

Proyecto 2 - Analisis Exploratorio

Manuel Valenzuela 15072, Davis Álvarez 15842, Jorge SÚchite 15293

23 de agosto de 2018

Descripción del Tema

Para este segundo proyecto propone observar la población de los países del mundo a lo largo de los años y así mismo su área superficial, con estos datos obtener una densidad poblacional y ver si esta relacionada de alguna manera con la felicidad de las personas. Con estos datos se tratará de hacer 2 pronósticos. El primero de la población en base a su propio comportamiento; y el segundo, de la felicidad del mundo en base a su propio comportamiento y así mismo a la densidad poblacional a lo largo del tiempo.

Para hacer esto se utilizarán 3 datasets, el primero es World Happiness Report, el cual es un dataset que contiene una puntuación de felicidad anotada según producción económica, apoyo social, etc. El segundo es Countries Population, el cual contiene la población de cada país a lo largo de los años, y por último, el tercero sería Countries of the World, del cual se obtendría el área superficial de cada país.

Problema Científico

Lo que se buscará realizar con este proyecto es observar la densidad poblacional y la felicidad de las personas de cada país a lo largo del tiempo, ver si tienen algún tipo de relación entre sí y tratar de pronosticar que tan felices serán las personas en los años próximos. Así mismo, observando el comportamiento de la población a lo largo de los años se buscará pronosticar la población de los países en los próximos años.

Objetivos

Generales

- Encontrar la Densidad Poblacional de los países del mundo a lo largo de los años.
- Encontrar si hay una relación entre la densidad poblacional y la felicidad de las personas.

Específicos

- Describir el dataset de Countries Population, encontrar un patrón del comportamiento de la población y pronosticarla en los próximos años.
- Describir el dataset de World Happiness Report, encontrar las variables que más inciden en la felicidad de las personas y pronosticarla en los próximos años.
- Comparar los pronósticos y observar si efectivamente hay una relación entre la densidad poblacional y la felicidad de las personas.

Descripción de los Datos

A continuación se describen los datasets y así mismo cada una de las variables que estos tienen. Luego de eso se explicará el proceso de limpieza que se llevó a cabo para tomar en cuenta solo las variables que serán de utilidad para cumplir los objetivos planteados.

World Happiness Report

El World Happiness Report es una encuesta histórica sobre el estado de la felicidad global. El primer informe se publicó en 2012, el segundo en 2013, el tercero en 2015 y el cuarto en la Actualización de 2016. The World Happiness 2017, que clasifica a 155 países por su nivel de felicidad, fue lanzado en las Naciones Unidas en un evento que celebra el Día Internacional de la Felicidad el 20 de marzo.

Los rankings y puntajes de felicidad utilizan data de una encuesta de Gallup World. Las puntuaciones son basadas en las respuestas a las preguntas de evaluación de vida contenidas en esta encuesta. Estas preguntas se conocen como la Escalera de Cantril, esta les pide a los encuestados que piensen en una escalera con la mejor vida posible para ellos siendo un 10 y la peor vida posible siendo un 0. A continuación se describen las variables contenidas en este dataset.

No.	Variable	Tipo	Descripción
1	Country	Factor	Nombre del país
2	Region	Factor	Region a la que pertenece
3	Happiness.Rank	int	Rango del país basado en el Happiness Score
4	Happiness.Score	num	Puntuación obtenida de Encuesta
5	Standard.Error	num	La desviación Estándar del Happiness Score
6	Economy..GDP.per.Capita	num	El grado en que el Producto Interno Bruto del país contribuye al cálculo del puntaje de felicidad.
7	Family	num	El grado en que la familia contribuye al cálculo del puntaje de felicidad.
8	Health..Life.Expectancy	num	el grado en que la esperanza de vida contribuye al cálculo del puntaje de felicidad
9	Freedom	num	La medida en que la libertad contribuye al cálculo de la puntuación de la felicidad
10	Trust..Government.Corruption.	num	El grado en que la percepción de la corrupción contribuye al puntaje de felicidad
11	Generosity	num	El grado en que generosidad contribuye al cálculo del puntaje de felicidad
12	Dystopia.Residual	num	La medida en que la distopía residual contribuyó al cálculo de la puntuación de la felicidad
13	Year	num	Año en el que fue hecha la encuesta
14	Lower.Confidence.Interval	num	Intervalo de confianza más bajo del puntaje de felicidad
15	Upper.Confidence.Interval	num	Intervalo de confianza superior del puntaje de felicidad
16	Whisker.high	num	Bigote Superior
17	Whisker.low	num	Bigote Inferior

De las variables descritas en la tabla anterior se tenían 3 datasets, uno por cada año. El proceso de limpieza que se hizo con estos datos fue cuadrar las variables en todos los datasets, es decir, que se hizo que todos tuviesen las mismas variables con el mismo nombre. Luego se unieron los 3 datasets dejando al año como una variable más.

```
#Juntamos Datasets de Happiness en una nueva Variable
h2015["year"] <- 2015
h2016["year"] <- 2016
h2017["year"] <- 2017

h2015[c("Lower.Confidence.Interval", "Upper.Confidence.Interval", "Whisker.high", "Whisker.low")] <- NA
h2016[c("Standard.Error", "Whisker.high", "Whisker.low")] <- NA
h2017[c("Standard.Error", "Lower.Confidence.Interval", "Upper.Confidence.Interval", "Region")] <- NA

hWorld <- rbind(h2015, h2016, h2017)
```

Countries Population

Este Dataset contiene la población de 217 países a lo largo de 56 años, desde 1960 hasta el 2016. A continuación se detallan las variables de este dataset.

No.	Variable	Tipo	Descripción
1	i..Country	Factor	ID del país
2	Country.Code	Factor	Código del país
3	Indicator.Name	Factor	Nombre del indicador
4	Indicator.Code	Factor	Código del indicador
5	year	chr	Año de medida
6	population	num	Población

El proceso de limpieza que se hizo con este dataset fue poner los años como valor en una fila ya que estos venían como columnas. Obteniendo 6 variables con 12,369 observaciones.

```
#Ordenamos pWorld
colnames(pWorld)[5:61] <- c("1960","1961","1962","1963","1964","1965","1966","1967","1968","1969","1970")
pWorld <- gather(pWorld, year, population, 5:61)
pWorld <- subset( pWorld, select = -X)
```

Countries of the World

Este dataset contiene información sobre población, región, tamaño del área, mortalidad infantil y más. Todos estos conjuntos de datos están formados por datos del gobierno de EE. UU. Las variables contenidas se detallan a continuación.

No.	Variable	Tipo	Descripción
1	Country	Factor	Nombre del país
2	Region	Factor	Región en la que se encuentra el país
3	Population	int	Población Actual
4	Area..sq.mi..	int	Área superficial en mi^2
5	Pop..Density..per.sq.mi..	num	Densidad Poblacional en $personas/mi^2$
6	Coastline..coast.area.ratio.	num	línea costera
7	Net.migration	num	Migración neta
8	Infant.mortality..per.1000.births.	num	Mortalidad infantil (por 1000 nacimientos)
9	GDP...per.capita.	int	Producto Interno Bruto
10	Literacy....	num	Porcentaje de Alfabetismo
11	Phones..per.1000.	num	Teléfonos por cada 1000 personas
12	Arable....	num	Porcentaje de tierra cultivable
13	Crops....	num	Porcentaje de cultivos
14	Other....	num	Otros
15	Climate	num	Clima
16	Birthrate	num	Tasa de Nacimiento
17	Deathrate	num	Tasa de Mortalidad
18	Agriculture	num	Agricultura
19	Industry	num	Industria
20	Service	num	Servicio

Para este dataset fue necesario convertir algunas columnas de tipo Factos a Numeric, por lo que se reemplazó la “,” por “.”. Obteniendo un dataset con 227 observaciones y 20 variables.

```
#Convertimos Factors a num/int reemplazando "," por "."
cWorld$Coastline..coast.area.ratio.<- as.numeric(gsub(",", ".", gsub("\\.", "", cWorld$Coastline..coast
cWorld$Pop..Density..per.sq..mi.. <- as.numeric(gsub(",", ".", gsub("\\.", "", cWorld$Pop..Density..per
cWorld$Net.migration <- as.numeric(gsub(",", ".", gsub("\\.", "", cWorld$Net.migration)))
cWorld$Infant.mortality..per.1000.births. <- as.numeric(gsub(",", ".", gsub("\\.", "", cWorld$Infant.mo
cWorld$Literacy.... <- as.numeric(gsub(",", ".", gsub("\\.", "", cWorld$Literacy....)))
cWorld$Phones..per.1000. <- as.numeric(gsub(",", ".", gsub("\\.", "", cWorld$Phones..per.1000.)))
cWorld$Crops.... <- as.numeric(gsub(",", ".", gsub("\\.", "", cWorld$Crops....)))
cWorld$Other.... <- as.numeric(gsub(",", ".", gsub("\\.", "", cWorld$Other....)))
cWorld$Climate <- as.numeric(gsub(",", ".", gsub("\\.", "", cWorld$Climate)))
cWorld$Arable.... <- as.numeric(gsub(",", ".", gsub("\\.", "", cWorld$Arable....)))
cWorld$Birthrate <- as.numeric(gsub(",", ".", gsub("\\.", "", cWorld$Birthrate)))
cWorld$Deathrate <- as.numeric(gsub(",", ".", gsub("\\.", "", cWorld$Deathrate)))
cWorld$Agriculture <- as.numeric(gsub(",", ".", gsub("\\.", "", cWorld$Agriculture)))
cWorld$Industry <- as.numeric(gsub(",", ".", gsub("\\.", "", cWorld$Industry)))
cWorld$Service <- as.numeric(gsub(",", ".", gsub("\\.", "", cWorld$Service)))

View(cWorld)
View(hWorld)
```

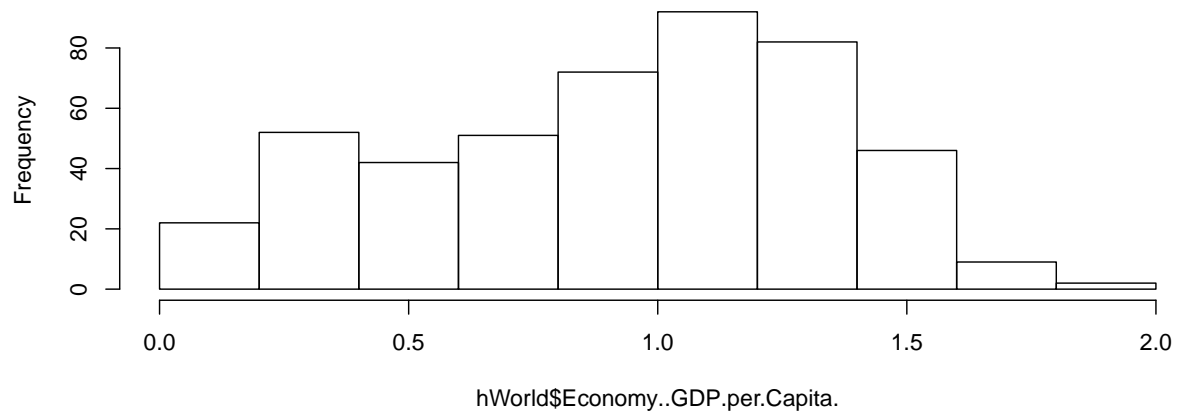
Exploración de Datos

World Happiness Report

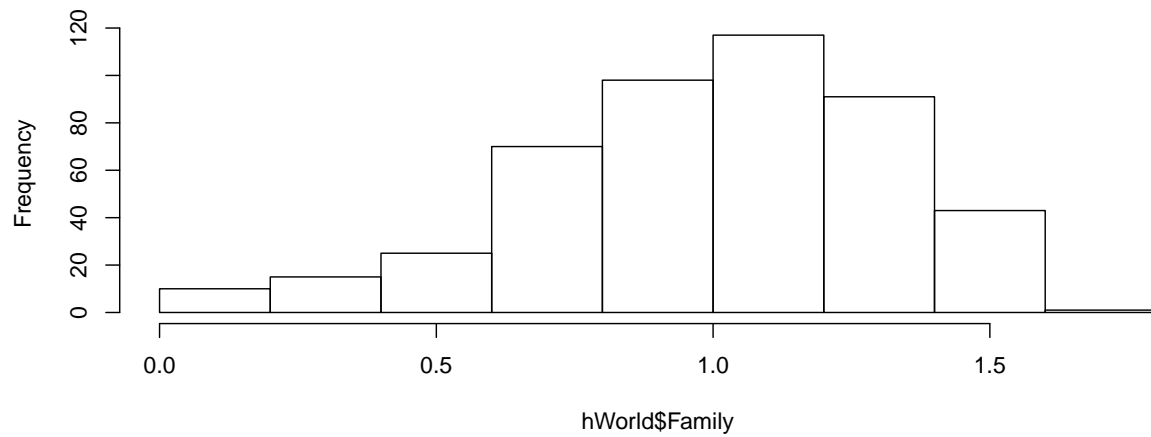
A continuación se muestran los histogramas de cada variable de *World Happiness Report*



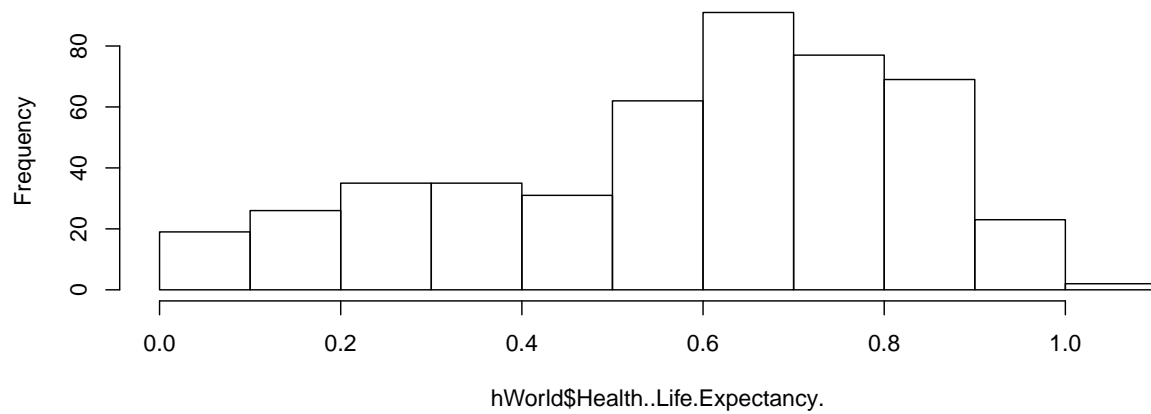
Histogram of hWorld\$Economy..GDP.per.Capita.

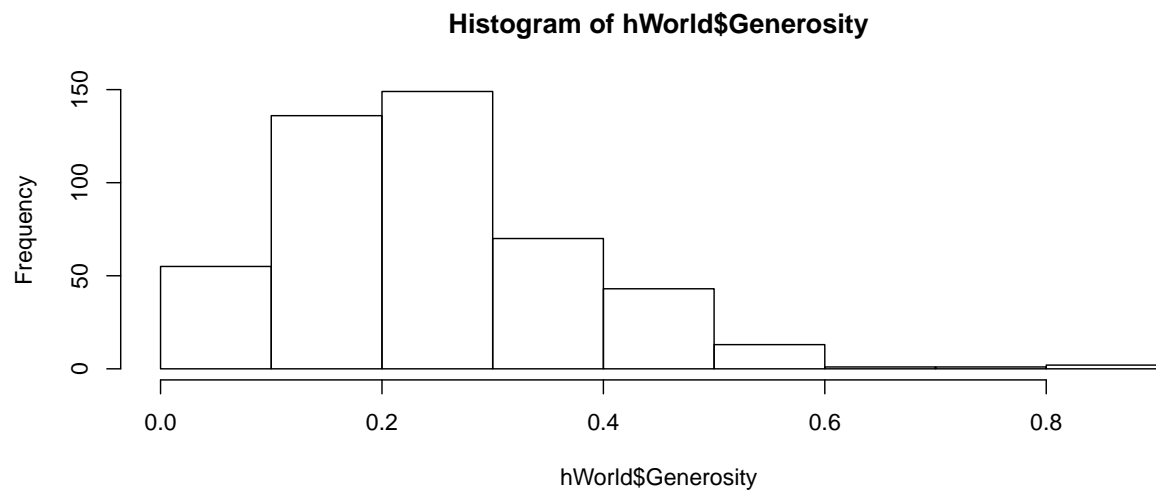
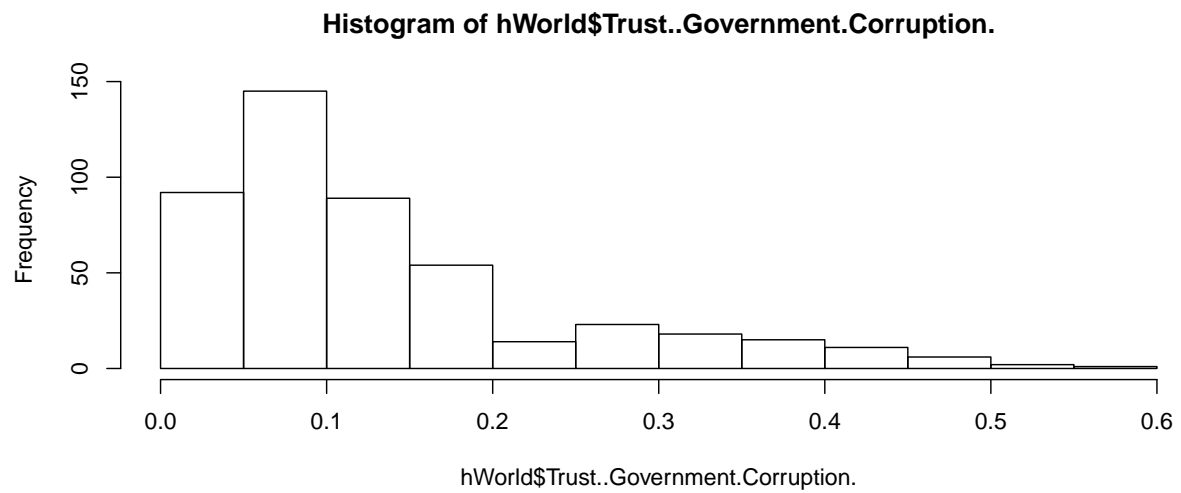
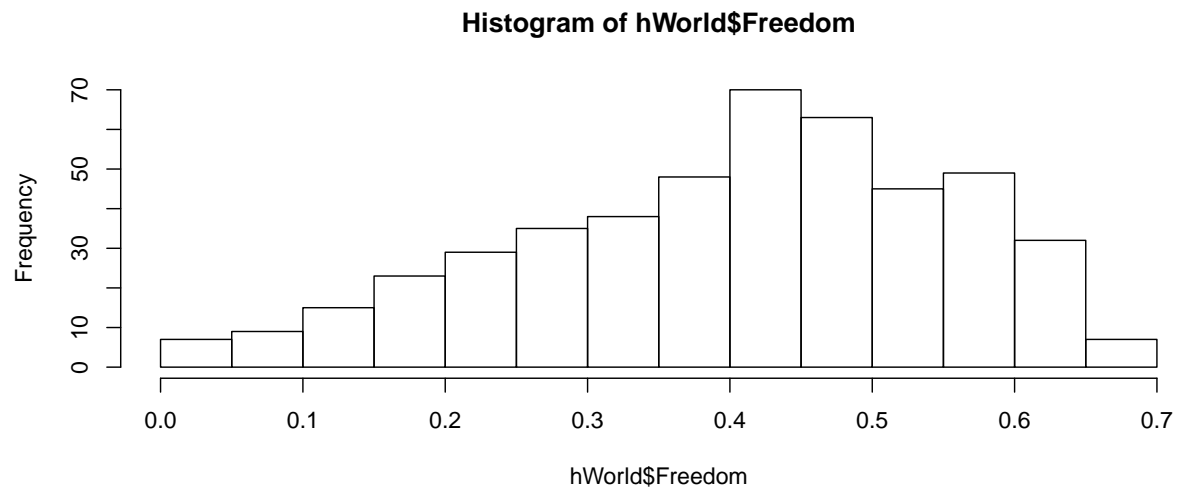


Histogram of hWorld\$Family



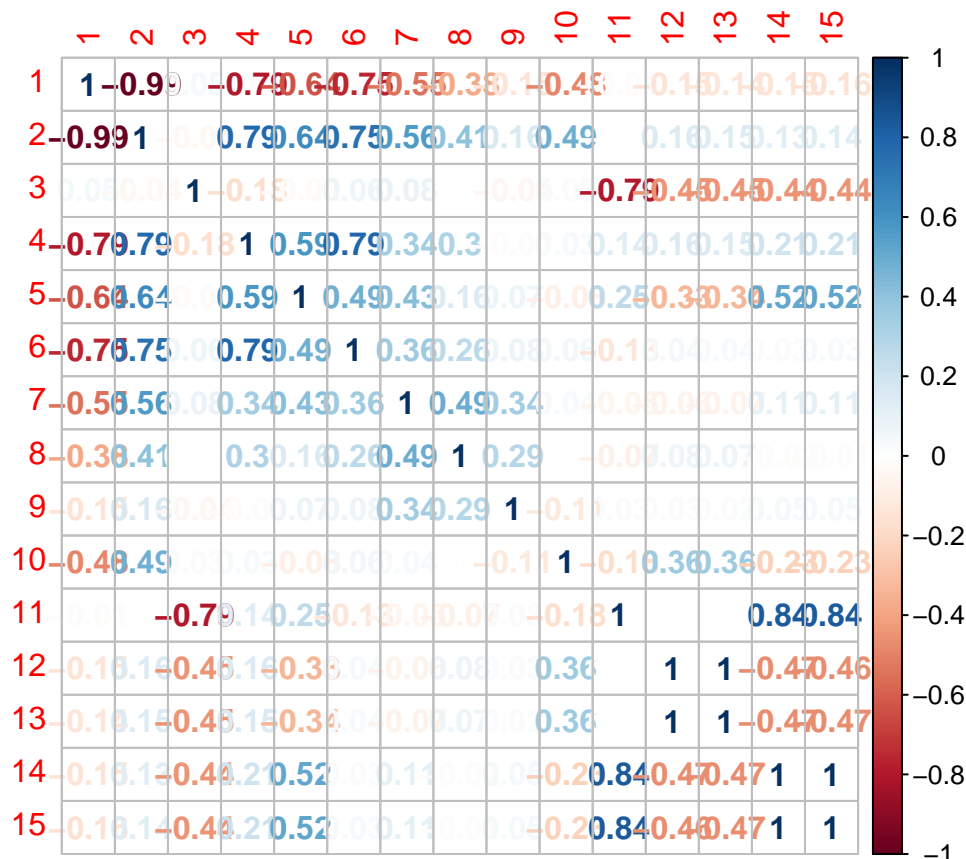
Histogram of hWorld\$Health..Life.Expectancy.





