lower.tri(cormat)]<- NA
    return(cormat)
  }

  upper_tri <- get_upper_tri(cormat)
upper_tri
```

```
##                          area birth_rate death_rate infant_mortality_rate
## area                        1      -0.04       0.09                 -0.01
## birth_rate                 NA       1.00       0.15                  0.87
## death_rate                 NA         NA       1.00                  0.36
## infant_mortality_rate      NA         NA         NA                  1.00
## internet_users             NA         NA         NA                    NA
## life_exp_at_birth          NA         NA         NA                    NA
## maternal_mortality_rate    NA         NA         NA                    NA
## net_migration_rate         NA         NA         NA                    NA
## population                 NA         NA         NA                    NA
## population_growth_rate     NA         NA         NA                    NA
## third_world                NA         NA         NA                    NA
##                          internet_users life_exp_at_birth
## area                               0.57             -0.01
## birth_rate                        -0.15             -0.83
## death_rate                         0.01             -0.48
## infant_mortality_rate             -0.12             -0.88
## internet_users                     1.00              0.13
## life_exp_at_birth                    NA              1.00
## maternal_mortality_rate              NA                NA
## net_migration_rate                   NA                NA
## population                           NA                NA
## population_growth_rate               NA                NA
## third_world                          NA                NA
##                          maternal_mortality_rate net_migration_rate
population
## area                                       -0.01               0.03
0.46
## birth_rate                                  0.73              -0.13      -
0.02
## death_rate                                  0.35              -0.07
0.00
## infant_mortality_rate                       0.78              -0.10
0.04
## internet_users                             -0.11               0.01
0.76
## life_exp_at_birth                          -0.67               0.15      -
0.02
## maternal_mortality_rate                     1.00              -0.02      -
0.01
```

```
## net_migration_rate                              NA                1.00
0.00
## population                                      NA                  NA
1.00
## population_growth_rate                          NA                  NA
NA
## third_world                                     NA                  NA
NA
##                          population_growth_rate third_world
## area                                      -0.02       -0.06
## birth_rate                                 0.56        0.70
## death_rate                                -0.17        0.25
## infant_mortality_rate                      0.45        0.72
## internet_users                            -0.09       -0.11
## life_exp_at_birth                         -0.35       -0.61
## maternal_mortality_rate                    0.41        0.61
## net_migration_rate                         0.70       -0.04
## population                                -0.01       -0.06
## population_growth_rate                     1.00        0.39
## third_world                                  NA        1.00
```

```r
# Melt the correlation matrix
library(reshape2)
melted_cormat <- melt(upper_tri, na.rm = TRUE)
# Heatmap
library(ggplot2)
ggplot(data = melted_cormat, aes(Var2, Var1, fill = value))+
 geom_tile(color = "white")+
 scale_fill_gradient2(low = "blue", high = "red", mid = "white",
   midpoint = 0, limit = c(-1,1), space = "Lab",
   name="Pearson\nCorrelation") +
  theme_minimal()+
 theme(axis.text.x = element_text(angle = 45, vjust = 1,
    size = 12, hjust = 1))+
 coord_fixed()
```



We can observe there may be correlations with life_exp_at_birth and population_growth_rate. There may be a potential association between all of these variables. What is particularly interesting is that life expectancy at birth may have the strongest correlation out of all of the variables in the dataset as it has a single star next to it, which will be important when creating our regression model.

Now, let's move onto creating our full model to observe any statistically significant variables in our dataset.

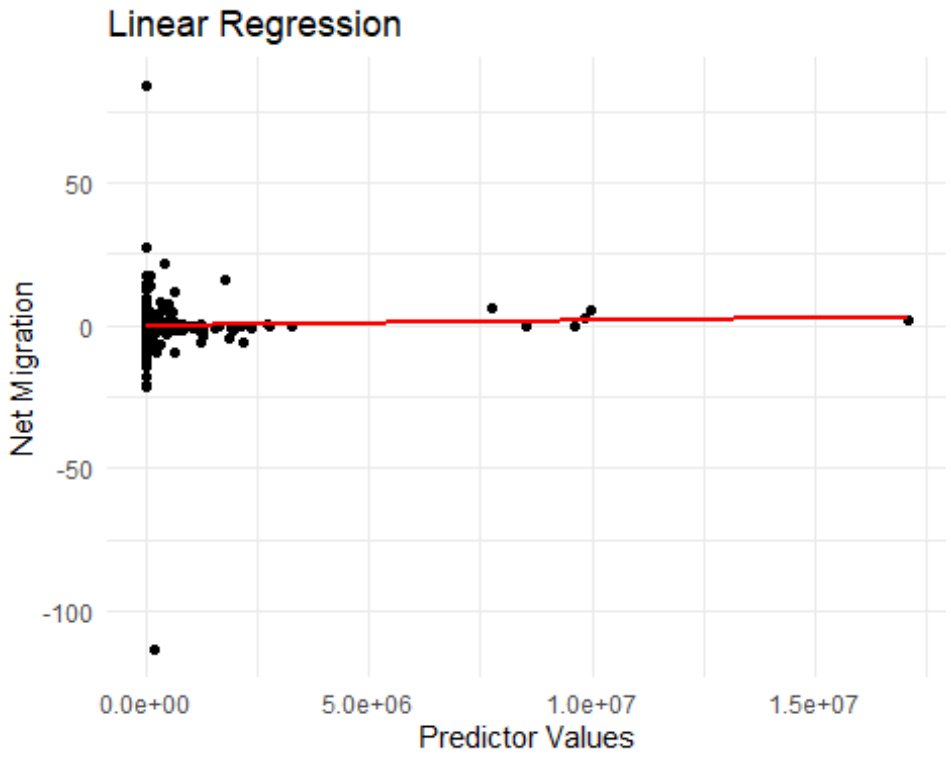
Creating the Full Regression model

