Project Final

Carl Kawodza

2023-12-04

library(socviz)  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.4 ✔ 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

data(boomer)

str(boomer)

## tibble [1,644 × 6] (S3: tbl\_df/tbl/data.frame)  
## $ date : Date[1:1644], format: "1938-01-01" "1938-02-01" ...  
## $ month : int [1:1644] 1 2 3 4 5 6 7 8 9 10 ...  
## $ n\_days : int [1:1644] 31 28 31 30 31 30 31 31 30 31 ...  
## $ births : int [1:1644] 51820 47421 54887 54623 56853 53145 53214 50444 50545 50079 ...  
## $ total\_pop: int [1:1644] 41215000 41215000 41215000 41215000 41215000 41215000 41215000 41215000 41215000 41215000 ...  
## $ country : chr [1:1644] "England and Wales" "England and Wales" "England and Wales" "England and Wales" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. year = col\_integer(),  
## .. month = col\_integer(),  
## .. n\_days = col\_integer(),  
## .. births = col\_integer(),  
## .. total\_pop = col\_integer(),  
## .. births\_pct = col\_double(),  
## .. births\_pct\_day = col\_double(),  
## .. date = col\_date(format = ""),  
## .. seasonal = col\_double(),  
## .. trend = col\_double(),  
## .. remainder = col\_double(),  
## .. country = col\_character()  
## .. )

missing\_values <- colSums(is.na(boomer))  
print(missing\_values)

## date month n\_days births total\_pop country   
## 0 0 0 0 0 0

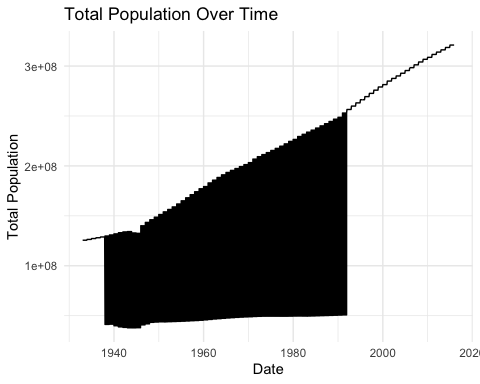
duplicates <- sum(duplicated(boomer))  
print(paste("Number of duplicates:", duplicates))

## [1] "Number of duplicates: 0"

#Handling missing values by imputation  
boomer\_imputed <- boomer%>%  
 mutate\_all(~ifelse(is.na(.), mean(., na.rm = TRUE), .))

#Removing duplicates  
boomer\_no\_duplicates <- distinct(boomer)

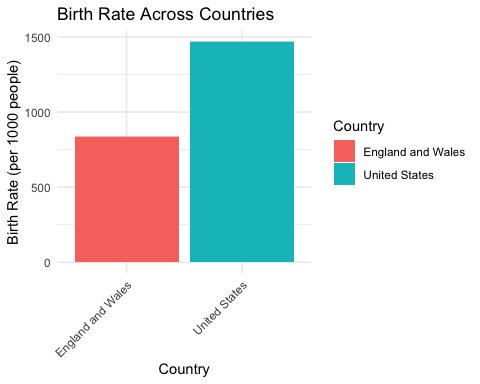
#How does the total population vary over time? Are there specific periods of notable population growth or decline?  
boomer$date <- as.Date(boomer$date)  
  
# Plotting the total population over time  
ggplot(boomer, aes(x = date, y = total\_pop)) +  
 geom\_line() +  
 labs(title = "Total Population Over Time",  
 x = "Date",  
 y = "Total Population") +  
 theme\_minimal()



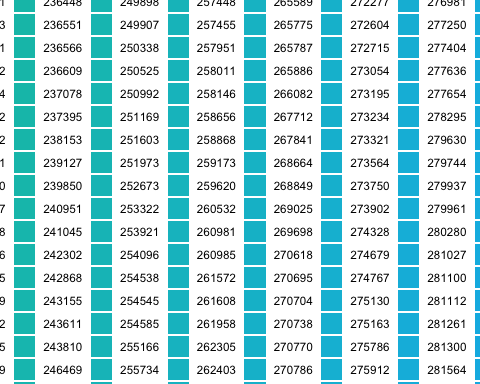
#summary statistics  
summary(boomer[c('total\_pop', 'births')])