Correlación entre variables

Para poder observar si existe algún tipo de relación entre las variables numéricas del dataset, se realizó un gráfico de correlación en donde cada variable se puso contra cada variable del dataset. En el resultado podemos observar la correlación existente entre las variables:



En el diagrama de correlación se observó que existen coeficientes de correlaciones de varios valores, sin embargo, para este experimento los valores arriba de un 0.4 fueron considerados como significativos. Por lo que las relaciones que más se destacan son las siguientes:

- Happiness.Rank - Happiness.Score: 0.99
- Happiness.Rank - Economy..GDP.per.Capita: 0.79
- Happiness.Rank - Family: 0.64
- Happiness.Rank - Health..Life.Expectancy : 0.75
- Happiness.Rank - Freedom: 0.55
- Happiness.Rank - Dystopia.Residual: 0.55
- Happiness.Score - Economy..GDP.per.Capita: 0.79
- Happiness.Score - Family: 0.64
- Happiness.Score - Health..Life.Expectancy: 0.75
- Happiness.Score - Freedom: 0.56

- Happiness.Score - Trust..Government.Corruption: 0.41
- Happiness.Score - Dystopia.Residual: 0.49
- Economy..GDP.per.Capita - Family: 0.59
- Economy..GDP.per.Capita - Health..Life.Expectancy: 0.79
- Family - Health..Life.Expectancy: 0.49
- Family - Freedom: 0.43
- Freedom - Trust..Government.Corruption.: 0.49

Dentro de las correlaciones más significativas, podemos resaltar que la existente entre *Happiness.Rank-Happiness.Score* es de 0.99, esto se debe a que el puesto del rank depende directamente de la puntuación obtenida en el test. Así mismo podemos observar que la familia, la economía y la esperanza de vida tienen gran influencia en el rank de felicidad de los países.

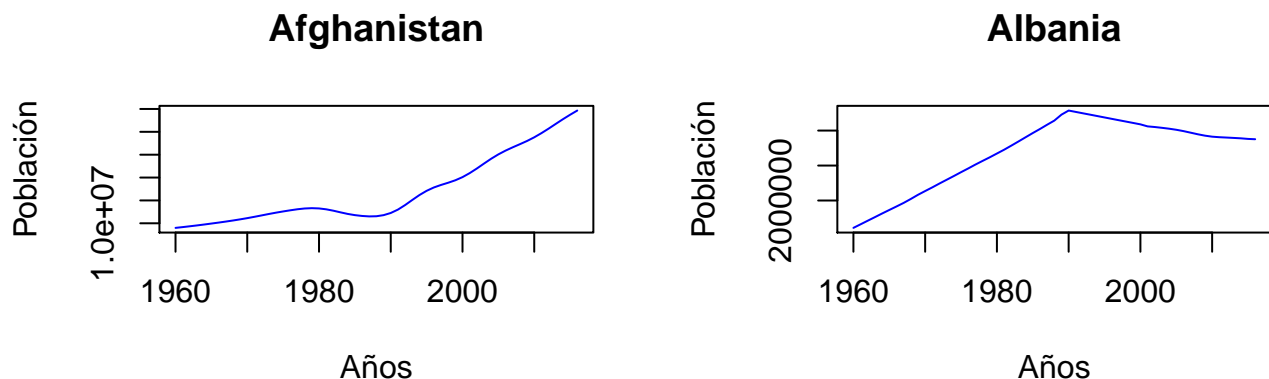
También es importante aclarar que las correlaciones existentes entre una variable y las variables *Lower.Confidence.Interval*, *Upper.Confidence.Interval*, *Whisker.high* y *Whisker.low* fueron omitidas debido a que estas variables no se encuentran registradas en más de un dataset.

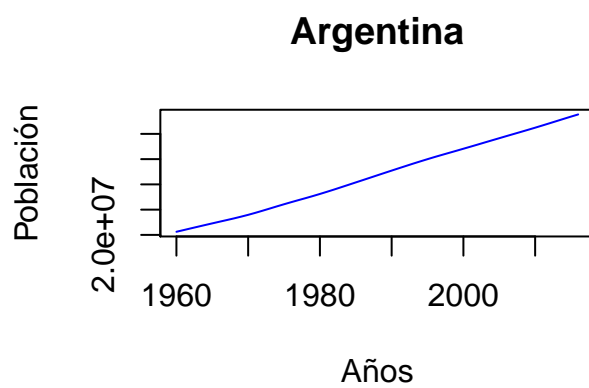
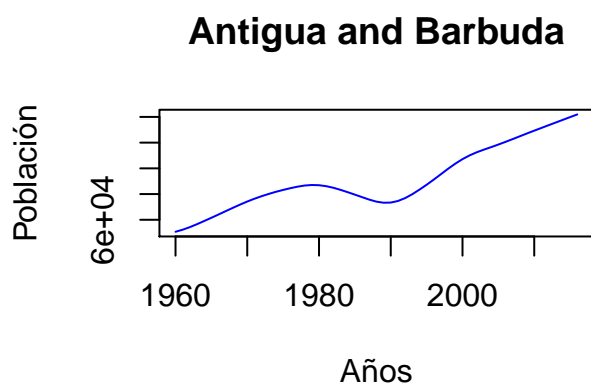
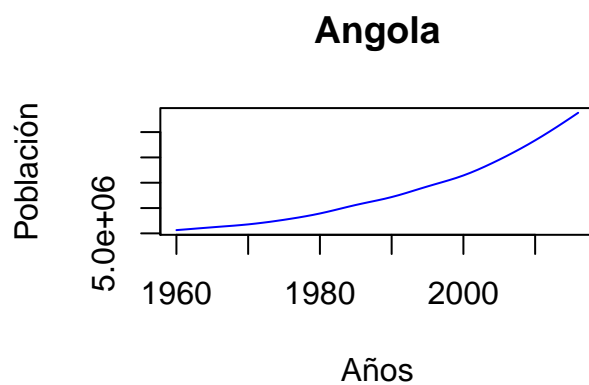
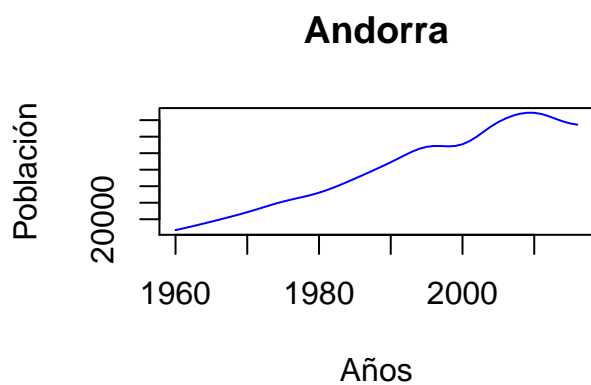
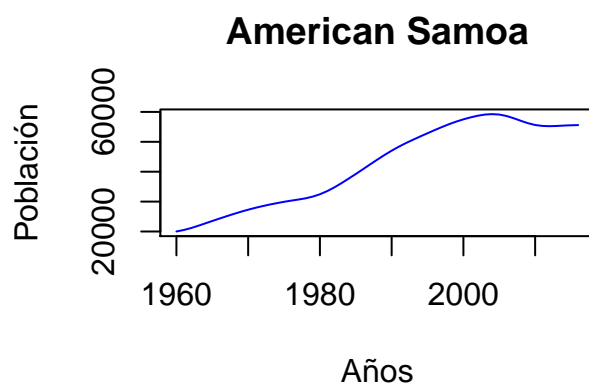
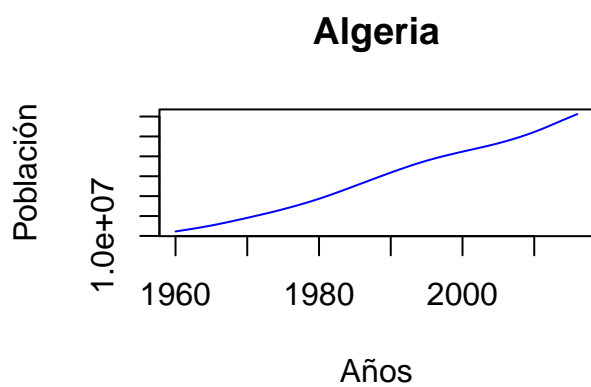
Countries Population

Los hallazgos encontrados en el dataset de Countries Population son simples, debido a que se encontró que la variable population, que describe la población de cada país, depende únicamente de la variable year. Por ende, crear un modelo que describa o pronostique la población de cada país consiste en tratar de encontrar un patrón y un modelo entre la variable population y la variable year. Debido a que solo se necesita de la variable year, tomada en diferentes unidades de tiempo, para describir cómo se comporta la población en cada país.

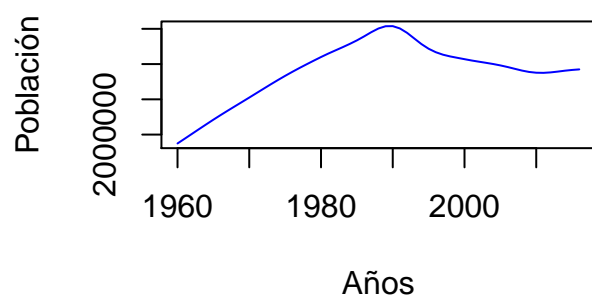
Gráficas de Población

A continuación se muestran las gráficas de población de algunos países:

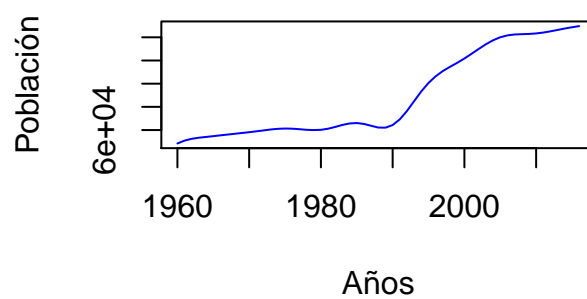




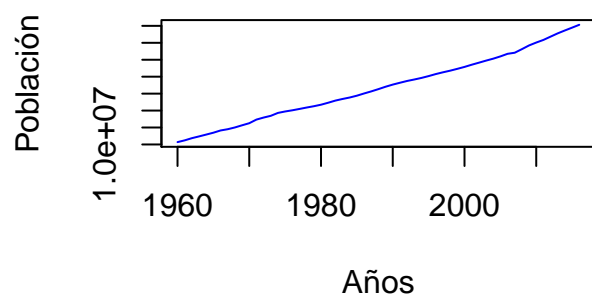
Armenia



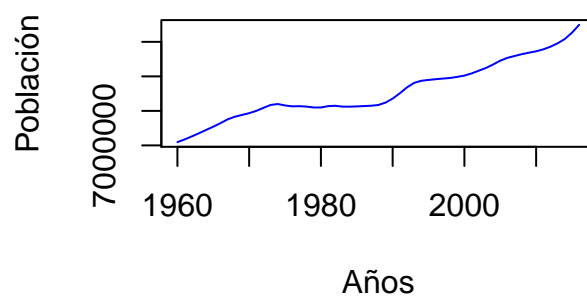
Aruba



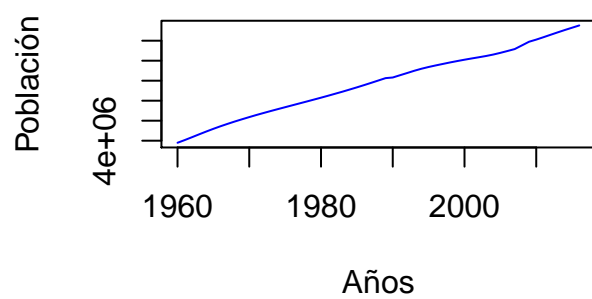
Australia



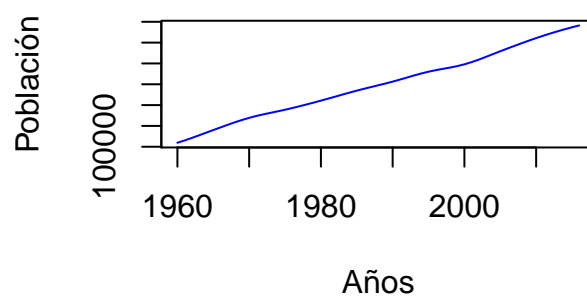
Austria

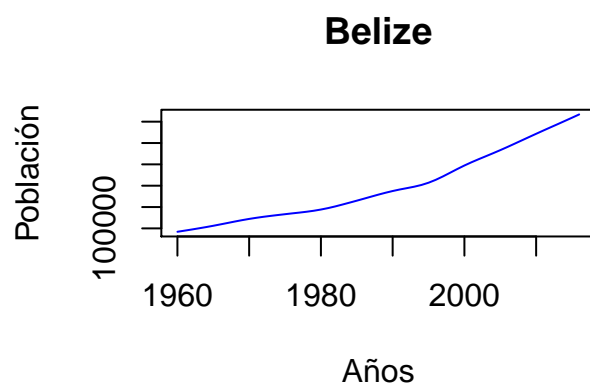
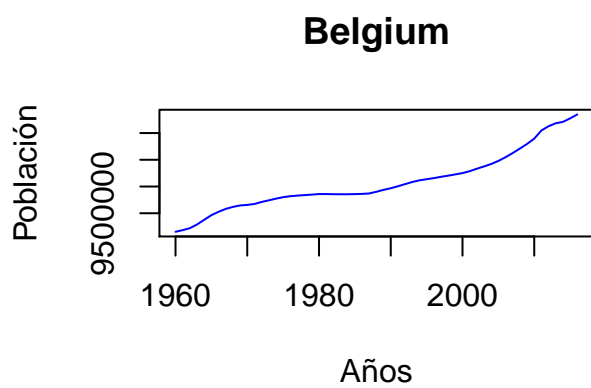
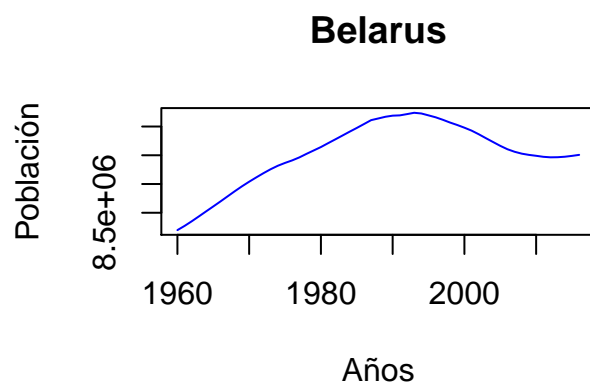
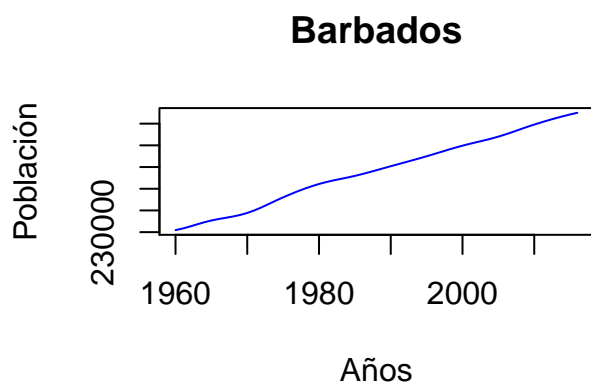
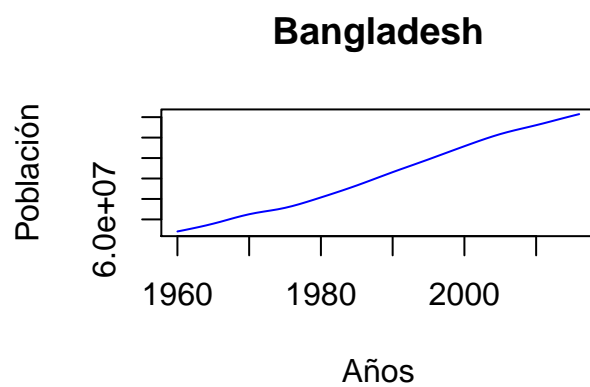
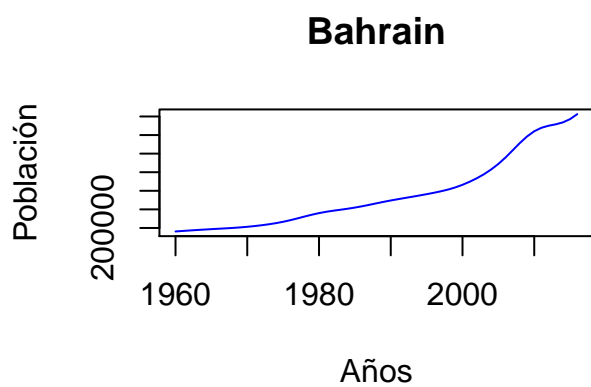


