1.Análisis Exploratorio de Datos

February 21, 2019

I.1 Estadística. Análisis Exploratorio de Datos

¿Qué es la estadística?

La **estadística** es el campo de las matemáticas que estudia cómo construir modelos, establecer hipótesis y tomar decisiones a partir de **datos empíricos** provenientes de medidas, experimentos u observaciones.

Los elementos objeto de estudio, de los que nos interesa conocer cierta **información**, conforman la **población estadística**. Tales poblaciones, así como la información que nos interesa extraer, pueden tener naturalezas muy variadas. Como veremos, *la información de interés la representamos mediante variables*.

Veamos algunos ejemplos:

- Población: Todos los residentes de un país.
  - Información: voto en próximas elecciones
  - Información: relación entre género, altura y peso
- Población: Señales recibidas en un receptor, contaminadas por ruido
  - Información: señales originales, sin ruido
  - Información: Símbolos envíados
- Población: Todos los países del mundo
  - Información: número de habitantes de cada país
  - Información: producto interior bruto de cada uno
  - Información: renta per cápita de cada país
  - Información: esperanza de vida en cada uno
  - Información: relación entre renta per cápita y esperanza de vida

Variables y observaciones

Una **variable estadística** o **variable muestral** es una magnitud de interés, relativa a una población, que puede observarse. Por ejemplo, en la población formada por todos los países del mundo, tenemos magnitudes observables tales como el número de habitantes, el producto interior bruto, la renta per cápita o la esperanza de vida. O, en una población formada por una red de sensores, podemos considerar la temperatura, la humedad y el nivel de CO2, por ejemplo.

El conjunto de valores que puede tomar una variable estadística se llama **espacio muestral**. Las variables estadísticas pueden ser:

- **Variables continuas**: toman *valores numéricos reales*. Por ejemplo, la altura de los individuos de una población, el producto interior bruto de los países del mundo o la lectura de la señal medida por un receptor.
- **Variables discretas**: toman valores *numéricos enteros*. Por ejemplo, el número de habitantes de cada uno de los países del mundo, o nivel detectado por un receptor de una señal digital multinivel.
- **Variables categóricas**: sus valores *no son numéricos*, sino *pertenecientes a clases*, a veces llamadas *niveles*. Por ejemplo, la ciudadanía de un individuo, el grupo sanguíneo o los símbolos recuperados por un receptor digital.
  - Los **intervalos** son variables categóricas que agrupan rangos de valores numéricos.
  - A veces es posible establecer una **relación de orden** en los valores de una variable categórica. Por ejemplo, podemos ordenar personas por altura, aunque no sepamos el valor numérico.

Con frecuencia, sólo tendremos acceso a una subpoblación. Por ejemplo, si la población es muy numerosa o infinita, todo lo más que podremos hacer es recoger los datos de una **subpoblación muestral** elegida de tal forma que pensemos que representa adecuadamente a la población general. Estudiándola esperamos poder generalizar ciertas características que sean aplicables a toda la población. Por ejemplo,

- **Observando** mediante una encuesta la intención de voto de unos pocos miles de personas intentamos **inferir** qué resultado van a tener las elecciones.
- Midiendo señales y ruido en **experimentos** controlados esperamos **inferir** modelos de comportamiento que nos permitan tratar nuevas señales recibidas.

En otros caso, podemos disponer de la totalidad de la población. Piénsese, por ejemplo, en la población estadística formada por todos los países del mundo, que son unos 200.

En cualquier caso, sea cual fuere la aplicación, la primera tarea es la **recogida de datos**, sea por pura observación o mediante experimentación.

En nuestro caso, vamos a ilustrar los conceptos principales de la **estadística descriptiva** mediante datos reales, provenientes del Banco Mundial, correspondientes a ciertos indicadores de todos los países del mundo así como de agregaciones de países en el periodo de tiempo 1960 - 2017. En particular trabajaremos con los siguientes indicadores:

- Número de habitantes, que nos permite analizar los países por el tamaño de sus poblaciones
- Producto interior bruto, que nos permite hacerlo por el tamaño de sus economías
- Renta per cápita, que nos permite analizar los países por su riqueza
- Esperanza de vida, que nos proporciona una indicación de la salud y, en cierta manera, de la calidad de vida

También analizaremos relaciones estadísticas entre estas **variables**. Por ejemplo, ¿el tamaño de los países tiene una relación con la riqueza de los mismos? ¿y la riqueza con la esperanza de vida?

Etapas en el procesado estadístico de datos

De lo expuesto podemos intuir que el procesado de datos se estructura en tres etapas sucesivas, íntimamente relacionadas, y con un cierto solapamiento entre ellas:

1. **Manipulación de datos**: los datos provienen de fuentes diversas, por lo que frecuentemente estan recogidos en formatos diferentes y, con casi total seguridad, de modo que directamente no podemos trabajar con ellos. Además, es habitual que algunos de ellos falten o estén contaminados por imprecisiones en su adquisición. En ingeniería la presencia de **ruido** es omnipresente, pero en otras disciplinas también nos encontramos con incorrecciones en los mismos.
2. **Análisis de datos**: corresponde al **anaĺisis exploratorio de datos** que nos permite representar y entender los datos disponibles, preparándolos para ulteriores análisis. En esta etapa podemos considerar que *los datos se convierten en información*.
3. **Ciencia de datos**: se utilizan las herramientas de la estadística matemática y del aprendizaje artificial para extrar inferencias, estimaciones, clasificaciones y generalizaciones a partir de la información presente en los datos. Podemos considerar que, en esta etapa, *la información se convierte en conocimiento*.

Secuencias, tablas y series temporales

Una **observación** es una captura del valor de una variable correspondiente a una **muestra** de la población estadística. Normalmente tendremos muchas observaciones para cada variable, conjunto que denominamos **muestreo**. Por ejemplo, en la población de países del mundo tendremos un **muestreo** consistente en una observación o muestra correspondiente a cada país, de cada una de las variables que estemos estudiando (número de habitantes, producto interior bruto, etc). El **tamaño** N **de un muestreo** (a veces se dice de una muestra) indica el **número de observaciones** o muestras que se toman de la variable bajo estudio.

La **dimensionalidad** del conjunto de datos viene determinado por el número de variables que se observan para cada miembro de la subpoblación muestral. Por ejemplo, si para cada país sólo se considera el número de habitantes, el conjunto de datos es **univariado**. Si se considera el número de habitantes y el producto interior bruto, el conjunto de datos es **bivariado**. Por lo general, se denomina **multivariada** a la población muestral de la que se miden tres o más características, en nuestro caso, por ejemplo, porque añadimos la renta per cápita y la esperanza de vida.

Cuando tenemos varias variables y estamos interesados en entender cómo se relacionan, es importante distinguir entre **variables independientes**, que reflejan causas, y **variables dependientes**, que reflejan efectos. Por ejemplo, podemos considerar que una renta per cápita elevada es causa de una mayor esperanza de vida. Pero, ¿podría ser al revés?, esto es, que una mayor esperanza de vida sea la causa de una mayor renta per cápita.

Si las subpoblaciones muestrales son grandes y multivariadas no es sencillo interpretar los datos disponibles. Es necesario disponer de técnicas que lo hagan posible, como veremos en los próximos apartados. Una cuestión crucial es la debida ordenación de los datos en series y tablas, cuestión que se enmarca dentro de la etapa de manipulación de datos:

- **Secuencias o series**: son tablas unidimensionales en las que se recogen las observaciones correspondientes a una única variable. Si las observaciones se realizan a lo largo del tiempo, estamos ante **series temporales**.
- **Tablas**: en inglés denominadas ***data frames***. Son ordenaciones de muestreos multivariados, por lo general, *asociando cada columna a una variable y cada fila a una observación*. Cuando se

incluye una secuencia temporal de medidas, hay que reflexionar si cada instante de tiempo debe considerarse una variable o una observación.

¿Qué es el análisis exploratorio de datos?

Se trata de un conjunto de técnicas para sistematizar la debida interpretación de los datos disponibles. Entre ellas, la **estadística descriptiva** permite obtener parámetros que describen la subpoblación muestral, tanto obteniendo **sumarios numéricos** de los mismos, como recurriendo a **descripciones gráficas**. Pero ello no es suficiente. Hay que aproximarse críticamente a los datos disponibles, intentando entenderlos conforme sea el proceso subyacente que los genera.

El análisis exploratorio de datos incluye la estadística descriptiva, con sus sumarios numéricos y sus representaciones gráficas, y la interpretación crítica que se haga de los datos para su debida comprensión. Sólo entonces puede plantearse la inferencia de las propiedades del conjunto de la población muestral. No lo olvidemos: **la estadística suele trabajar con subpoblaciones muestrales, que se describen y analizan, de forma previa a inferir una caracterización del conjunto de la población**.

Lectura de datos del Banco Mundial

Nuestra **población estadística** son todos los países del mundo, **que muestreamos en su totalidad**. Para ello vamos a trabajar con datos obtenidos por el Banco Mundial y disponibles en su portal web, con periodicidad anual desde 1960, hasta 2017. Para facilitar las cosas vamos a utilizar datos previamente descargados en formato CSV (*Comma Separated Values*), aunque podríamos hacerlo también "*online*".

- **Variables estadísticas o muestrales**: población (número de habitantes, en millones de personas), producto interior bruto (en miles de millones de dólares), renta per cápita (en miles de dólares) y esperanza de vida (en años) para cada uno de los años entre 1960 y 2017. Son las columnas de la tabla.
- **Observaciones**: las mediciones de las variables estadísticas para cada uno de los países. Cuando no tengamos una medida, utilizaremos el símbolo *NA* (Not Available) o el símbolo *NaN* (Not a Number).

Como queremos inferir conclusiones para el conjunto de los paídes del mundo, consideramos cada uno de ellos una muestra u observación, en vez de considerar a cada año, que será una variable distinta. Por ejemplo, una variable es la población en 1970, y otra variable es la esperanza de vida en 1990. Las muestras u observaciones son las medidas para cada país de cada una de las variables.

Análisis preliminar de la población mundial

Tenemos una tabla para cada indicador. En cada tabla tenemos 217 países (filas) y 58 variables (columnas), correspondientes éstas a cada uno de los años de los que se dispone de medidas. Para facilitar la visualización, con frecuencia *transpondremos* las tablas, de modo que los años (variables) correspondan a las filas y los países (observaciones o muestras) correspondan a las columnas. Esto no debe confundirnos, y se hace sólo a efectos de visualización.

```
<class 'pandas.core.frame.DataFrame'>
Index: 217 entries, ABW to ZWE
```

```
Columns: 58 entries, 1960 to 2017
Freq: A-DEC
dtypes: float64(58)
memory usage: 100.0+ KB
```

Países con poblaciones máximas y mínimas. Rango de la población Para cada variable (población en un año), nos interesa, en primer lugar, saber si nos faltan valores, y cuáles son los valores máximos y mínimos. Podemos visualizarlo fácilmente, poniendo en rojo las observaciones que faltan (NaN) y resaltando en amarillo los valores máximos y mínimos.

El **valor máximo** y el **valor mínimo** de las muestras de una variable estadística numérica ofrecen una interesante información preliminar.

El **rango** es la diferencia entre los valores máximo y mínimo de tal variable numérica.

Adviértase que China se ha mantenido desde 1960 como el país de mayor población del mundo, y que los países más pequeños tienen una población minúscula.

Identificación de países a partir de sus códigos {'CHN': 'China', 'MAF': 'St. Martin (French part)', 'NRU': 'Nauru', 'TUV': 'Tuvalu', 'ERI': 'Eritrea', 'PSE': 'West Bank and Gaza', 'SRB': 'Serbia', 'SXM': 'Sint Maarten (Dutch part)'}

Rango de poblaciones

	1960	1970	1980	1990	2000	2010	2015	2017
idxmax	CHN	CHN	CHN	CHN	CHN	CHN	CHN	CHN
max	667.070	818.315	981.235	1135.185	1262.645	1337.705	1371.220	1386.395
idxmin	MAF	MAF	NRU	TUV	TUV	NRU	TUV	TUV
min	0.004	0.005	0.007	0.009	0.009	0.010	0.011	0.011

Ordenación de países por población. Estadísticos de orden

El **estadístico de orden k** de una variable muestral de tamaño n (con n muestras) corresponde al k-ésimo menor valor. Por tanto, **el estadístico de orden 1 corresponde al mínimo**, mientras que el **estadístico de orden n corresponde al máximo**.

Podemos también invertir el orden de los estadísticos, de forma que el orden 1 corresponda al máximo y el orden n al mínimo.

A continuación prepararemos una tabla, en la que cada variable corresponde al puesto por población, en orden descendente (máximo = 1), ocupado por un país, en relación al total mundial en un año determinado.

VER TRANSPARENCIA ONLINE

Evolución en el orden de población de una selección de países Country Code CHN IND USA RUS NGA JPN DEU GBR FRA ESP Year 1960 1.0 2.0 3.0 4.0 13.0 5.0 7.0 9.0 12.0 18.0 1970 1.0 2.0 3.0 4.0 11.0 6.0 8.0 12.0 14.0 22.0 1980 1.0 2.0 3.0 5.0 11.0 7.0 9.0 14.0 15.0 24.0 1990 1.0 2.0 3.0 6.0 10.0 7.0 12.0 17.0 15.0 26.0 2000 1.0 2.0 3.0 6.0 10.0 9.0 12.0 21.0 20.0 28.0 2010 1.0 2.0 3.0 9.0 7.0 10.0 16.0 22.0 20.0 27.0 2015 1.0 2.0 3.0 9.0 7.0 10.0 16.0 22.0 21.0 30.0 2017 1.0 2.0 3.0 9.0 7.0 11.0 16.0 22.0 21.0 30.0

Análisis de los países mayores.

```
['BGD', 'BRA', 'CHN', 'DEU', 'GBR', 'IDN', 'IND', 'ITA', 'JPN', 'MEX', 'NGA', 'PAK', 'RUS', 'USA
```

1 2 3 4 5 6 7 8 9 10 Year 1960 CHN IND USA RUS JPN IDN DEU BRA GBR ITA 1970 CHN IND USA RUS IDN JPN BRA DEU BGD PAK 1980 CHN IND USA IDN RUS BRA JPN BGD DEU PAK 1990 CHN IND USA IDN BRA RUS JPN PAK BGD NGA 2000 CHN IND USA IDN BRA RUS PAK BGD JPN NGA 2010 CHN IND USA IDN BRA PAK NGA BGD RUS JPN 2015 CHN IND USA IDN BRA PAK NGA BGD RUS JPN 2017 CHN IND USA IDN BRA PAK NGA BGD RUS MEX

{'BGD': 'Bangladesh', 'BRA': 'Brazil', 'CHN': 'China', 'DEU': 'Germany', 'GBR': 'United Kingdom', 'IDN': 'Indonesia', 'IND': 'India', 'ITA': 'Italy', 'JPN': 'Japan', 'MEX': 'Mexico', 'NGA': 'Nigeria', 'PAK': 'Pakistan', 'RUS': 'Russian Federation', 'USA': 'United States'}

Comparación de la población agregada de los países mayores con la mundial Year 1960 1970 1980 1990 2000 2010 2015 2017 LC Total 1845.0 2223.2 2667.5 3166.7 3650.6 4069.6 4268.4 4350.8 WLD Total 3014.9 3664.3 4414.3 5267.9 6099.5 6909.7 7329.3 7501.7 LC/WLD (%) 61.2 60.7 60.4 60.1 59.9 58.9 58.2 58.0

Algunas descripciones gráficas

Representemos gráficamente las poblaciones de los países mayores del mundo.

Adviértase que estos países pueden ser representados por una **variable categórica** cuyos valores posibles son los "códigos de los países".

Dado que cada país tiene una población, podemos considerar que *la variable categórica juega el papel de* ***variable independiente***, mientras que la población, *una variable numérica, juega el papel de* ***variable dependiente***.

Diagrama de barras o de columnas (bar plot, bar graph, bar chart) Un diagrama de barras o de columnas es un gráfico que representa datos recogidos en una variable numérica (dependiente), por ejemplo poblaciones, frente una variable categórica independiente, por ejemplo países. Esto se hace mediante barras horizontales o verticales, cuyas *longitudes son proporcionales a los datos* (variables numéricas dependientes) que representan, asignando una barra a cada variable categórica independiente. Ello permite a los diagramas de barras comparar datos numéricos asociados a categorías discretas.