```
fullmodel <- lm(net_migration_rate ~ ., data = cia_factbook_no_country) #
creates full model
summary(fullmodel) # prints summary of the model

##
## Call:
## lm(formula = net_migration_rate ~ ., data = cia_factbook_no_country)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.322 -0.471 -0.222  0.028 34.272
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.421e+00  5.537e+00   0.257    0.798
## area                    -1.221e-08  1.288e-07  -0.095    0.925
```

```
## birth_rate               -9.966e-01  5.490e-02 -18.154   <2e-16 ***
## death_rate                9.303e-01  9.445e-02   9.849   <2e-16 ***
## infant_mortality_rate     1.615e-03  2.311e-02   0.070    0.944
## internet_users           -2.219e-09  1.003e-08  -0.221    0.825
## life_exp_at_birth        -3.200e-03  5.989e-02  -0.053    0.957
## maternal_mortality_rate   6.532e-04  1.335e-03   0.489    0.625
## population                2.611e-11  2.382e-09   0.011    0.991
## population_growth_rate    9.437e+00  1.692e-01  55.763   <2e-16 ***
## third_world               2.136e-01  7.255e-01   0.294    0.769
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.817 on 213 degrees of freedom
## Multiple R-squared:  0.9389, Adjusted R-squared:  0.936
## F-statistic:   327 on 10 and 213 DF,  p-value: < 2.2e-16
```

Looking at the statistics here, the full model is already excellent at predicting the values of the data judging by the high r-squared and extremely low p-value. Most of the variables do not pass the significance test, and according to the correlation matrix they don't necessarily correlate with the net migration rate.

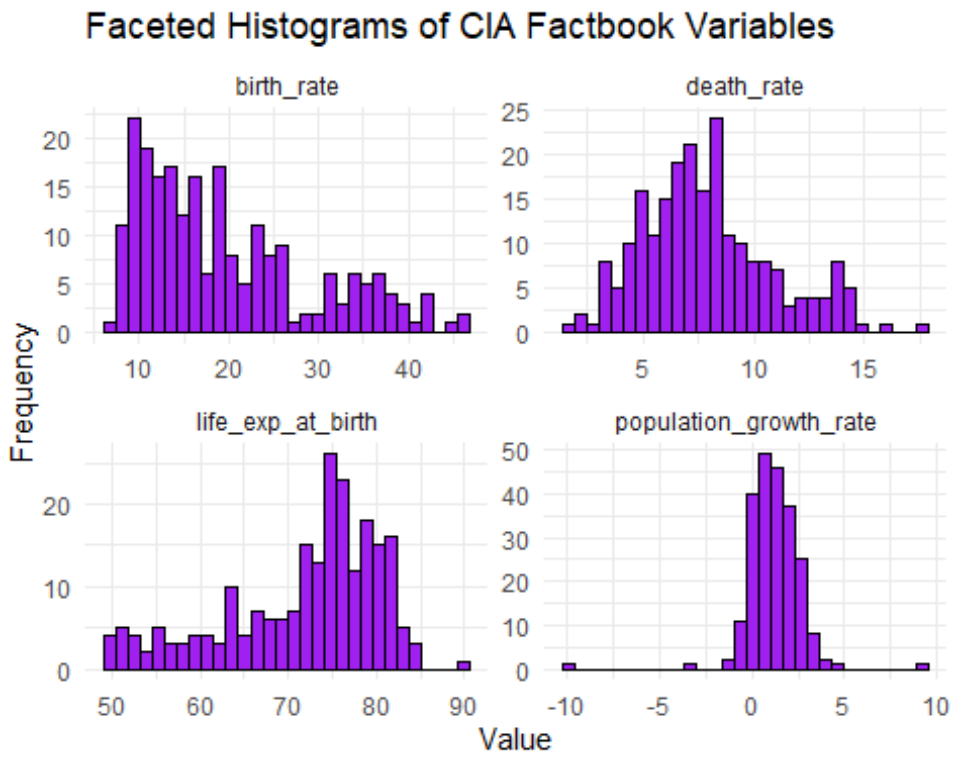
```
# creates linear regression plot