Azerbaijan

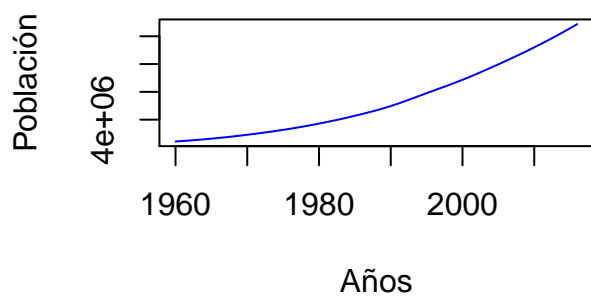


Bahamas, The

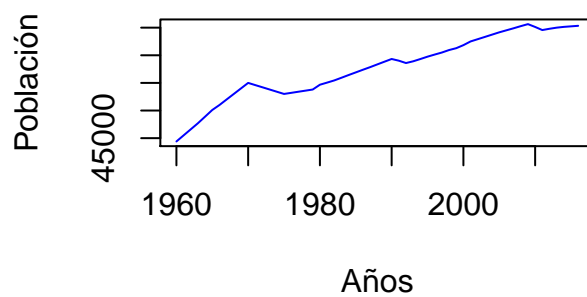




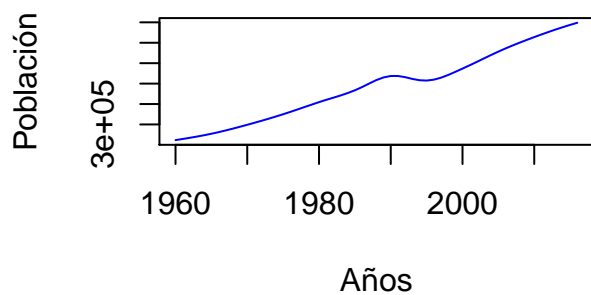
Benin



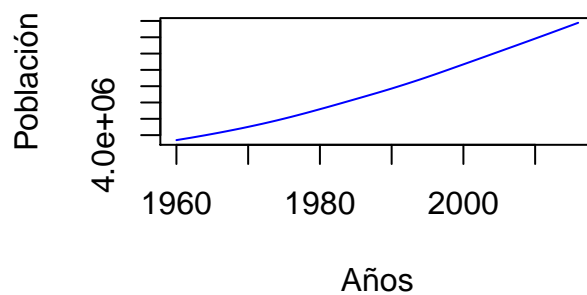
Bermuda



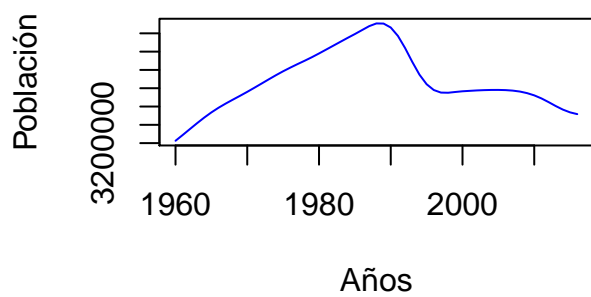
Bhutan



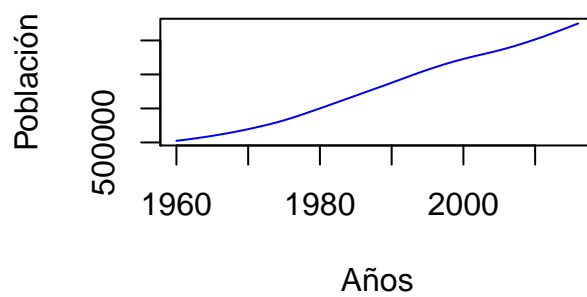
Bolivia



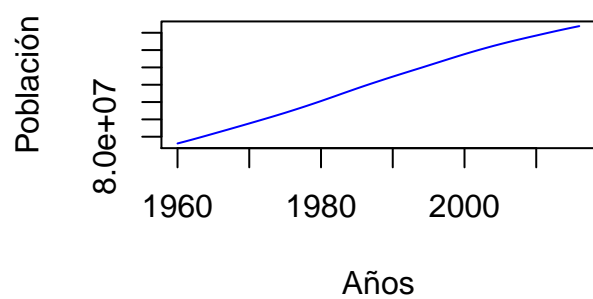
Bosnia and Herzegovina



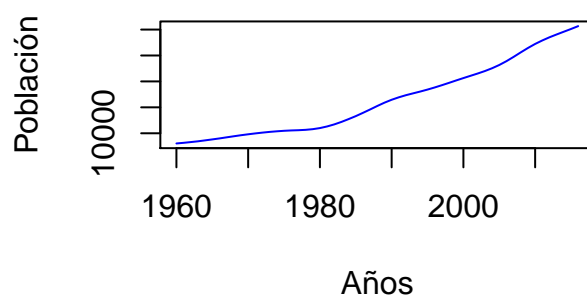
Botswana



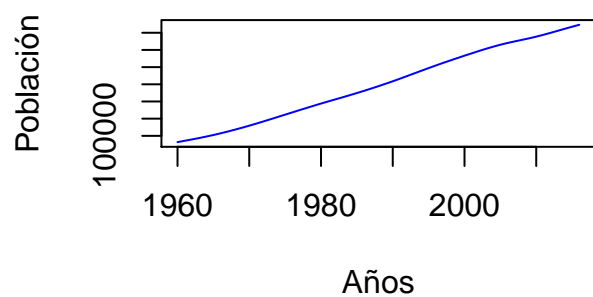
Brazil



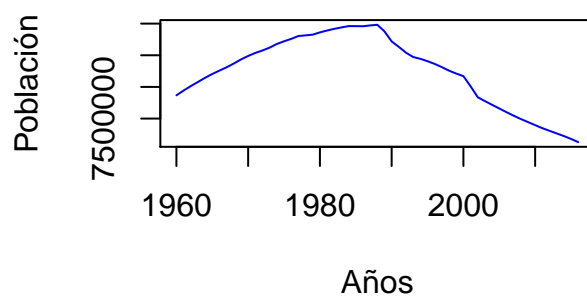
British Virgin Islands



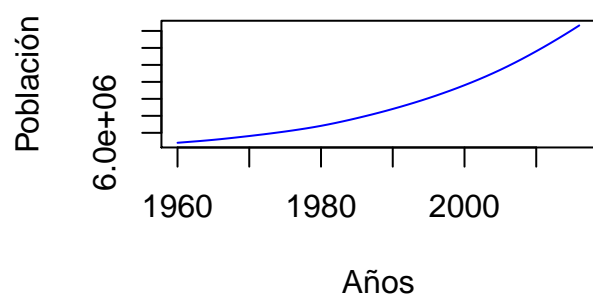
Brunei Darussalam



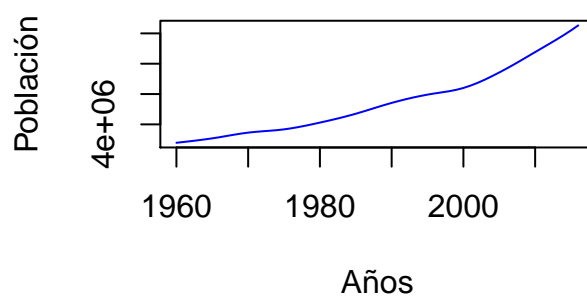
Bulgaria



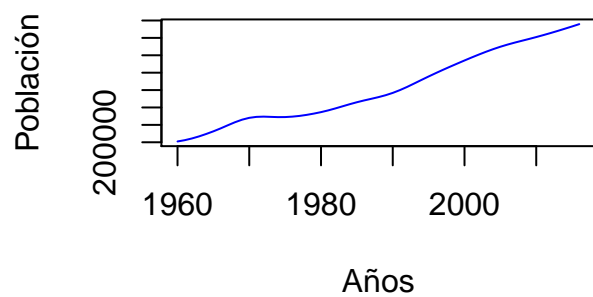
Burkina Faso



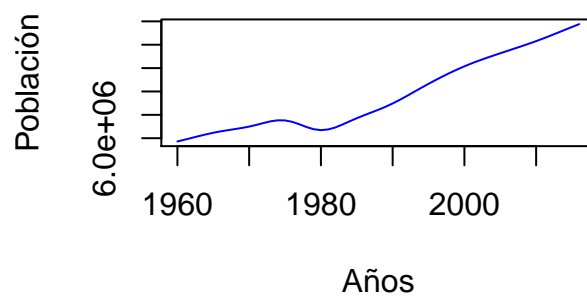
Burundi



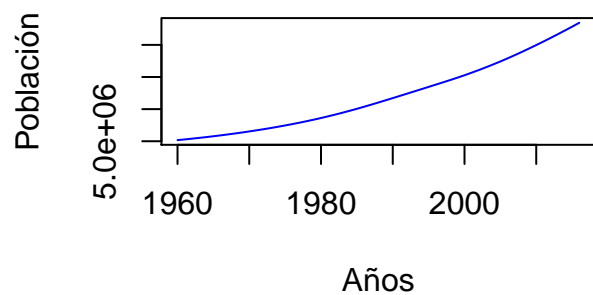
Cabo Verde



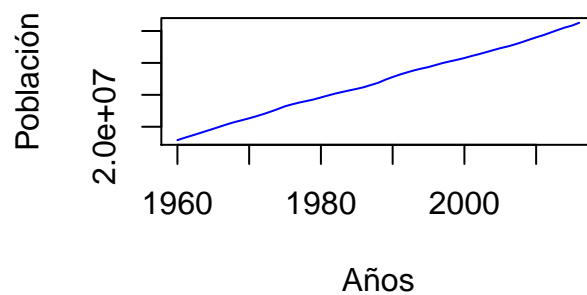
Cambodia



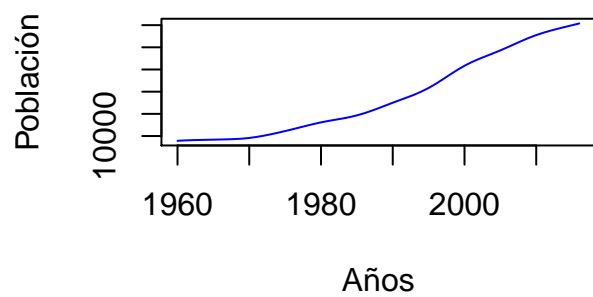
Cameroon



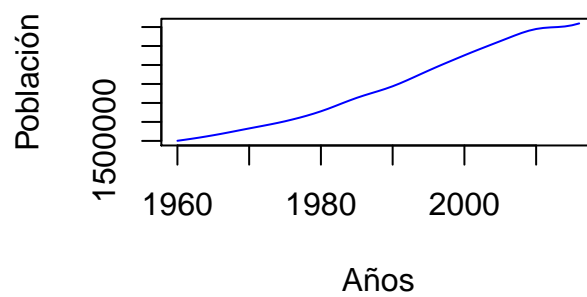
Canada



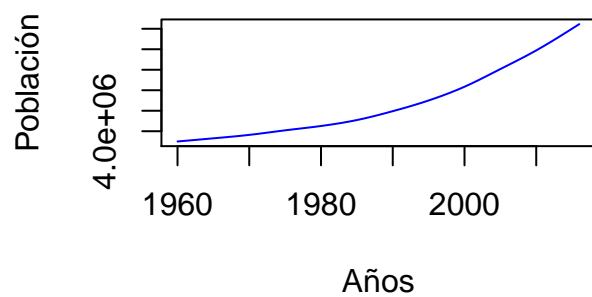
Cayman Islands



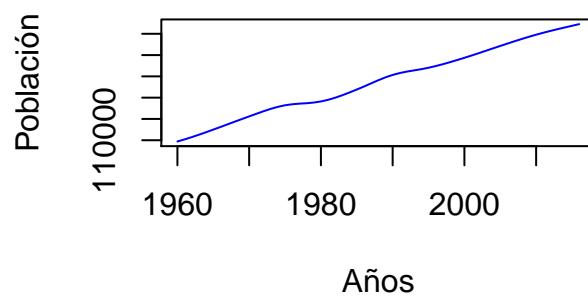
Central African Republic



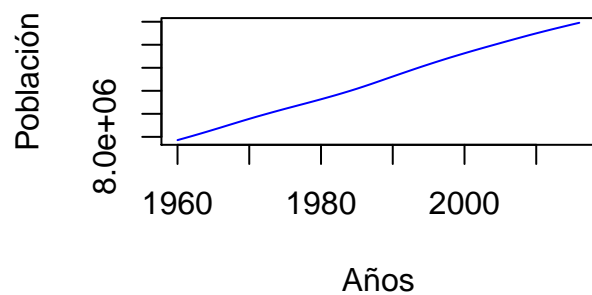
Chad



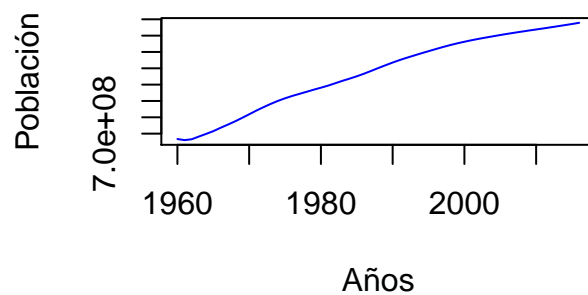
Channel Islands



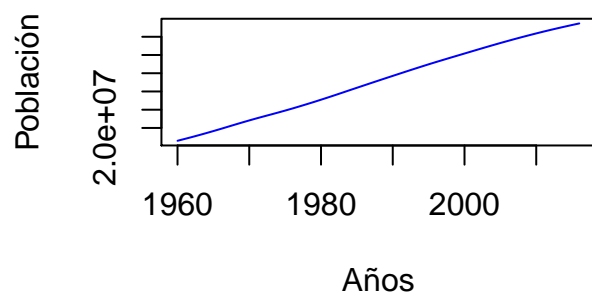
Chile



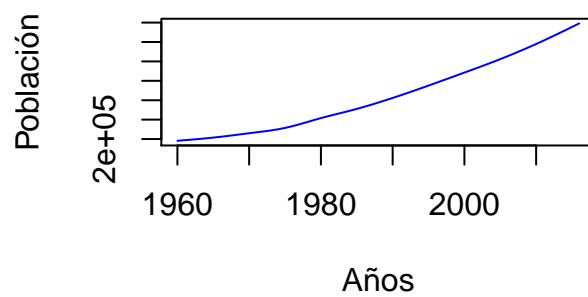
China



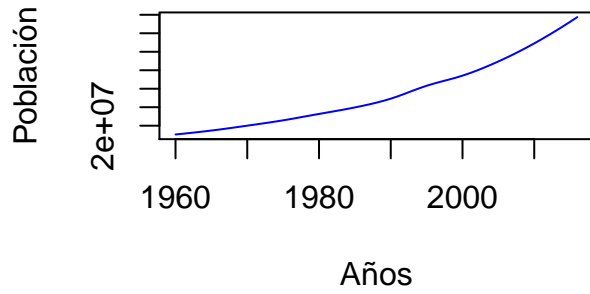
Colombia



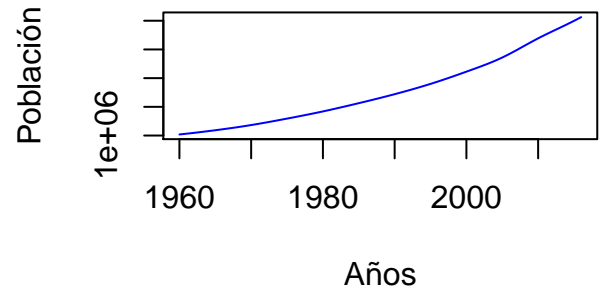
Comoros



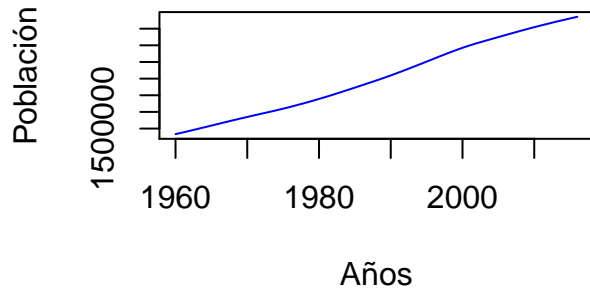
Congo, Dem. Rep.



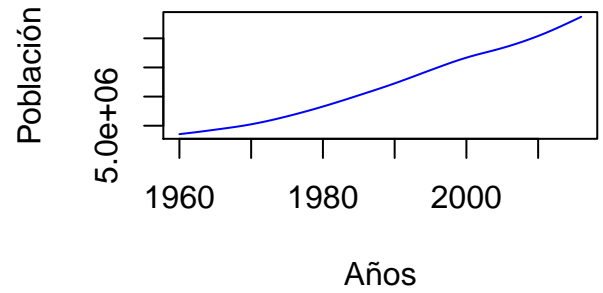
Congo, Rep.