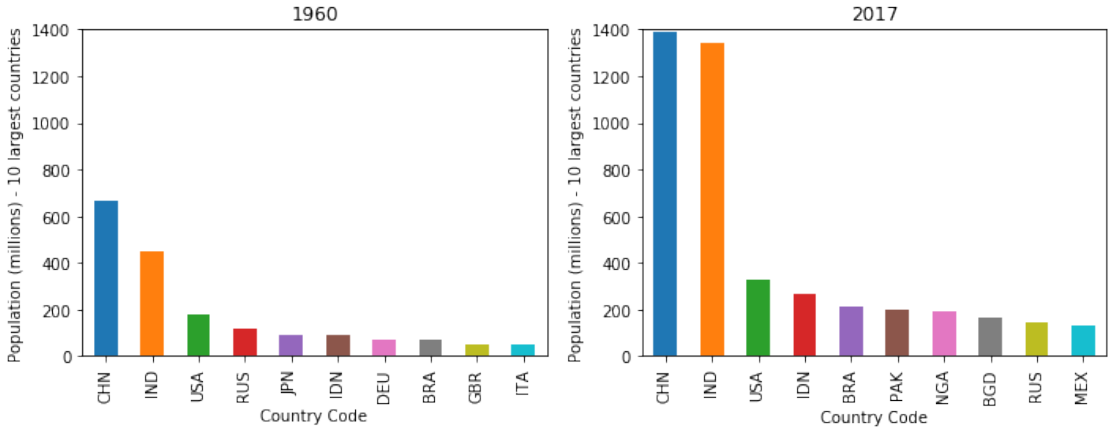


Diagrama de barras horizontales

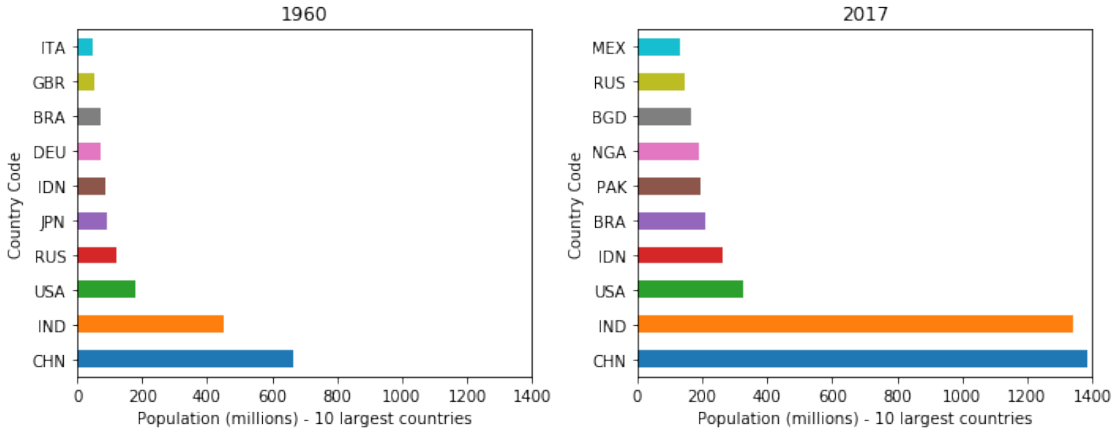
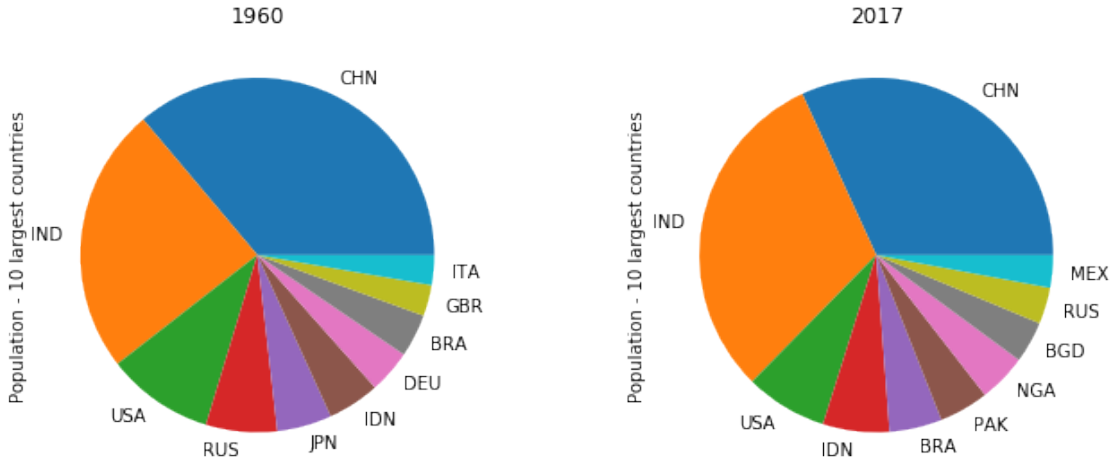
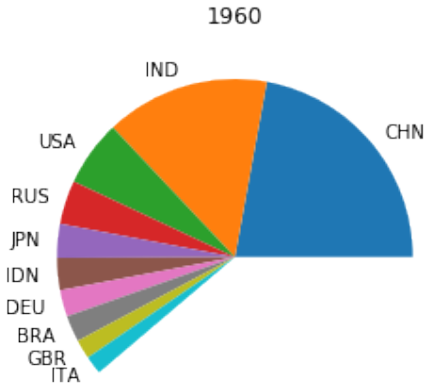


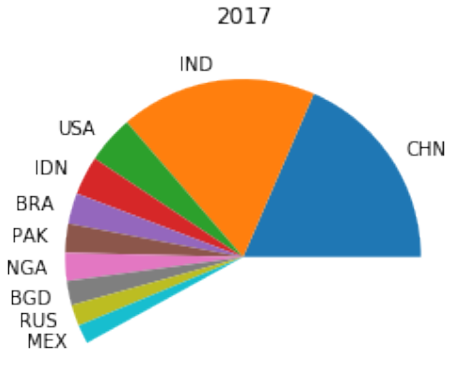
Diagrama circular o de tarta (pieplot) Un diagrama circular o de tarta es un gráfico que representa datos recogidos en una variable numérica (dependiente), por ejemplo poblaciones, frente una variable categórica independiente, por ejemplo países. En este caso los datos (variables numéricas dependientes) se representan mediante sectores circulares, cuyos ángulos son proporcionales a los datos que representan, asignando un sector a cada variable categórica independiente. Ello permite a los diagramas circulares o de tarta, al igual que a los de barras, comparar datos numéricos asociados a categorías discretas.



Podemos también ver cuánto suponen los países más poblados en relación a la población mundial. Para ello normalizamos las cantidades, considerando que la totalidad de la población

mundial ocupa un sector de 360º, esto es el círculo completo. Por tanto, en blanco queda la población del mundo no incluida en los países mayores.





Histogramas

Consideremos que nuestra muestra consiste en N observaciones x_i , $i = 1 \dots N$ de una variable muestral.

La **frecuencias absoluta** F_i indica el número de veces que se repite el valor i-ésimo de la variable muestral.

Lógicamente, la suma de frecuencias absolutas de todos los valores posibles de la variable coincide con el número de observaciones: $\sum_i F_i = N$.

La **frecuencia relativa** f_i es la frecuencia absoluta dividida por el número de muestras: $f_i = F_i/N$.

La suma de las frecuencias relativas es la unidad: $\sum_i f_i = 1$.

Podemos representar tanto las frecuencias absolutas como las relativas mediante un diagrama de barras.

Este cálculo de las frecuencias absolutas y relativas tiene pleno sentido

- para variables categóricas
- para variables numéricas discretas, si el número de valores que toma es pequeño en relación al tamaño de la muestra

Por ejemplo, si la variable estadística es el color de los coches, los valores que puede tomar será los colores, blanco, rozo, azul, amarillo, negro,... y la frecuencia absoluta es el número de coches de cada color que hemos observado. Si sumamos el número de coches de cada color que hemos observado, obtenemos el total de coches observados.

Sin embargo, si la variable estadística puede tomar valores sobre un rango continuo, difícilmente coincidirán los valores de distintas observaciones. Lo mismo paso aún pudiendo tomar un conjunto discreto de valores, si los valores posibles son muchos en relación al número de observaciones.

En tales casos las observaciones se agrupan en intervalos, mediante un procedimiento conocido como *binning* o *bucketing*. Para ello, se realiza una partición del espacio muestral ocupado por las observaciones en K intervalos, cuyos límites son $l_0, l_1 \dots l_N$.

El intervalo k-ésimo es $I_k = (l_{k-1}, l_k]$, que se convierte en una variable categórica que se observa tantas veces como muestras x_i caigan dentro del mismo. Por tanto, la **frecuencia absoluta** F_k **del intervalo** I_k es igual al número de observaciones que caigan dentro de él, y su **frecuencia relativa** $f_k = F_k/\mathrm{N}$.

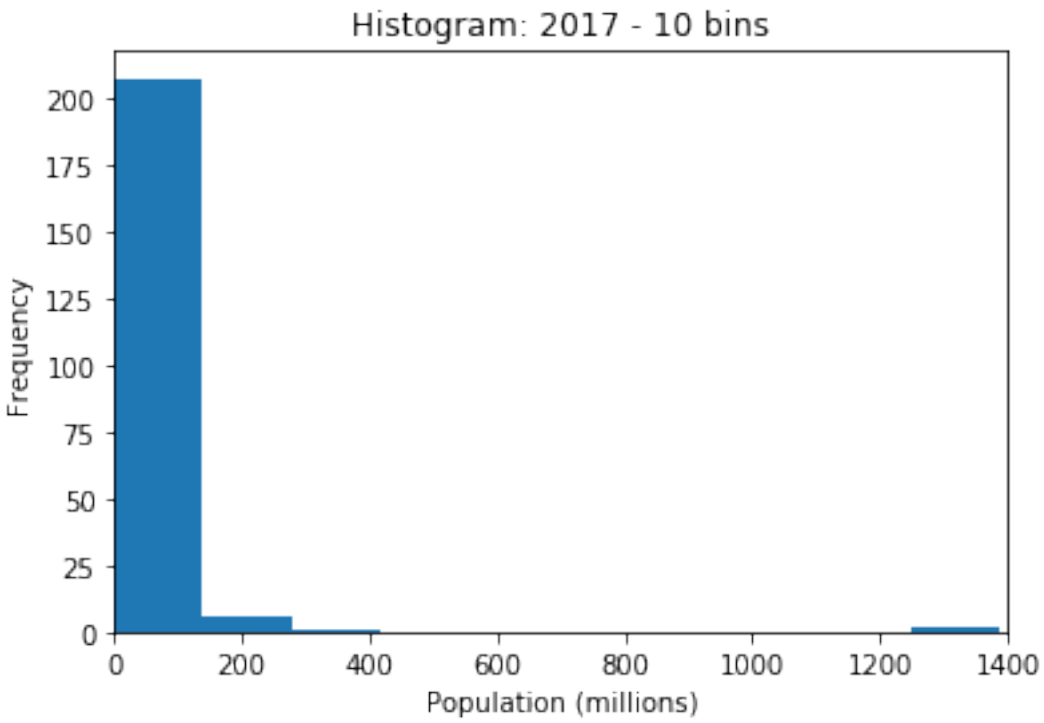
Los intervalos no tienen por qué tener igual longitud. **La partición del espacio muestral se diseña según se distribuyan las observaciones**.

Asignemos las poblaciones de 2017 de los países del mundo a una **partición de 10 intervalos idénticos**, y veamos las **frecuencias absolutas de cada intervalo** (cuántos países tiene cada uno). Los valores negativos se deben a que el software busca los límites automáticamente, sin que se le haya especificado que no puede haber valores de población negativos.

VER TRANSPARENCIA ONLINE

Country Code ABW AFG AGO \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code ALB AND ARE \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code ARG ARM ASM \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code ATG AUS AUT \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code AZE BDI BEL \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code BEN BFA BGD \ 2017 (-1.375, 138.65] (-1.375, 138.65] (138.65, 277.288]
Country Code BGR BHR BHS \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code BIH BLR BLZ \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code BMU BOL BRA \ 2017 (-1.375, 138.65] (-1.375, 138.65] (138.65, 277.288]
Country Code BRB BRN BTN \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code BWA CAF CAN \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code CHE CHI CHL \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code CHN CIV CMR \ 2017 (1247.757, 1386.395] (-1.375, 138.65] (-1.375, 138.65]
Country Code COD COG COL \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code COM CPV CRI \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code CUB CUW CYM \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code CYP CZE DEU \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code DJI DMA DNK \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code DOM DZA ECU \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code EGY ERI ESP EST \ 2017 (-1.375, 138.65] NaN (-1.375, 138.65] (-1.375, 138.65]
Country Code ETH FIN FJI \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code FRA FRO FSM \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code GAB GBR GEO \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code GHA GIB GIN \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code GMB GNB GNQ \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code GRC GRD GRL \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code GTM GUM GUY \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code HKG HND HRV \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code HTI HUN IDN \ 2017 (-1.375, 138.65] (-1.375, 138.65] (138.65, 277.288]
Country Code IMN IND IRL \ 2017 (-1.375, 138.65] (1247.757, 1386.395] (-1.375, 138.65]
Country Code IRN IRQ ISL \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code ISR ITA JAM \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code JOR JPN KAZ \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code KEN KGZ KHM \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]

Country Code KIR KNA KOR \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code KWT LAO LBN \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code LBR LBY LCA \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code LIE LKA LSO \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code LTU LUX LVA \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code MAC MAF MAR \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code MCO MDA MDG \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code MDV MEX MHL \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code MKD MLI MLT \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code MMR MNE MNG \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code MNP MOZ MRT \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code MUS MWI MYS \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code NAM NCL NER \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code NGA NIC NLD \ 2017 (138.65, 277.288] (-1.375, 138.65] (-1.375, 138.65]
Country Code NOR NPL NRU \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code NZL OMN PAK \ 2017 (-1.375, 138.65] (-1.375, 138.65] (138.65, 277.288]
Country Code PAN PER PHL \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code PLW PNG POL \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code PRI PRK PRT \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code PRY PSE PYF \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code QAT ROU RUS \ 2017 (-1.375, 138.65] (-1.375, 138.65] (138.65, 277.288]
Country Code RWA SAU SDN \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code SEN SGP SLB \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code SLE SLV SMR \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code SOM SRB SSD \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code STP SUR SVK \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code SVN SWE SWZ \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code SXM SYC SYR \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code TCA TCD TGO \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code THA TJK TKM \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code TLS TON TTO \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code TUN TUR TUV \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code TZA UGA UKR \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code URY USA UZB \ 2017 (-1.375, 138.65] (277.288, 415.926] (-1.375, 138.65]
Country Code VCT VEN VGB \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code VIR VNM VUT \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code WSM XKX YEM \ 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
Country Code ZAF ZMB ZWE 2017 (-1.375, 138.65] (-1.375, 138.65] (-1.375, 138.65]
(-1.375, 138.65] (138.65, 277.288] (277.288, 415.926] \ 2017 207 6 1
(415.926, 554.565] (554.565, 693.203] (693.203, 831.841] \ 2017 0 0 0
(831.841, 970.48] (970.48, 1109.118] (1109.118, 1247.757] \ 2017 0 0 0
(1247.757, 1386.395] 2017 2
Veamos el **histograma** correspondiente



Asignemos ahora las poblaciones a una partición de **100** y **200 intervalos** idénticos y representemos los histogramas respectivos.

`Partición en 100 intervalos`

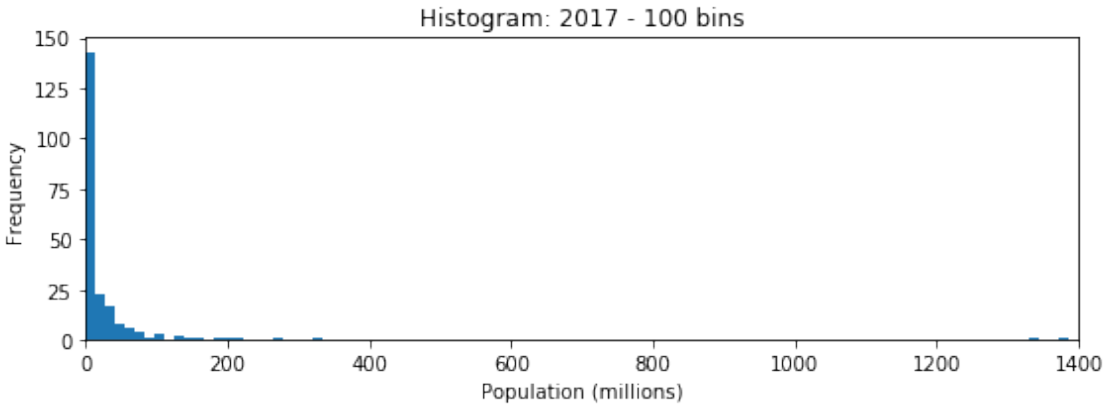
VER TRANSPARENCIA ONLINE
(-1.375, 13.875] (13.875, 27.739] (27.739, 41.603] (41.603, 55.467] \ 2017 143 23 17 8
(55.467, 69.33] (69.33, 83.194] (83.194, 97.058] (97.058, 110.922] \ 2017 6 4 1 3
(110.922, 124.786] (124.786, 138.65] (138.65, 152.513] \ 2017 0 2 1
(152.513, 166.377] (166.377, 180.241] (180.241, 194.105] \ 2017 1 0 1
(194.105, 207.969] (207.969, 221.833] (221.833, 235.696] \ 2017 1 1 0
(235.696, 249.56] (249.56, 263.424] (263.424, 277.288] \ 2017 0 0 1
(277.288, 291.152] (291.152, 305.016] (305.016, 318.879] \ 2017 0 0 0
(318.879, 332.743] (332.743, 346.607] (346.607, 360.471] \ 2017 1 0 0
(360.471, 374.335] (374.335, 388.199] (388.199, 402.062] \ 2017 0 0 0
(402.062, 415.926] (415.926, 429.79] (429.79, 443.654] \ 2017 0 0 0
(443.654, 457.518] (457.518, 471.382] (471.382, 485.246] \ 2017 0 0 0
(485.246, 499.109] (499.109, 512.973] (512.973, 526.837] \ 2017 0 0 0
(526.837, 540.701] (540.701, 554.565] (554.565, 568.429] \ 2017 0 0 0
(568.429, 582.292] (582.292, 596.156] (596.156, 610.02] \ 2017 0 0 0
(610.02, 623.884] (623.884, 637.748] (637.748, 651.612] \ 2017 0 0 0
(651.612, 665.475] (665.475, 679.339] (679.339, 693.203] \ 2017 0 0 0
(693.203, 707.067] (707.067, 720.931] (720.931, 734.795] \ 2017 0 0 0
(734.795, 748.658] (748.658, 762.522] (762.522, 776.386] \ 2017 0 0 0

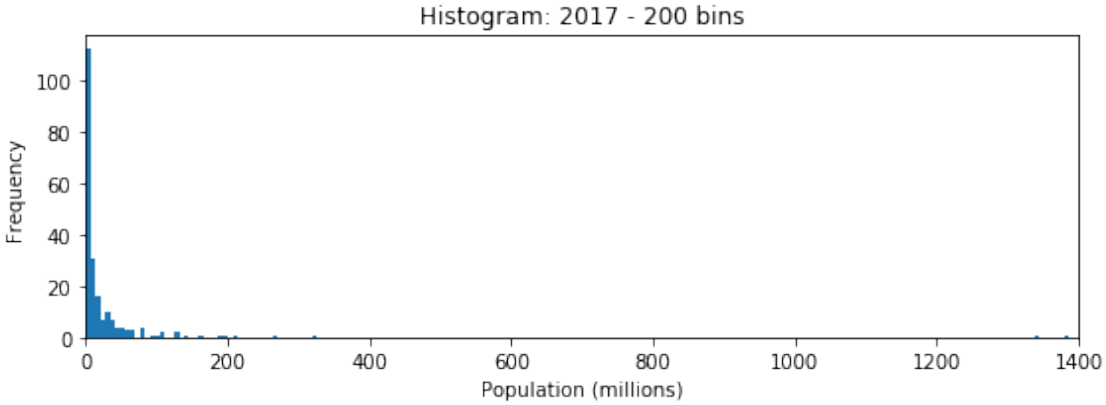
(776.386, 790.25] (790.25, 804.114] (804.114, 817.978] \ 2017 0 0 0
(817.978, 831.841] (831.841, 845.705] (845.705, 859.569] \ 2017 0 0 0
(859.569, 873.433] (873.433, 887.297] (887.297, 901.161] \ 2017 0 0 0
(901.161, 915.025] (915.025, 928.888] (928.888, 942.752] \ 2017 0 0 0
(942.752, 956.616] (956.616, 970.48] (970.48, 984.344] \ 2017 0 0 0
(984.344, 998.208] (998.208, 1012.071] (1012.071, 1025.935] \ 2017 0 0 0
(1025.935, 1039.799] (1039.799, 1053.663] (1053.663, 1067.527] \ 2017 0 0 0
(1067.527, 1081.391] (1081.391, 1095.254] (1095.254, 1109.118] \ 2017 0 0 0
(1109.118, 1122.982] (1122.982, 1136.846] (1136.846, 1150.71] \ 2017 0 0 0
(1150.71, 1164.574] (1164.574, 1178.437] (1178.437, 1192.301] \ 2017 0 0 0
(1192.301, 1206.165] (1206.165, 1220.029] (1220.029, 1233.893] \ 2017 0 0 0
(1233.893, 1247.757] (1247.757, 1261.62] (1261.62, 1275.484] \ 2017 0 0 0
(1275.484, 1289.348] (1289.348, 1303.212] (1303.212, 1317.076] \ 2017 0 0 0
(1317.076, 1330.94] (1330.94, 1344.803] (1344.803, 1358.667] \ 2017 0 1 0
(1358.667, 1372.531] (1372.531, 1386.395] 2017 0 1