ggplot(cia_factbook_no_country, aes(x = area + birth_rate +
infant_mortality_rate, internet_users + life_exp_at_birth +
maternal_mortality_rate + population + population_rate + third_world, y =
net_migration_rate)) +
  geom_point(color = "black", size = 1.5) +
  geom_smooth(method = "lm", formula = y ~ x, color = "red", se = FALSE) +
  labs(title = "Linear Regression",
       x = "Predictor Values",
       y = "Net Migration") +
  theme_minimal()
```



Let us explore the possibility of using variables which correlate stronger or are more statistically significant produces a more robust model, with a high adjusted r-squared. We will use the variables of birth rate, death rate, life expectancy at birth, and population growth rate as they either have a strong correlation and/or are statistically significant.

First, let's explore each variable individually.

```
cia_factbook_long <- cia_factbook %>%
  pivot_longer(cols = c(birth_rate, death_rate, life_exp_at_birth,
                        population_growth_rate),
               names_to = "variable", values_to = "value")

# creates faceted histogram plot
ggplot(cia_factbook_long, aes(x = value)) +
  geom_histogram(bins = 30, fill = "purple", color = "black") +
  facet_wrap(~variable, scales = "free") + # facet by variable, allow
different scales
  theme_minimal() +
  labs(title = "Faceted Histograms of CIA Factbook Variables",
       x = "Value",
       y = "Frequency")
```

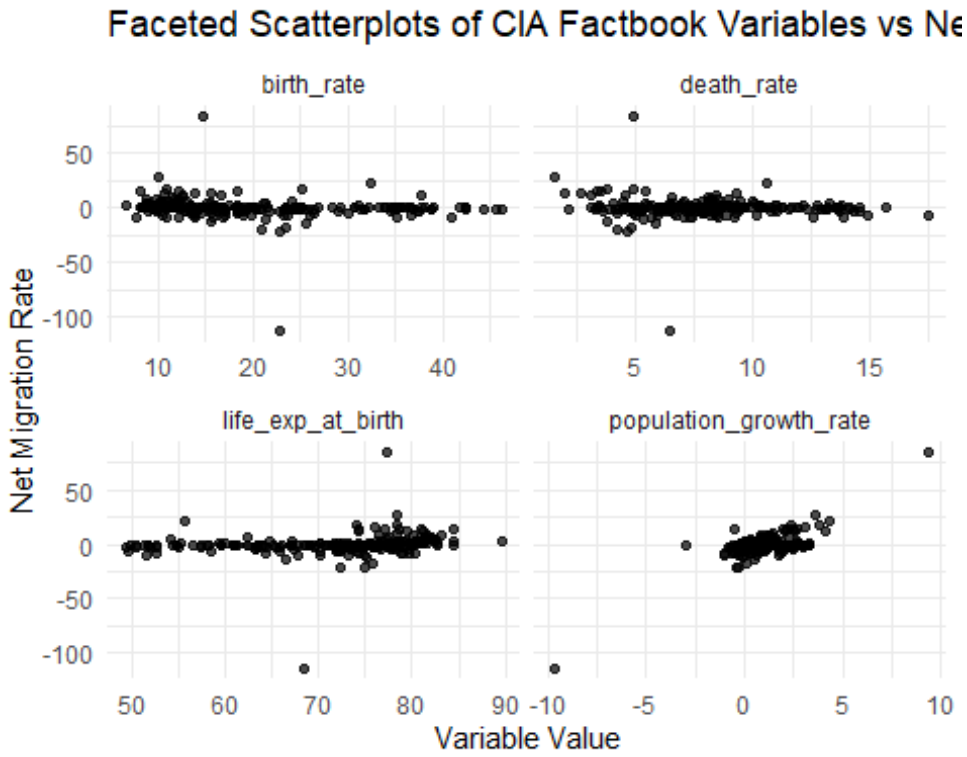


The majority of the variables we are looking at have a skewed spread. For birth rate, the data is skewed right which means that some countries have lots of babies being born but the vast majority is around 10-20%. The death rate is also skewed right and the majority of the data has a death rate of 5-10%, with some having death rates around 15%. Life expectancy at birth is skewed left, which makes sense considering that the numeric data is ordinal. The majority of countries have a life expectancy of 70-80 years old, which makes sense. The population growth rate is more bell shaped than the other variables but is still skewed right. The majority of the countries have a population growth rate around 1%.

Now let's see how each variable individually correlates with the net migration rate.