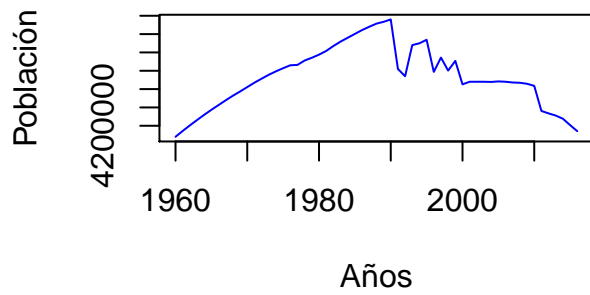
Costa Rica



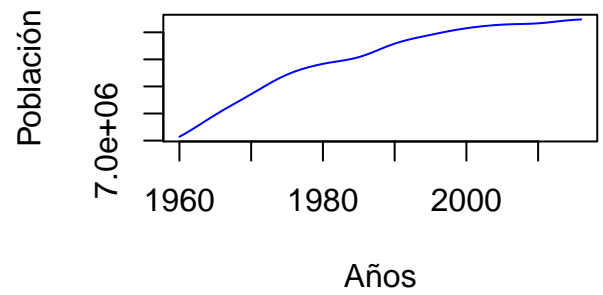
Cote d'Ivoire



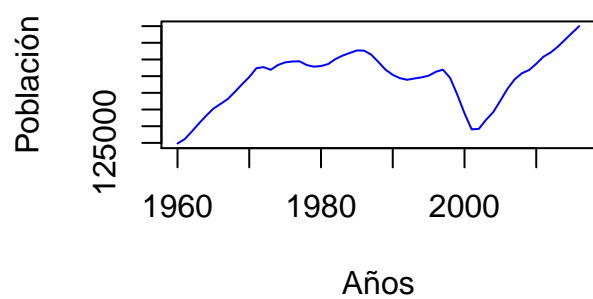
Croatia



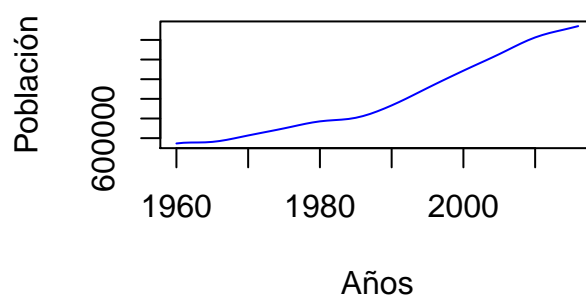
Cuba



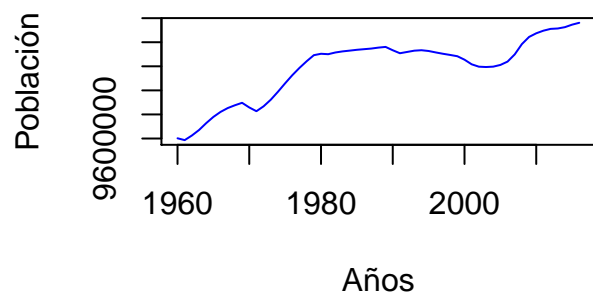
Curacao



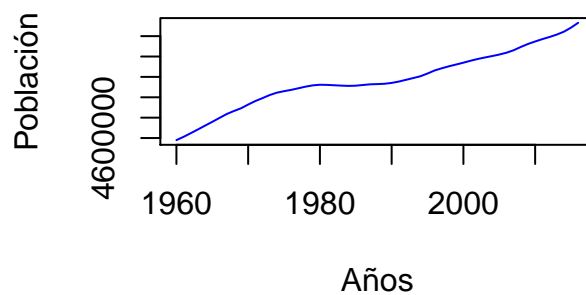
Cyprus



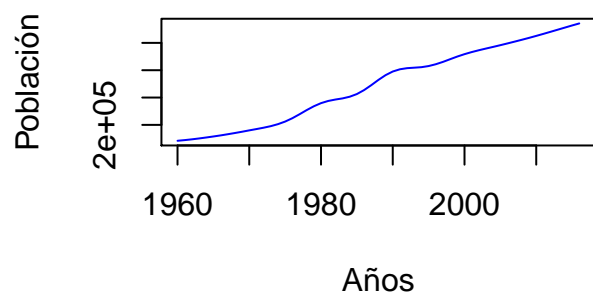
Czech Republic



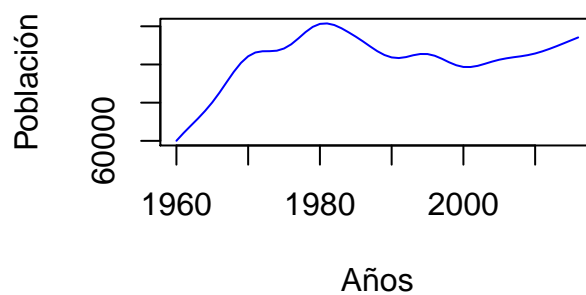
Denmark



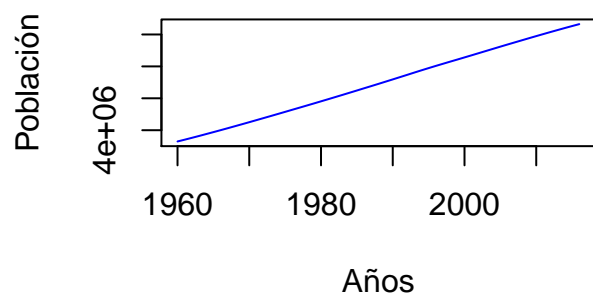
Djibouti



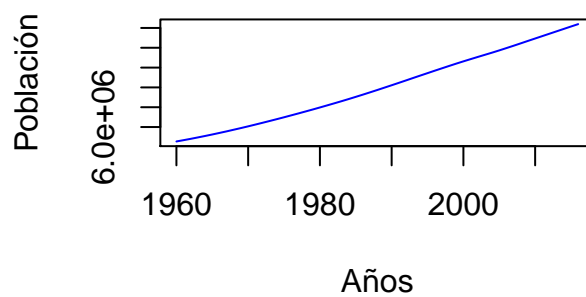
Dominica



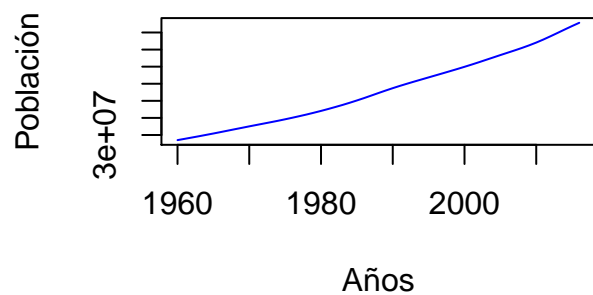
Dominican Republic



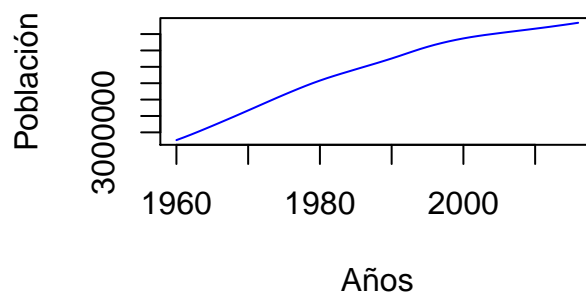
Ecuador



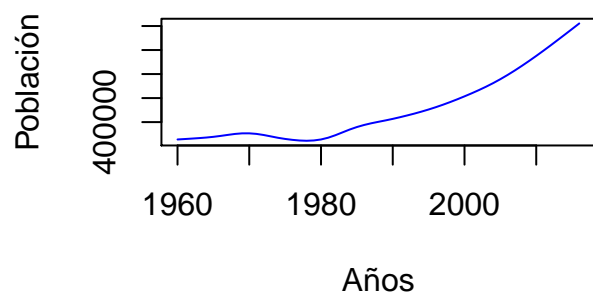
Egypt



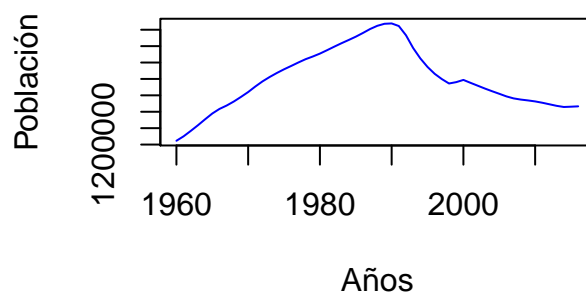
El Salvador

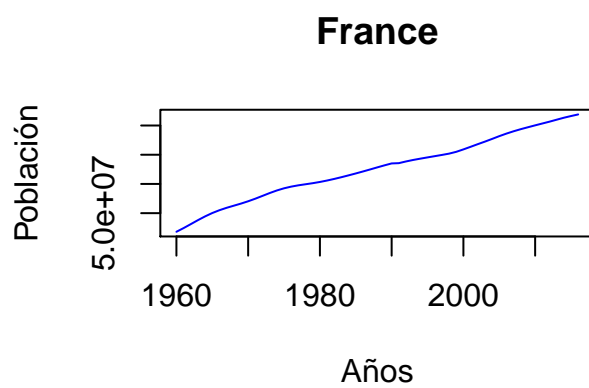
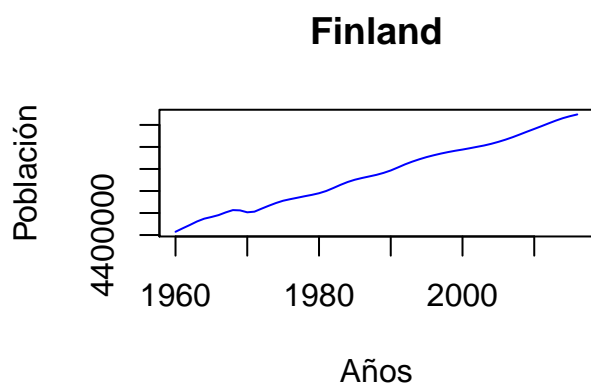
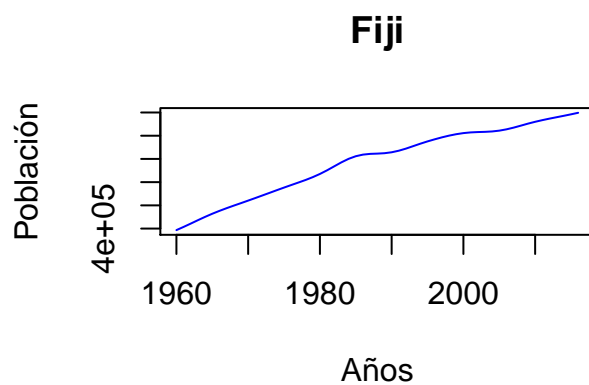
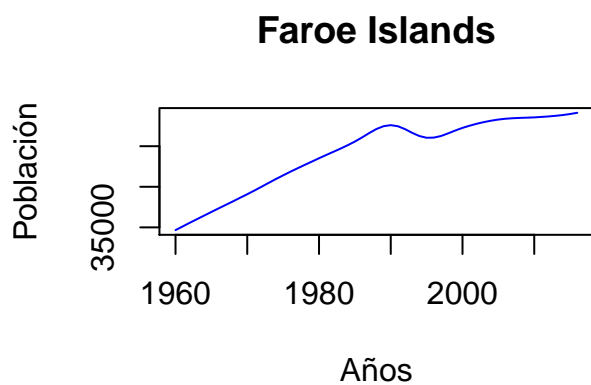
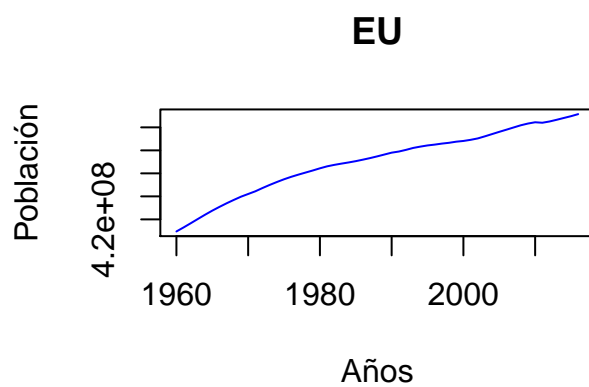
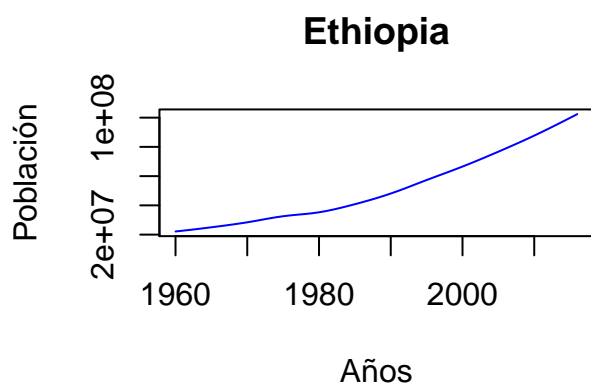


Equatorial Guinea

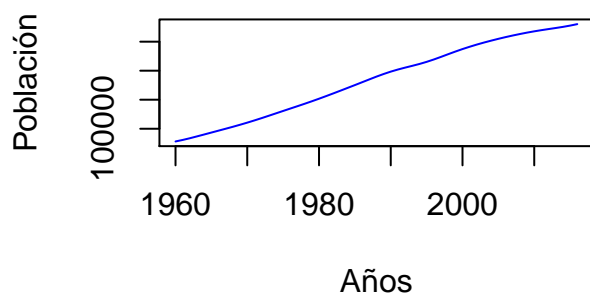


Estonia

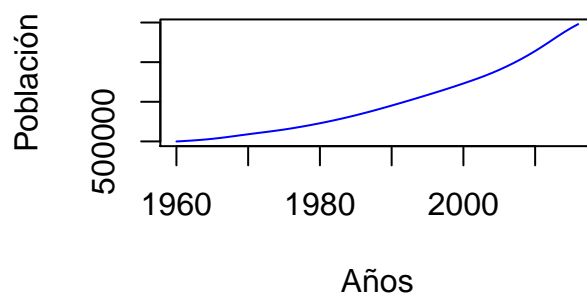




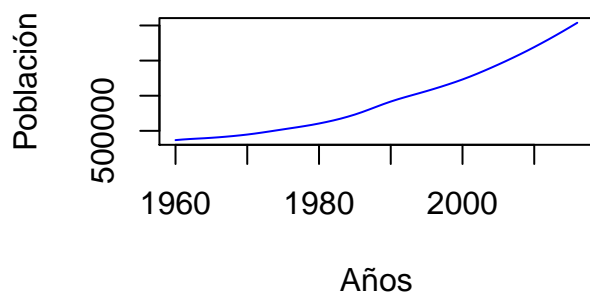
French Polynesia



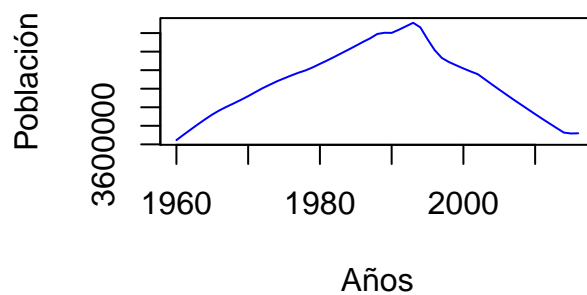
Gabon



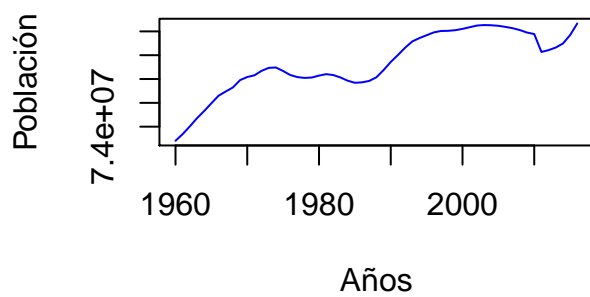
Gambia, The



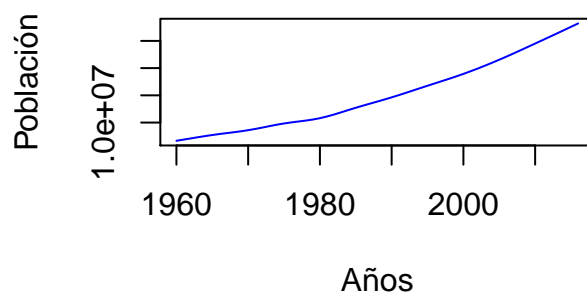
Georgia



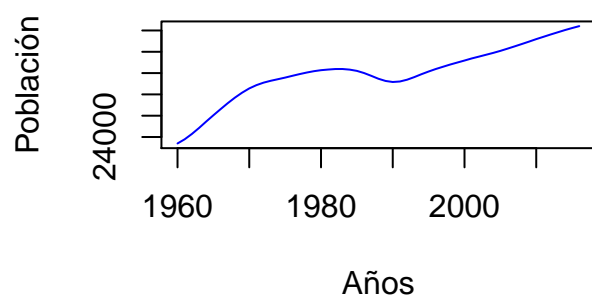
Germany



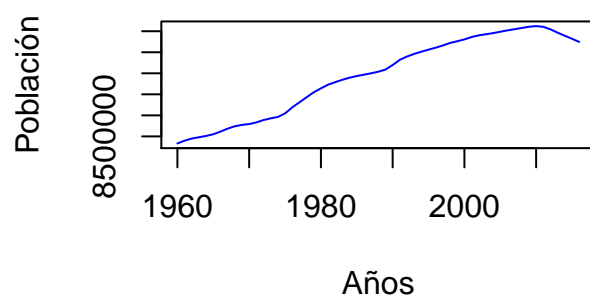
Ghana



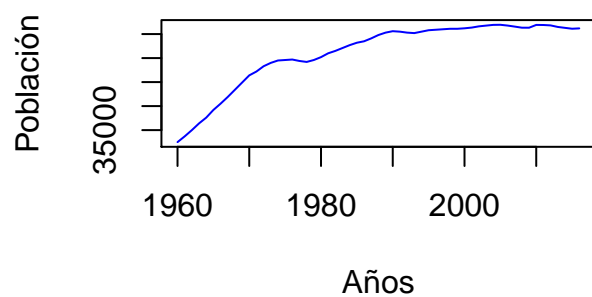
Gibraltar



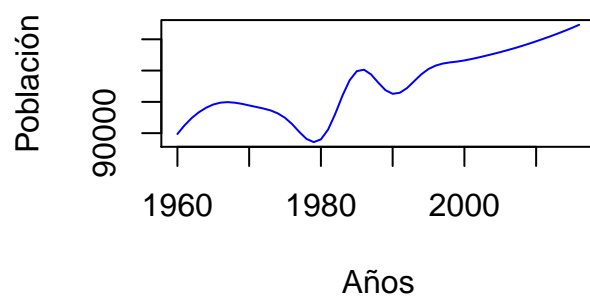
Greece



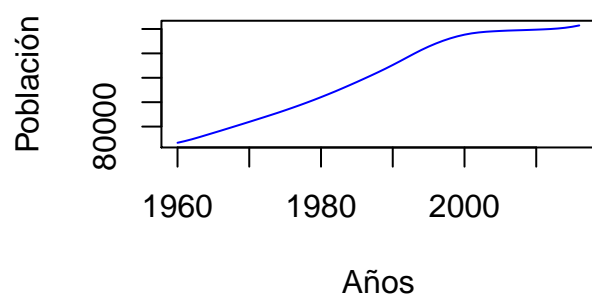
Greenland



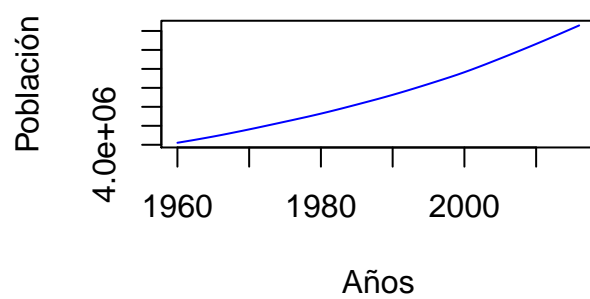
Grenada



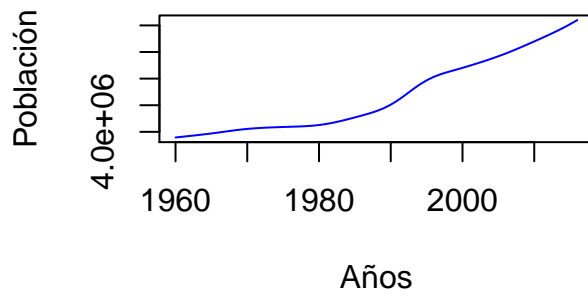
Guam



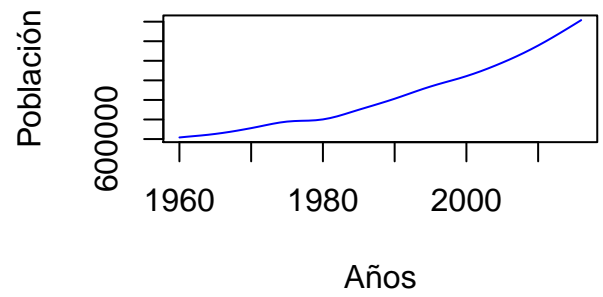
Guatemala



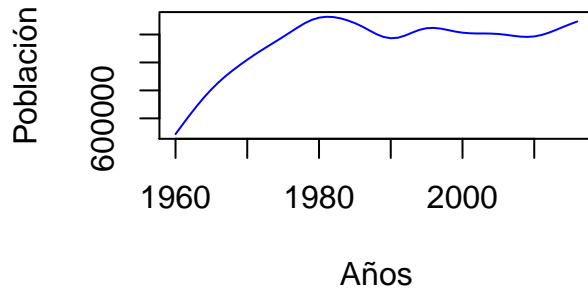
Guinea



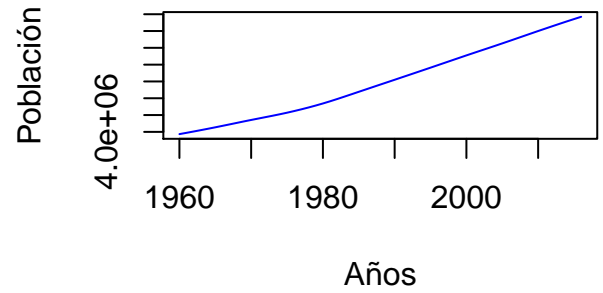
Guinea-Bissau



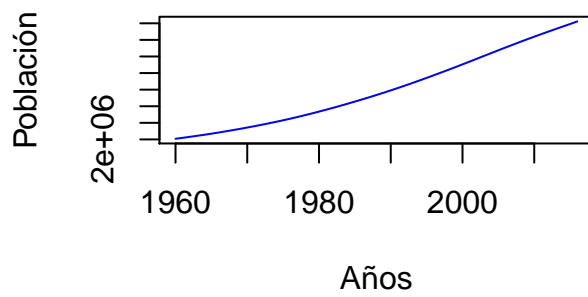
Guyana



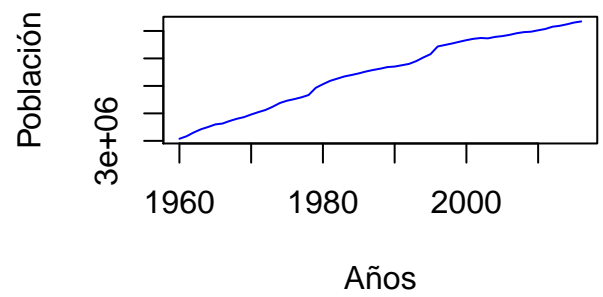
Haiti



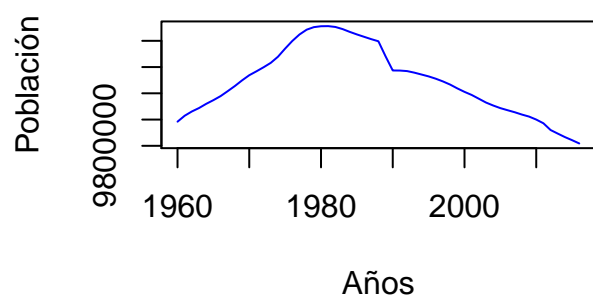
Honduras



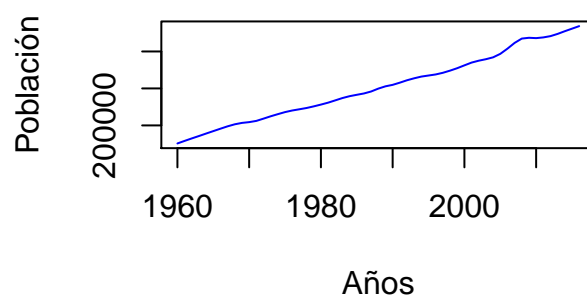
Hong Kong SAR, China



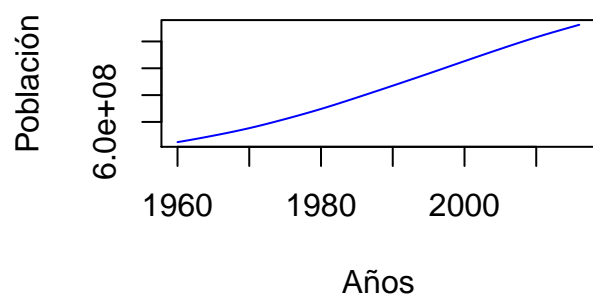
Hungary



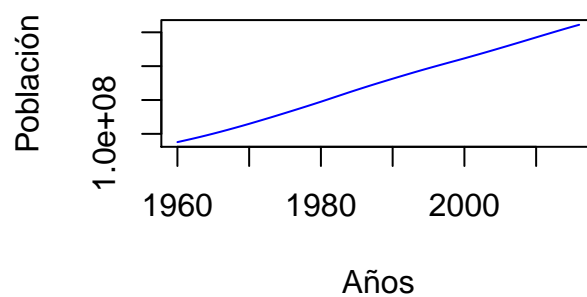
Iceland



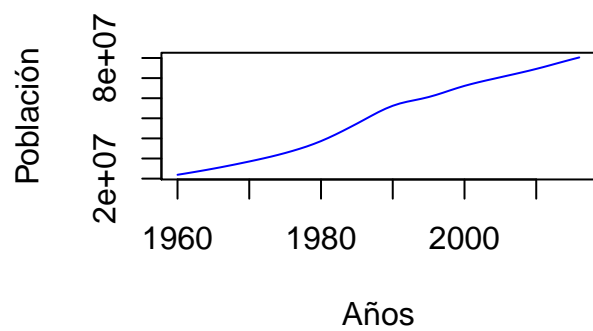
India



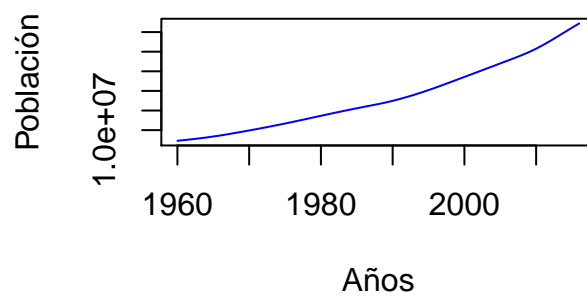
Indonesia

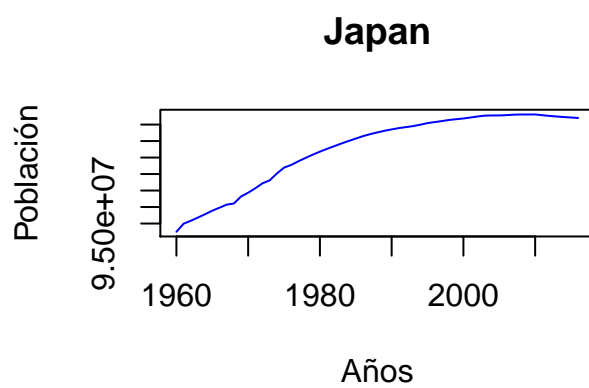
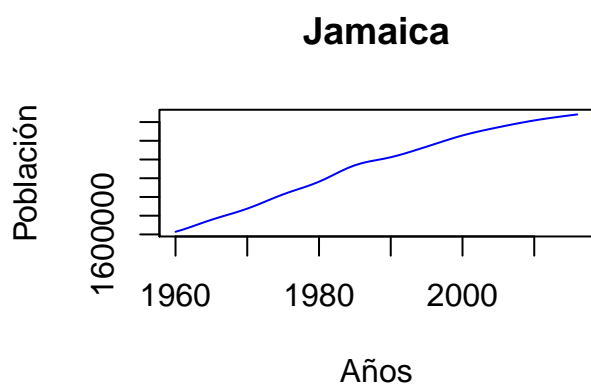
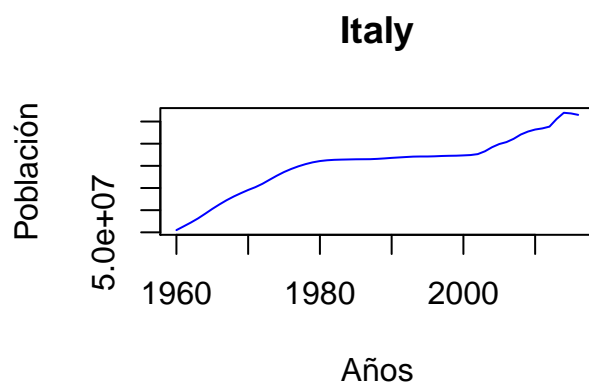
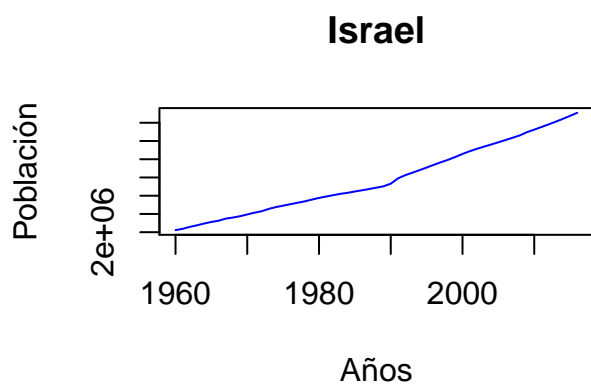
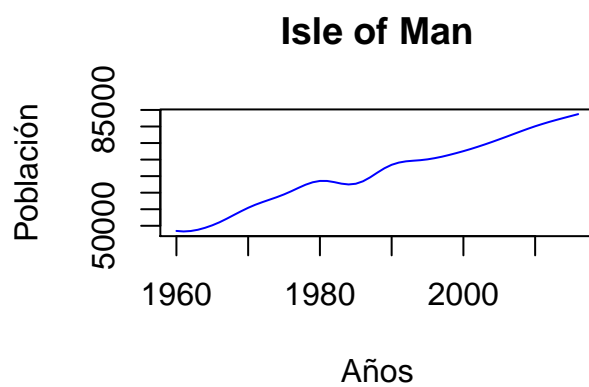
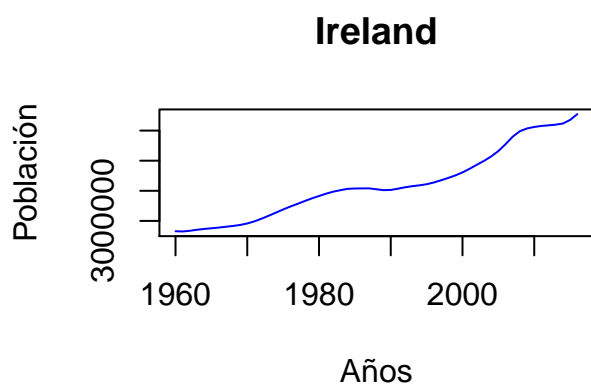