Partición en 200 intervalos

VER TRANSPARENCIA ONLINE
(-1.375, 6.943] (6.943, 13.875] (13.875, 20.807] (20.807, 27.739] \ 2017 112 31 16 7
(27.739, 34.671] (34.671, 41.603] (41.603, 48.535] (48.535, 55.467] \ 2017 10 7 4 4
(55.467, 62.398] (62.398, 69.33] (69.33, 76.262] (76.262, 83.194] \ 2017 3 3 0 4
(83.194, 90.126] (90.126, 97.058] (97.058, 103.99] (103.99, 110.922] \ 2017 0 1 1 2
(110.922, 117.854] (117.854, 124.786] (124.786, 131.718] \ 2017 0 0 2
(131.718, 138.65] (138.65, 145.581] (145.581, 152.513] \ 2017 0 1 0
(152.513, 159.445] (159.445, 166.377] (166.377, 173.309] \ 2017 0 1 0
(173.309, 180.241] (180.241, 187.173] (187.173, 194.105] \ 2017 0 0 1
(194.105, 201.037] (201.037, 207.969] (207.969, 214.901] \ 2017 1 0 1
(214.901, 221.833] (221.833, 228.765] (228.765, 235.696] \ 2017 0 0 0
(235.696, 242.628] (242.628, 249.56] (249.56, 256.492] \ 2017 0 0 0
(256.492, 263.424] (263.424, 270.356] (270.356, 277.288] \ 2017 0 1 0
(277.288, 284.22] (284.22, 291.152] (291.152, 298.084] \ 2017 0 0 0
(298.084, 305.016] (305.016, 311.948] (311.948, 318.879] \ 2017 0 0 0
(318.879, 325.811] (325.811, 332.743] (332.743, 339.675] \ 2017 1 0 0
(339.675, 346.607] (346.607, 353.539] (353.539, 360.471] \ 2017 0 0 0
(360.471, 367.403] (367.403, 374.335] (374.335, 381.267] \ 2017 0 0 0
(381.267, 388.199] (388.199, 395.131] (395.131, 402.062] \ 2017 0 0 0
(402.062, 408.994] (408.994, 415.926] (415.926, 422.858] \ 2017 0 0 0
(422.858, 429.79] (429.79, 436.722] (436.722, 443.654] \ 2017 0 0 0
(443.654, 450.586] (450.586, 457.518] (457.518, 464.45] \ 2017 0 0 0
(464.45, 471.382] (471.382, 478.314] (478.314, 485.246] \ 2017 0 0 0
(485.246, 492.177] (492.177, 499.109] (499.109, 506.041] \ 2017 0 0 0
(506.041, 512.973] (512.973, 519.905] (519.905, 526.837] \ 2017 0 0 0
(526.837, 533.769] (533.769, 540.701] (540.701, 547.633] \ 2017 0 0 0
(547.633, 554.565] (554.565, 561.497] (561.497, 568.429] \ 2017 0 0 0
(568.429, 575.36] (575.36, 582.292] (582.292, 589.224] \ 2017 0 0 0
(589.224, 596.156] (596.156, 603.088] (603.088, 610.02] \ 2017 0 0 0

(610.02, 616.952] (616.952, 623.884] (623.884, 630.816] \ 2017 0 0 0
(630.816, 637.748] (637.748, 644.68] (644.68, 651.612] \ 2017 0 0 0
(651.612, 658.544] (658.544, 665.475] (665.475, 672.407] \ 2017 0 0 0
(672.407, 679.339] (679.339, 686.271] (686.271, 693.203] \ 2017 0 0 0
(693.203, 700.135] (700.135, 707.067] (707.067, 713.999] \ 2017 0 0 0
(713.999, 720.931] (720.931, 727.863] (727.863, 734.795] \ 2017 0 0 0
(734.795, 741.727] (741.727, 748.658] (748.658, 755.59] \ 2017 0 0 0
(755.59, 762.522] (762.522, 769.454] (769.454, 776.386] \ 2017 0 0 0
(776.386, 783.318] (783.318, 790.25] (790.25, 797.182] \ 2017 0 0 0
(797.182, 804.114] (804.114, 811.046] (811.046, 817.978] \ 2017 0 0 0
(817.978, 824.91] (824.91, 831.841] (831.841, 838.773] \ 2017 0 0 0
(838.773, 845.705] (845.705, 852.637] (852.637, 859.569] \ 2017 0 0 0
(859.569, 866.501] (866.501, 873.433] (873.433, 880.365] \ 2017 0 0 0
(880.365, 887.297] (887.297, 894.229] (894.229, 901.161] \ 2017 0 0 0
(901.161, 908.093] (908.093, 915.025] (915.025, 921.956] \ 2017 0 0 0
(921.956, 928.888] (928.888, 935.82] (935.82, 942.752] \ 2017 0 0 0
(942.752, 949.684] (949.684, 956.616] (956.616, 963.548] \ 2017 0 0 0
(963.548, 970.48] (970.48, 977.412] (977.412, 984.344] \ 2017 0 0 0
(984.344, 991.276] (991.276, 998.208] (998.208, 1005.139] \ 2017 0 0 0
(1005.139, 1012.071] (1012.071, 1019.003] (1019.003, 1025.935] \ 2017 0 0 0
(1025.935, 1032.867] (1032.867, 1039.799] (1039.799, 1046.731] \ 2017 0 0 0
(1046.731, 1053.663] (1053.663, 1060.595] (1060.595, 1067.527] \ 2017 0 0 0
(1067.527, 1074.459] (1074.459, 1081.391] (1081.391, 1088.322] \ 2017 0 0 0
(1088.322, 1095.254] (1095.254, 1102.186] (1102.186, 1109.118] \ 2017 0 0 0
(1109.118, 1116.05] (1116.05, 1122.982] (1122.982, 1129.914] \ 2017 0 0 0
(1129.914, 1136.846] (1136.846, 1143.778] (1143.778, 1150.71] \ 2017 0 0 0
(1150.71, 1157.642] (1157.642, 1164.574] (1164.574, 1171.506] \ 2017 0 0 0
(1171.506, 1178.437] (1178.437, 1185.369] (1185.369, 1192.301] \ 2017 0 0 0
(1192.301, 1199.233] (1199.233, 1206.165] (1206.165, 1213.097] \ 2017 0 0 0
(1213.097, 1220.029] (1220.029, 1226.961] (1226.961, 1233.893] \ 2017 0 0 0
(1233.893, 1240.825] (1240.825, 1247.757] (1247.757, 1254.689] \ 2017 0 0 0
(1254.689, 1261.62] (1261.62, 1268.552] (1268.552, 1275.484] \ 2017 0 0 0
(1275.484, 1282.416] (1282.416, 1289.348] (1289.348, 1296.28] \ 2017 0 0 0
(1296.28, 1303.212] (1303.212, 1310.144] (1310.144, 1317.076] \ 2017 0 0 0
(1317.076, 1324.008] (1324.008, 1330.94] (1330.94, 1337.872] \ 2017 0 0 0
(1337.872, 1344.803] (1344.803, 1351.735] (1351.735, 1358.667] \ 2017 1 0 0
(1358.667, 1365.599] (1365.599, 1372.531] (1372.531, 1379.463] \ 2017 0 0 0
(1379.463, 1386.395] 2017 1

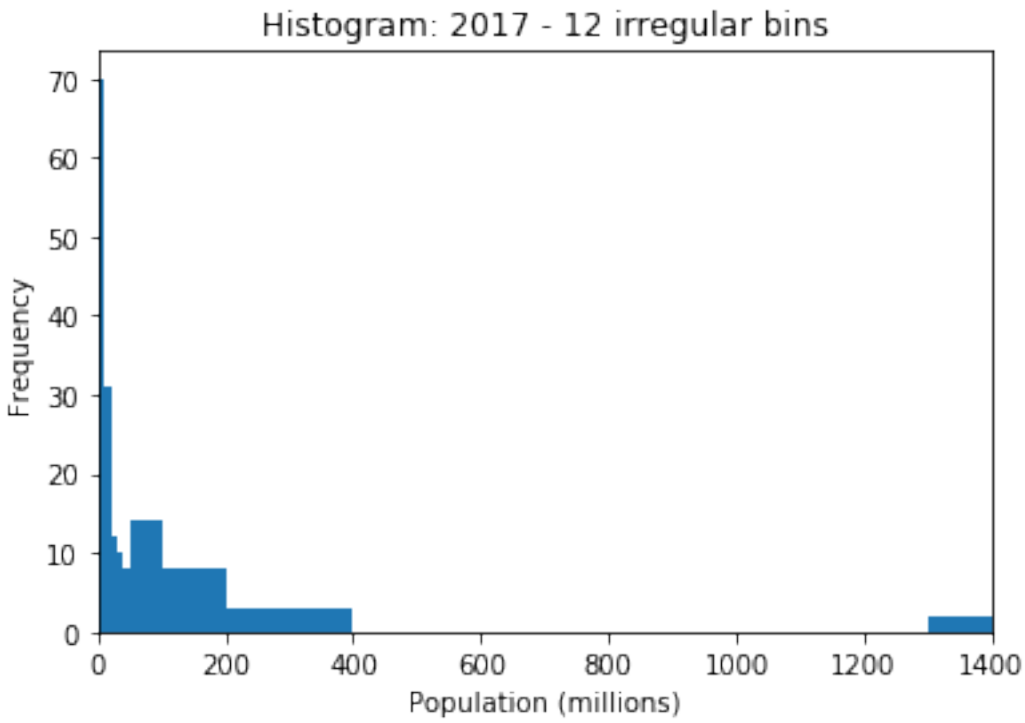




Hagamos ahora una partición en 12 intervalos irregulares, cuyos puntos de frontera son

$$\{0, 0.5, 1, 10, 20, 30, 40, 50, 100, 200, 400, 1300, 1400\}$$

VER TRANSPARENCIA ONLINE
(0.0, 0.5] (0.5, 1.0] (1.0, 10.0] (10.0, 20.0] (20.0, 30.0] \ 2017 47 11 70 31 12
(30.0, 40.0] (40.0, 50.0] (50.0, 100.0] (100.0, 200.0] \ 2017 10 8 14 8
(200.0, 400.0] (400.0, 1300.0] (1300.0, 1400.0] 2017 3 0 2



Histograma acumulativo

Las **frecuencias absolutas acumulativas** en el valor de cada observación correesponde a la suma de frecuencias absolutas acumuladas hasta dicha observación:

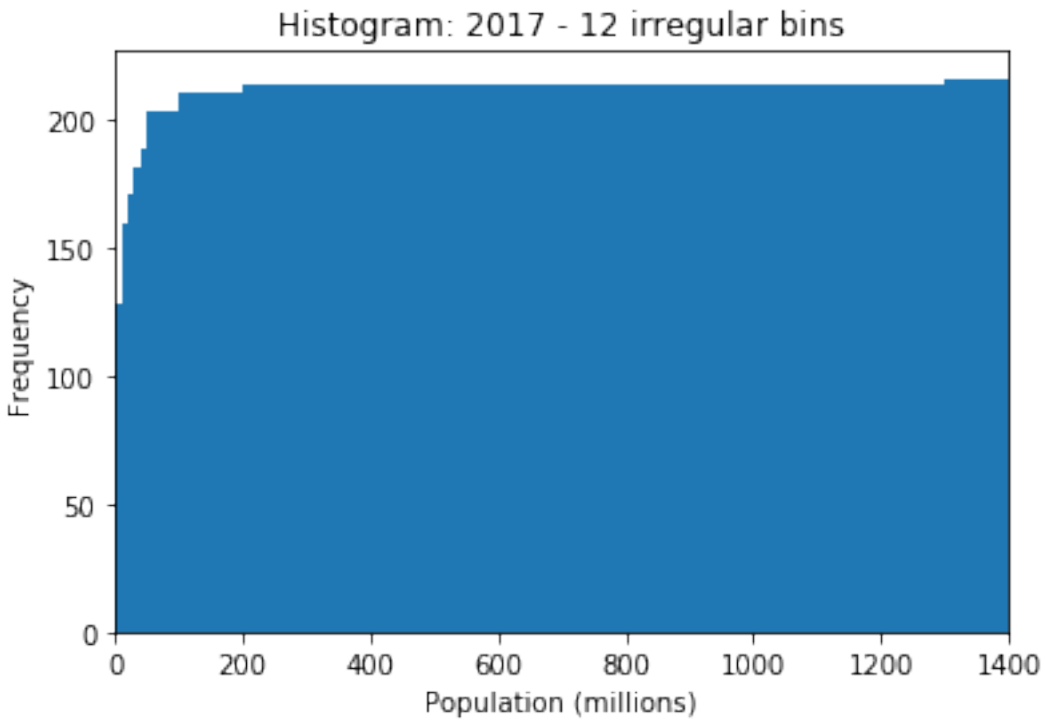
$$F_{cum}(x_n) = \sum_{i \leq n} F(x_i)$$

De forma semejante a como hemos hecho antes, podemos definir también:

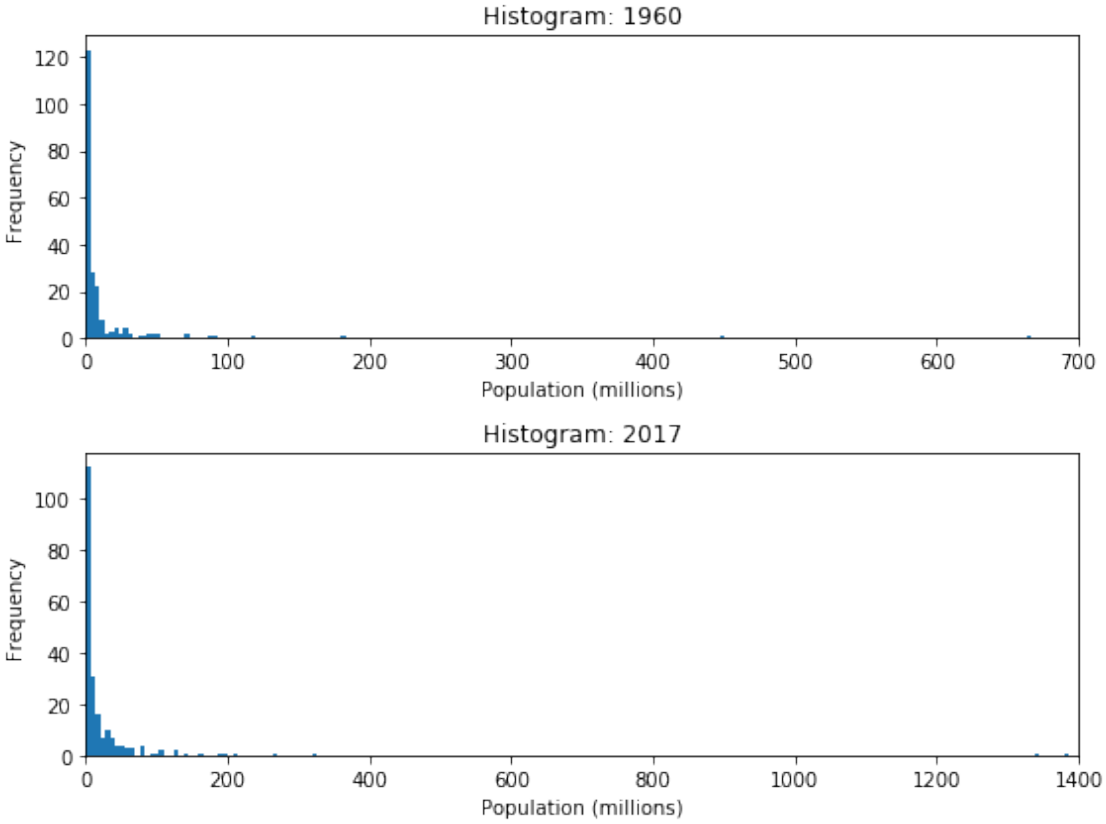
- La **frecuencia cumulativa relativa** sin más que dividir por el número de observaciones N.
- El **histograma acumulativo**

VER TRANSPARENCIA ONLINE

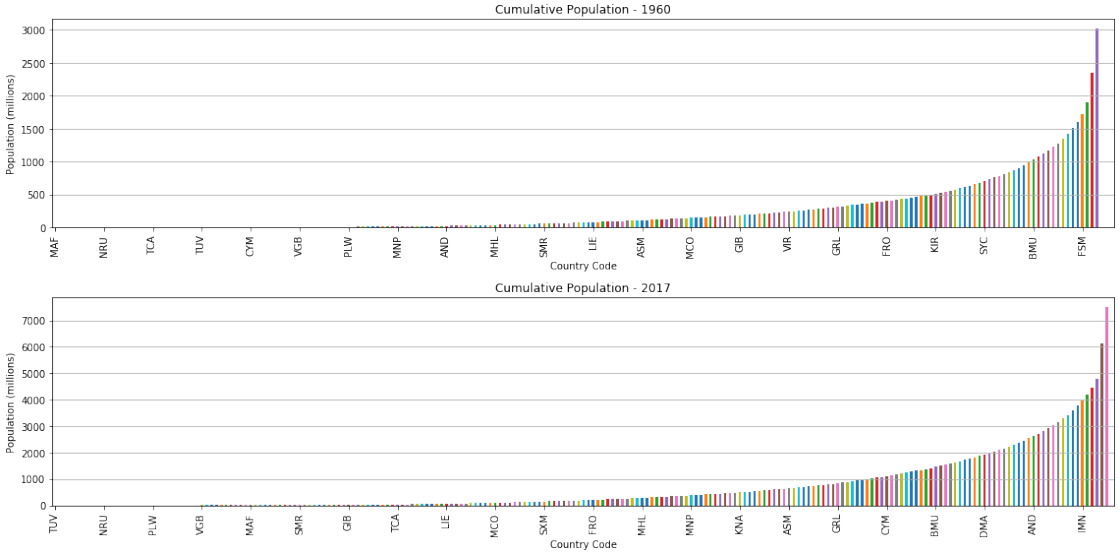
F Fcum (0.0, 0.5] 47 47 (0.5, 1.0] 11 58 (1.0, 10.0] 70 128 (10.0, 20.0] 31 159 (20.0, 30.0] 12 171 (30.0, 40.0] 10 181 (40.0, 50.0] 8 189 (50.0, 100.0] 14 203 (100.0, 200.0] 8 211 (200.0, 400.0] 3 214 (400.0, 1300.0] 0 214 (1300.0, 1400.0] 2 216



Como colofón, comparemos los histogramas de las poblaciones de los países del mundo en 1960 y 2017, ambos calculados con 200 intervalos iguales:



Y las poblaciones acumuladas de todos los países del mundo a lo largo del periodo 1960 a 2017:



VER TRANSPARENCIA ONLINE

Year 1960 1970 1980 1990 2000 2010 2015 2017 suma 3014.9 3664.3 4414.3 5267.9 6099.5 6909.7 7329.3 7501.7 WLD 3032.2 3685.8 4439.3 5288.1 6121.7 6932.9 7357.6 7530.4

Estadísticos descriptivos: cuantiles, forma, tendencia central, dispersión

Un **estadístico**, también llamado **estadístico muestral** es una función matemática de las observaciones o muestras de una **variable muestral**, que permite hacer una caracterización parcial de las mismas. Utilizando simultáneamente varios estadísticos es posible hacer una caractericación más completa de la muestra.

