Post 1 Investigating the Demographic Transition

Patrick Chao

Introduction

The world is growing a rapid pace, with 7.6 billion people on the world currently. However, just over 200 years ago we had under a billion. What was the change that led to such a massive increase in population? What will the world look like in the next tens or hundreds of years? Overpopulation has been a problem that has been discussed for many years, but now with more powerful tools, we can analyze the effects.

The purpose of this post is not to delve directly into population modeling as this has been done many times before (and quite a daunting task), but to look into population growths from the past and analyze trends. Specifically, I will look into visualizing the *demographic transition*.

The first thing to do is collect data. I will be using data from the world bank, where I will use birth rate, mortality rate, life expectancy, and fertility rate. These are defined in the following fashion:

- Birth Rate: the number of births per thousand people in a year
- Mortality Rate: the number of births per thousand people in a year
- Life Expectancy: the number of births per thousand people in a year
- Fertility Rate: the average number of children born per a woman in her lifetime

Data Cleaning

Let's initialize all our libraries.

```
library(ggplot2)
library(dplyr)
library(ggthemr)
library(tidyr)
library(readr)
library(broom)
library(magrittr)
library(ggrepel)
```

After downloading the data, we need to clean it.

```
##Read in files
birth <- read_csv("birthRate.csv",skip=4)
mortality <- read_csv("mortalityRate.csv",skip=4)
lifeExpectancy <- read_csv("lifeExpectancy.csv",skip=4)
fertility <- read_csv("fertilityRate.csv",skip=4)
population <- read_csv("populationRate.csv",skip=4)

##Reformat column names
colnames(population)[1] <- "CountryName"
colnames(population)[2] <- "CountryCode"
colnames(population)[3] <- "IndicatorName"
colnames(population)[4] <- "IndicatorCode"

colnames(fertility)[1] <- "CountryName"
colnames(fertility)[2] <- "CountryCode"
colnames(fertility)[3] <- "IndicatorName"
colnames(fertility)[4] <- "IndicatorCode"

colnames(mortality)[1] <- "CountryName"
colnames(mortality)[2] <- "CountryCode"
colnames(mortality)[3] <- "IndicatorName"
colnames(mortality)[4] <- "IndicatorCode"

colnames(lifeExpectancy)[1] <- "CountryName"
colnames(lifeExpectancy)[2] <- "CountryCode"
colnames(lifeExpectancy)[3] <- "IndicatorName"
colnames(lifeExpectancy)[4] <- "IndicatorCode"

colnames(birth)[1] <- "CountryName"
colnames(birth)[2] <- "CountryCode"
colnames(birth)[3] <- "IndicatorName"
colnames(birth)[4] <- "IndicatorCode"

#Establish feature types
fertility$FeatureType <- 'FertilityRate'
mortality$FeatureType <- 'MortalityRate'
birth$FeatureType <- 'BirthRate'
lifeExpectancy$FeatureType <- 'LifeExpectancy'
population$FeatureType <- 'PopulationRate'

#Combine the data frames into a single one
fullData <- rbind(fertility,mortality)
fullData <- rbind(fullData,lifeExpectancy)
fullData <- rbind(fullData,birth)
fullData <- rbind(fullData,population)

#Remove unnecessary columns
fullData$IndicatorName <- NULL
fullData$IndicatorCode <- NULL
fullData$'2016' <- NULL
fullData$X62 <- NULL
fullData$CountryCode <- NULL

fullData <- fullData[rowSums(is.na(fullData))<1,]

#Convert to easily plottable time series format
timeSeries <- fullData %>% gather('Year', 'Value', "1960":"2015")
```

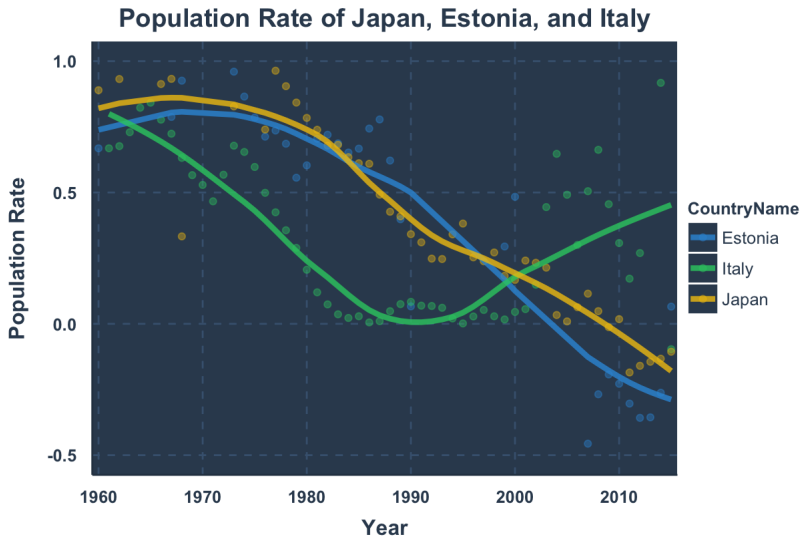
A Brief Background

In general, population rates have been increasing rapidly for the past hundred years. The peak for population growth rates occurred in 1970, with a whopping **2.1%**, with the current global growth rate around **1%**.