```
cia_factbook_long <- cia_factbook %>%
  pivot_longer(cols = c(birth_rate, death_rate, life_exp_at_birth,
                        population_growth_rate),
               names_to = "variable", values_to = "x_value") # Renaming for
clarity

# Creating the faceted scatterplot
ggplot(cia_factbook_long, aes(x = x_value, y = net_migration_rate)) +
  geom_point(color = "black", alpha = 0.7) +
  facet_wrap(~variable, scales = "free_x") + # Facet by variable, allow free
x-scale
  theme_minimal() +
  labs(
    title = "Faceted Scatterplots of CIA Factbook Variables vs Net Migration
Rate",
```

```
    x = "Variable Value",
    y = "Net Migration Rate"
  )
```



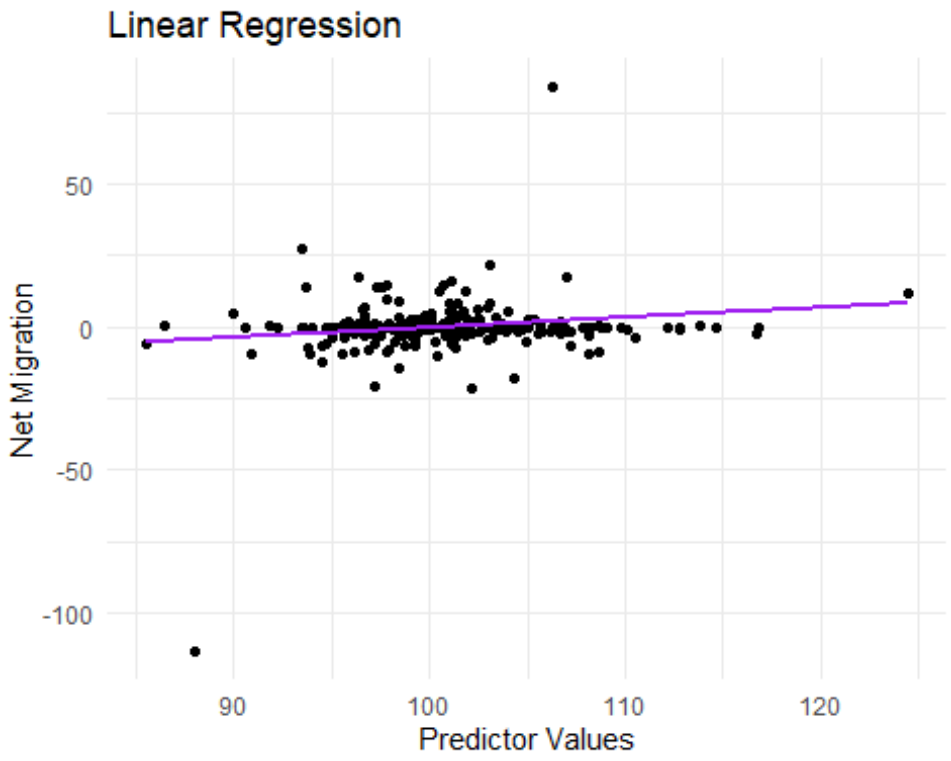
As shown above, there seems to be a very strong linear pattern between net migration and birth rate, life expectancy, and death rate. For population growth rate, there appears to be some bunchiness with the variables.

```
model1 <- lm(net_migration_rate ~ birth_rate + death_rate +
population_growth_rate + life_exp_at_birth, data = cia_factbook) # creates a
model with statistically significant values from the correlation matrix and
the full model summary
summary(model1) # provides a summary for model1

##
## Call:
## lm(formula = net_migration_rate ~ birth_rate + death_rate +
population_growth_rate +
##     life_exp_at_birth, data = cia_factbook)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.399 -0.453 -0.235  0.002 34.343
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.962298   4.868867   0.198    0.844
```

```
## birth_rate              -0.975848   0.044557 -21.901   <2e-16 ***
## death_rate               0.951361   0.085439  11.135   <2e-16 ***
## population_growth_rate   9.451968   0.165377  57.154   <2e-16 ***
## life_exp_at_birth       -0.002834   0.051298  -0.055    0.956
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.782 on 219 degrees of freedom
## Multiple R-squared:  0.9387, Adjusted R-squared:  0.9375
## F-statistic: 837.7 on 4 and 219 DF,  p-value: < 2.2e-16
```

This version of the model is slightly more robust as the adjusted r-squared has increased by .0015. It is also just as statistically significant as the original model looking at the p-value of 2.2e-16.

```
ggplot(cia_factbook, aes(x = birth_rate + death_rate + population_growth_rate
+ life_exp_at_birth, y = net_migration_rate)) +
  geom_point(color = "black", size = 1.5) +
  geom_smooth(method = "lm", formula = y ~ x, color = "purple", se = FALSE) +
  labs(title = "Linear Regression",
       x = "Predictor Values",
       y = "Net Migration") +
  theme_minimal()
```



Is this model the best model we can get for our target variable? To confirm this we will use the step() function.