Dado que las observaciones o muestras estadísticas representan aspectos de una población, los estadísticos calculados a partir de las muestras también caracterizan parcialmente a la población.

Consideremos que nuestra muestra consiste en N observaciones x_i , $i = 1 \dots N$ de una variable muestral. Llamamos **estadístico** tanto a una función f definida sobre tales observaciones, como al valor que tomo la función al ser evaluada $f(x_1, \dots, x_N)$. Salvo que indiquemos otra cosa, entenderemos que estadístico se refiere a esta segunda acepción (el valor calculado por la función a partir de las observaciones).

Adviértase que **un estadístico es aleatorio**, pues está calculado a partir de observaciones que también lo son.

Hay distintos tipos de estadístico, dependiendo del uso que se les quiera dar. Entre ellos:

- **Estadísticos descriptivos**: se utilizan para realizar un resumen descriptivo de la muestra, y se asocian generalmente al anáĺisis exploratorio de datos y a la estadística descriptiva.
- **Estimadores**: se utilizan para estimar parámetros de poblaciones estadísticas.
- **Estadísticos de test**: se utilizan para hacer contrastes de hipótesis a partir de las observaciones.

Los estadísticos son **robustos** si se comportan bien en presencia de observaciones atípicas (*outliers*).

De momento nos centraremos en los estadísticos descriptivos más comunes, que suelen denominarse **estadísticos descriptivos de resumen** (*summary statistics*), pues nos permiten obtener un rápido resumen de las características de la muestra.

Estadísticos descriptivos más habituales:

- **Estadísticos cuantiles**: entre ellos, la **mediana**, **cuartiles**, **deciles** y **centiles**.
- **Estadísticos de forma de la distribución**: son el **coeficiente de apuntamiento** (*skewness*) y el **coeficiente de kurtosis**.
- **Estadísticos de tendencia central**: son, entre otros, el **rango medio** (*midrange*), el **rango medio intercuartil** (*midhinge*), la **media**, la **mediana**, la **media ponderada**, las **medias recortadas** y la **moda**.
- **Estadísticos de dispersión**: son, entre otros, la **desviación típica** y la **varianza** y los distintos **rangos**.
- **Estadísticos de dependencia**: por ejemplo, el **coeficiente de correlación**.

Los estadísticos de orden, entre ellos el mínimo y el máximo, y las frecuencias de las onservaciones e intervalos son también estadísticos.

Cuantiles

Los **cuantiles** agrupan las observaciones de una variable estadística *ordenada* (numérica o categórica con relación de orden) en intervalos con *aproximadamente* igual número de muestras. Estos intervalos se ordenan sucesivamente, tocándose en un único punto de frontera, sin solaparse, y el conjunto de todos ellos contiene todas las observaciones.

Los cuantiles son los puntos de frontera entre intervalos sucesivos. Si q es el número de intervalos de la partición, tendremos $q-1$ puntos frontera que llamaremos **q-quantiles**. Los intervalos contendrán idéntico número de muestras si el número total de observaciones N es divisible por el número de intervalos q.

Algunos cuantiles tienen nombres especiales:

- Partición en 2 intervalos: hay un único 2-cuantil, que se llama **mediana**. Cada intervalo contiene el 50% de las observaciones.
- Particiones en 4 intervalos: hay tres 4-cuantiles, Q_1, Q_2 y Q_3, que se llaman **cuartiles**. Cada intervalo contiene el 25% de las observaciones. El cuartil Q_2 coincide con la mediana.
- Particiones en 10 intervalos: hay nueve 10-cuantiles, $D_1 \dots D_9$, que se llaman **deciles**. Cada intervalo contiene el 10% de las observaciones.
- Particiones en 100 intervalos: hay noventa y nueve 100-cuantiles, $C_1 \dots C_{99}$, que se llaman **centiles**. Cada intervalo contiene el 1% de las observaciones.

Los cuantiles proporcionan una útil interpetación en términos de la **frecuencia acumulada relativa**, esto es del *porcentaje de observaciones que quedan por debajo*.

- La mediana tiene *aproximadamente* el 50% de las observaciones por debajo de su valor.
- En cuanto a los cuartiles:
- El cuartil $Q_1 \equiv Q_{25\%}$ tiene *aproximadamente* el 25% de las muestras por debajo.
- El cuartil $Q_2 \equiv Q_{50\%}$ coincide con la mediana y tiene *aproximadamente* el 50% de las observaciones por debajo.
- El cuartil $Q_3 \equiv Q_{75\%}$ coincide con la mediana y tiene *aproximadamente* el 75% de las observaciones por debajo.
- A veces se habla del cuartil Q_0, valor mínimo sin ninguna observación por debajo, y del cuartil Q_4, valor máximo con todas las observaciones por debajo.
- El planteamiento es el mismo con otros cuatiles. Por ejemplo:
- El decil $D_3 \equiv D_{30\%}$ tiene el 30% de las observaciones por debajo.
- El centil $C_{55} \equiv C_{55\%}$ tiene el 55% de las observaciones por debajo.

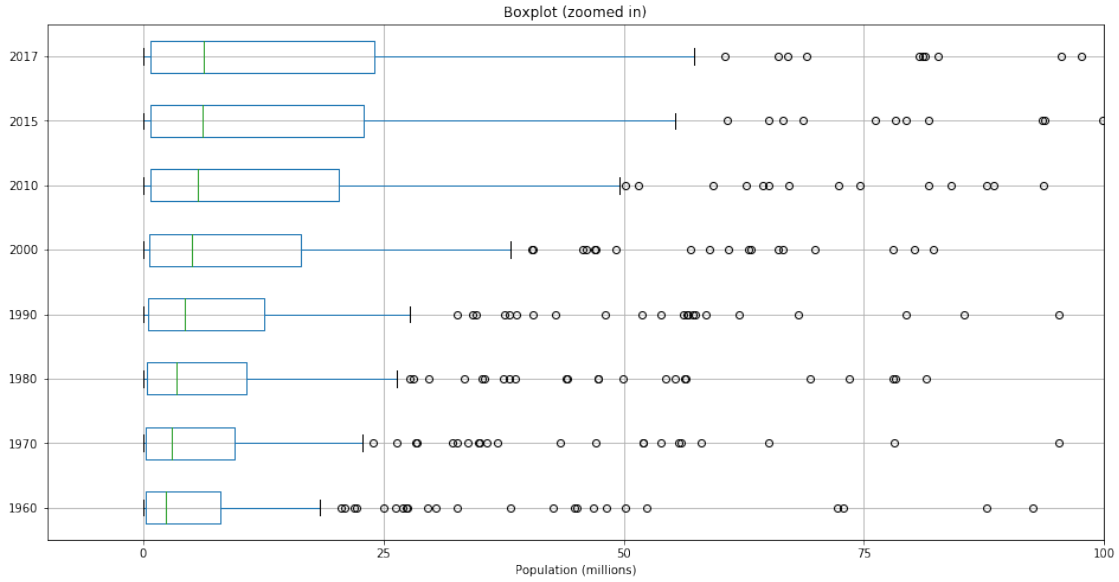
Veamos, por ejemplo, los **deciles** de las poblaciones de los países del mundo en 2017:
VER TRANSPARENCIA ONLINE
(0.0102, 0.0901] 22 (0.0901, 0.395] 22 (0.395, 1.431] 21 (1.431, 3.717] 22 (3.717, 6.296] 21 (6.296, 10.294] 22 (10.294, 17.113] 21 (17.113, 31.977] 22 (31.977, 63.287] 21 (63.287, 1386.395] 22 Name: 2017, dtype: int64
Igualmente, los **cuartiles** del año 2017:
(0.0102, 0.8] 54 (0.8, 6.296] 54 (6.296, 24.114] 54 (24.114, 1386.395] 54 Name: 2017, dtype: int64

Diagrama de caja y bigote (box and whisker plot)

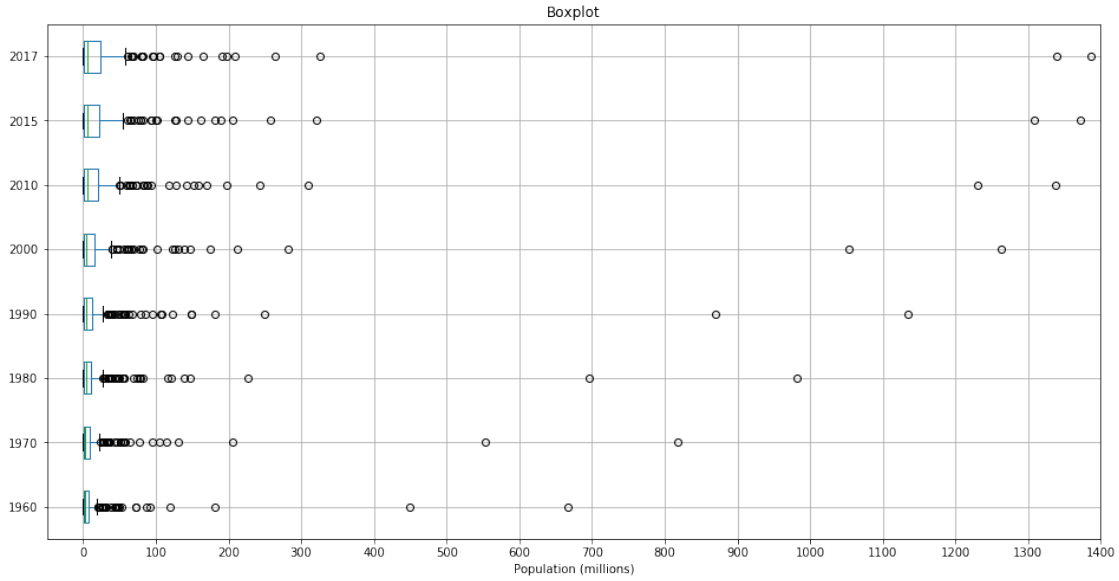
La distribución de las observaciones en el especio muestral puede ilustrarse mediante un diagrama de cajas (*boxplot*) y bigote (*whisker*):

- La **mediana** o segundo intercuantil es, como veremos, un *indicador de la tendencia central o localización de la muestra* o población estadística.
- La **caja** se forma con el primer cuartil $Q1$, la mediana o segundo cuartil $Q2$ y el tercer cuartil.
- La cantidad $IQR = Q_3 - Q_1$ se llama **rango intercuantil**, y, como veremos, es un *indicador de la dispersión o escala de la muestra* o población estadística.
- El **bigote** se forma extendiendo los extremos de la caja un máximo de $1.5 \times IQR$, si bien el coeficiente 1.5 puede variarse dependiendo de la representación.
- Las observaciones que quedan por debajo o por encima de los bigotes se consideran atípicas (*ouliers*) y se representan con círculos.

Representemos el diagramas de caja y bigote de las poblaciones de los países del mundo en un conjunto de años seleccionados, eliminando todos los que tengan población mayor de 100 millones para mejorar visualización.



Y ahora considerando todos los países, sin exclusión:



Coeficientes de forma

La distribución de observaciones se caracteriza con los diversos estadísticos descriptivos, y se analiza gráficamente, entre otras posibilidades, con los histogramas y diagramas de caja y bigotes.

- Localización o tendencia central: valor en torno al cual se distribuyen las observaciones. Por ejemplo, la mediana o segundio cuartil, el rango medio (*midrange*) $\frac{mn+mx}{2}$, el rango medio intercuartil (*midhinge*) $\frac{Q_1+Q_3}{2}$ o la trimedia (*trimean*) $\frac{Q_1+2Q_2+Q_3}{4}$.
- Escala o dispersión: medida de cuánto se alejan las observaciones entre sí o de su localización central. Por ejemplo, el rango $(mx - mn)$ o el rango intercuantil $IQR = Q_3 - Q_1$.
- Asimetría: si el histograma se extiende hacia la derecha o hacia la izquierda
- Apuntamiento (*kurtosis*): indica si las observaciones se distribuyen principalmente en torno a la localización central o, por el contrario, si el histograma muestra colas laterales gruesas.

Coeficiente de asimetría (*skewness*) El coeficiente de asimetría es:

- **nulo** si las observaciones se distribuyen simétricamente en torno a la localización central. El histograma muestra un eje vertical de simetría.
- **negativo** si las observaciones se extienden hacia la izquierda y están más concentradas hacia la derecha. El histograma muestra una cola más pronunciada hacia la izquierda.
- **positivo** si las observaciones se extienden hacia la derecha y están más concentradas hacia la izquierda. El histograma muestra una cola más pronunciada hacia la derecha.

Hay varios estadísticos para hacer el cómputo del coeficiente de asimetría, aunque no vamos a verlos aquí.

Coeficiente de curtosis Vamos a trabajar con el **exceso de curtosis** que es el coeficiente *habitual* de curtosis, que por definición es siempre positivo, minorado en 3 para que la distribución Gaussiana o normal resulte con exceso de curtosis nulo:

- **nulo** si las observaciones siguen una *distribución normal*. Digamos que es una situación intermedia. Estrictamente, la curtosis de la distribución normal es 3 pero, como se ha dicho, se suele trabajar con el exceso de curtosis, restando 3.
- **positivo** si las observaciones se concentran en torno a la localización central, con colas gruesas y menos frecuencias de valores intermedios que en la distribución normal. Cuanto mayor sea el número mayor es el apuntamiento, esto es, mayor concentración en torno a la localización central y, al tiempo, colas más gruesas.
- **negativo** si las observaciones se concentran menos en torno a la localización central, con colas estrechas o sin ellas y con más fecuencias de valores intermedios en que la distribución normal. Cuanto menor sea el número menor es el apuntamiento, esto es, menor concentración en torno a la localización central, colas más estrechas y más valores intermedios.

Los valores negativos se deben a que se trata de un exceso de curtosis, pues estricamente la curtosis siempre es positiva.

VER TRANSPARENCIA ONLINE

Year 1960 1970 1980 1990 2000 2010 2015 2017 midrange 333.5 409.2 490.6 567.6 631.3 668.9 685.6 693.2 median 2.4 2.9 3.5 4.4 5.0 5.7 6.2 6.3 midhinge 4.1 4.9 5.6 6.6 8.5 10.6 11.8 12.5 trimean 3.3 3.9 4.5 5.5 6.8 8.2 9.0 9.4 range 667.1 818.3 981.2 1135.2 1262.6 1337.7 1371.2 1386.4 iqr 7.8 9.2 10.4 12.2 15.8 19.7 22.1 23.3 skew 9.1 9.2 9.2 9.2 9.1 9.0 8.9 8.9 kurt 92.4 94.4 93.8 91.9 89.2 86.3 84.8 84.5

1960 1970 1980 1990 2000 2010 2015 2017 median 2.4 2.9 3.5 4.4 5.0 5.7 6.2 6.3 skew 9.1 9.2 9.2 9.2 9.1 9.0 8.9 8.9 kurt 92.4 94.4 93.8 91.9 89.2 86.3 84.8 84.5

Estadísticos de tendencia central

Calculan una localización o valor típico en torno al cual se distribuyen las observaciones o muestras. Ya hemos visto algunes estadísticos de tendencia central:

- El **rango medio** (*midrange*) o valor medio de las observaciones mínima y máxima $\frac{mn+mx}{2}$.
- La **mediana** o segundo cuartil.
- El **rango medio intercuartil** (*midhinge*) o valor medio del primer y tercer cuartil $\frac{Q_1+Q_3}{2}$.

Veamos ahora:

- La **media**
- Las **medias recortadas** o **truncadas** (*trimmed / truncated means*)
- La **media ponderada** (*weighted mean*)
- La **moda**
- El **valor cuadrático medio** (*Root Mean Square - RMS*)

Ilustraremos las definiciones con dos secuencias de observaciones:

- Número impar de observaciones: $S_1 = (2, 1, 2, 6, 4)$
- Número par de observaciones: $S_2 = (2, 1, 2, 6, 4, 99)$. Puede verse que esta secuencia es la misma que la anterior, con una observación más.

Advirtamos que el último valor es muy grande en relación al resto de muestras. Se trata de un **valor atípico** (*outlier*), que debe manejarse con cuidado. ¿Habrá habido un error de medida o es correcta la muestra? Como veremos, los valores atípicos pueden producir efectos indeseados en el cómputo de algunos estadísticos.

Rango medio Como hemos visto, es el promedio de los valores mínimo y máximo.

Calculemos el rango medio de las dos secuencias de ejemplo:

- $min(S_1) = 1$, $max(S_1) = 6 \implies range(S_1) = \frac{1+6}{2} = 3.5$
- $min(S_2) = 1$, $max(S_2) = 99 \implies range(S_2) = \frac{1+99}{2} = 50$

```
range(S1): 3.5
range(S2): 50.0
```

Rango medio intercuartil Como también hemos visto, se trata del promedio del primer y tercer cuartiles.

Hagamos el cálculo con las dos secuencias de ejemplo, cuyos valores debemos ordenar y asignar a cuatro intervalos.

- $S_1 : (1, \mathbf{2}, \mathbf{2}, 4, \mathbf{6}) \implies min = 1, Q_1 = 2, Q_2 = 2, Q_3 = 4, max = 6$

$$midhinge(S_1) = \frac{2+4}{2} = 3$$

- $S_2 : (1, \mathbf{2}, \mathbf{2}, 4, 6, \mathbf{99}) \implies mn = 1, Q_1 = 2, Q_2 = 3, Q_3 = 5.5, mx = 99$

$$midhinge(S_2) = \frac{2+5.5}{2} = 3.75$$

Como puede verse hay una ambigüedad para asignar observaciones a los intervalos y calcular los cuartiles. Mostramos el cáculo obtenido en *Pandas* con interpolación lineal.

```
In [39]: pd.DataFrame({'S_1': S_1.quantile([0,.25,.50,.75,1]),
                        'S_2': S_2.quantile([0,.25,.50,.75,1])})
```

VER TRANSPARENCIA ONLINE

S_1 S_2 0.0 1.0 1.0 0.2 2.0 2.0 0.5 2.0 3.0 0.8 4.0 5.5 1.0 6.0 99.0

Mediana muestral Es el valor intermedio de la variable, que tiene tantas observaciones por encima como por debajo de ella. Para calcularla, debemos ordenar las observaciones de menor a mayor.

Si el número de observaciones es impar, la mediana muestral se obtiene sin ambigüedad, pues coincide con la observación que está justo en el centro de la secuencia ordenada.

Si es el número de observaciones es par hay que obtener la mediana a partir de las dos observaciones centrales, generalmente por interpolación lineal.