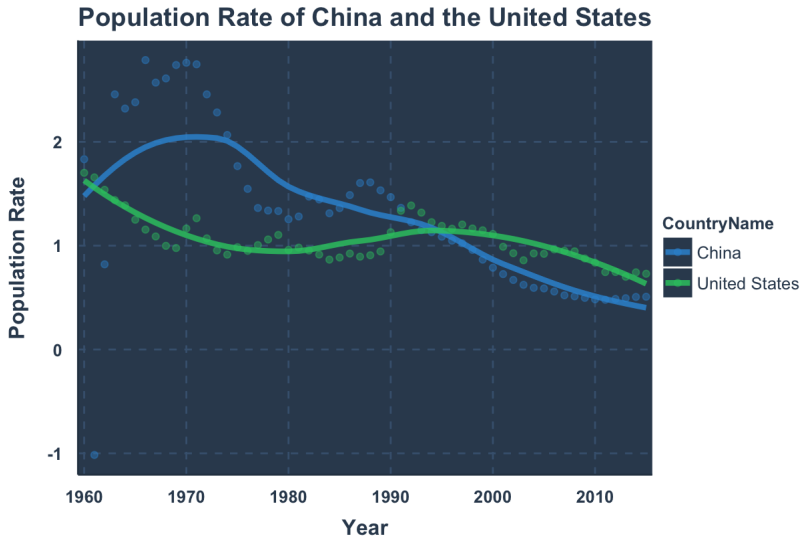
One may assume that successful countries continually grow and expand their population. However, this is not the case. Many countries have actually plateaued with respect to population growth rates, or have actually decreased.

For example, Japan, Estonia, Germany, and Italy have begun population decline or will begin to do so in the near future. Some countries such as the United States and China have significantly decreased population growth rates, and have begun to steady.

```
ggthemr('flat dark')

#Constants for plotting
lineAlpha <- 0.8
pointAlpha <- 0.4
lineSize=1.5

#Plot Japan, Estonia, and Italy
timeSeries %>% filter(CountryName %in% c("Japan","Estonia","Italy") &
FeatureType == "PopulationRate")  %>% ggplot(aes(Year, Value, col = CountryName)) +
geom_point(alpha=pointAlpha) +stat_smooth(geom='line',alpha=lineAlpha,aes(group=CountryName),
size=lineSize) +scale_x_discrete(breaks=seq(1960, 2020, 10))+ylim(-0.5,1)+
labs(y="Population Rate", x="Year",title="Population Rate of Japan, Estonia, and Italy")
```




```
#Plot China and the United States
timeSeries %>% filter(CountryName %in% c("China","United States") & FeatureType == "PopulationRate")  %>%
ggplot(aes(Year, Value, col = CountryName)) + geom_point(alpha=pointAlpha) +
stat_smooth(geom='line', alpha=lineAlpha,aes(group=CountryName),size=lineSize) + scale_x_discrete(breaks=seq(1960,
2020, 10))+labs(y="Population Rate", x="Year",title="Population Rate of China and the United States")
```

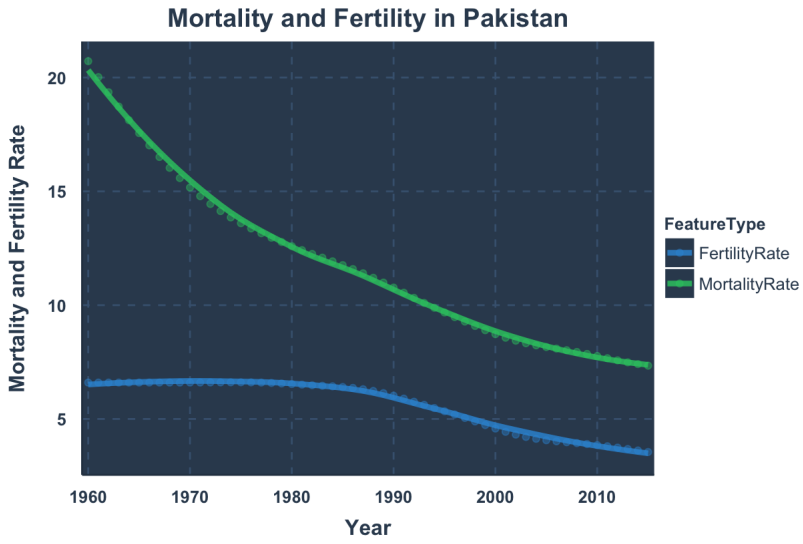


Demographic Transition

There is a general trend in population growths due to technological and societal improvements, dubbed the *demographic transition*.