Iran

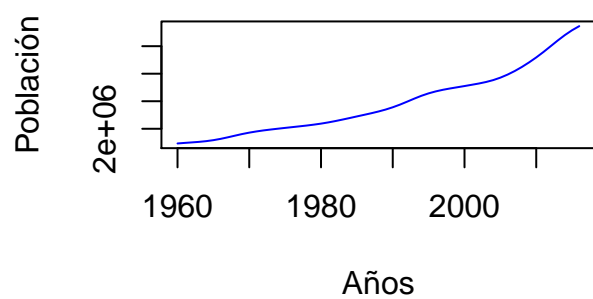


Iraq

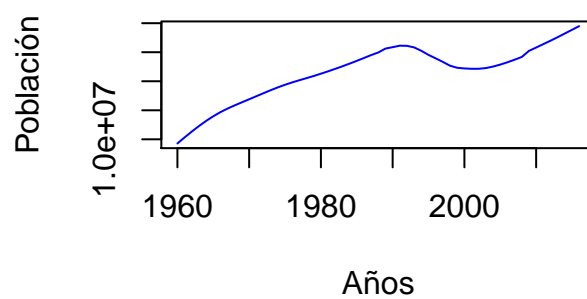




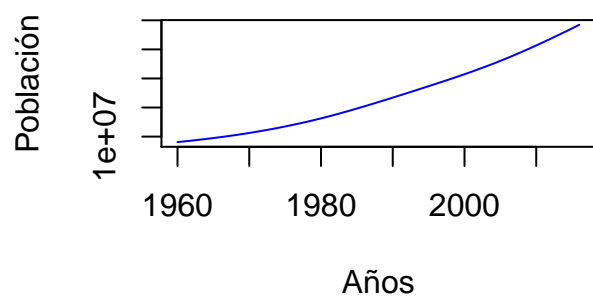
Jordan



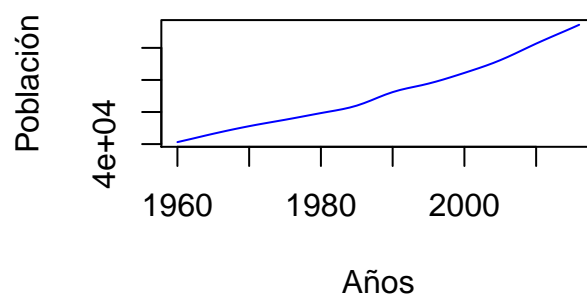
Kazakhstan



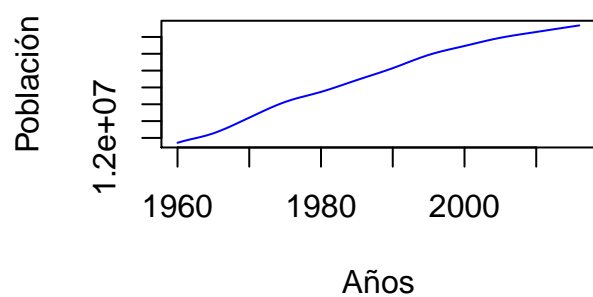
Kenya



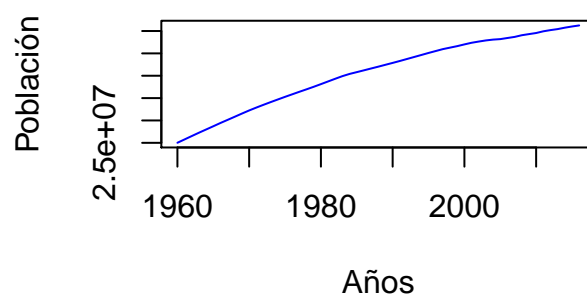
Kiribati



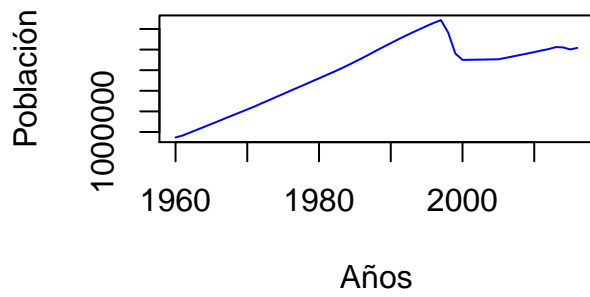
Korea, Dem. People Rep.



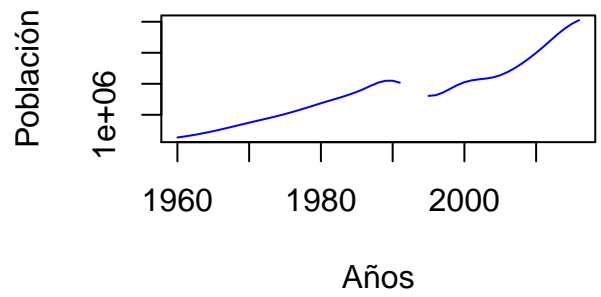
Korea, Rep.



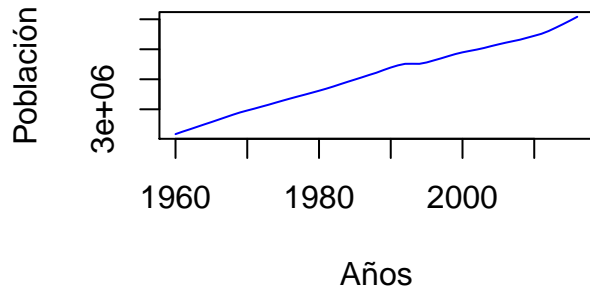
Kosovo



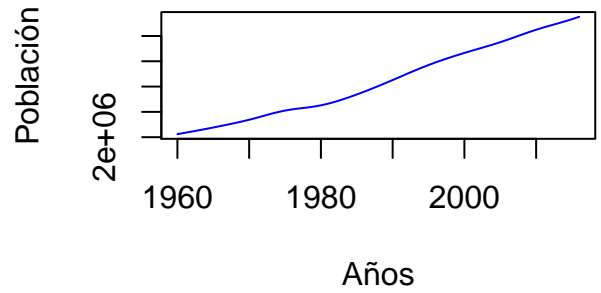
Kuwait



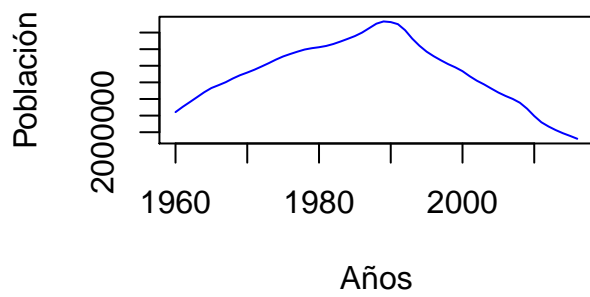
Kyrgyz Republic



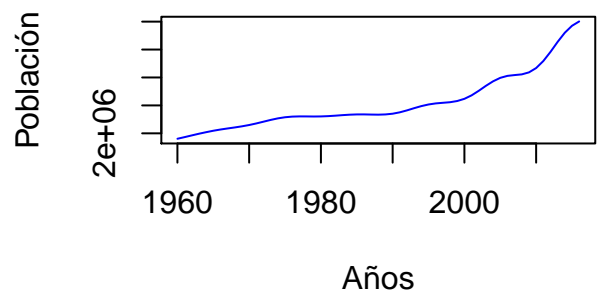
Lao PDR



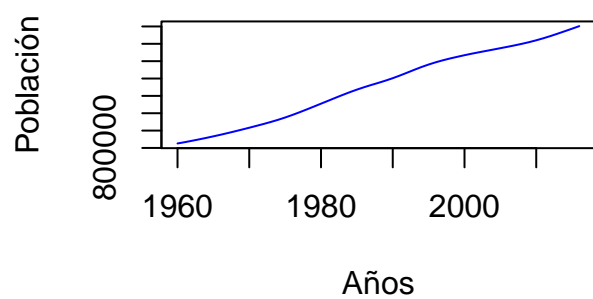
Latvia



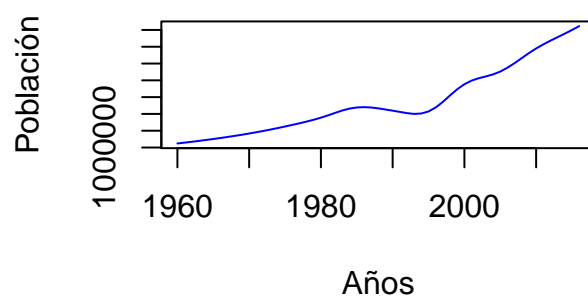
Lebanon



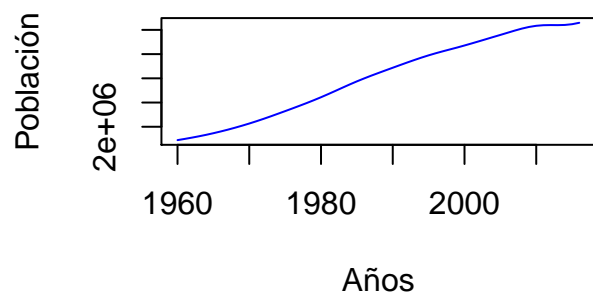
Lesotho



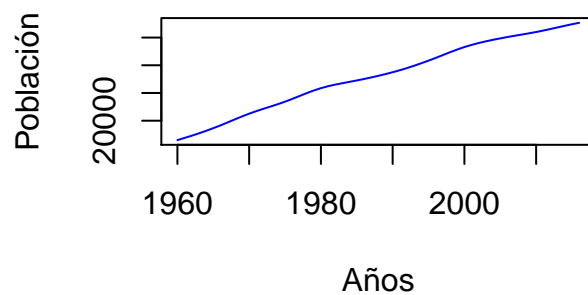
Liberia



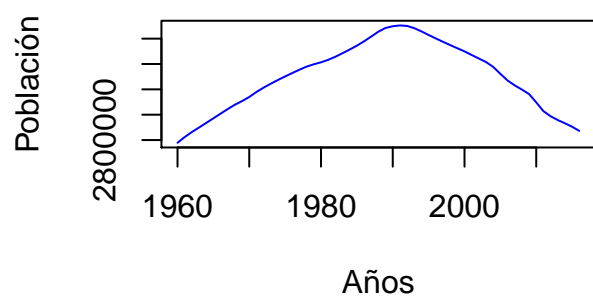
Libya



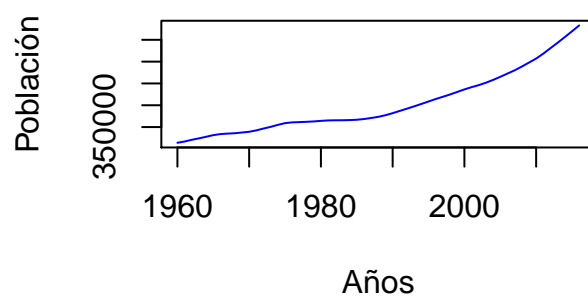
Liechtenstein



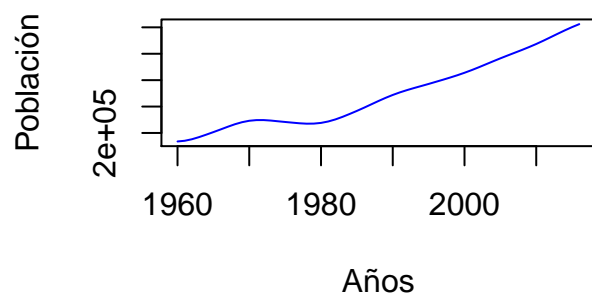
Lithuania



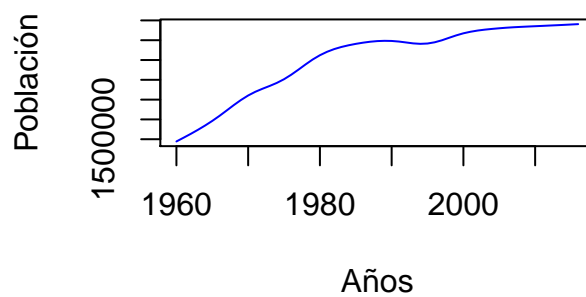
Luxembourg



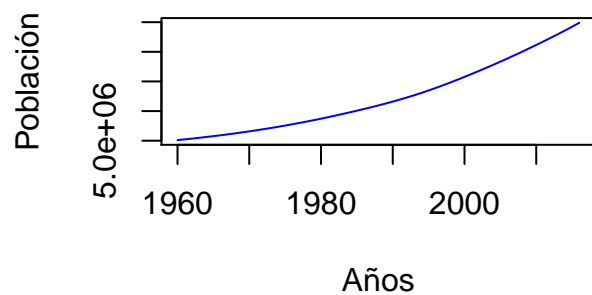
Macao SAR, China



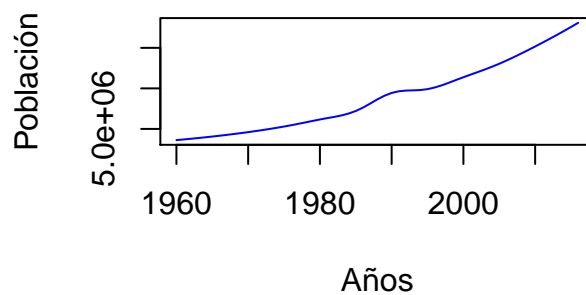
Macedonia, FYR



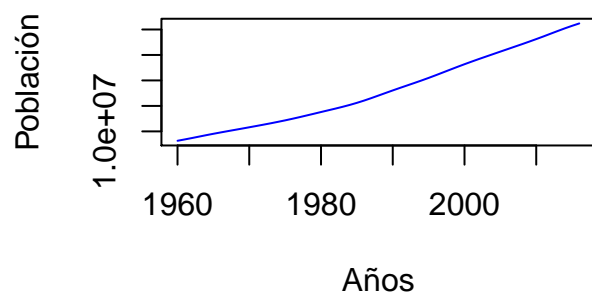
Madagascar



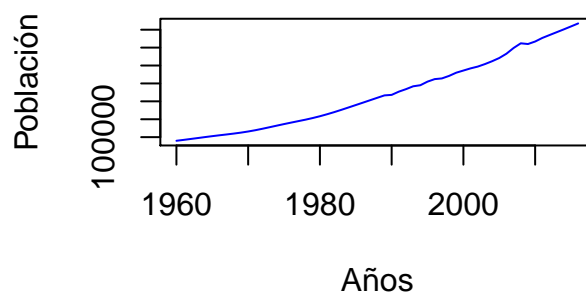
Malawi



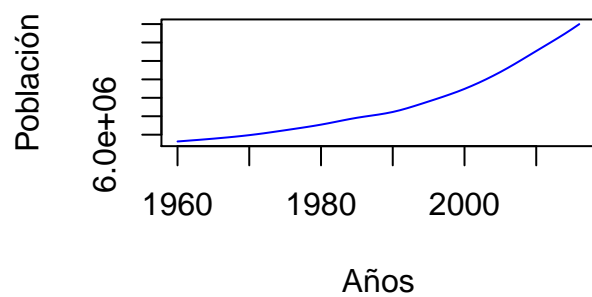
Malaysia



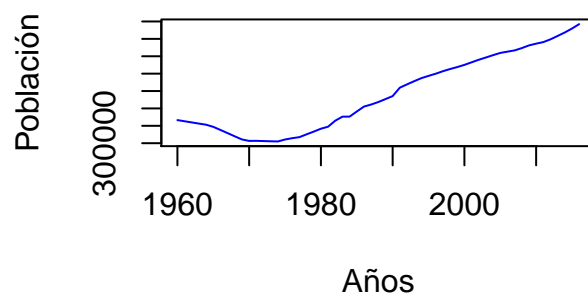
Maldives



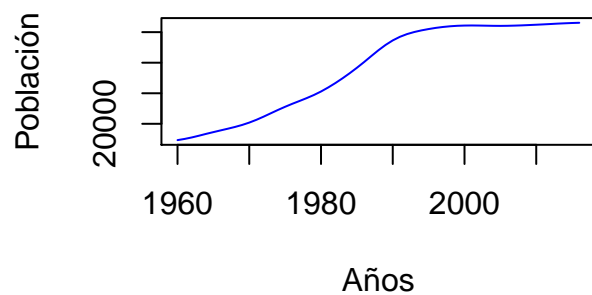
Mali



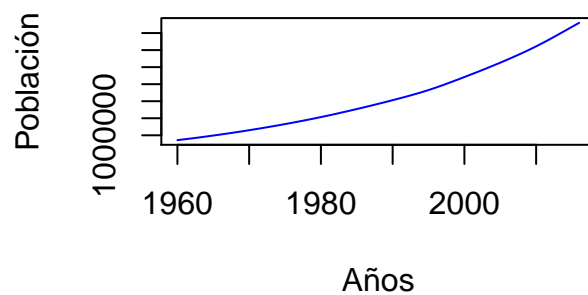
Malta



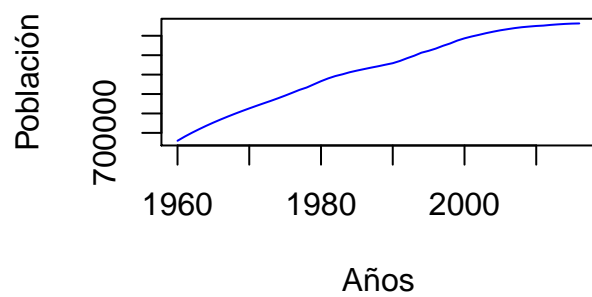
Marshall Islands



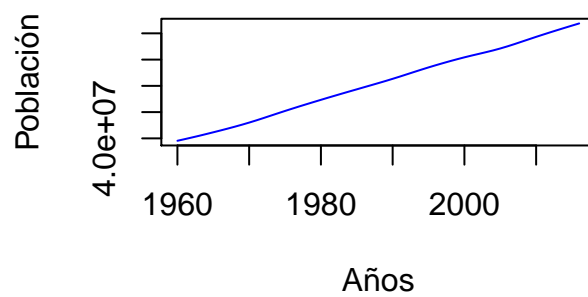
Mauritania



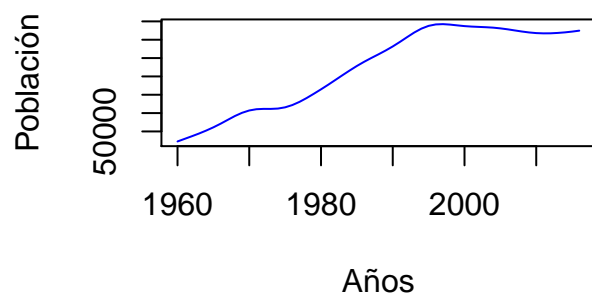
Mauritius



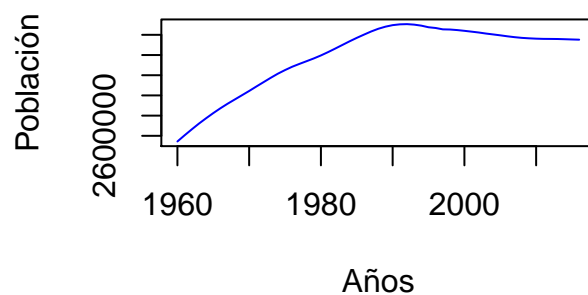
Mexico



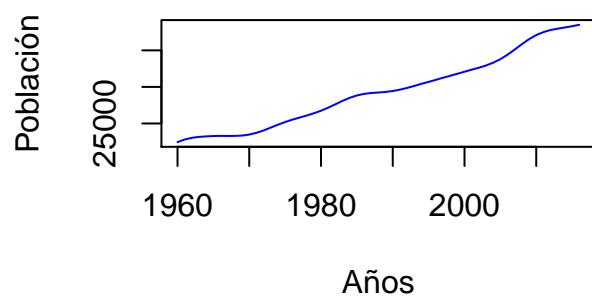
Micronesia, Fed. Sts.