Veámoslo para las dos secuencias anteriores, cuyas muestras ahora hemos de ordenar:

- $S_1 = (2, 1, 2, 5, 3) \rightarrow (1, 2, \mathbf{2}, 4, 6)$: El valor que está en medio es el 2.
- $S_2 = (2, 1, 2, 5, 3, 99) \rightarrow (1, 2, \mathbf{2}, \mathbf{4}, 6, 99)$: Los valores que hay en medio son el 2 y el 4, que debemos promediar.

$$median(S_1) = 2$$

$$median(S_2) = \frac{2+4}{2} = 3$$

Adviértase que **la mediana muestral puede calcularse con variables categóricas, siempre que sus valores admitan una relación de orden**. Por ejemplo, si la variable categórica son personas, y puedo ordenarlas por estatura, puedo determinar quien es la persona con estatura mediana aunque no conozca los valores numéricos de las estaturas.

Media muestral Es el *promedio de las observaciones*, calculado como la suma de todas las observaciones dividida por el número total de las mismas

$$\hat{x} = mean(x_1, \ldots, x_N) = \frac{1}{N} \sum_{1=1}^{N} x_i$$

La media muestral de la primera secuencia, S_1 es:

$$mean(S_1) = \frac{1}{5}(2 + 1 + 2 + 6 + 4) = 15/5 = 3$$

Adviértase como varía la media muestral de la segunda secuencia, S_2. Esto es debido al **valor atípico** 114:

$$mean(S_2) = \frac{1}{6}(2 + 1 + 2 + 6 + 4 + 99) = 114/6 = 19$$

Medias muestrales recortadas o truncadas (trimmed / truncated means) Se trata de un conjunto de técnicas que intentan proporcionar **estadísticos robustos de localización central**, eliminando valores atípicos conforme a algún criterio:

- **Media truncada al n%**: se descartan el n% de las observaciones menores y el n% de las observaciones mayores, calculándose la media muestral de las observaciones restantes.
- **Media modificada**: se descartan los valores menor y mayor de la muestra.
- **Medias interdecil** e **intercuartil**: la media muestral se restringe a las observaciones que están, respectivamente, entre el primer y noveno decil, y entre el primer y tercer cuartil, descartando las restantes. Corresponden, respectivamente a las medias truncadas al 10% y al 25%.
- **Medias *winsorizadas***: en vez de descartar las observaciones, éstas se sustituyen con los valores extremos de las que se consideran para el cálculo.

Veamos, por ejemplo, las **media truncada** al 17% de la secuencia S_2. Esta secuencia tiene 6 muestras, de modo que el 17% es aproximadamente 1 muestra que habría que quitar de ambos extremos. Por tanto, coincide con la **media modificada**:

$$mean_{17\%}(S_2) = \frac{1}{4}(2 + 2 + 6 + 4) = 14/4 = 3.5$$

Podemos también calcular la versión *winsorizada*:

$$winsormean_{17\%}(S_2) = \frac{1}{4}(2+2+2+6+4+6) = 22/6 = 11/3 \approx 3.667$$

Media muestral *ponderada* Ponderamos cada observación con un coeficiente α_i, que expresa la confianza que tenemos en cada una de ellas. Si confiamos igualmente en todas las observaciones, todos los coeficientes son idénticos e iguales a $\alpha_i = 1/N$ y resulta la media muestral o promedio anterior.

$$\hat{x} = mean_w(x_1, \ldots, x_N) = \sum_{1=1}^{N} \alpha_i x_i \qquad 0 \leq \alpha_i \leq 1 \qquad \sum_{i=1}^{N} \alpha_i = 1$$

Supongamos que en las secuencias anteriores tenemos plena confianza en nuestras primeras cinco muestras, pero que es nula en la sexta, pues nos proporciona un valor atípico. Podemos plantear una ponderación como sigue:

$$\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0.20 \quad \alpha_6 = 0 \qquad \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + \alpha_5 = 1$$

$$mean_w(S_2) = 0.20 \times 2 + 0.20 \times 1 + 0.20 \times 2 + 0.20 \times 6 + 0.20 \times 4 + 0 \times 99 = 3$$

Tal vez sí oterguemos cierta confianza a la última observación pero, en todo caso, inferior a las anteriores. Por ejemplo:

$$\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0.19 \quad \alpha_6 = 0.05 \qquad \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + \alpha_5 = 1$$

$$mean_w(S_2) = 0.19 \times 2 + 0.18 \times 1 + 0.18 \times 2 + 0.18 \times 6 + 0.18 \times 4 + 0.10 \times 99 = 7.8$$

Moda muestral Es el valor de la variable con mayor frecuencia absoluta. En el ejemplo anterior, sería el color con mayor número de coches.

Podemos también obtener la moda de nuestras secuencias numéricas de ejemplo, por ejemplo de la $S_1 = (2, 1, 2, 5, 3)$. Veamos las frecuencias absolutas de cada valor:

- El 1 aparece 1 vez: $F_1 = 1$
- El 2 aparece 2 veces: $F_2 = 2$
- El 3 aparece 1 vez: $F_3 = 1$
- El 4 no aparece: $F_4 = 0$
- El 5 aparece 1 vez: $F_5 = 1$

Por tanto, la moda corresponde al valor 2. Sin embargo, si tuviéramos la secuencia $S_3 = (2, 1, 2, 5, 5)$ tendríamos dos modas, el 2 y el 5, pues ambas aparecen dos veces

Medias de valores absolutos y media cuadrática Si las observaciones pueden ser positivas y negativas las medidas de tendencia central compensan ambos tipos de valores, pudiéndose desplazar hacia el cero.

Considérese, por ejemplo, la secuencia $S_4 = (-3, -1, 2, 4, 3, 1, -2, -4)$, cuyo valor medio es:

$$mean(S_4) = \frac{(-3) + (-1) + 2 + 4 + 3 + 1 + (-2) + (-4)}{8} = 0$$

Una media muy pequeña, o nula, puede deberse tanto a que las observaciones sean muy pequeñas, o nulas, o a que se compensen valores positivos y negativos.

En ocasiones nops interesa distinguir ambas situaciones, para lo que las observcaiones con valores negativos se convierten en otras con valores positivos.

Definimos la **media de valores absolutos** como:

$$mean_{abs}(x_1, \ldots, x_N) = \frac{1}{N} \sum_{1=1}^{N} \|x_i\|$$

.

Y la **media cuadrática** (*root mean square* ó *rms*) como:

$$rms(x_1, \ldots, x_N) = \sqrt{\frac{1}{N} \sum_{1=1}^{N} x_i^2}$$

.

```
mean(S4) = 0.0
mean_abs(S4) = 2.5
rms(S4) = 2.7386127875258306
```

Estadísticos de dispersión

Ya hemos visto algunos estadísticos importantes de dispersión:

- El **rango** que es la diferecnia entre los valores máximo y mínimos observados.
- El **rango intercuartil** $IQR = Q_3 - Q_1$. Podemos utilizar otros rangos intercuantiles. Por ejemplo, el **rango D1-D9** o el **rango C15-C85**.

Estos estadísticos cuantifican la dispersión fijándose directamente en la extensión de los valores que toman las observaciones, o rango de las mismas.

Una forma alternativa de cuantificar la dispersión es considerar cuánto se separan las observaciones de una medida de localización o tendencia central $< x >$ (media, mediana, moda,...) de las observaciones x_i.

Para ello se contruye, a partir de la secuencia de observaciones y de la medida de localización central elegida, una nueva **secuencia de distancias absolutas**

$$S_{AD}(x_i) = \|x_i - < x > \| \ , \ i = 1 \ldots N$$

La dispersión se mide computando una medida de tendencia central la secuencia de distancias absolutas.

Hay muchas posibles medidas de la dispersión:

- **Desviación absoluta media (*mean absolute deviation* ó *MAD*)** alrededor de la media, de la mediana, de la moda o de cualquier otra medida de localización central. Se calcula con la **media** de tales distancias absolutas.
- **Desviación absoluta mediana** alrededor de la media, de la mediana, de la moda o de cualquier otra medida de localización central. Se calcula con la **mediana** de tales distancias absolutas.
- **Desviación absoluta máxima** alrededor de la media, de la mediana, de la moda o de cualquier otra medida de localización central. En este caso simplemente se coge la distancia absoluta mayor.

Desviación típica (*standard deviation* ó *sdv*) y varianza muestrales Primeramente se calcula el valor medio de las observaciones, y la distancia de todas ellas al mismo. La **desviación típica muestral** es simplemente la media cuadrática de las distancias de cada observación al valor medio $\hat{x} = \frac{1}{N}\sum x_i$.

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{x})^2}$$

La división por $N-1$ asegura un **estimador insesgado**, pues se utiliza una media muestral computada con las mismas observaciones.

La varianza muestral es el cuadrado de la desviación típica muestral.

Coeficiente de variación La dispersión de las observaciones debe entenderse en relación a la localización o tendencia central de las mismas. Para ello se definen distintos coeficientes de variuación, como razón entre la medida de dispersión y la de tendencia central. El más habitual utiliza la desviación típica y la media

$$c_v = \frac{s}{\hat{x}}$$

Podemos utilizar una definición del coeficiente que sea **robusta** (poca influencia de valores atípicos), por ejemplo:

$$c_v = \frac{IQR}{midhinge}$$

Veamos cómo resulta la desviación típica y el coeficiente de variación para las cuatro secuencias de ejemplo:

```
S1:        mean = 3.0         std = 2.00         cv = 0.67
S2:        mean = 19.0         std = 39.23         cv = 2.06
S3:        mean = 3.0         std = 1.87         cv = 0.62
S4:        mean = 0.0         std = 2.93         cv = 0.98
```

Estadísticos de resumen

En la práctica, inicialmente se selecciona un número pequeño de estadísticos de resumen que nos permita hacer una primera evaluación de las observaciones, junto con el histograma.

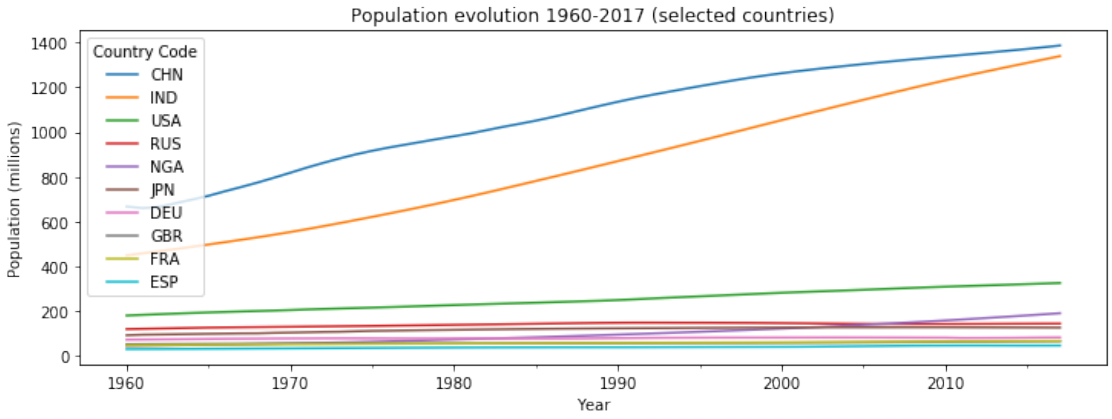
Veamos los estadísticos de resumen para las poblaciones d elos países del mundo:

VER TRANSPARENCIA ONLINE

Series temporales

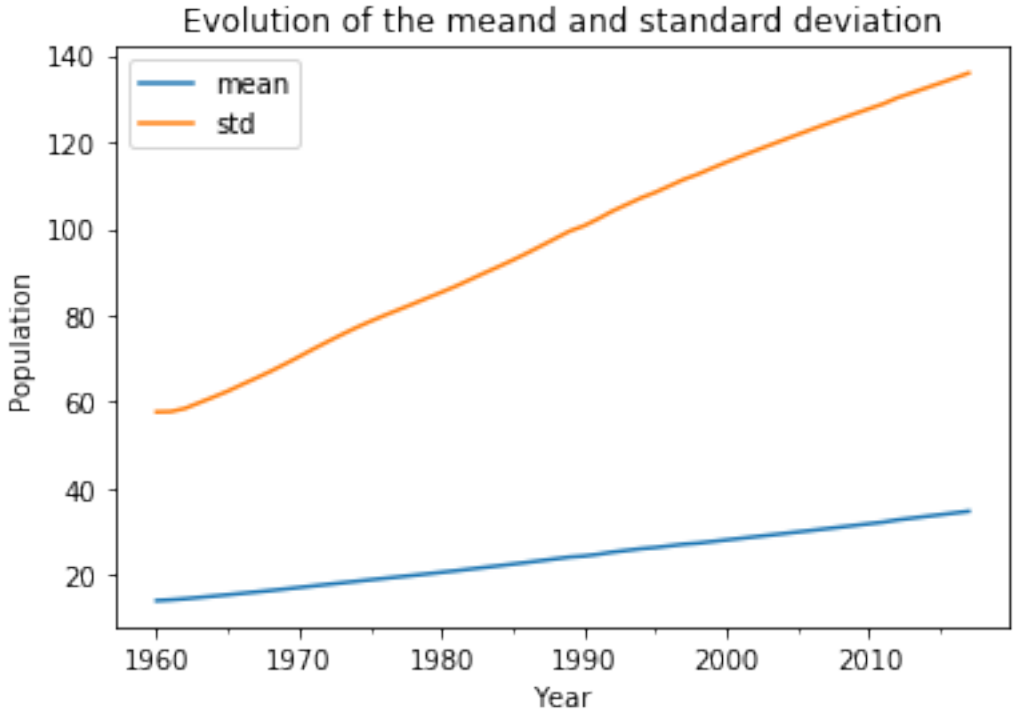
Estudiemos ahora la evolución temporal de la población de los países del mundo. La población de cada año supone una variable estadística de la que se toma una observación para cada país. Estamos interesados en estudiar el comportamiento de la población muestral a lo largo del tiempo.

Veamos primero la evolución de las observaciones correspondientes a una selección de países:

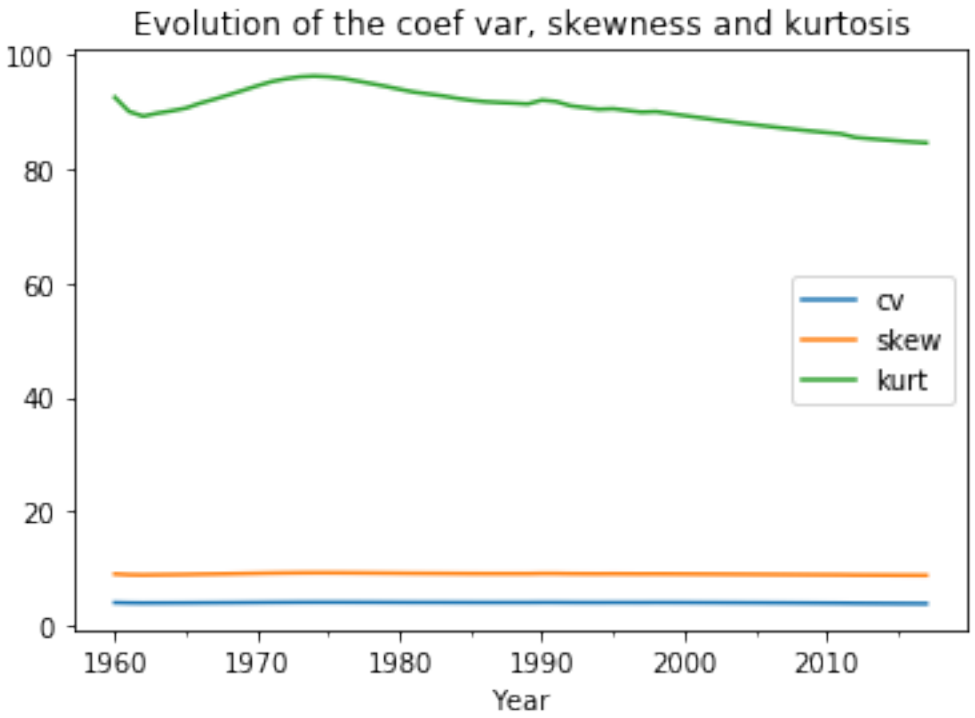


Dado que para cada año tenemos una variable estadística, podemos calcular estadísticos para las observaciones de cada uno de ellos.

Veamos primero cómo evoluciona el valor medio y la desviación típica de las poblaciones de los países del mundo:



Y veamos ahora la evolución del coeficiente de variación, el coeficiente de asimetría y el coeficiente de curtosis de las poblaciones de los países del mundo:



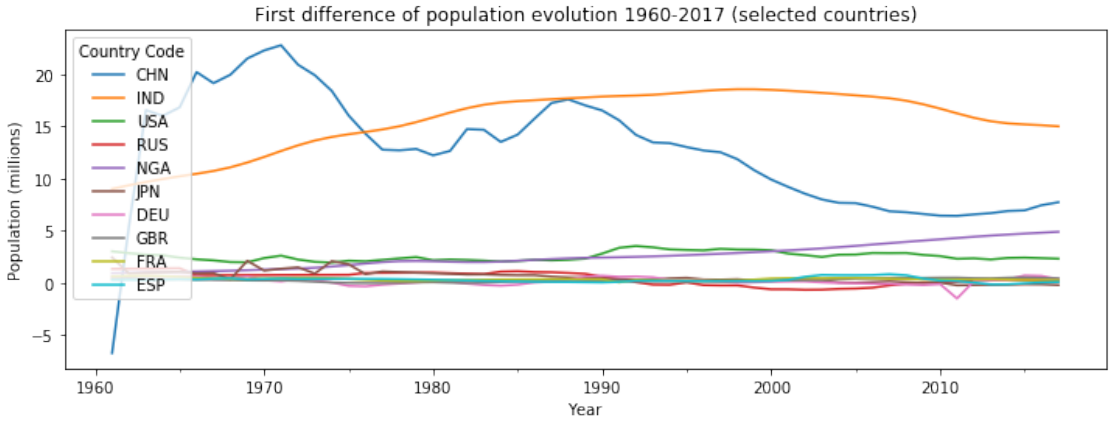
Estacionariedad

Un conjunto de series temporales de variables estadísticas, que, por tanto, para cada instante de tiempo proporciona una subpoblación muestral de observaciones, se dice que es **estacionaria** si sus estadísticos se mantienen constantes a lo largo del tiempo.