- Stage One: In the past, both population and mortality rates were high, around 40 births and deaths per thousand individuals each year.
- Stage Two: With improvements in technology and medicine, the mortality rate sharply declines while birthrates stay high. This causes a sudden boom in population rate, and the overall population increases dramatically.
- Stage Three: Birth rates steadily fall as economic conditions improve and access to contraception. Population growth continues at a decreased rate.
- Stage Four: Birth rates and mortality rates are essentially equivalent and the population stabilizes.

```
#Plotting Pakistan
 timeSeries %>% filter(CountryName %in% c("Pakistan") & FeatureType %in% c("FertilityRate","MortalityRate"))  %>%
ggplot(aes(Year, Value, col = FeatureType)) + geom_point(alpha=pointAlpha) +
stat_smooth(geom='line', alpha=lineAlpha,aes(group=FeatureType),size=lineSize) + scale_x_discrete(breaks=seq(1960,
2020, 10))+labs(y="Mortality and Fertility Rate", x="Year",title="Mortality and Fertility in Pakistan")
```

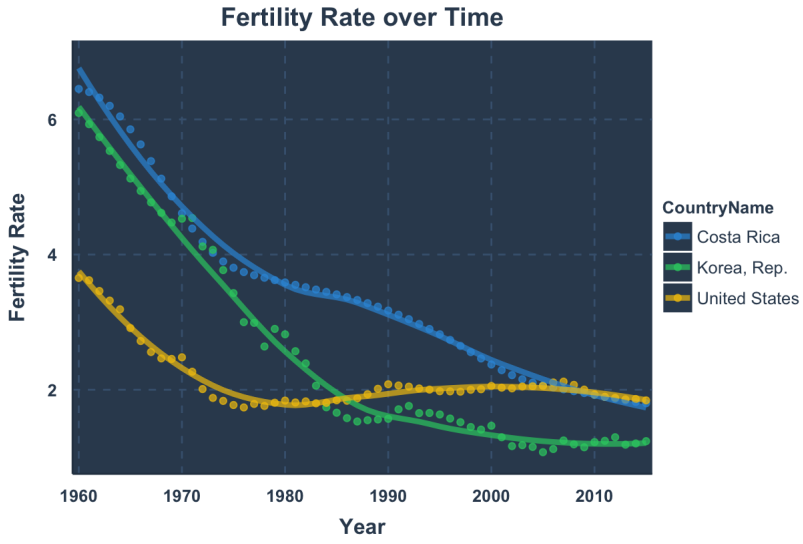


For example, this is a plot of Pakistan, a country in Stage 3. There is a steep drop in birth rates as expected, and a lesser decrease in mortality rate. Within the 20 years, these rates will converge and the country will stabilize.

Elementary Plots

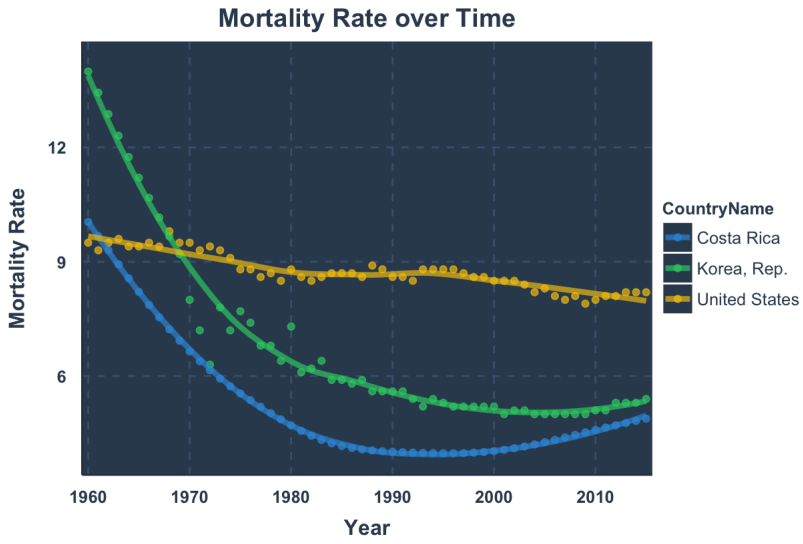
A first elementary step is to begin plotting some graphs.

```
ggthemr('flat dark')
lineAlpha <- 0.7
pointAlpha <- 0.7
lineSize=1.5
#Plotting Fertility Rate
 timeSeries %>% filter(CountryName %in% c("United States", "Korea, Rep.","Costa Rica") &
FeatureType == "FertilityRate")  %>% ggplot(aes(Year, Value, col = CountryName)) +
geom_point(alpha=pointAlpha) + stat_smooth(geom='line',alpha=lineAlpha,
aes(group=CountryName),size=lineSize) + scale_x_discrete(breaks=seq(1960, 2020, 10))+
labs(y="Fertility Rate", x="Year",title="Fertility Rate over Time")
```