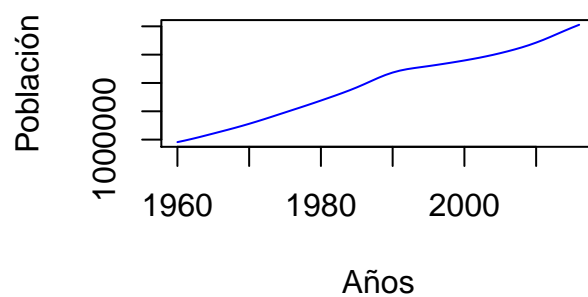
Moldova



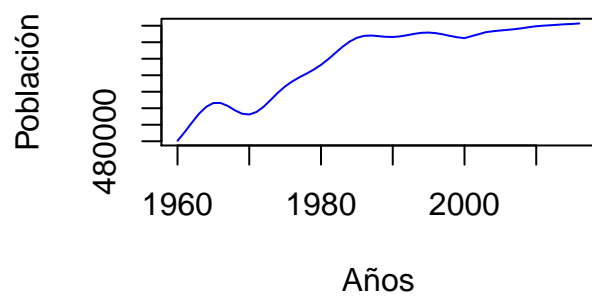
Monaco



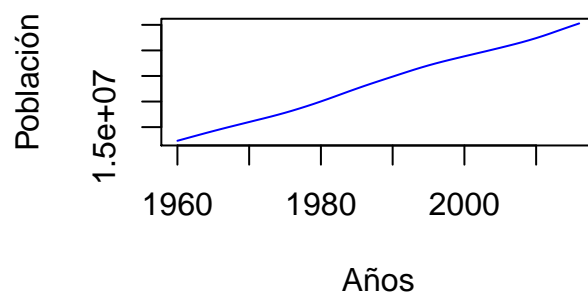
Mongolia



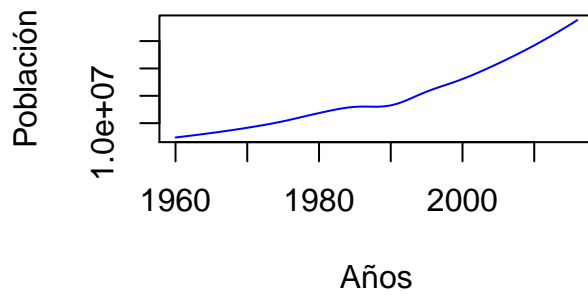
Montenegro



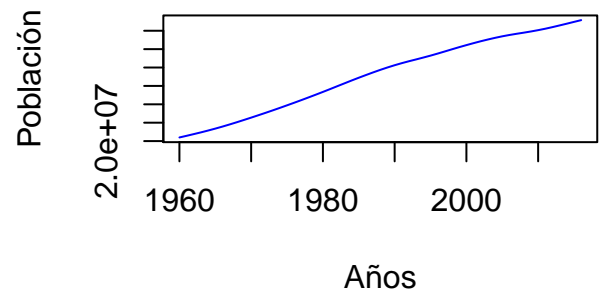
Morocco



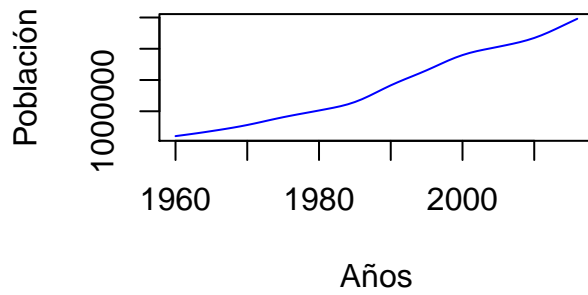
Mozambique



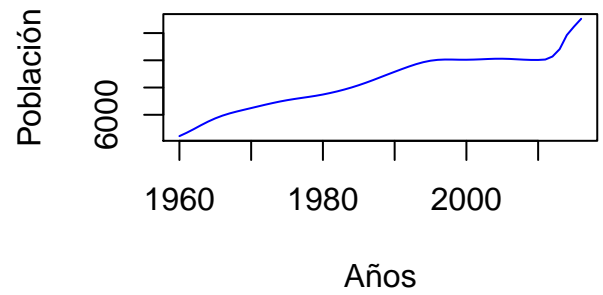
Myanmar



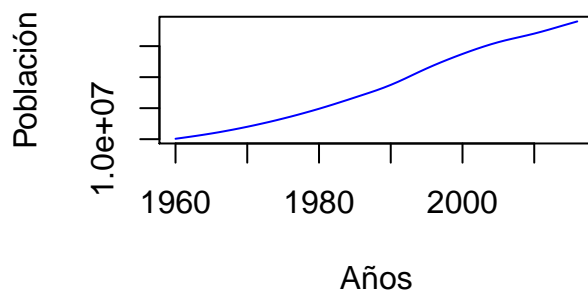
Namibia



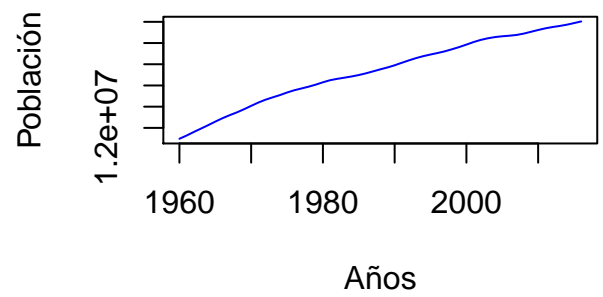
Nauru



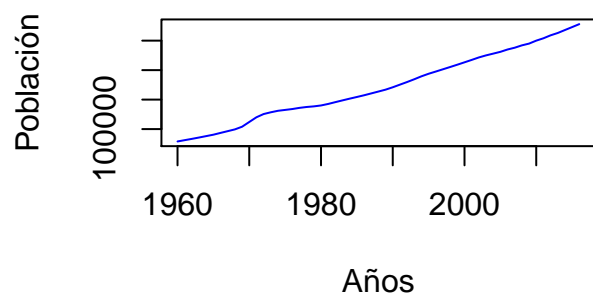
Nepal



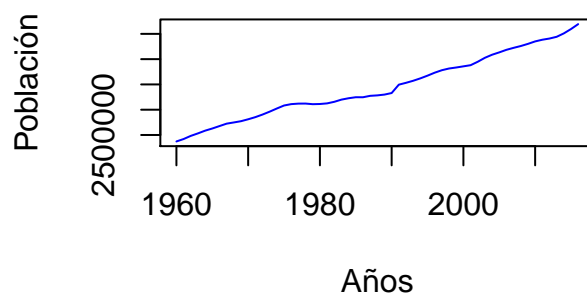
Netherlands



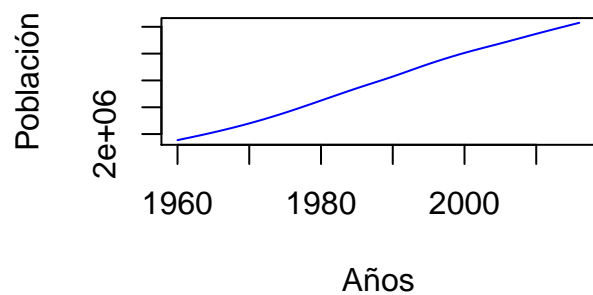
New Caledonia



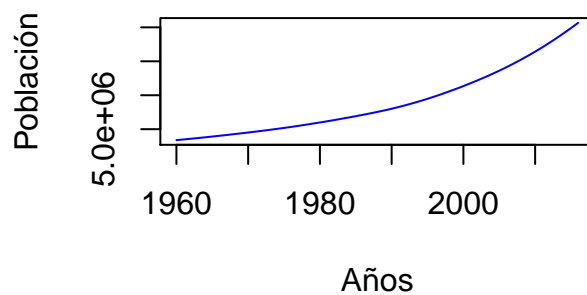
New Zealand



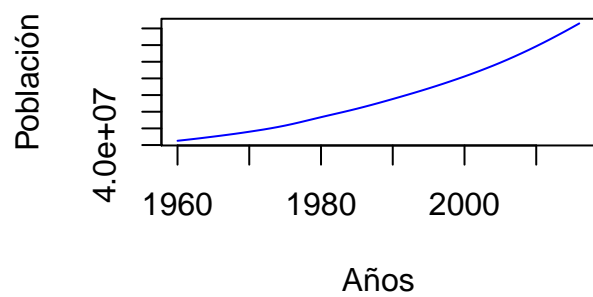
Nicaragua



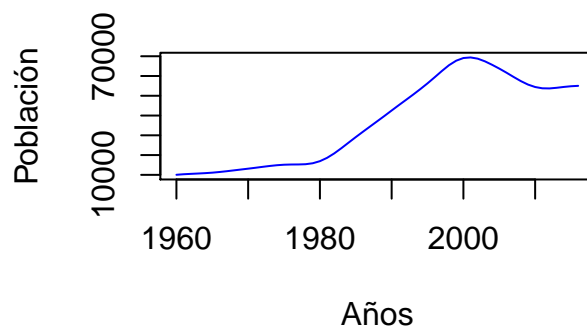
Niger

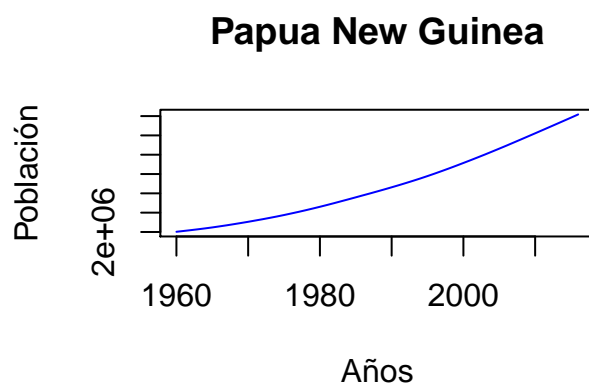
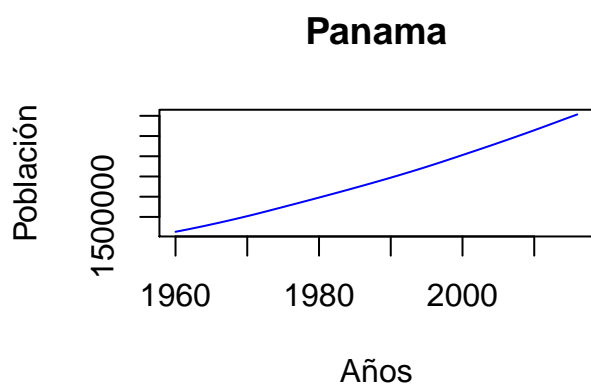
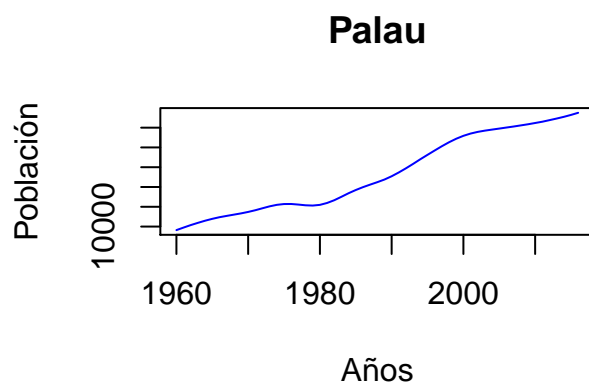
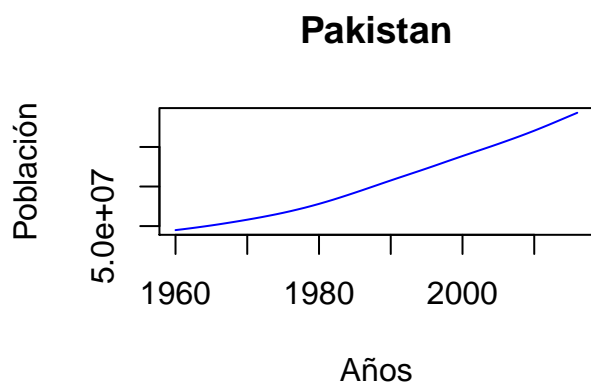
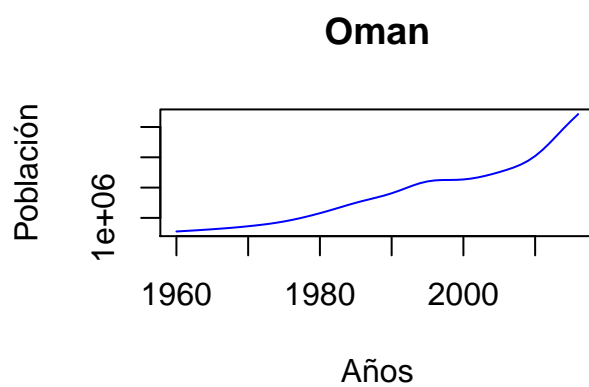
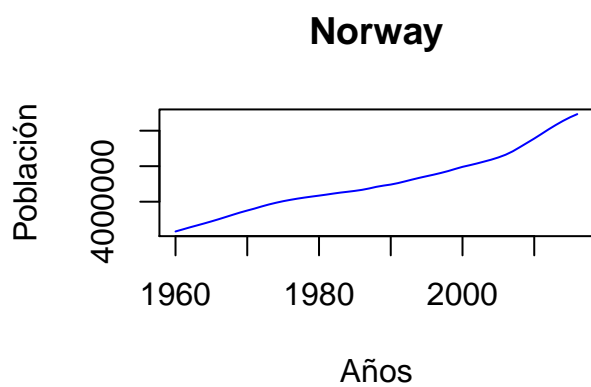


Nigeria

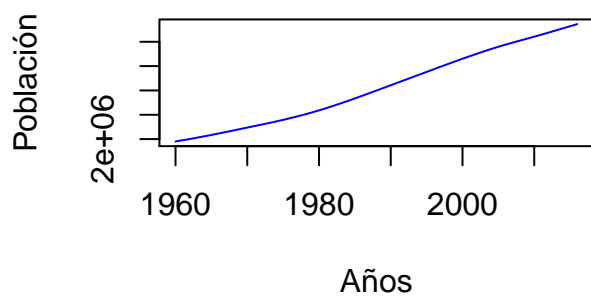


Northern Mariana Islands

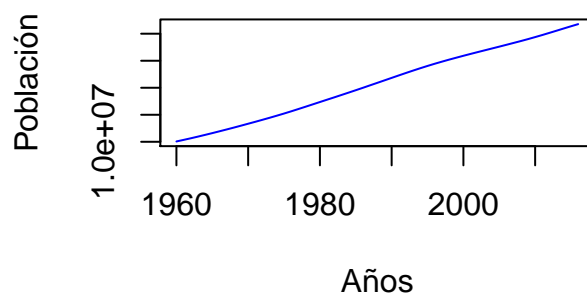




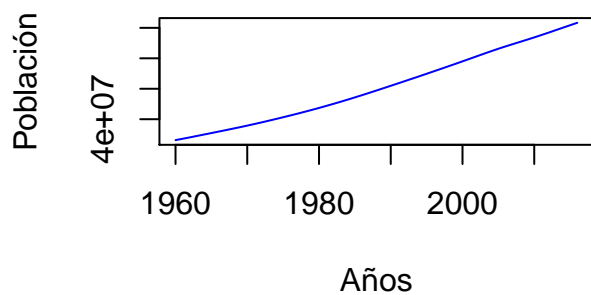
Paraguay



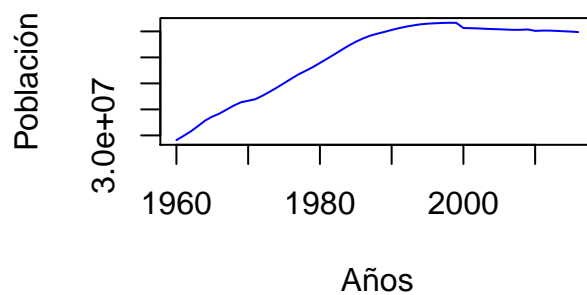
Peru



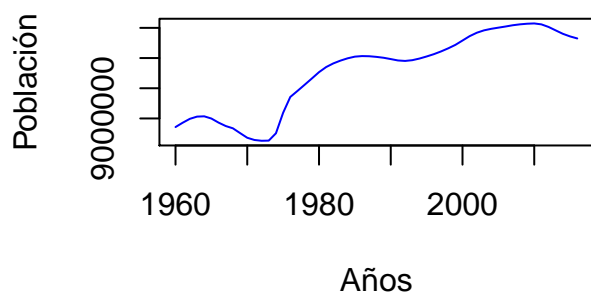
Philippines



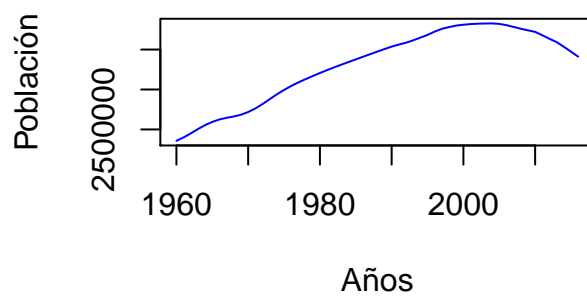
Poland

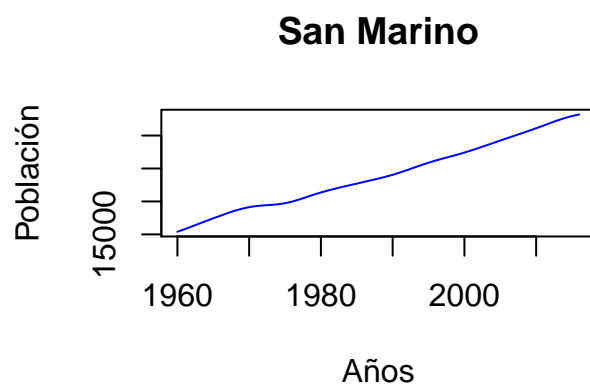
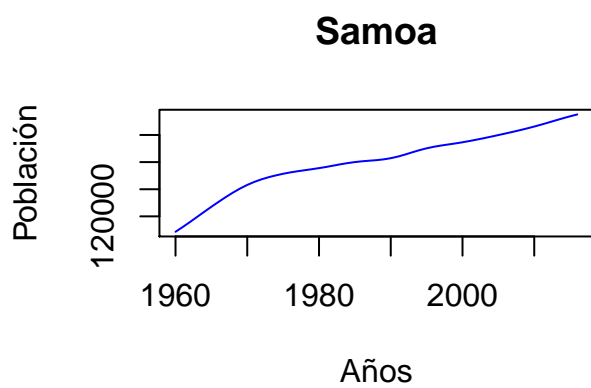
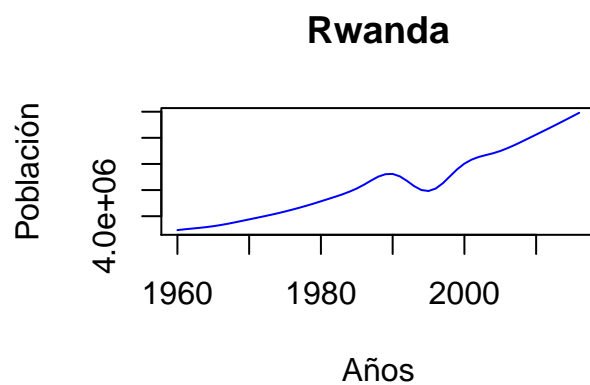
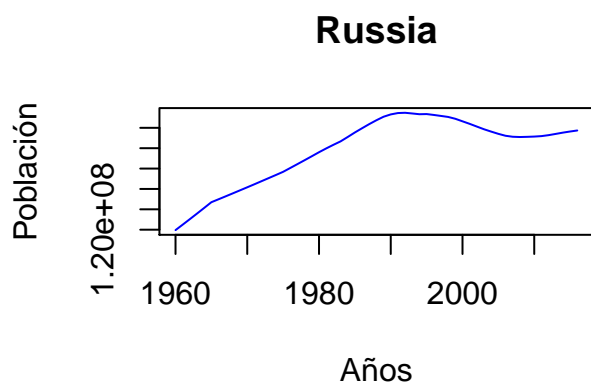
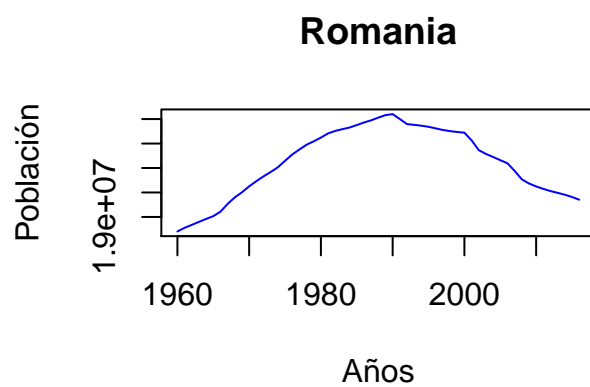
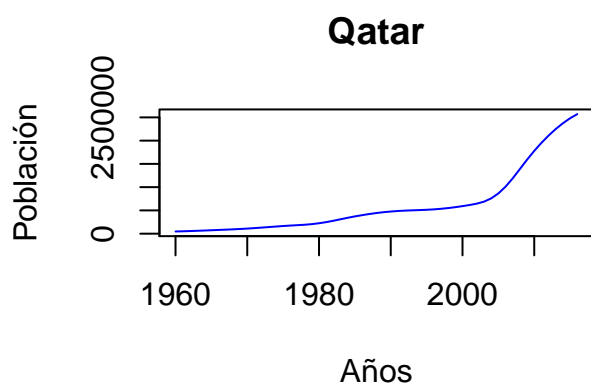