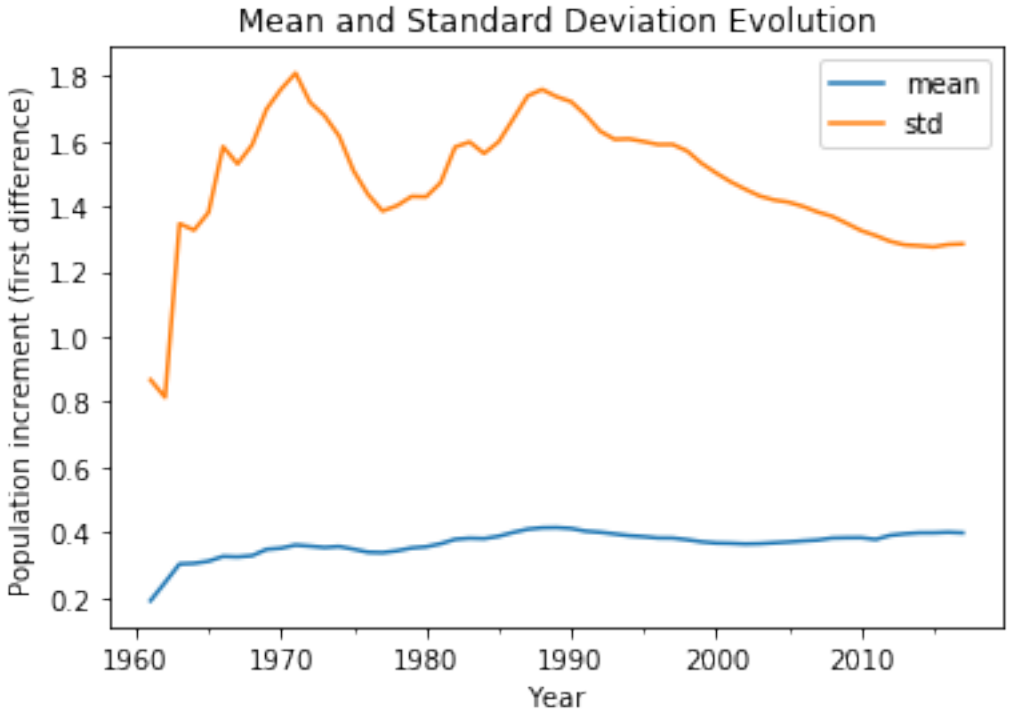
En sentido amplio, podemos referirnos a la estacionariedad en relación a un estadístico. Por ejemplo, **estacionariedad en la media** o **estacionariedad en la varianza**, si dichos estadísticos se mantienen constantes.

En las representaciones anteriores vemos que la población de los países del mundo no es estacionaria ni en la media ni en la varianza.

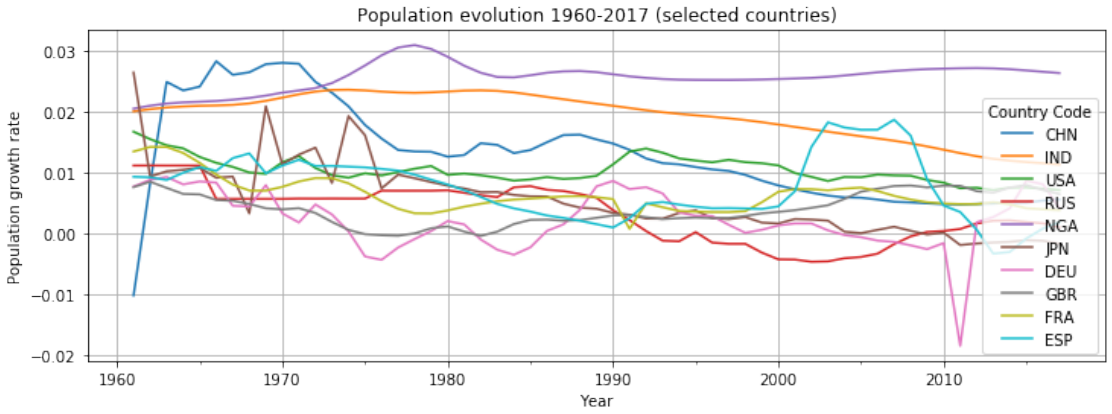
Generalmente estamos interesados en trabajar con series estacionarias. Por ello a veces se transforman las series originales para aproximar un comportamiento estacionario. Una transformación que se usa habitualmente es la primera diferencia, o incremento entre dos instantes de tiempo sucesivos:



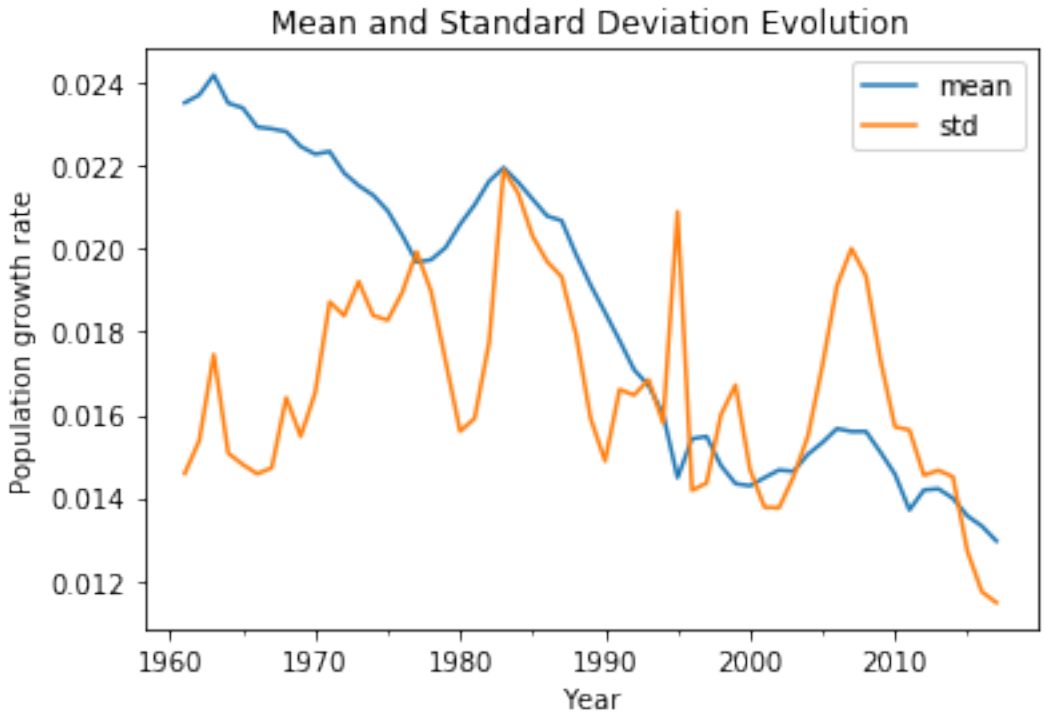
Veamos cómo evolucionan la media y la desviación típica de la primera diferencia $x[n+1] - x[n]$:



También suele resultar útil calcular la tasa de variación $\frac{x[n+1]-x[n]}{x[n]}$:
<pandas.io.formats.style.Styler at 0x7fef9d2104e0>



Veamos cómo evolucionan la media y la desviación típica de la tasa de variación:



0.1 Relación entre variables estadísticas

Dos variables estadísticas pueden variar independientemente entre sí, o tener una relación.

El **diagrama de dispersión** muestra observaciones correspondientes de ambas variables sobre el plano, asignando la coordonada x a una y la y a la otra. Permite obtener rápidamente una intuación de si existe relación entre ambas variables y cómo es.

El **coeficiente de correlación** es un estadístico que muestra la **relación lineal** entre ambas variables, con valores entre -1 y 1. Un valor de 0 indica que no hay relación lineal. Un valor de -1 indica una relación lineal perfecta negativa (cuando una variable crece la otra decrece) y un valor de 1 una relación lineal perfecta positiva.

Diagrama de dispersión y coeficiente de correlación entre las observaciones de las poblaciones de todos los países del mundo en los años 2015 y 2017:

Year 2015 2017 Year 2015 1.0 1.0 2017 1.0 1.0

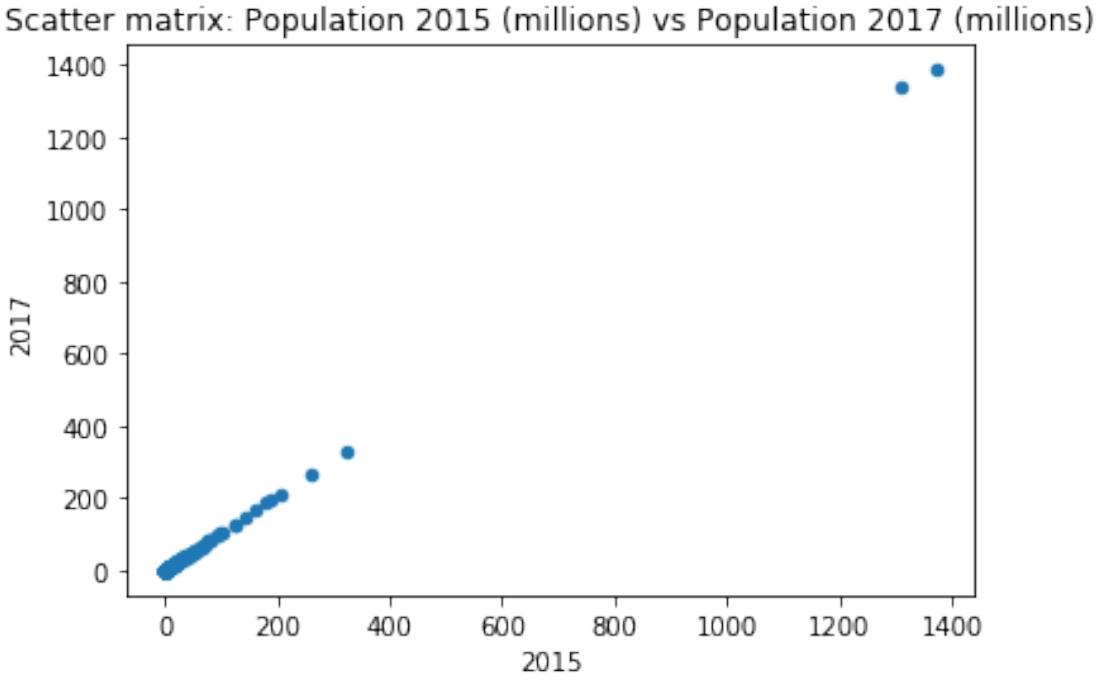
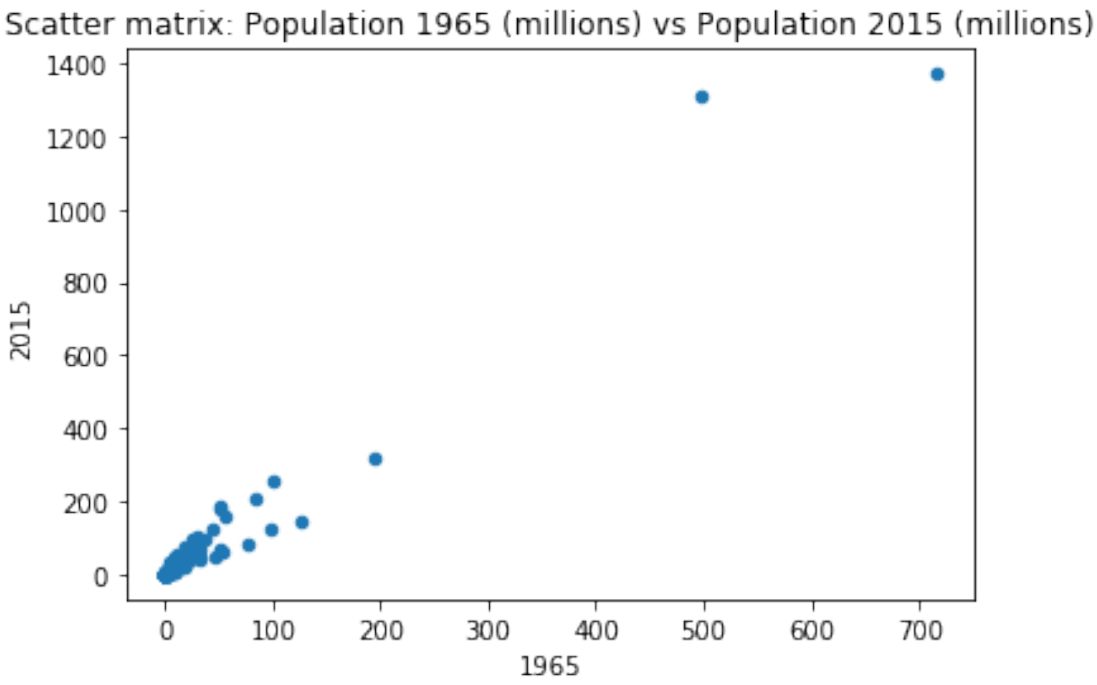


Diagrama de dispersión y coeficiente de correlación entre las observaciones de las poblaciones de todos los países del mundo en los años 1965 y 2015:

Year 1965 2015 Year 1965 1.0 1.0 2015 1.0 1.0



Análisis del producto interior bruto (PIB) mundial

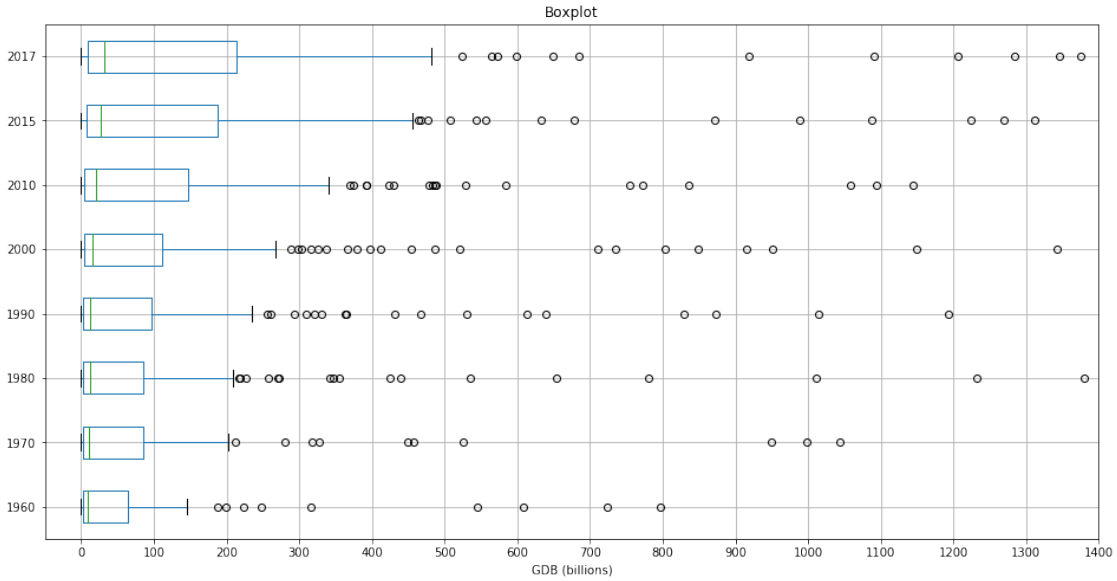
<pandas.io.formats.style.Styler at 0x7fef9cfb1a90>
<pandas.io.formats.style.Styler at 0x7fef9cecac8>

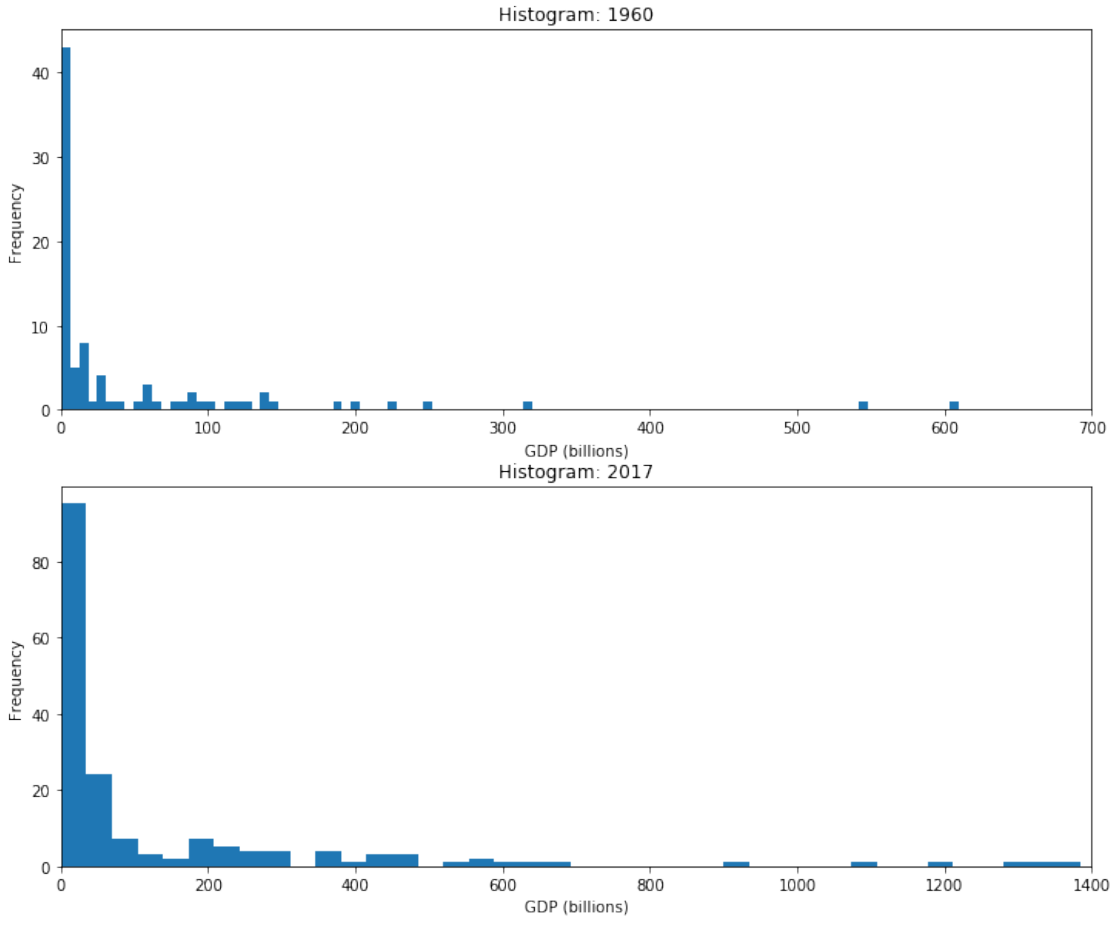
```
['AUS', 'BRA', 'CAN', 'CHN', 'DEU', 'ESP', 'FRA', 'GBR', 'IND', 'ITA', 'JPN', 'MEX', 'NLD', 'RUS
```

1 2 3 4 5 6 7 8 9 10 Year 1960 USA JPN GBR FRA ITA CAN BRA ESP AUS NLD 1970 USA JPN DEU FRA GBR ITA CAN ESP BRA AUS 1980 USA JPN DEU FRA ITA GBR BRA CAN ESP MEX 1990 USA JPN DEU FRA ITA GBR RUS BRA CAN ESP 2000 USA JPN DEU FRA CHN GBR ITA BRA CAN ESP 2010 USA CHN JPN DEU FRA GBR BRA ITA IND CAN 2015 USA CHN JPN DEU FRA GBR BRA IND ITA CAN 2017 USA CHN JPN DEU FRA GBR IND BRA ITA CAN