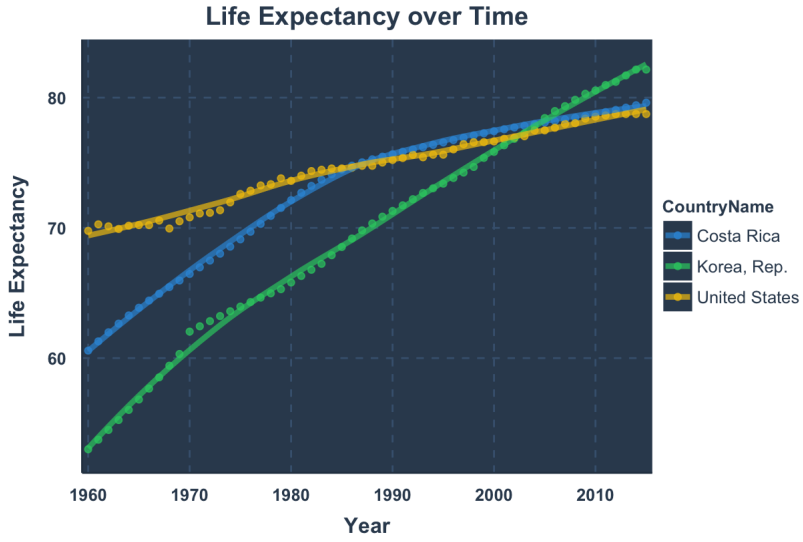
Looking at fertility rate, we notice that Costa Rica and Korea began with quite high fertility rates, with about 6 births per woman in 1960. This sharply dropped to about 2 by 2015. The United States started off lower and had a much slower transition. This is most likely because the United States was already farther along stage 3 than Costa Rica and Korea, causing a smaller change overall.

```
timeSeries %>% filter(CountryName %in% c("United States", "Korea, Rep.","Costa Rica") &
FeatureType == "MortalityRate")  %>% ggplot(aes(Year, Value, col = CountryName)) +
geom_point(alpha=pointAlpha) +  stat_smooth(geom='line',
alpha=lineAlpha,aes(group=CountryName),size=lineSize)+
scale_x_discrete(breaks=seq(1960, 2020, 10))+
labs(y="Mortality Rate", x="Year",title="Mortality Rate over Time")
```



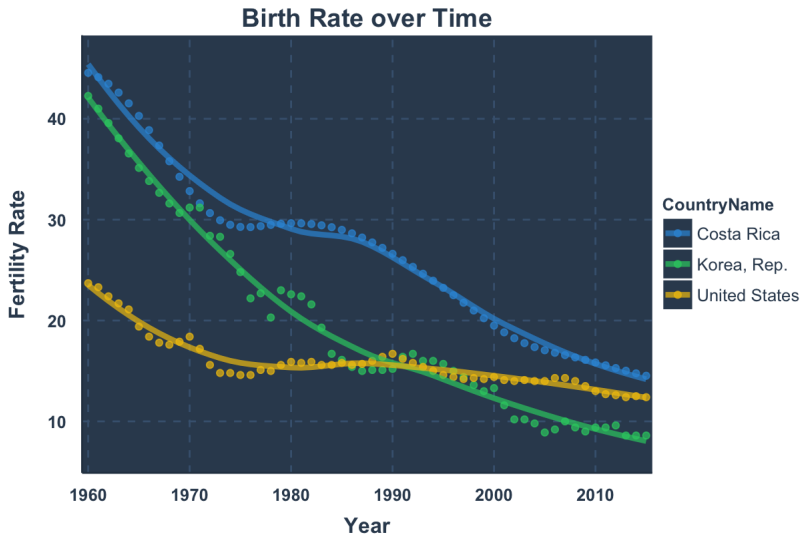
Mortality rate tells a slightly different story. The United States began with the lowest mortality rate, but the value has hardly changed over 50 years. Costa Rica and Korea experienced huge drops in mortality rate, demonstrating the improvements in technology and health. For some reason, Costa Rica's mortality rate actually slightly increased in recent years. However, this does not necessarily imply that Costa Rica is more "unsafe". This could merely be an effect of a higher proportion of older individuals passing away, due to a higher than average birth rate ≈ 70 years ago.

```
timeSeries %>% filter(CountryName %in% c("United States", "Korea, Rep.","Costa Rica")
& FeatureType == "LifeExpectancy")  %>% ggplot(aes(Year, Value, col = CountryName)) +
geom_point(alpha=pointAlpha) +  stat_smooth(geom='line', alpha=lineAlpha,aes(group=CountryName),size=lineSize)+
scale_x_discrete(breaks=seq(1960, 2020, 10))+
labs(y="Life Expectancy", x="Year",title="Life Expectancy over Time")
```