## total\_pop births   
## Min. : 37785000 Min. : 41560   
## 1st Qu.: 49327100 1st Qu.: 60578   
## Median :143446000 Median :254876   
## Mean :148257771 Mean :205888   
## 3rd Qu.:229465714 3rd Qu.:324270   
## Max. :321039839 Max. :392230

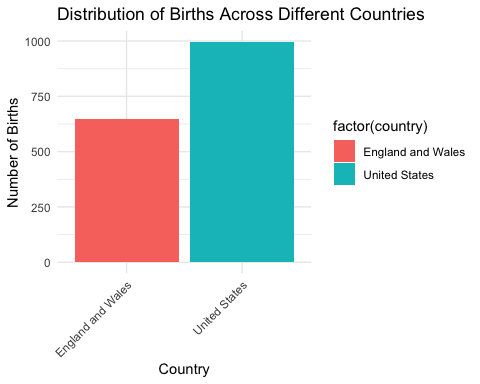
ggplot(boomer, aes(x = country, y = births / total\_pop \* 1000, fill = country)) +  
 geom\_bar(stat = "identity") +  
 labs(title = "Birth Rate Across Countries",  
 x = "Country",  
 y = "Birth Rate (per 1000 people)",  
 fill = "Country") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



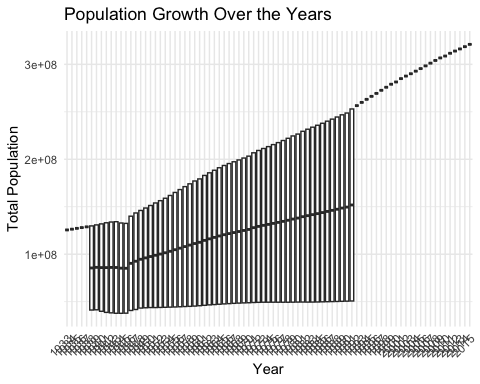
#Are there any seasonal variations in birth rates throughout the year?  
  
library(ggplot2)  
  
#Plotting the birth rates across different months  
ggplot(boomer, aes(x = factor(births), fill = factor(births))) +  
 geom\_bar() +  
 labs(title = "Seasonal Variations in Birth Rates",  
 x = "Month",  
 y = "Number of Births") +  
 scale\_x\_discrete(labels = month.name) +   
 theme\_minimal()



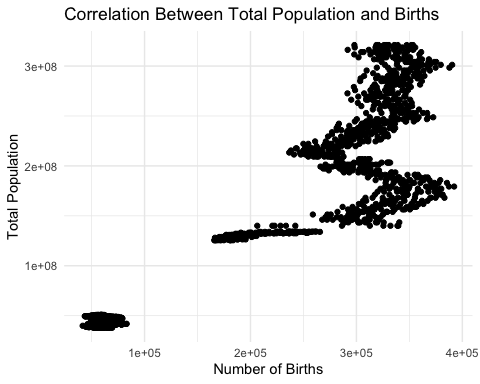
#For a specific country, what is the distribution of births across different months?  
  
library(ggplot2)  
  
#Plotting the distribution of births across different countries  
ggplot(boomer, aes(x = factor(country), fill = factor(country))) +  
 geom\_bar() +  
 labs(title = "Distribution of Births Across Different Countries",  
 x = "Country",  
 y = "Number of Births") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



#What is the overall trend in population growth or decline over the years?  
library(ggplot2)  
boomer$date <- as.Date(boomer$date)  
  
# Extract the year from the date  
boomer$year <- lubridate::year(boomer$date)  
  
# Plotting a boxplot to show the overall trend in population growth or decline over the years  
ggplot(boomer, aes(x = factor(year), y = total\_pop)) +  
 geom\_boxplot() +  
 labs(title = "Population Growth Over the Years",  
 x = "Year",  
 y = "Total Population") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



#Is there any correlation between the total population and the number of births within a given time frame?  
library(ggplot2)  
boomer$date <- as.Date(boomer$date)  
  
#Scatter plot to visualize the relationship  
ggplot(boomer, aes(x = births, y = total\_pop)) +  
 geom\_point() +  
 labs(title = "Correlation Between Total Population and Births",  
 x = "Number of Births",  
 y = "Total Population") +  
 theme\_minimal()



#Calculating correlation coefficient  
correlation\_coefficient <- cor(boomer$total\_pop, boomer$births)  
print(paste("Correlation Coefficient: ", round(correlation\_coefficient, 2)))

## [1] "Correlation Coefficient: 0.92"