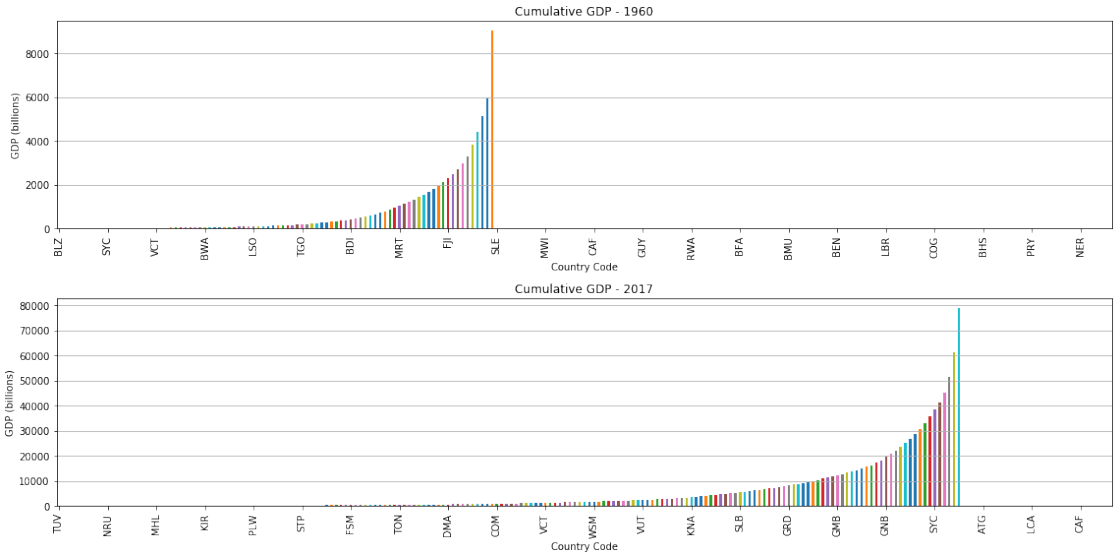
Country Code CHN IND USA RUS NGA JPN DEU GBR FRA ESP Year 1960 14.0 12.0 1.0 nan 25.0 2.0 nan 3.0 4.0 8.0 1970 17.0 14.0 1.0 nan 27.0 2.0 3.0 5.0 4.0 8.0 1980 15.0 16.0 1.0 nan 29.0 2.0 3.0 6.0 4.0 9.0 1990 11.0 15.0 1.0 7.0 38.0 2.0 3.0 6.0 4.0 10.0 2000 5.0 14.0 1.0 11.0 40.0 2.0 3.0 6.0 4.0 10.0 2010 2.0 9.0 1.0 11.0 30.0 3.0 4.0 6.0 5.0 12.0 2015 2.0 8.0 1.0 11.0 26.0 3.0 4.0 6.0 5.0 12.0 2017 2.0 7.0 1.0 11.0 26.0 3.0 4.0 6.0 5.0 12.0

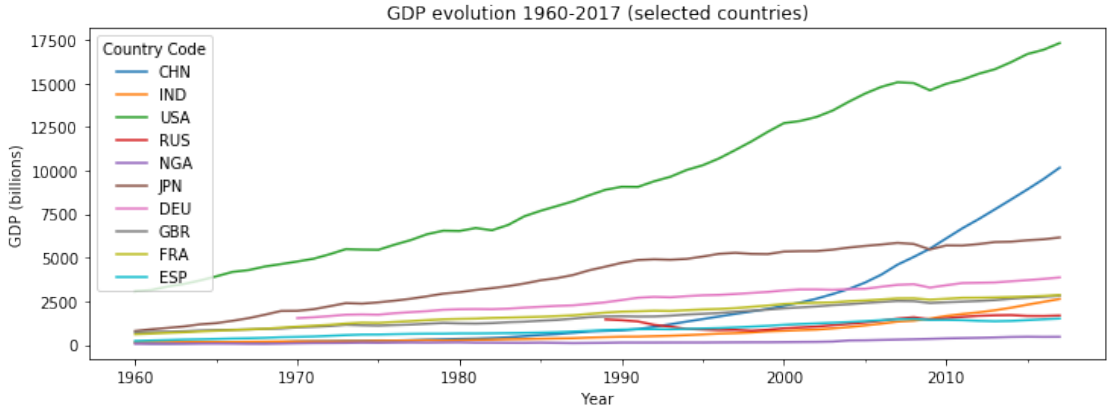
Year 1960 1970 1980 1990 2000 2010 2015 2017 LC Total 6927.5 13015.8 18670.8 26126.6 33952.3 42870.0 49271.2 52056.2 WLD Total 9025.8 17209.5 25612.0 37286.8 49611.5 65390.2 74628.2 78710.0 LC/WLD (%) 76.8 75.6 72.9 70.1 68.4 65.6 66.0 66.1

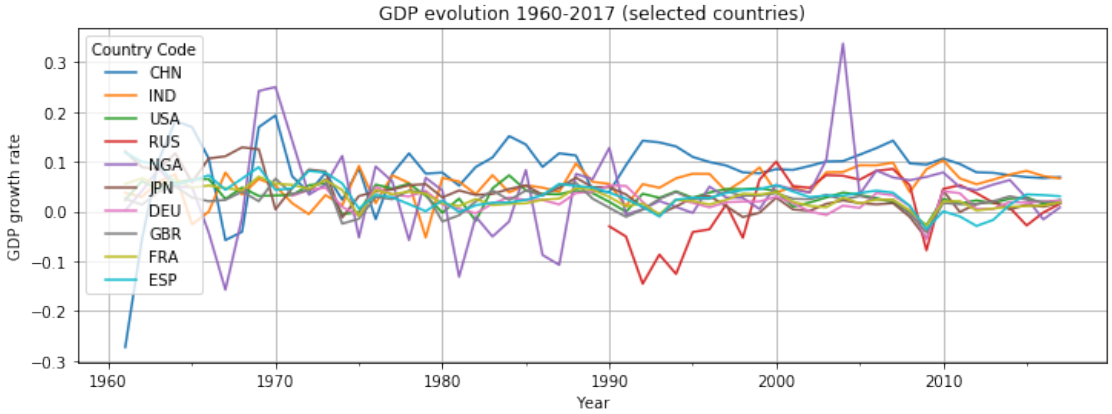

Boxplot
2017
2015
2010
2000
1990
1980
1970
1960
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
GDB (billions)

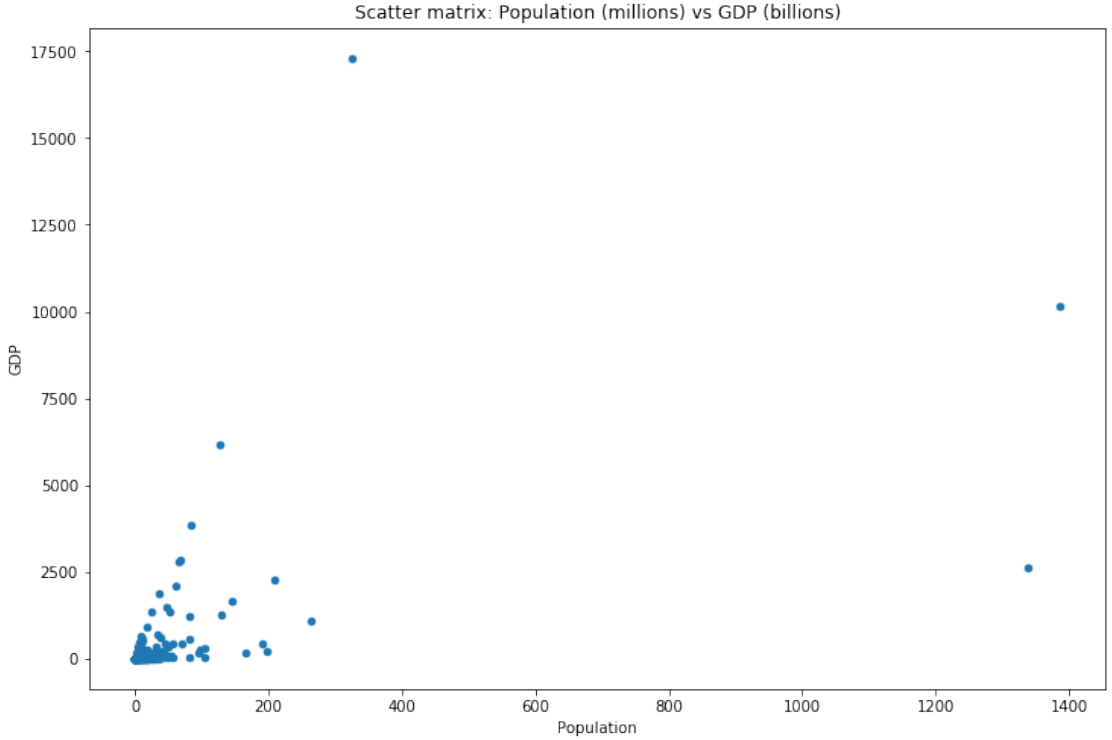

Histogram: 1960
Frequency
0
10
20
30
40
0
100
200
300
400
500
600
700
GDP (billions)
Histogram: 2017
Frequency
0
20
40
60
80
0
200
400
600
800
1000
1200
1400
GDP (billions)

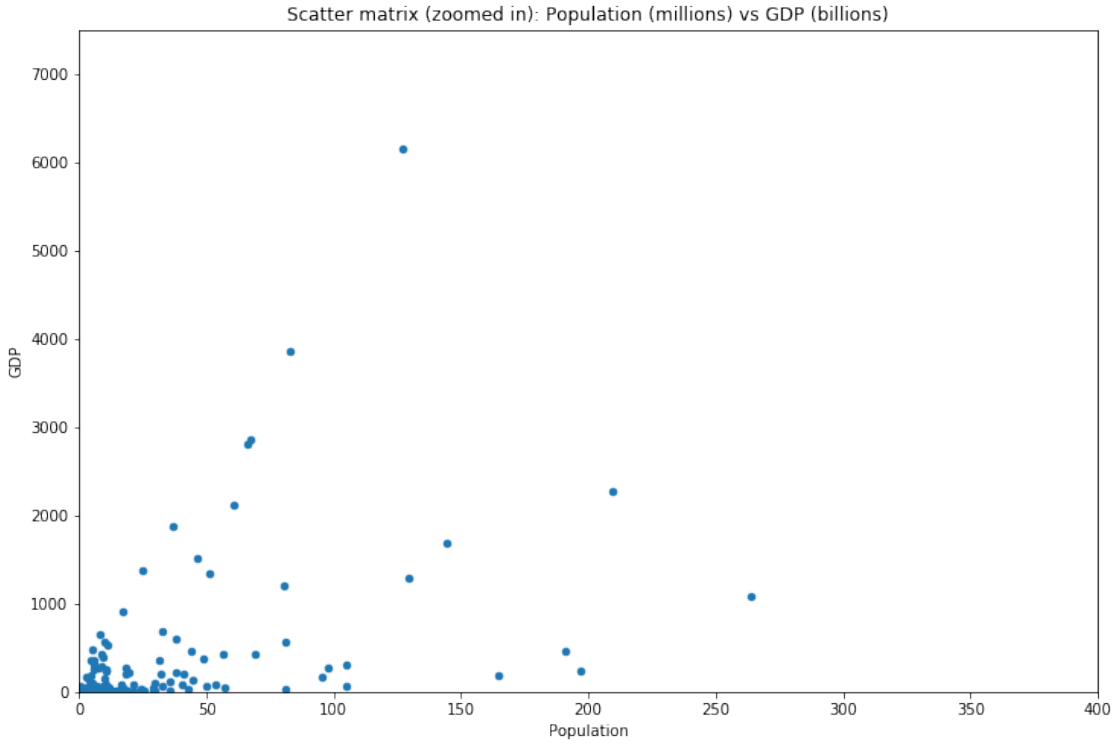
1960 1970 1980 1990 2000 2010 2015 2017 mean 100.3 153.7 188.3 223.3 259.7 322.1 382.7 423.2 median 8.9 10.2 11.7 12.0 15.3 20.3 26.3 32.2 skew 7.3 6.9 7.2 7.9 9.1 8.7 8.2 7.9 kurt 60.8 55.5 62.3 74.4 99.0 91.3 80.5 73.6

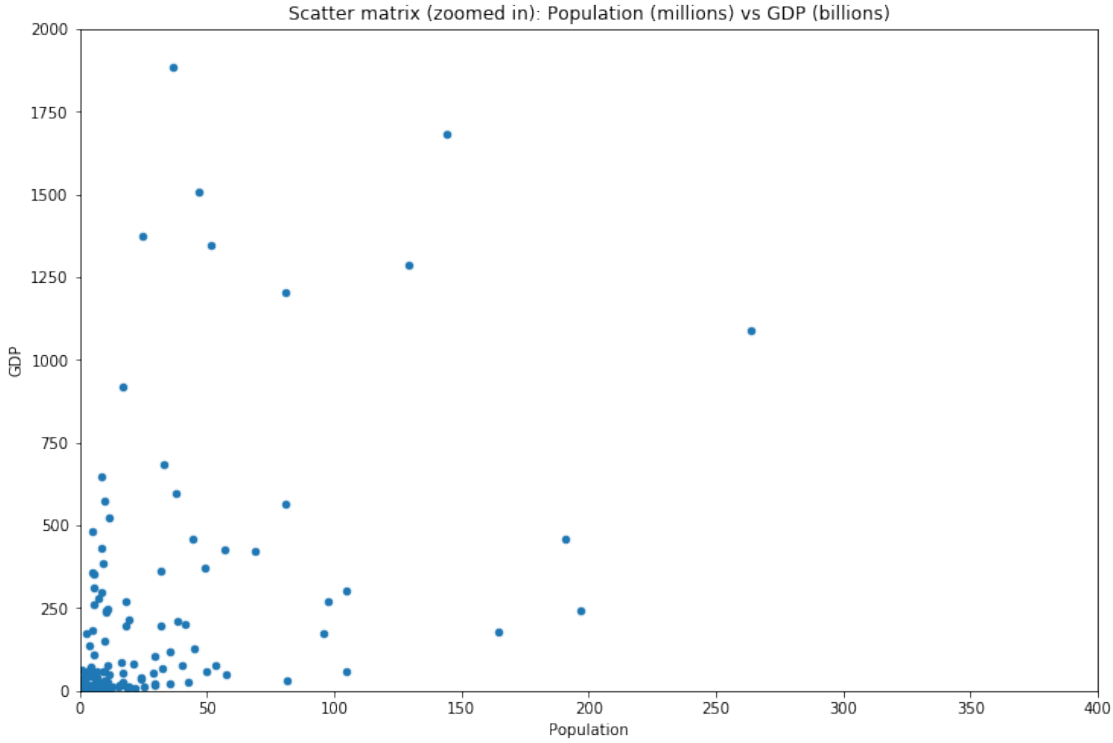





GDP evolution 1960-2017 (selected countries)
Country Code
CHN
IND
USA
RUS
NGA
JPN
DEU
GBR
FRA
ESP
GDP growth rate
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
1960
1970
1980
1990
2000
2010
Year


Scatter matrix: Population (millions) vs GDP (billions)
GDP
17500
15000
12500
10000
7500
5000
2500
0
0
200
400
600
800
1000
1200
1400
Population


Scatter matrix (zoomed in): Population (millions) vs GDP (billions)
GDP
Population
0
1000
2000
3000
4000
5000
6000
7000
0
50
100
150
200
250
300
350
400


Scatter matrix (zoomed in): Population (millions) vs GDP (billions)
GDP
Population
0
250
500
750
1000
1250
1500
1750
2000
0
50
100
150
200
250
300
350
400

Population GDP Population 1.0 0.5 GDP 0.5 1.0

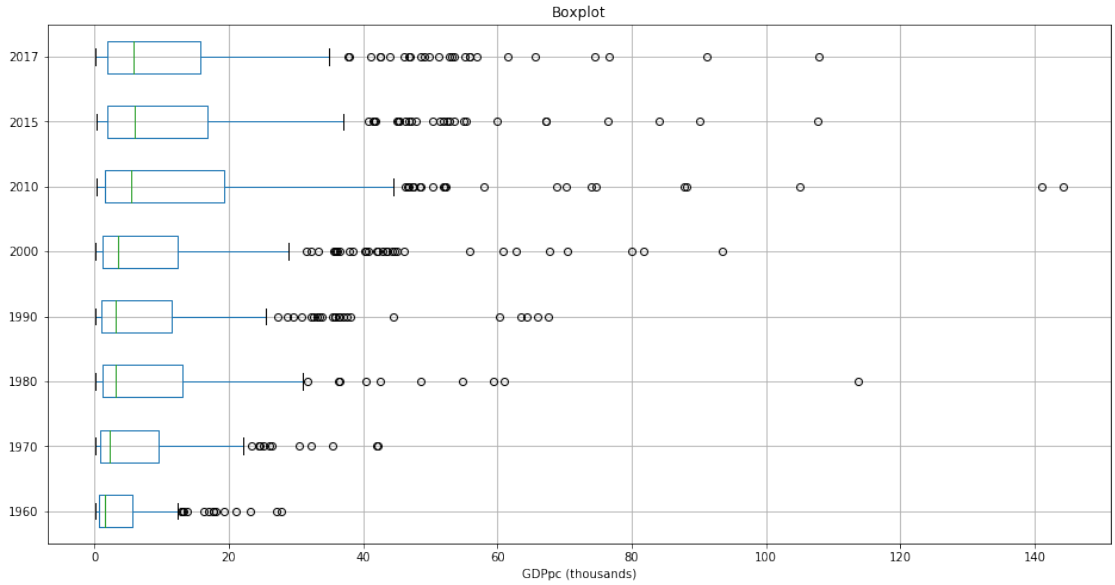
Análisis de la renta per cápita (PIBpc) mundial

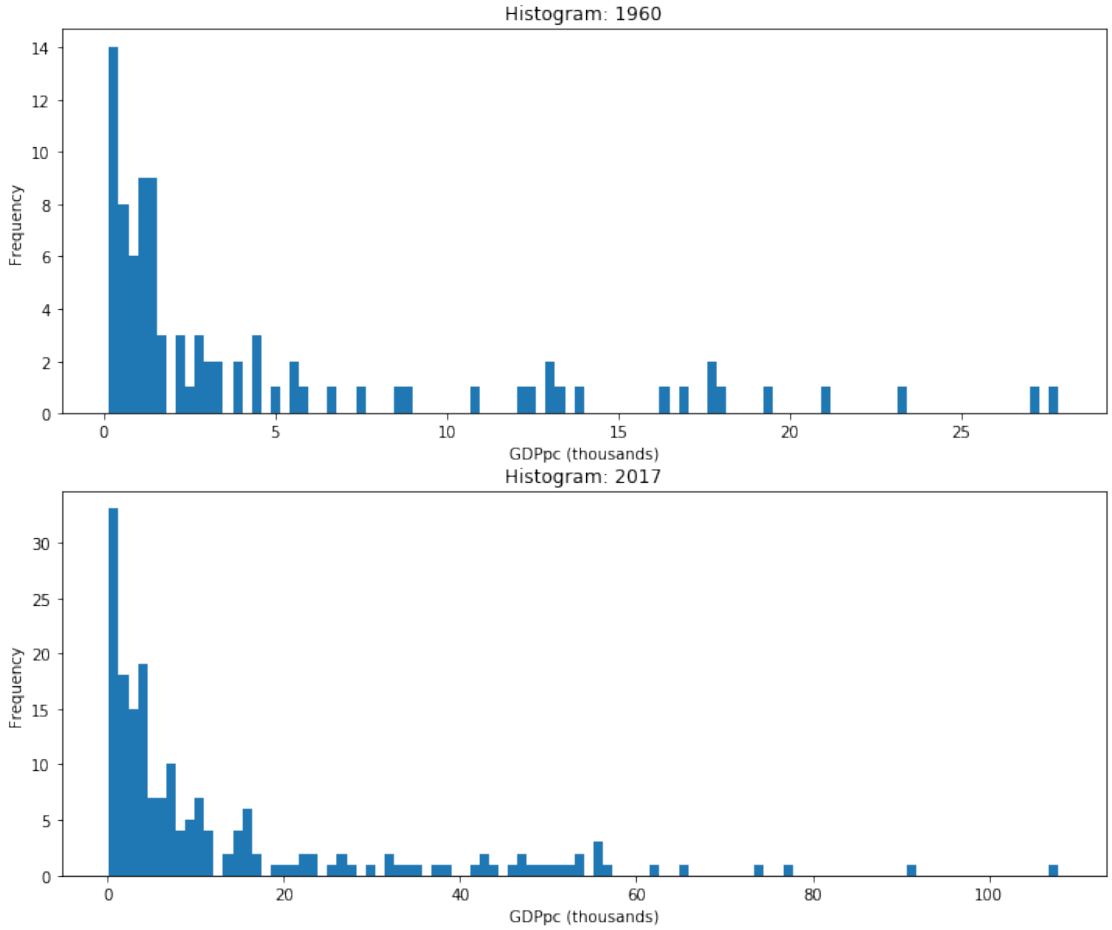
<pandas.io.formats.style.Styler at 0x7fef9c9e12b0>
<pandas.io.formats.style.Styler at 0x7fef9c71bb70>

```
['AND', 'ARE', 'AUS', 'BHS', 'BMU', 'BRN', 'CAN', 'CHE', 'DNK', 'IMN', 'IRL', 'JPN', 'KWT', 'LIE
```