```r
model1 <- step(fullmodel) # uses stepwise function
```

```
## Start:  AIC=474.67
## net_migration_rate ~ area + birth_rate + death_rate +
infant_mortality_rate +
##     internet_users + life_exp_at_birth + maternal_mortality_rate +
##     population + population_growth_rate + third_world
##
##                            Df Sum of Sq     RSS     AIC
## - population                1       0.0  1690.0  472.67
## - life_exp_at_birth         1       0.0  1690.0  472.67
## - infant_mortality_rate     1       0.0  1690.0  472.67
## - area                      1       0.1  1690.1  472.68
## - internet_users            1       0.4  1690.4  472.72
## - third_world               1       0.7  1690.7  472.76
## - maternal_mortality_rate   1       1.9  1691.9  472.92
## <none>                                   1690.0  474.67
## - death_rate                1     769.6  2459.6  556.73
## - birth_rate                1    2614.9  4304.9  682.11
## - population_growth_rate    1   24671.3 26361.3 1088.03
##
## Step:  AIC=472.67
## net_migration_rate ~ area + birth_rate + death_rate +
infant_mortality_rate +
##     internet_users + life_exp_at_birth + maternal_mortality_rate +
##     population_growth_rate + third_world
##
##                            Df Sum of Sq     RSS     AIC
## - life_exp_at_birth         1       0.0  1690.0  470.67
## - infant_mortality_rate     1       0.0  1690.0  470.67
## - area                      1       0.1  1690.1  470.68
## - third_world               1       0.7  1690.7  470.76
## - internet_users            1       0.7  1690.7  470.76
## - maternal_mortality_rate   1       1.9  1691.9  470.92
## <none>                                   1690.0  472.67
## - death_rate                1     786.5  2476.5  556.26
## - birth_rate                1    2648.7  4338.6  681.86
## - population_growth_rate    1   24671.3 26361.3 1086.03
##
## Step:  AIC=470.67
## net_migration_rate ~ area + birth_rate + death_rate +
infant_mortality_rate +
##     internet_users + maternal_mortality_rate + population_growth_rate +
##     third_world
##
##                            Df Sum of Sq     RSS     AIC
## - area                      1       0.1  1690.1  468.68
## - infant_mortality_rate     1       0.1  1690.1  468.68
## - third_world               1       0.7  1690.7  468.76
## - internet_users            1       0.7  1690.7  468.77
```

```
## - maternal_mortality_rate  1       1.9  1691.9  468.92
## <none>                                  1690.0  470.67
## - death_rate               1    1117.7  2807.7  582.38
## - birth_rate               1    3675.1  5365.1  727.43
## - population_growth_rate   1   24793.5 26483.5 1085.07
##
## Step:  AIC=468.68
## net_migration_rate ~ birth_rate + death_rate + infant_mortality_rate +
##     internet_users + maternal_mortality_rate + population_growth_rate +
##     third_world
##
##                            Df Sum of Sq     RSS     AIC
## - infant_mortality_rate    1       0.1  1690.2  466.69
## - third_world              1       0.7  1690.8  466.77
## - internet_users           1       1.5  1691.6  466.88
## - maternal_mortality_rate  1       1.9  1692.0  466.93
## <none>                                  1690.1  468.68
## - death_rate               1    1129.0  2819.1  581.28
## - birth_rate               1    3681.1  5371.2  725.68
## - population_growth_rate   1   24850.8 26540.9 1083.55
##
## Step:  AIC=466.69
## net_migration_rate ~ birth_rate + death_rate + internet_users +
##     maternal_mortality_rate + population_growth_rate + third_world
##
##                            Df Sum of Sq     RSS     AIC
## - third_world              1       0.9  1691.1  464.81
## - internet_users           1       1.5  1691.7  464.89
## - maternal_mortality_rate  1       2.4  1692.6  465.01
## <none>                                  1690.2  466.69
## - death_rate               1    1292.8  2983.0  591.94
## - birth_rate               1    6441.5  8131.7  816.58
## - population_growth_rate   1   24854.0 26544.2 1081.58
##
## Step:  AIC=464.81
## net_migration_rate ~ birth_rate + death_rate + internet_users +
##     maternal_mortality_rate + population_growth_rate
##
##                            Df Sum of Sq     RSS     AIC
## - internet_users           1       1.5  1692.6  463.01
## - maternal_mortality_rate  1       2.9  1694.0  463.19
## <none>                                  1691.1  464.81
## - death_rate               1    1334.6  3025.7  593.13
## - birth_rate               1    7844.9  9536.0  850.26
## - population_growth_rate   1   24956.2 26647.2 1080.45
##
## Step:  AIC=463.01
## net_migration_rate ~ birth_rate + death_rate + maternal_mortality_rate +
##     population_growth_rate
##
```