Portugal

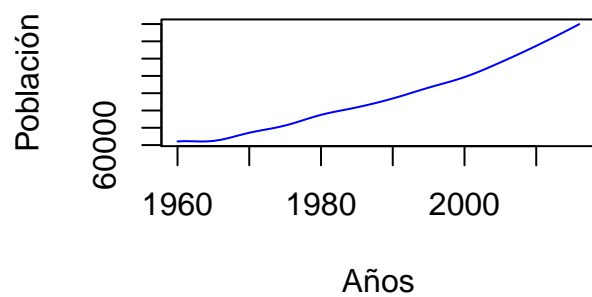


Puerto Rico

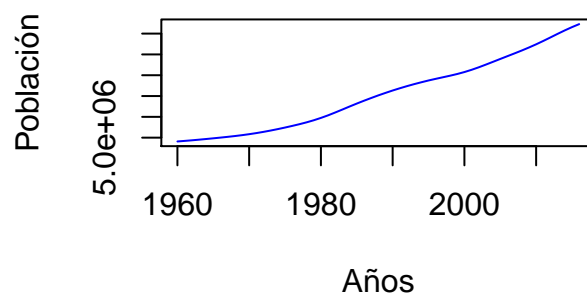




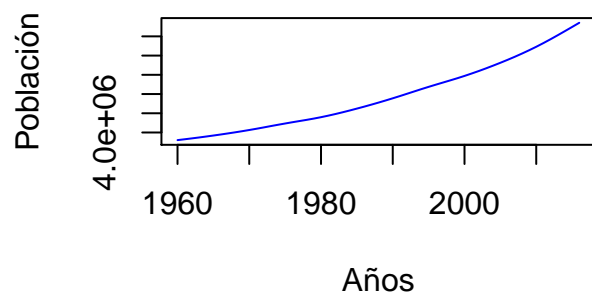
Sao Tome and Principe



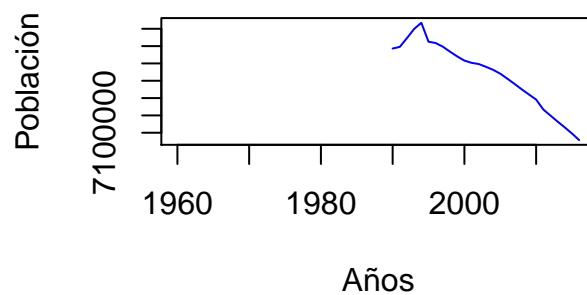
Saudi Arabia



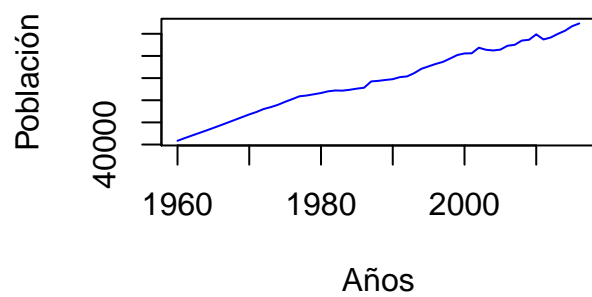
Senegal



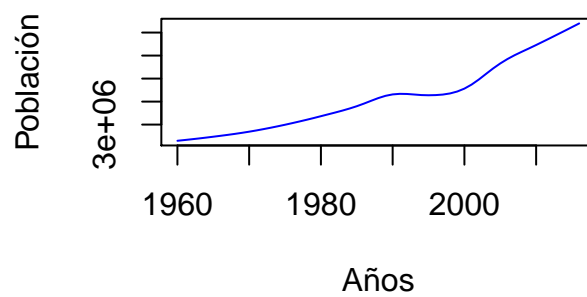
Serbia



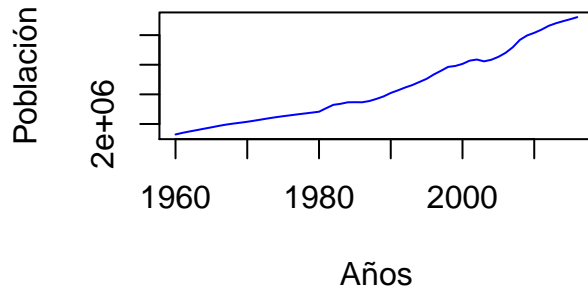
Seychelles



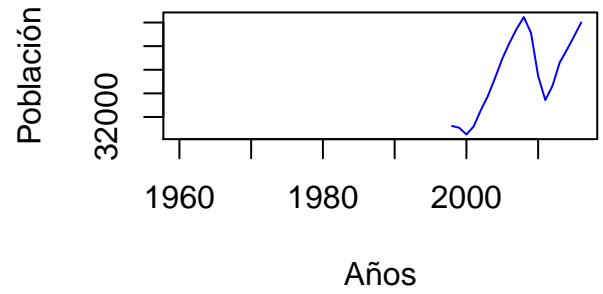
Sierra Leone



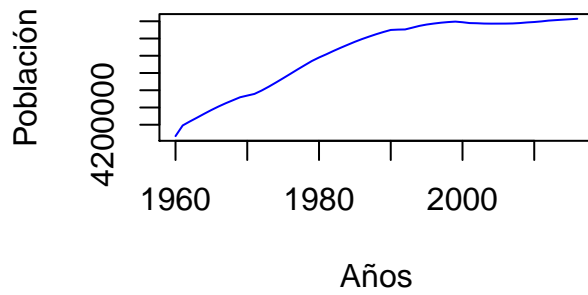
Singapore



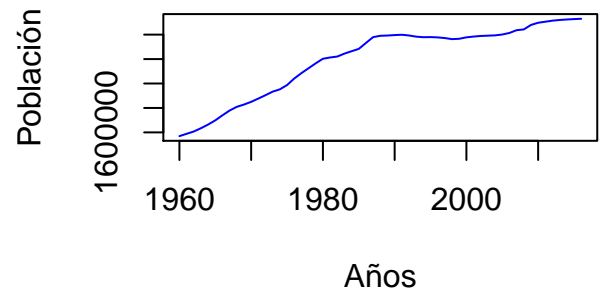
Sint Maarten (Dutch part)



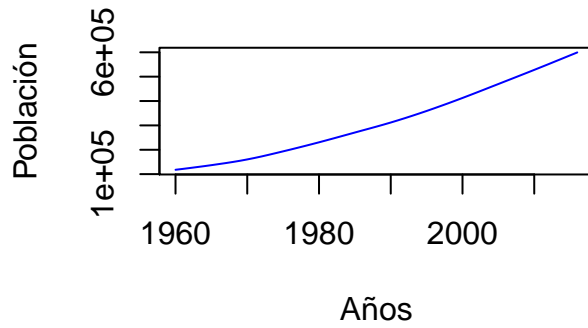
Slovak Republic



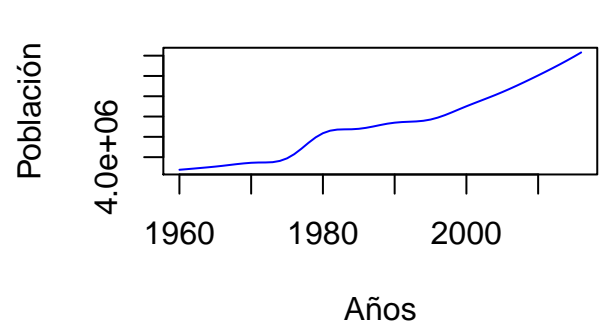
Slovenia



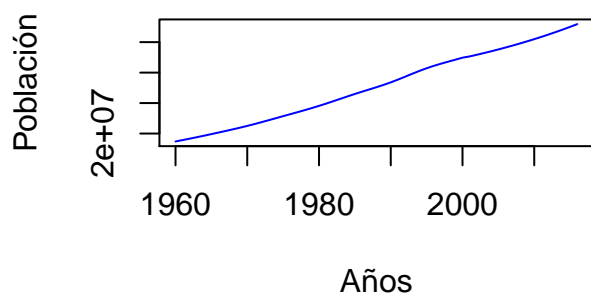
Solomon Islands



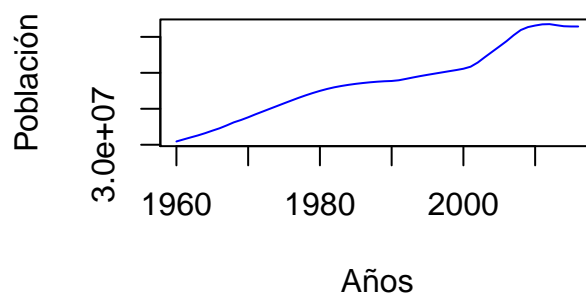
Somalia



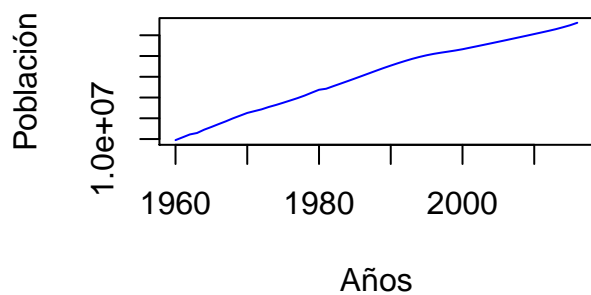
South Africa



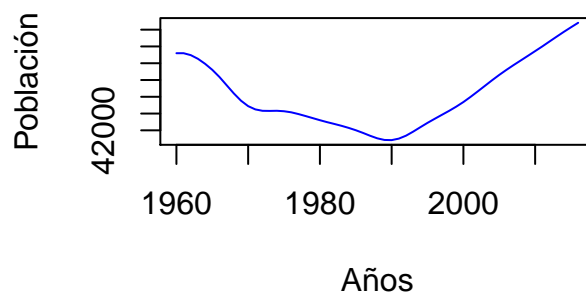
Spain



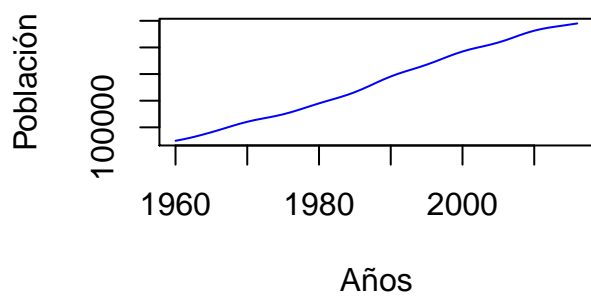
Sri Lanka



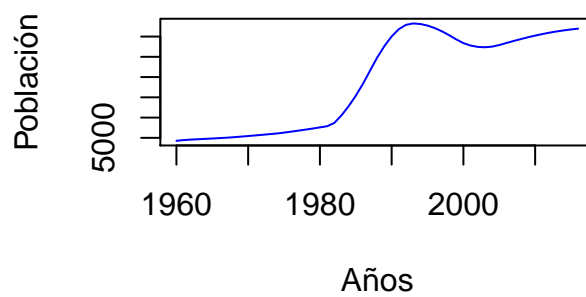
St. Kitts and Nevis



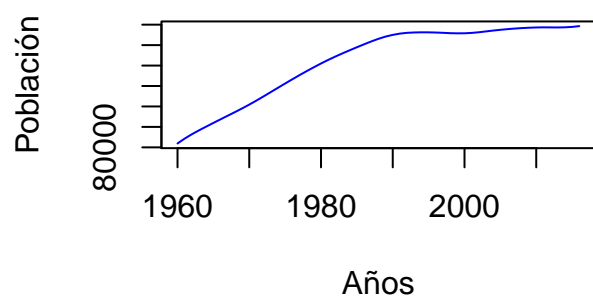
St. Lucia



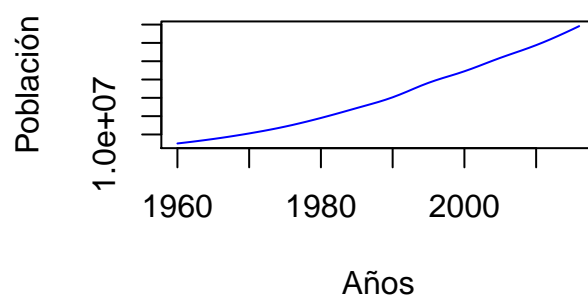
St. Martin (French part)