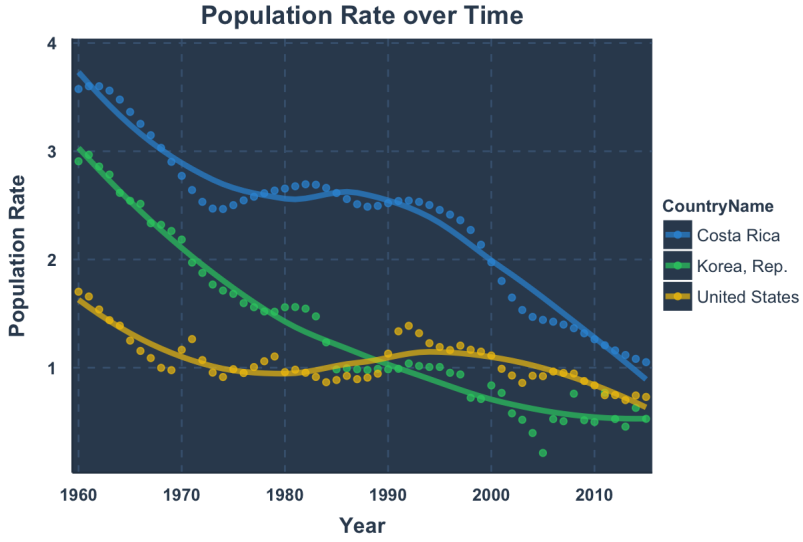
Life expectancy moves at a predictable trend. All countries move upwards with respect to life expectancy, although Costa Rica and Korea move at much faster rate. The United States also began at a higher life expectancy, which is again logical due to starting stage 3 earlier. South Korea has a massive shift in life expectancy, starting from 53 in 1960 and climbing to 82, while the United States moves from 69 to 78 over 50 years.

```
timeSeries %>% filter(CountryName %in% c("United States", "Korea, Rep.","Costa Rica") &
FeatureType == "BirthRate")  %>% ggplot(aes(Year, Value, col = CountryName)) +
geom_point(alpha=pointAlpha) +  stat_smooth(geom='line', alpha=lineAlpha,aes(group=CountryName),
size=lineSize)+scale_x_discrete(breaks=seq(1960, 2020, 10))+
labs(y="Fertility Rate", x="Year",title="Birth Rate over Time")
```



Birth rate is very similar to fertility rate, so I will not go into specific analysis.

```
timeSeries %>% filter(CountryName %in% c("United States", "Korea, Rep.","Costa Rica") &
FeatureType == "PopulationRate")  %>% ggplot(aes(Year, Value, col = CountryName)) +
geom_point(alpha=pointAlpha) +  stat_smooth(geom='line', alpha=lineAlpha,aes(group=CountryName),
size=lineSize)+scale_x_discrete(breaks=seq(1960, 2020, 10))+
labs(y="Population Rate", x="Year",title="Population Rate over Time")
```



The population rate graph also reveals interesting information. The population rates all seem to converge at around 0.8, which represents quite a stable population. The change in population rate represents the progression in the stage 3 phase. The United States decreased a bit but remained relatively constant. Costa Rica greatly decreased, and South Korea dropped even past the United States in terms of population.

Principal Components

An interesting way to analyze the relationships between countries is to investigate the principal components.

By applying PCA, we may plot the countries based on these components to visualize any trends. A huge advantage is PCA in R contains the "scale" variable which allows us to easily scale the data. This prevents some features like life expectancy seeming overly important compared to fertility because life expectancy is ≈ 80 while fertility is only 2 to 6.

Note that I only apply PCA on the four components Fertility Rate, Life Expectancy, Birth Rate, and Mortality Rate. Including Population Rate would be similar to cheating, since we essentially can just use population growth rates to illustrate demographic transitions.

```
countriesByRates <- spread(timeSeries,FeatureType,Value)

na_countries <- countriesByRates %>% filter_all(any_vars(is.na(.)))%>% pull(CountryName)%>%unique()

countriesByRates <- countriesByRates %>% filter(! CountryName %in% na_countries)
pca <- countriesByRates %>% select(contains("Rate"),-PopulationRate, LifeExpectancy) %>% prcomp(scale=TRUE)
summary(pca)
```