```
##                            Df Sum of Sq     RSS     AIC
## - maternal_mortality_rate  1       3.0  1695.5  461.40
## <none>                                  1692.6  463.01
## - death_rate               1    1333.1  3025.7  591.13
## - birth_rate               1    7896.2  9588.8  849.50
## - population_growth_rate   1   24954.7 26647.3 1078.45
##
## Step:  AIC=461.4
## net_migration_rate ~ birth_rate + death_rate + population_growth_rate
##
##                           Df Sum of Sq     RSS     AIC
## <none>                                 1695.5  461.40
## - death_rate               1    1607.3  3302.8  608.76
## - birth_rate               1   12397.6 14093.1  933.76
## - population_growth_rate   1   25406.4 27102.0 1080.24
```

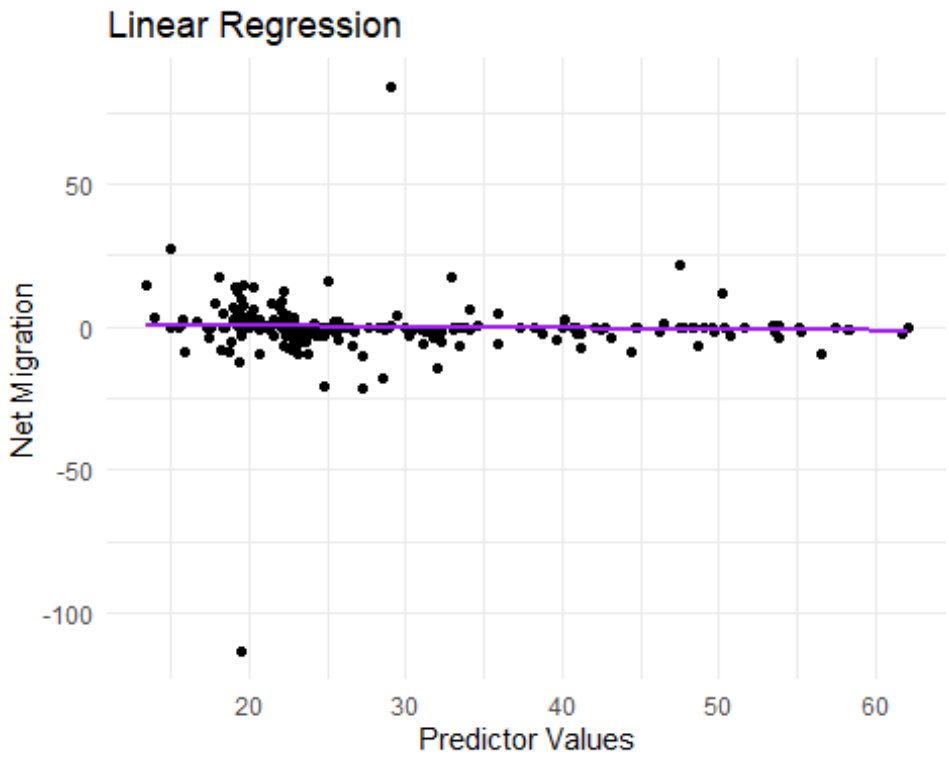
The most optimal model to predict net migration rate contains the variables birth rate, death rate, and population growth rate. All of the values pass the significance test and the R-squared is slightly improved from the previous model. This means that this is the best model to predict the net migration rate.

```
summary(model1)
```

```
##
## Call:
## lm(formula = net_migration_rate ~ birth_rate + death_rate +
population_growth_rate,
##     data = cia_factbook_no_country)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.395 -0.447 -0.234  0.013 34.341
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.69542    0.61311   1.134    0.258
## birth_rate              -0.97379    0.02428 -40.108   <2e-16 ***
## death_rate               0.95434    0.06608  14.441   <2e-16 ***
## population_growth_rate  9.45134    0.16461  57.416   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.776 on 220 degrees of freedom
## Multiple R-squared:  0.9387, Adjusted R-squared:  0.9378
## F-statistic:  1122 on 3 and 220 DF,  p-value: < 2.2e-16
```

```
ggplot(cia_factbook, aes(x = birth_rate + death_rate + population_growth_rate
, y = net_migration_rate)) +
  geom_point(color = "black", size = 1.5) +
  geom_smooth(method = "lm", formula = y ~ x, color = "purple", se = FALSE) +
```