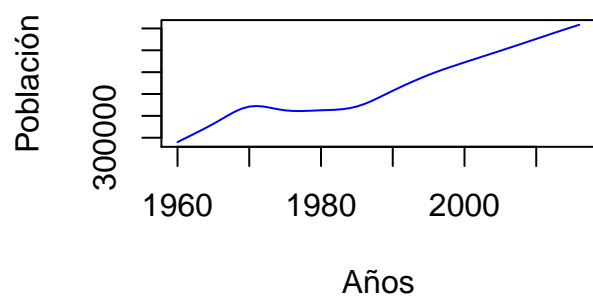
St. Vincent and the Grenadines



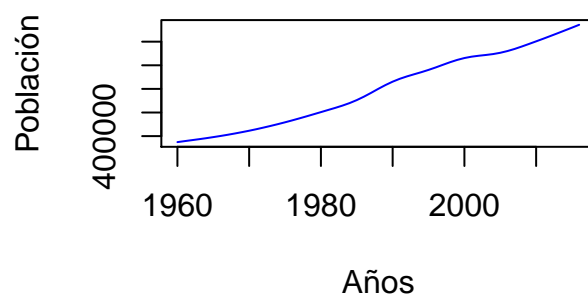
Sudan



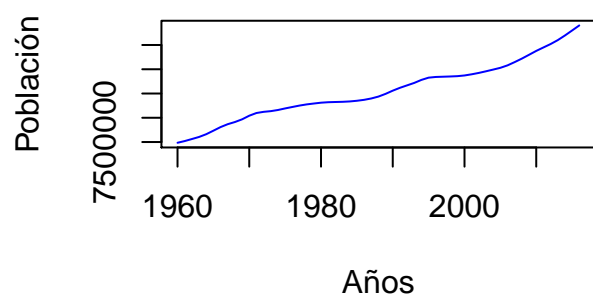
Suriname



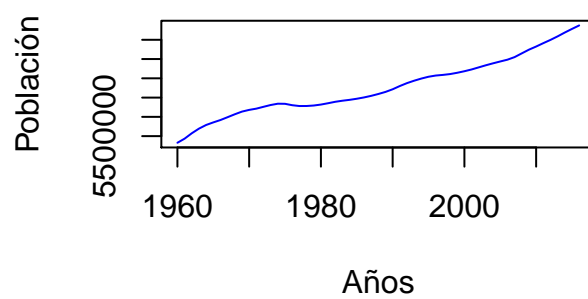
Swaziland



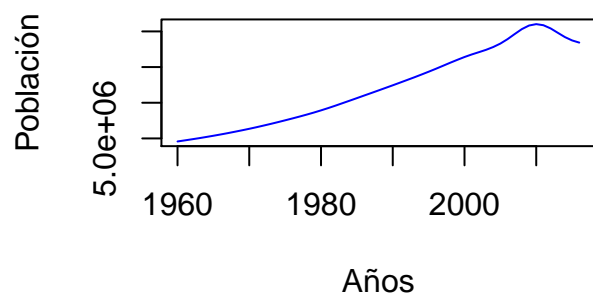
Sweden



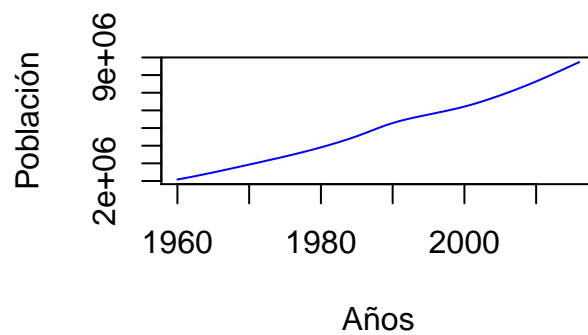
Switzerland



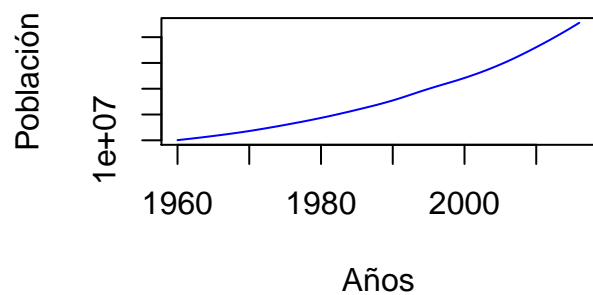
Syrian Arab Republic



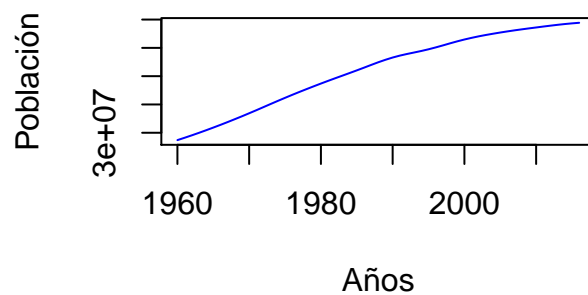
Tajikistan



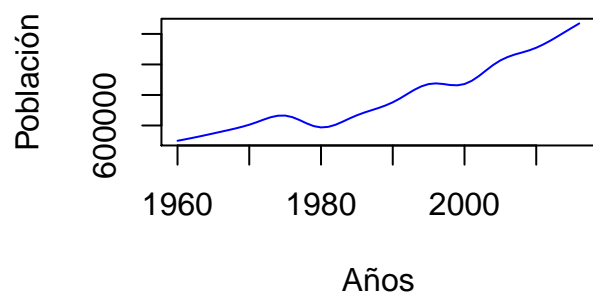
Tanzania



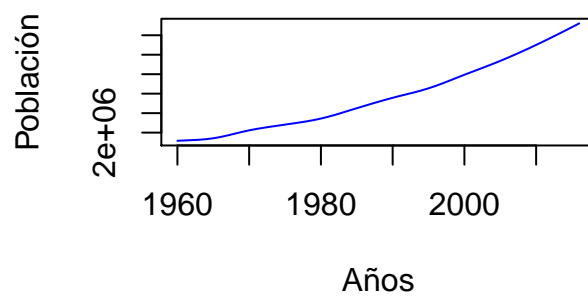
Thailand



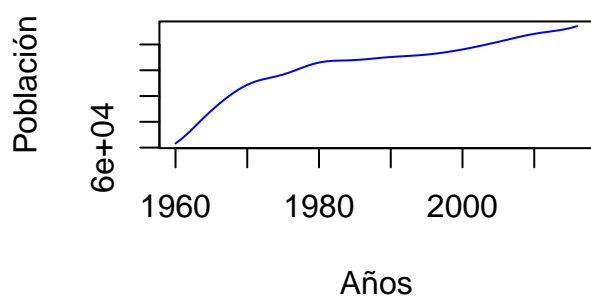
Timor-Leste



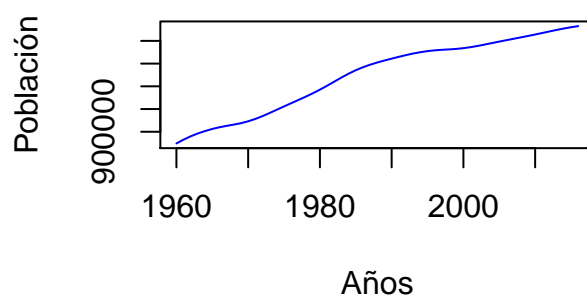
Togo



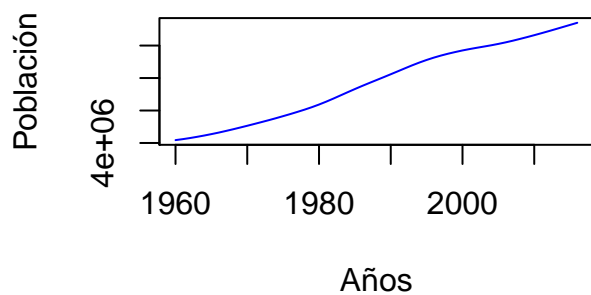
Tonga



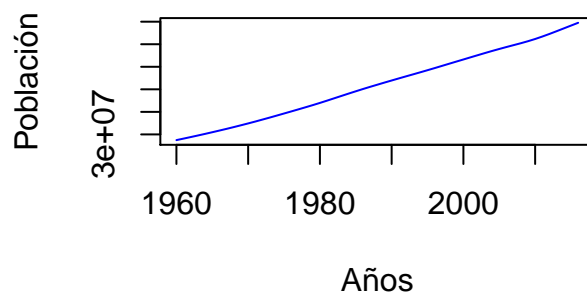
Trinidad and Tobago



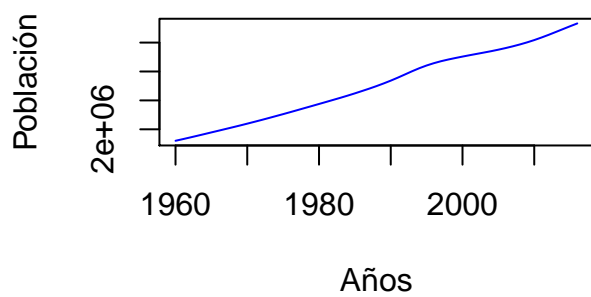
Tunisia



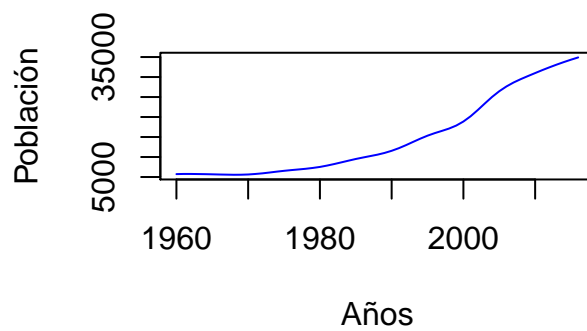
Turkey

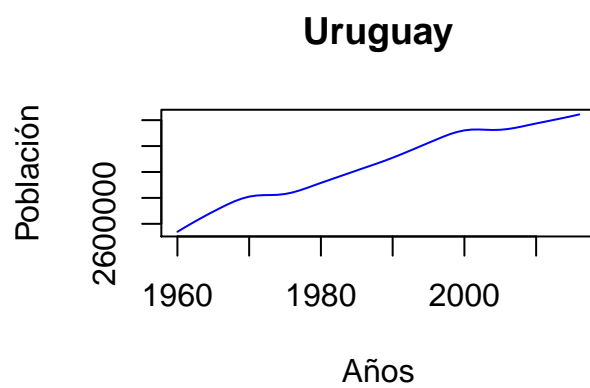
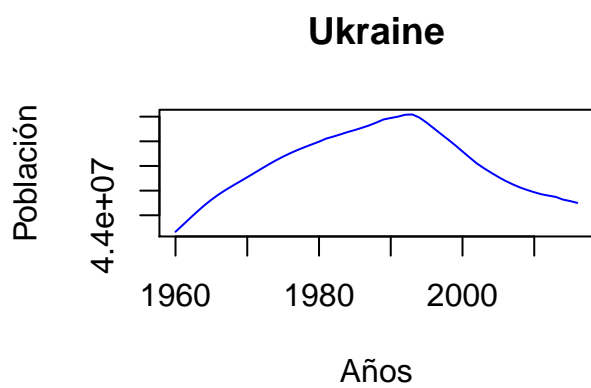
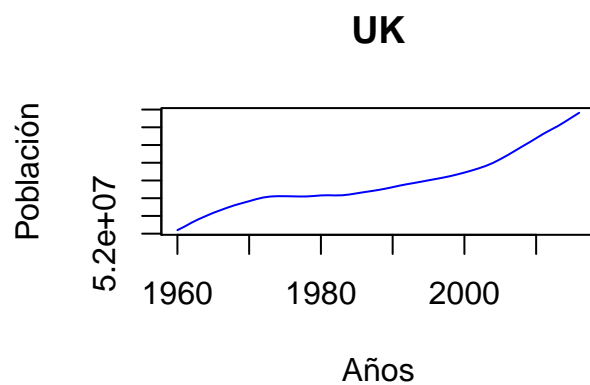
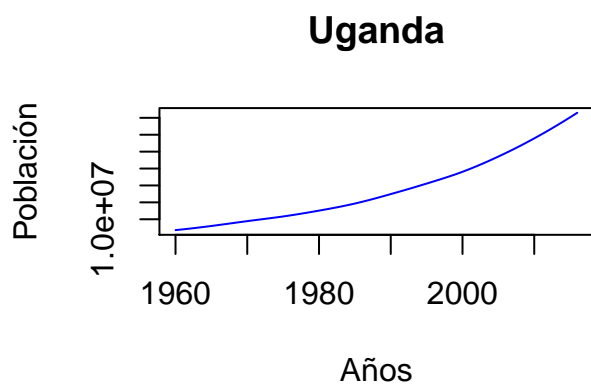
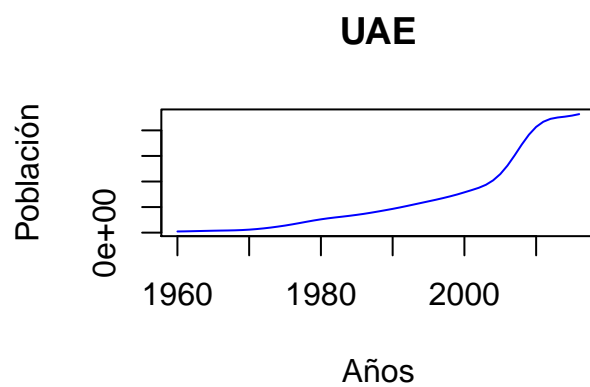
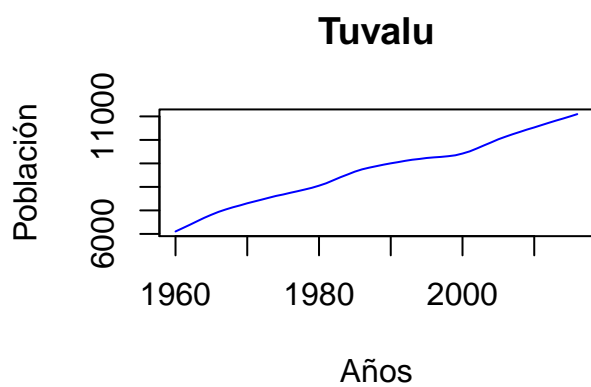


Turkmenistan

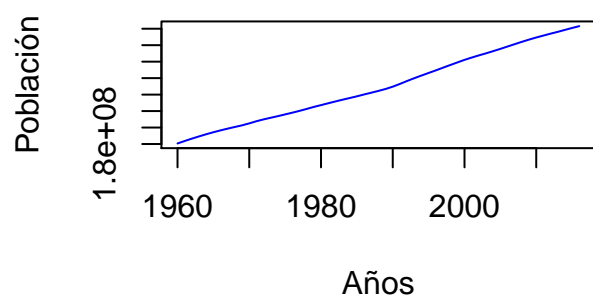


Turks and Caicos Islands

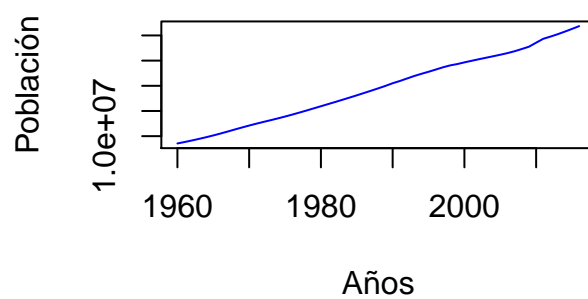




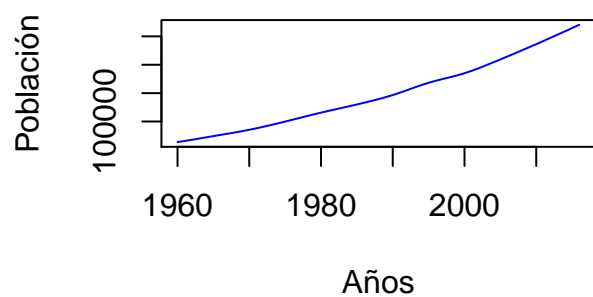
USA



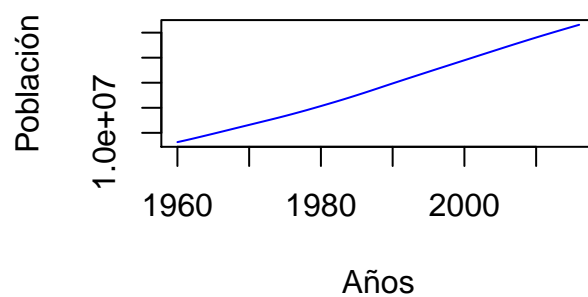
Uzbekistan



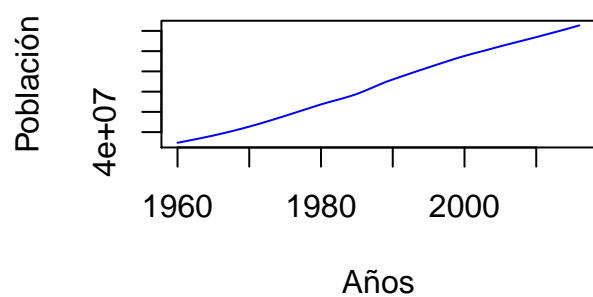
Vanuatu



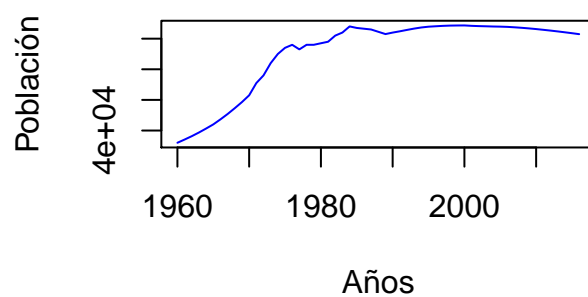
Venezuela, RB

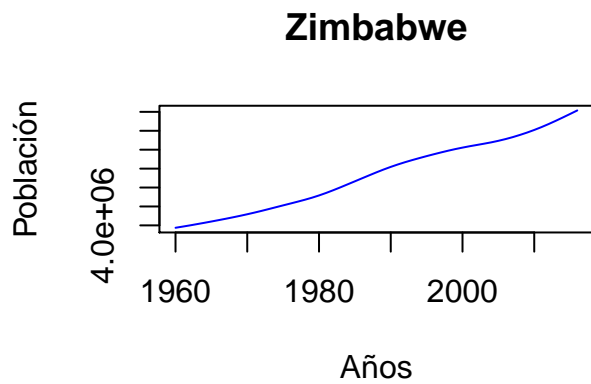
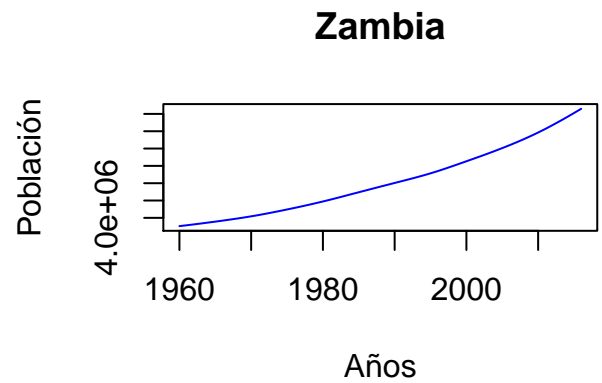
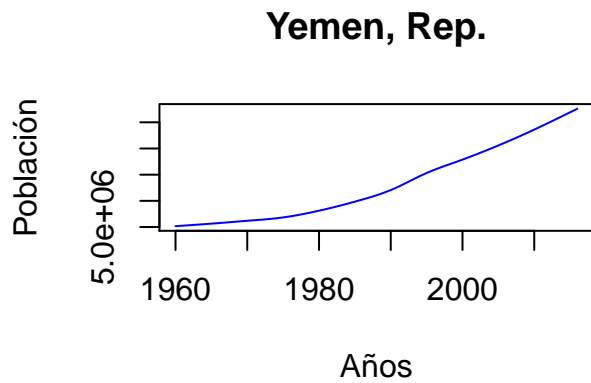
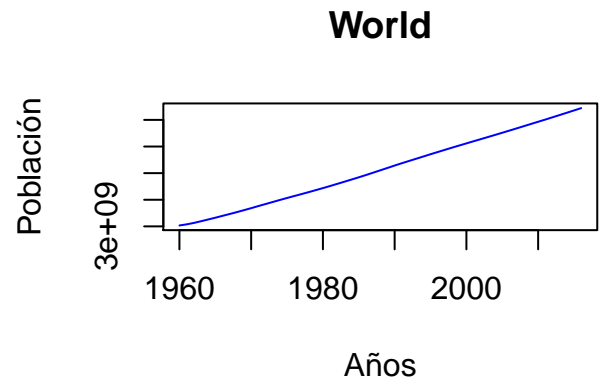
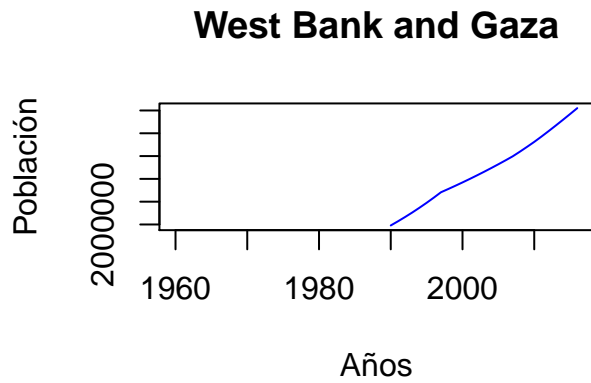


Vietnam



Virgin Islands (U.S.)





Frecuencias entre las Variables

Countries of the World

Como se mencionó anteriormente este dataset contiene distintos datos de los países del mundo. A continuación se muestra un resumen de cada variable de este dataset:

Country	Region	Population	Area..sq..mi..
Afghanistan : 1	SUB-SAHARAN AFRICA :51	Min. :7.026e+03	Min. : 2
Albania : 1	LATIN AMER. & CARIB :45	1st Qu.:4.376e+05	1st Qu.: 4648
Algeria : 1	ASIA (EX. NEAR EAST) :28	Median :4.787e+06	Median : 86600
American Samoa : 1	WESTERN EUROPE :28	Mean :2.874e+07	Mean : 598227
Andorra : 1	OCEANIA :21	3rd Qu.:1.750e+07	3rd Qu.: 441811
Angola : 1	NEAR EAST :16	Max. :1.314e+09	Max. :17075200
(Other) :221	(Other) :38	NA	NA

Pop..Density..per.sq..mi..	Coastline..coast.area.ratio.	Net.migration	Infant.mortality..per.1000.births.
13,8 : 2	0,00 : 44	0 : 62	: 3
2,7 : 2	0,09 : 4	: 3	9,95 : 3
372,5 : 2	0,13 : 4	-0,07 : 2	12,62 : 2
49,6 : 2	0,10 : 3	-0,31 : 2	4,39 : 2
66,6 : 2	0,15 : 3	-0,39 : 2	10,03 : 1
69,8 : 2	0,21 : 3	-0,71 : 2	10,09 : 1
(Other):215	(Other):166	(Other):154	(Other):215

GDP....per.capita.	Literacy....	Phones..per.1000.	Arable....
Min. : 500	: 18	: 4	0 : 9
1st Qu.: 1900	99,0 : 13	2,3 : 2	20 : 3
Median : 5550	97,0 : 11	2,7 : 2	: 2
Mean : 9690	98,0 : 10	255,6 : 2	1,64 : 2
3rd Qu.:15700	100,0 : 7	26,8 : 2	1,67 : 2
Max. :55100	98,6 : 4	269,5 : 2	10 : 2
NA's :1	(Other):164	(Other):213	(Other):207

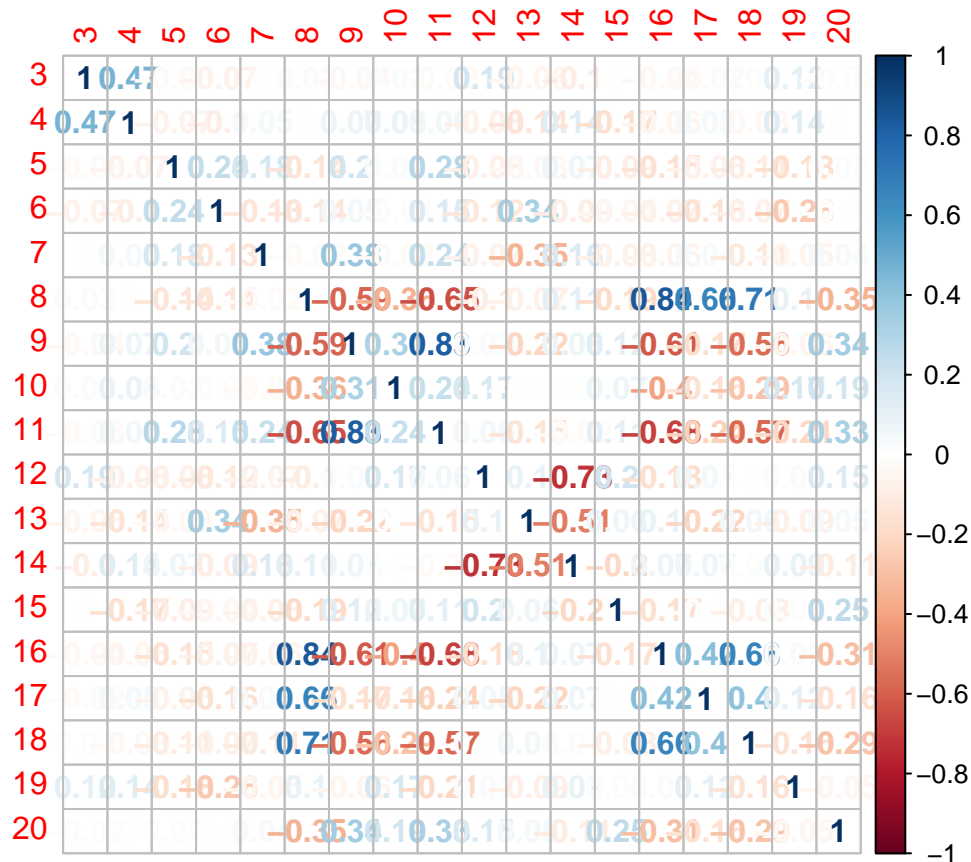
Crops....	Other....	Climate	Birthrate
0 : 28	100 : 8	: 22	: 3
0,03 : 5	75 : 3	1 : 29	12,56 : 2
0,01 : 4	: 2	1,5: 8	18,02 : 2
0,19 : 4	70 : 2	2 :111	18,79 : 2
0,4 : 4	70,44 : 2	2,5: 3	20,48 : 2
0,14 : 3	73,33 : 2	3 : 48	10 : 1
(Other):179	(Other):208	4 : 6	(Other):215

Deathrate	Agriculture	Industry	Service
: 4	: 15	: 16	: 15
10,31 : 2	0,01 : 8	0,11 : 4	0,684 : 5
12,25 : 2	0,04 : 6	0,17 : 4	0,55 : 4
14,02 : 2	0,03 : 5	0,1 : 3	0,62 : 4
3,92 : 2	0,018 : 4	0,12 : 3	0,46 : 3
5,28 : 2	0,06 : 4	0,18 : 3	0,549 : 3
(Other):213	(Other):185	(Other):194	(Other):193

Correlación entre variables

Para poder observar si existe algún tipo de relación entre las variables numéricas del dataset , se realizó un gráfico de correlación en donde cada variable se puso contra cada variable del dataset. En el resultado

podemos observar la correlación existente entre las variables:



En el diagrama de correlación se observó que existen coeficientes de correlaciones de varios valores, sin embargo, para este experimento los valores arriba de un 0.4 fueron considerados como significativos. Por lo que las relaciones que más se destacan son las siguientes:

- Population - Area..sq.mi.: 0.47
- Infant.mortality..per.1000.births. - GDP....per.capita.: 0.59
- Infant.mortality..per.1000.births. - Phones..per.1000.: 0.65
- Infant.mortality..per.1000.births. - Birthrate: 0.84
- Infant.mortality..per.1000.births. - Deathrate: 0.66
- Infant.mortality..per.1000.births. - Agriculture: 0.71
- GDP....per.capita. - Phones..per.1000.: 0.83
- GDP....per.capita. - Birthrate: 0.61
- GDP....per.capita. - Agriculture: 0.56
- Literacy.... - Birthrate: 0.4
- Phones..per.1000. - Birthrate: 0.68
- Phones..per.1000. - Agriculture: 0.57
- Arable.... - Other....: 0.73

- Crops.... - Other.....: 0.51
- Birthrate - Deathrate: 0.42
- Birthrate - Agriculture: 0.66
- Deathrate - Agriculture: 0.4

En los resultados obtenidos en el diagrama de correlación, se logró observar que la variable *Infant.mortality..per.1000.births* es la variable que tiene mayor relación con otras variables dentro del dataset, seguida de *GDP...per.capita..*. También, la correlación de mayor magnitud es la existente entre las variables *Infant.mortality..per.1000.births.* y *Birthrate.*

Análisis de PCA

Dados los datasets que se tienen *World Happiness Report*, *World Population* y *Countries of the World*; el único dataset en el que valdría la pena analizar sus componentes principales es el de *World Happiness Report*; eso es porque es el único dataset en el que tiene sentido una reducción de dimensionalidad ya que tiene bastantes variables que podrían estar relacionadas con la felicidad de la gente. Por otro lado, los otros 2 Datasets no tienen una dimensión muy alta (en el caso de *World Population*) o no se están relacionando las variables (en el caso de *Countries of the world*).

Previo a realizar el análisis de PCA se midió la adecuación muestral utilizando los coeficientes de Kaiser-Meyer-Olkin, la prueba de esfericidad de Bartlett y el estadístico de Barlett. Los resultados fueron los siguientes:

KMO	Barlett	RMS
0.329113	21154	0.045415

Debido a que la adecuación muestral es inaceptable no se llevó a cabo el analisis de PCA ni un análisis factorial.

Conclusiones y Hallazgos

Para el dataset World Happiness Report, se puede comenzar mencionando que las variables *Area..sq..mi..*, *Pop..Density..per.sq..mi..* y *Crops...* presentan tener multicolinealidad. Por lo que se recomienda que sean removidas en el caso que se planea calcular un modelo lineal múltiple o que esta característica sea tomada en cuenta al realizar algún análisis con el dataset.

Continuamos mencionando la relación existente entre variables como *Happiness.Rank*, la cual describe la posición en la que se encuentra cada país según su felicidad, en donde el tema familiar, la expectativa de vida y la economía tiene impacto directo. Así como que el grado que impacta el producto interno bruto en la felicidad es bastante parecido al grado de impacto de la expectativa de vida.

Los hallazgos encontrados en el dataset de *Countries Population* son simples, debido a que se encontró que la variable *population*, que describe la población de cada país, depende únicamente de la variable *year*. Ya que se encontró en el diagrama de correlación que cada año presenta tener una relación estrecha con los demás años.

Por ende, crear un modelo que describa o pronostique la población de cada país consiste en realizar una Regresión Lineal Simple entre la variable *population* y la variable *year*. Debido a que solo se necesita de la variable *year*, tomada en diferentes unidades de tiempo, para describir cómo se comporta la población en cada país.