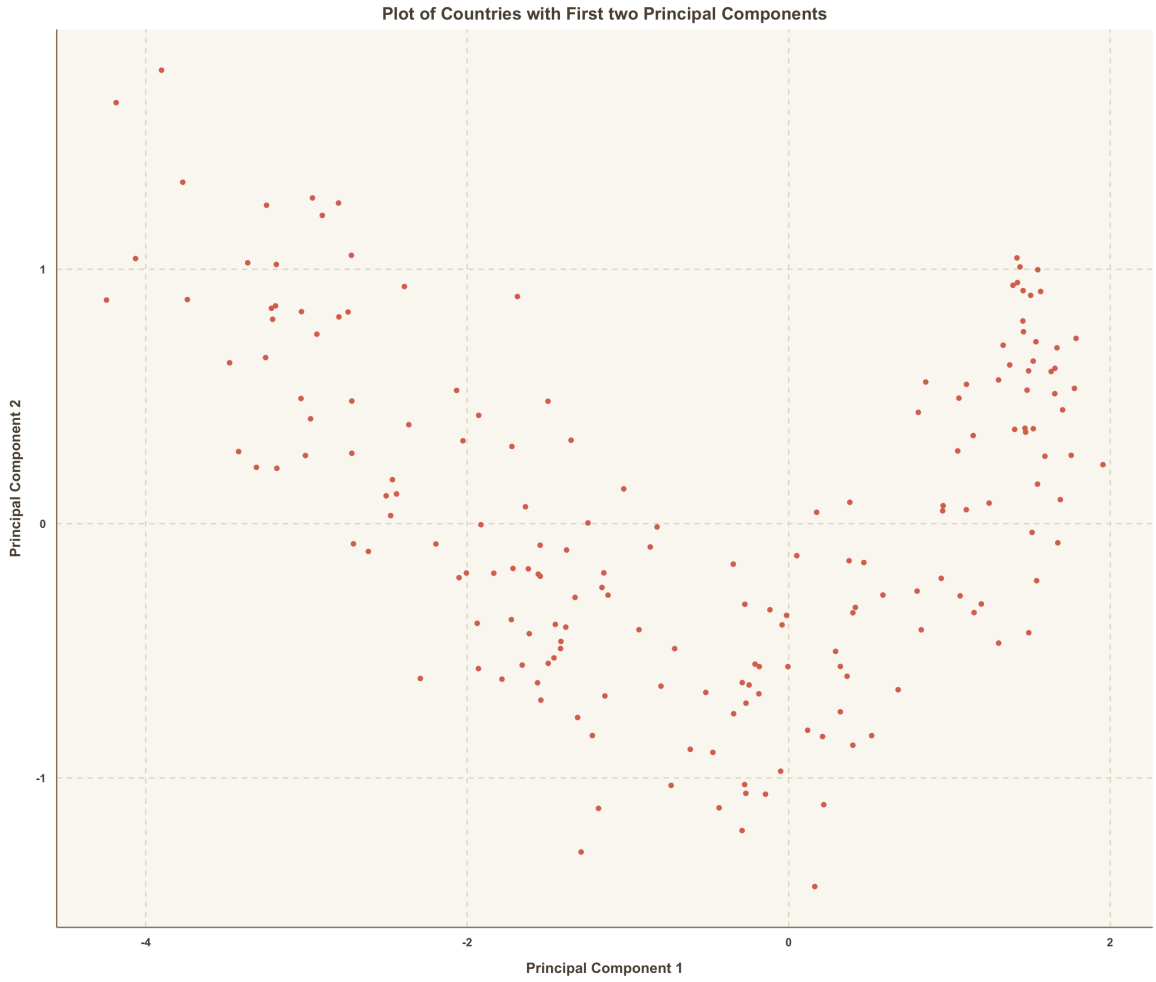
```
## Importance of components%s:
##                           PC1    PC2     PC3     PC4
## Standard deviation     1.8429 0.7276 0.24707 0.11530
## Proportion of Variance 0.8491 0.1323 0.01526 0.00332
## Cumulative Proportion  0.8491 0.9814 0.99668 1.00000
```

By looking at the proportion of the principal components, the first principal component represents 84.91% of the variance, meaning the first component essentially characterizes the points. The second component takes up a remaining 13.23%, meaning the first two components characterize a whopping 98.14 percent of the data.

Now we may project each country to a principal component. To do this, we need to accumulate each country's data into a single row. This is a bit tricky since each country has around 56 rows of data, one data point per 5 features per each year from 1960 to 2010. To address this issue, I compute a weighted average of the years. Each year receives a weight corresponding to how recent it is, beginning at 1 and ending at 0.04. The weights follow a geometric progression with a decay rate 0.95, which was chosen since intuitively I would like the midpoint of 30 years to receive a much lower weight, and I decided on about 0.2.