```
  labs(title = "Linear Regression",
       x = "Predictor Values",
       y = "Net Migration") +
  theme_minimal()
```

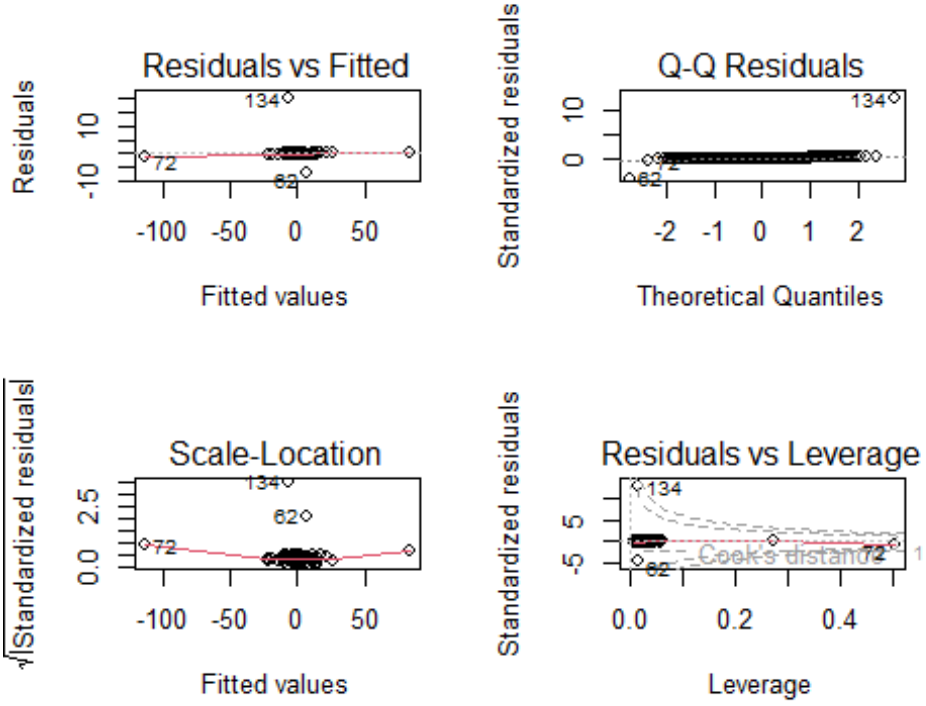


As shown, the values fit slightly better on the plot. Let's move onto training our model.

```
#Creating a train/test partition using random sampling.

set.seed(1231)#use a seed for reproducibility

#use 80% of dataset as training set and 20% as test set
sample <- sample(c(TRUE, FALSE), size = nrow(cia_factbook_no_country),
replace = TRUE, prob = c(0.8, 0.2))
ciaTrain  <- cia_factbook_no_country[sample, ]
ciaTest   <- cia_factbook_no_country[!sample, ]

fullTrain = lm(net_migration_rate ~ birth_rate + death_rate +
population_growth_rate, data = ciaTrain) #Note selection of dataset
names(fullTrain) #reminder of all the elements contained in this object

##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"

summary(fullTrain)#Print the model summary
```

```
##
## Call:
## lm(formula = net_migration_rate ~ birth_rate + death_rate +
population_growth_rate,
##     data = ciaTrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7170 -0.1983 -0.0735  0.0474 19.8829
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.24254    0.40562   0.598    0.551
## birth_rate              -1.00936    0.01504 -67.114   <2e-16 ***
## death_rate               1.01357    0.04370  23.193   <2e-16 ***
## population_growth_rate  9.92339    0.09884 100.398   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.59 on 177 degrees of freedom
## Multiple R-squared:  0.9831, Adjusted R-squared:  0.9828
## F-statistic:  3424 on 3 and 177 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2)) #format the plot
plot(fullTrain) #print the plots of the main diagnostics for your model fit
to the training data
```



The diagnostic plots show some issues with the regression model. Outliers, especially observations 134, 62, and 72, stand out in multiple plots. The residuals may not have constant variance (heteroscedasticity), and there are slight deviations from normality. Observation 134 also has high leverage, indicating it might strongly influence the model. These findings suggest the need to address outliers and check the model's assumptions.

However, these outliers may originate from the Syria and Lebanon outliers from the original dataset so we should take that into account.

```
predFull<-predict(object = fullTrain,     # The regression model fit with
training data
        newdata = ciaTest)           #  creates a new dataframe with the test
data

summary(model1)$coefficients # prints coefficient

##                            Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)              0.6954198 0.61310965   1.13425  2.579239e-01
## birth_rate              -0.9737851 0.02427918 -40.10783 3.892621e-103
## death_rate               0.9543434 0.06608360  14.44146  1.072226e-33
## population_growth_rate  9.4513397 0.16461203  57.41585 2.174566e-134

fullRMSE <- RMSE(predFull, ciaTest$net_migration_rate) ## calculates the RMSE
and R2
c(RMSE = fullRMSE, R2=summary(fullTrain)$r.squared)
```

```
##      RMSE        R2
## 5.5339733 0.9830616
```

The RMSE of the model is 5.5, which indicates there is a relatively large error meaning that the model does make incorrect predictions. Considering the migration rate's IQR is from around -2 to 1, the RMSE is high. However, the R-squared value of 0.983 means that 98.3% of the variation in the target variable is explained by the predictors of the model. This suggests an excellent fit!

Classification Model

```
model2 <- glm(third_world ~ .,
              data = cia_factbook_no_country,
              family = "binomial")

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(model2)

##
## Call:
## glm(formula = third_world ~ ., family = "binomial", data =
cia_factbook_no_country)
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              5.491e-01  1.398e+01   0.039  0.96867
## area                    -1.682e-06  8.173e-07  -2.058  0.03960 *
## birth_rate               1.579e-01  1.070e-01   1.475  0.14016
## death_rate              -5.137e-01  3.714e-01  -1.383  0.16670
## infant_mortality_rate    1.006e-01  3.573e-02   2.815  0.00488 **
## internet_users          -2.335e-07  2.893e-07  -0.807  0.41972
## life_exp_at_birth       -5.889e-02  1.586e-01  -0.371  0.71048
## maternal_mortality_rate  1.059e-03  2.098e-03   0.504  0.61392
## net_migration_rate       6.638e-02  9.607e-02   0.691  0.48957
## population              -4.869e-09  9.175e-09  -0.531  0.59563
## population_growth_rate  -6.594e-01  8.911e-01  -0.740  0.45929
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 219.098  on 223  degrees of freedom
## Residual deviance:  80.613  on 213  degrees of freedom
## AIC: 102.61
##
## Number of Fisher Scoring iterations: 11
```

The initial logistic regression model didn't converge, which means it struggled to make accurate predictions due to the complexity of the data. Several variables, including birth rate, death rate, and population growth rate, showed very large standard errors and p-values, indicating they weren't adding much to the model. The coefficients were unstable, which led to unreliable results.

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.3.3
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.3.3
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
vif(glm(third_world ~ ., data = cia_factbook_no_country, family =
"binomial"))
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##                     area              birth_rate              death_rate
##                 1.723803                8.150252               13.663896
##   infant_mortality_rate          internet_users       life_exp_at_birth
##                 5.511132                1.101262               17.044186
## maternal_mortality_rate      net_migration_rate              population
##                 2.596951                6.063134                1.346347
##  population_growth_rate
##                11.477789
```

I simplified the model by removing less relevant variables like infant mortality rate and internet users. This cleaner version included death rate, net migration rate, population, and population growth rate. However, the model still didn't converge, suggesting there might still be multicollinearity or other issues affecting the accuracy of predictions. To improve this, I could look into variable correlations or apply regularization techniques.

```
model3 <- glm(third_world ~ death_rate + net_migration_rate + population +
population_growth_rate,
                    data = cia_factbook_no_country,
                    family = "binomial")
summary(model3)
```

```
##
## Call:
## glm(formula = third_world ~ death_rate + net_migration_rate +
##     population + population_growth_rate, family = "binomial",
##     data = cia_factbook_no_country)
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -7.231e+00  9.827e-01  -7.358 1.87e-13 ***
## death_rate              2.778e-01  8.425e-02   3.297 0.000977 ***
## net_migration_rate     -2.275e-01  4.233e-02  -5.374 7.68e-08 ***
## population             -1.219e-08  8.714e-09  -1.399 0.161960
## population_growth_rate  2.308e+00  3.507e-01   6.580 4.71e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 219.10  on 223  degrees of freedom
## Residual deviance: 106.24  on 219  degrees of freedom
## AIC: 116.24
##
## Number of Fisher Scoring iterations: 7
```

Let's break down these results. The significant predictors of the model are death rate, net migration rate, and population growth rate. A higher death and population rate increases the likelihood of a country being underdeveloped. A higher migration rate out of a country would lower the likelihood of a country being a third world country. Also, the AIC value of 116.24 indicaters the model is more effective than the null model, or the assumption that there is no relationship with the predictors and the outcome.

However, one of thee variables is statistically insignificant (population) and we will remove this variable so we could produce a more robust classification model. There may need to be a more efficient version of the model which is supported by the number of Fisher Scoring iterations (7), indicating the model is potentially unstable and the algorithm did not fully converge.

We will try to produce a more robust model by removing the population variable.

```
model4 <- glm(third_world ~ death_rate + net_migration_rate
                      + population_growth_rate ,
                      data = cia_factbook_no_country,
                      family = "binomial")
summary(model4)
```

```
##
## Call:
## glm(formula = third_world ~ death_rate + net_migration_rate +
##     population_growth_rate, family = "binomial", data =
cia_factbook_no_country)
```

```
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -7.2772     0.9693  -7.508 6.03e-14 ***
## death_rate                0.2648     0.0827   3.201  0.00137 **
## net_migration_rate       -0.2245     0.0415  -5.411 6.27e-08 ***
## population_growth_rate    2.2672     0.3425   6.620 3.59e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 219.10  on 223  degrees of freedom
## Residual deviance: 108.64  on 220  degrees of freedom
## AIC: 116.64
##
## Number of Fisher Scoring iterations: 6
```

As observed from the output, the number of Fisher Scoring Interations has slightly improved signaling that the model is slightly more stable. Additionally, the p-values for the model have slightly improved indicating that the predictors are more strongly associated with whether a country is third world or not. Furthermore, the AIC (116.64) is slightly higher than the previous model's (116.24). Typically models with lower AICs are superior to those with higher AICs, but given the significan coefficients and the better Fister Scoring Interactions this model is overall more robust, but the difference between the models is quite minimal.

```
```