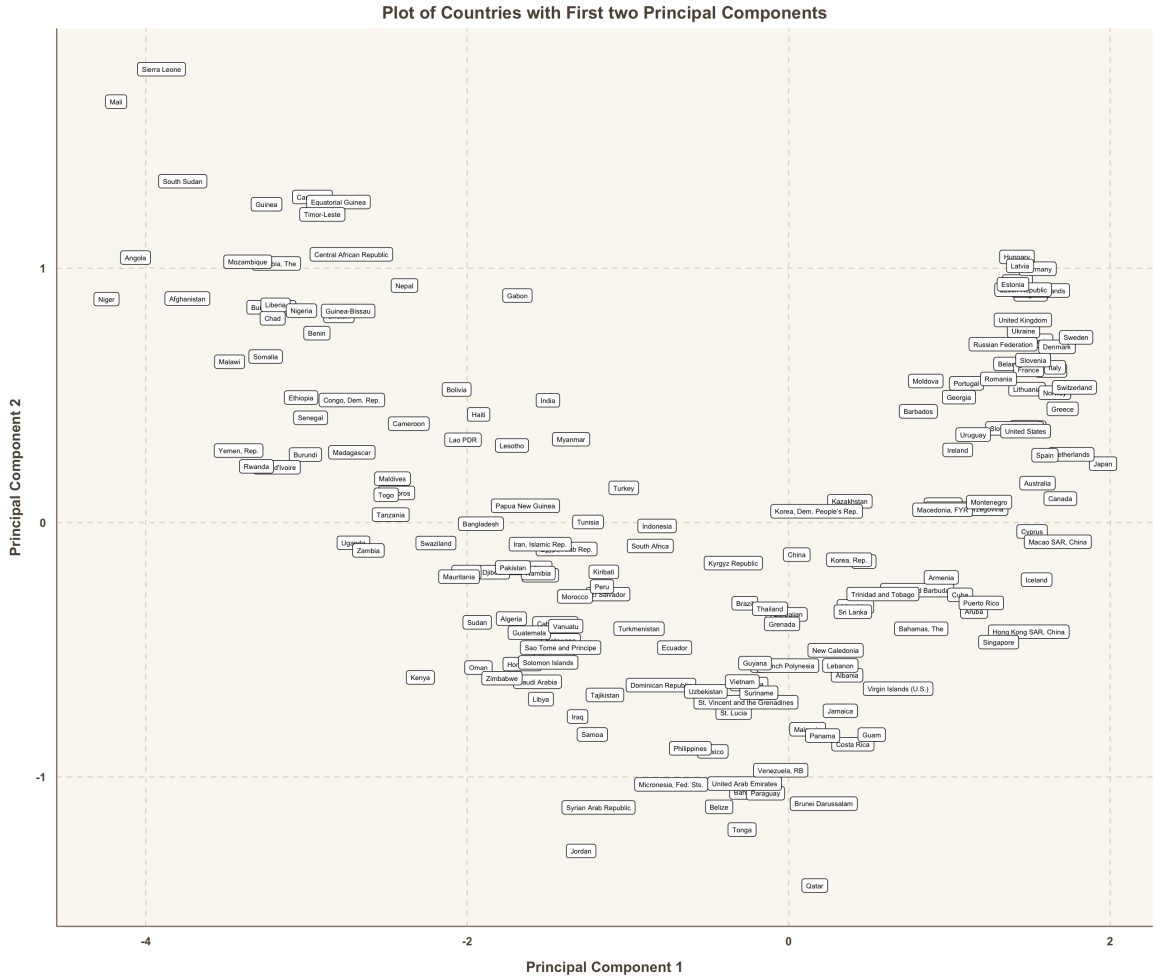
```
gamma = 0.95

pcDF<-data.frame(pca$x,Country=countriesByRates$CountryName) %>% group_by(Country) %>%
rev(.) %>%summarise_if(is.numeric,funs(weighted.mean(.,gamma^(0: (length(.)-1)))))

ggthemr("dust")
```

```
pcDF %>% ggplot() + geom_point(aes(PC1, PC2)) +
labs(y="Principal Component 2", x="Principal Component 1",
title="Plot of Countries with First two Principal Components")
```



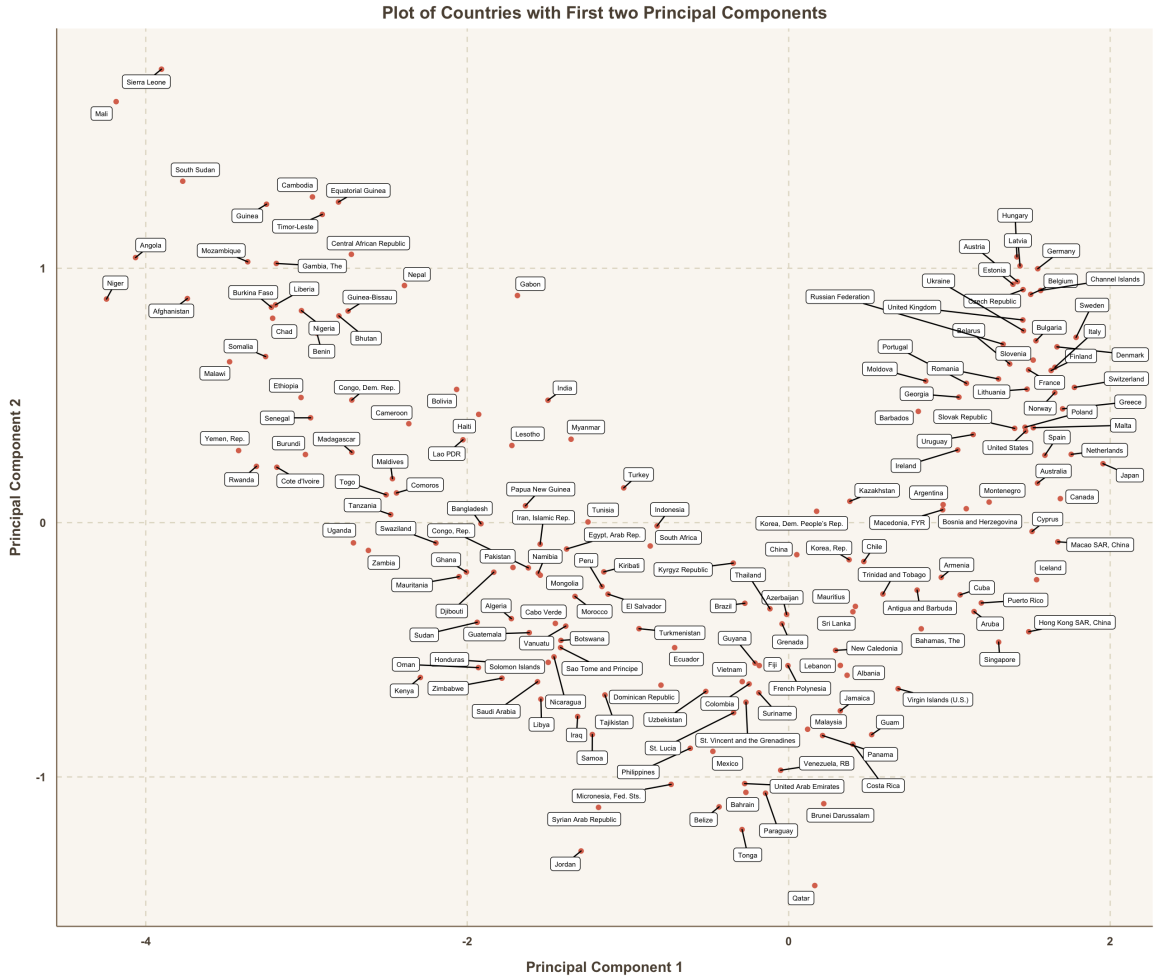
This is a plot of the first two principal components. This does not tell us that much since the data is not labeled. There is an interesting *U* shape to the countries.

```
pcDF %>% ggplot() + geom_label(aes(PC1, PC2,label=Country),size=2)+
labs(y="Principal Component 2", x="Principal Component 1",
title="Plot of Countries with First two Principal Components")
```



This is a plot of the first two principal components with the countries labeled. The countries are a bit difficult to read since they are slightly overlapped. By using *ggrepel* we can prevent them from overlapping, but it is still a bit busy.

```
pcDF %>% ggplot() + geom_point(aes(PC1, PC2))+
geom_label_repel(aes(x=PC1,y=PC2,label=Country),  size = 2)+
labs(y="Principal Component 2", x="Principal Component 1",
title="Plot of Countries with First two Principal Components")
```



Now by looking the individual countries, we can see a clear trend. The countries that seem to be more industrialized and farther in the demographic transition tend to be along the top right. This includes Germany, Sweden, Italy, Finland, the UK, Japan, and the US. As you approach the left side, there are more developing countries such as Chad, Nigeria, and Angola. Most of the variation may be represented as traveling through the demographic transition from left to right, which makes sense since the first principal component conveys a large majority of the information.

A future investigation may be to analyze the general trend in countries over time. A modest prediction would be that countries move from the top left to the lower center to the top right over the process of the demographic transition.

Conclusion

By considering elementary plots of fertility rate, mortality rate, life expectancy, and birth rates, we are able to view clear trends of stages in the demographic transition. By considering simple trends and stabilizations in populations, we observe that many countries are participating in Stage 3, where birth rates are high and mortality is low, contributing to the exploding world population. As time progresses, more countries then enter Stage 4, where population growth rates stabilize as fertility and mortality are equal.

Using more technical methods, it is very interesting how much information we may glean by simply plotting the principal components on a few features. This demonstrates the power and efficiency of the principal components to explain a data set. There are many future possibilities with this investigation, including analyzing how countries move over time, rather than considering a weighted average of the years.

Sources

Data Source
Demographic Transition
Demographic Transition 2
Negative Growth
Population Growth
PCA
PCA with Population
orld Population