1 2 3 4 5 6 7 8 9 10 Year 1960 BMU LUX NOR DNK AUS SWE BHS CAN USA NLD 1970 AND BMU LUX NOR DNK SWE AUS BHS CAN NLD 1980 ARE BMU BRN CHE NOR LUX AND SAU DNK CAN 1990 ARE LUX CHE BMU NOR DNK JPN SWE BRN CAN 2000 LUX NOR BMU SMR CHE ARE QAT DNK NLD USA 2010 MCO LIE LUX BMU NOR CHE IMN QAT SMR DNK 2015 LUX NOR IMN CHE QAT IRL DNK SWE AUS MAC 2017 LUX NOR CHE IRL QAT DNK SWE MAC AUS SGP

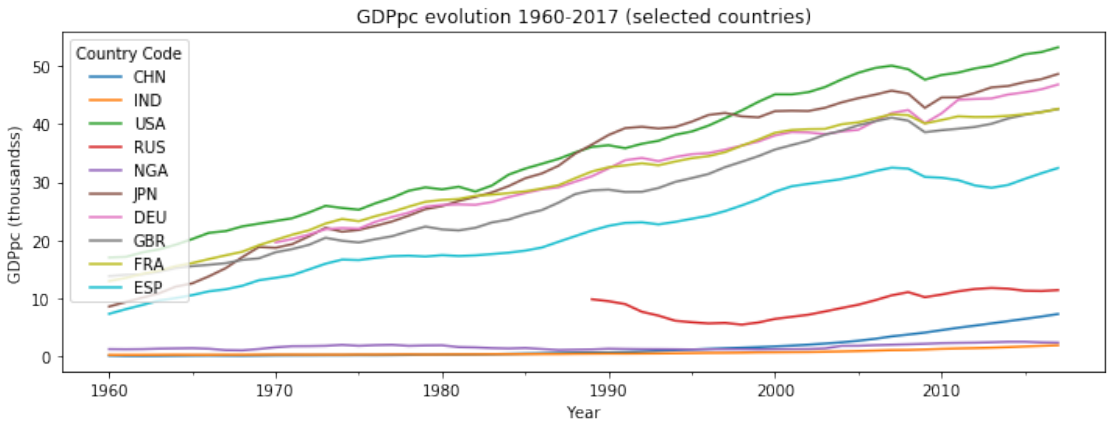
Year 1960 1970 1980 1990 2000 2010 2015 2017 Country Code CHN 89.0 111.0 131.0 138.0 129.0 110.0 93.0 81.0 IND 83.0 103.0 129.0 145.0 161.0 157.0 149.0 140.0 USA 9.0 11.0 14.0 11.0 10.0 16.0 13.0 12.0 RUS nan nan nan 45.0 72.0 72.0 66.0 61.0 NGA 52.0 68.0 88.0 117.0 143.0 145.0 141.0 134.0 JPN 19.0 18.0 21.0 7.0 16.0 22.0 17.0 17.0 DEU nan 16.0 20.0 19.0 22.0 26.0 20.0 19.0 GBR 11.0 20.0 25.0 22.0 27.0 30.0 25.0 23.0 FRA 14.0 15.0 18.0 18.0 21.0 27.0 24.0 22.0 ESP 20.0 25.0 29.0 27.0 32.0 38.0 34.0 30.0

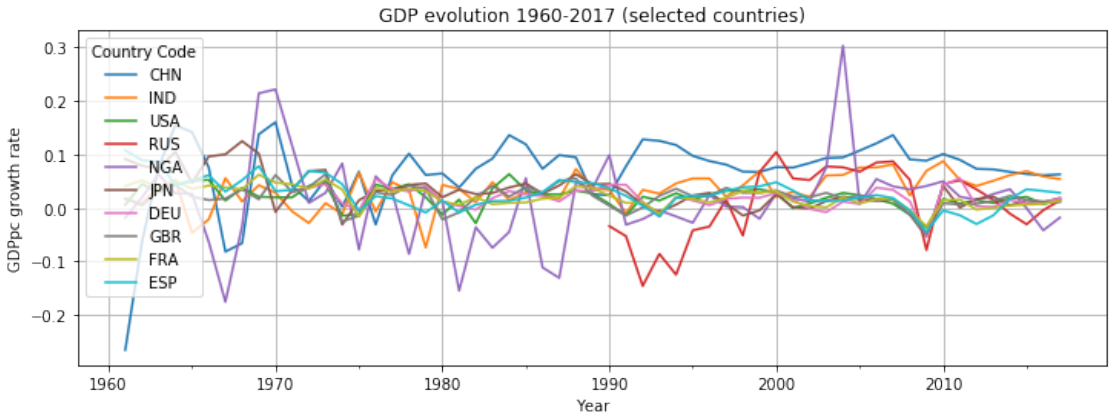


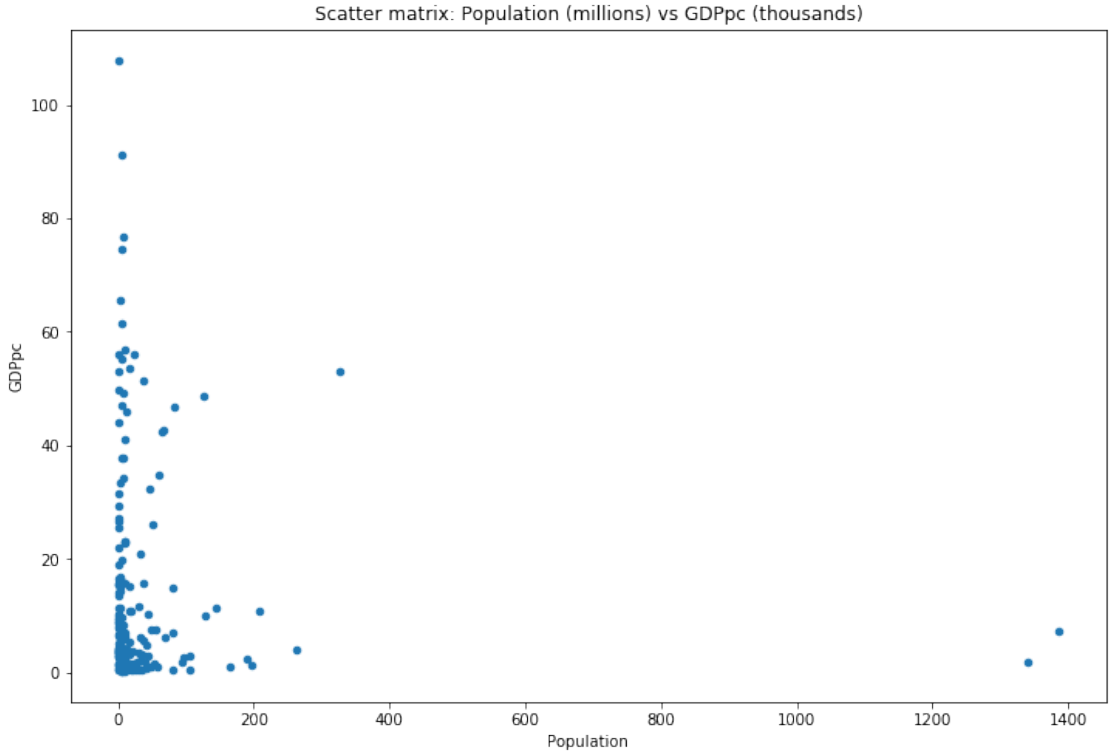


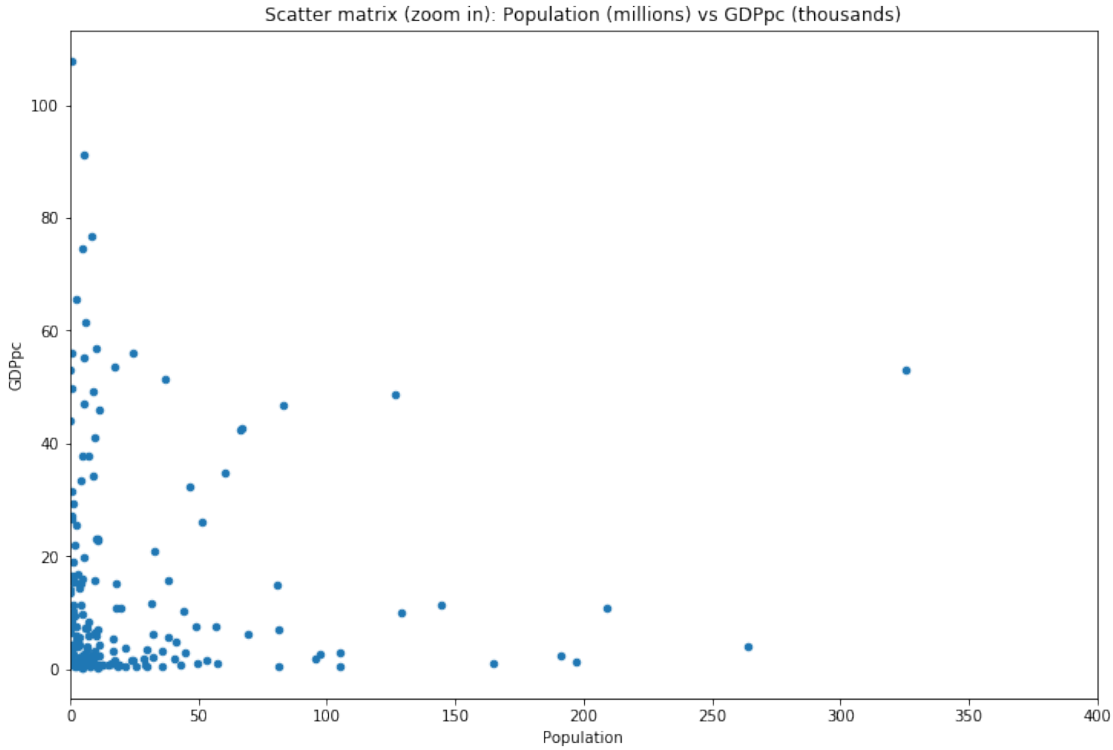
1960 1970 1980 1990 2000 2010 2015 2017 mean 4.9 7.2 10.1 9.8 12.0 15.5 14.4 14.2 median 1.5 2.2 3.1 3.1 3.5 5.4 6.1 5.8 skew 1.8 1.7 3.1 2.1 2.1 2.7 2.0 2.1 kurt 2.5 2.4 14.0 4.4 4.5 9.5 4.4 4.5

```
In [113]: gdp_pc_countries.T[selected_countries].plot(figsize=(12,4))
          plt.ylabel('GDPpc (thousandss)'), plt.title('GDPpc evolution 1960-2017 (selected count
```




GDP evolution 1960-2017 (selected countries)
Country Code
CHN
IND
USA
RUS
NGA
JPN
DEU
GBR
FRA
ESP
GDPpc growth rate
Year


Scatter matrix: Population (millions) vs GDPpc (thousands)
GDPpc
Population



Population GDPpc Population 1.0 -0.0 GDPpc -0.0 1.0

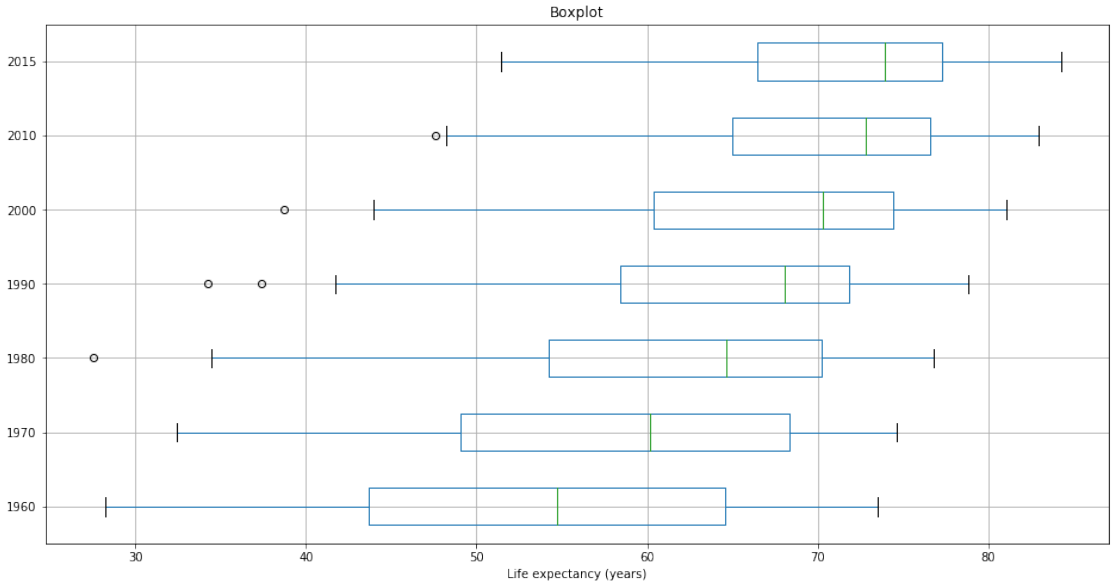
Análisis de la esperanza de vida de los países del mundo

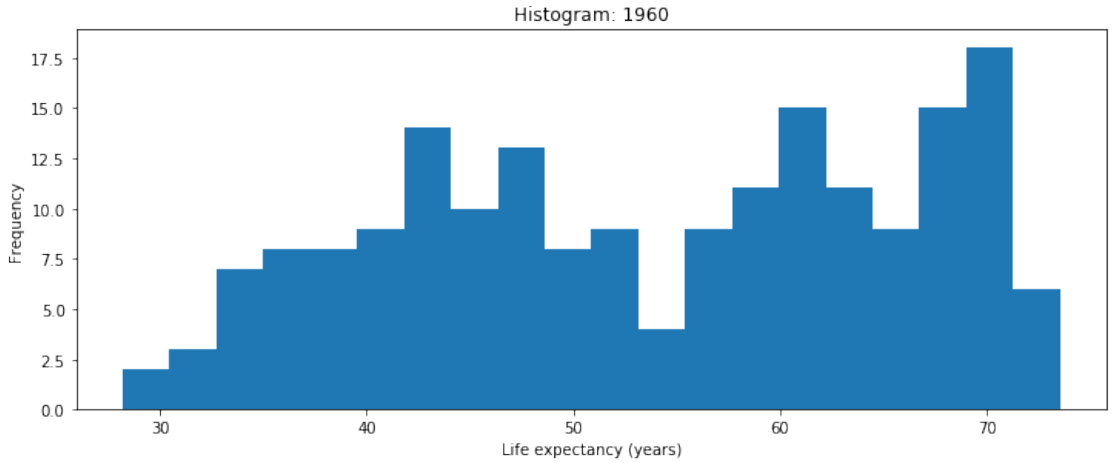
<pandas.io.formats.style.Styler at 0x7fefa38fb358>
<pandas.io.formats.style.Styler at 0x7fef9ebbbb00>

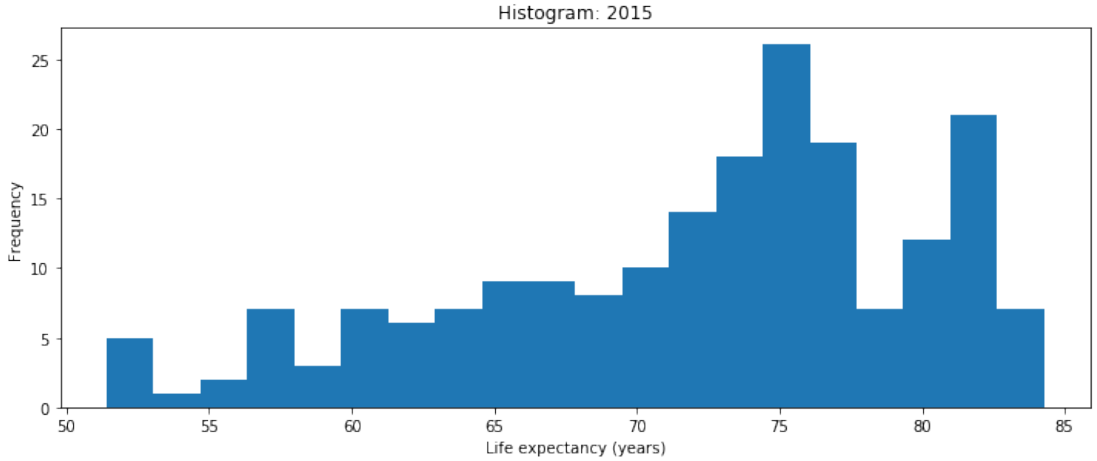
```
['AUS', 'BGR', 'CAN', 'CHE', 'CHI', 'CYM', 'CYP', 'DNK', 'ESP', 'FRA', 'FRO', 'GBR', 'GRC', 'HKG
```

1 2 3 4 5 6 7 8 9 10 Year 1960 NOR ISL NLD SWE DNK CHE NZL CAN GBR AUS 1970 SWE NOR ISL NLD DNK CHE CAN CYP ESP GBR 1980 ISL JPN NLD SWE NOR CHE ESP CAN CYP HKG 1990 JPN ISL SWE HKG CAN MAC CHE AUS ITA GRC 2000 JPN HKG MAC ITA CHE ISL SWE CAN AUS FRA 2010 HKG JPN MAC CHE CYM ITA ISL LIE AUS FRA 2015 HKG JPN MAC CHE ESP SGP LIE ITA ISL AUS

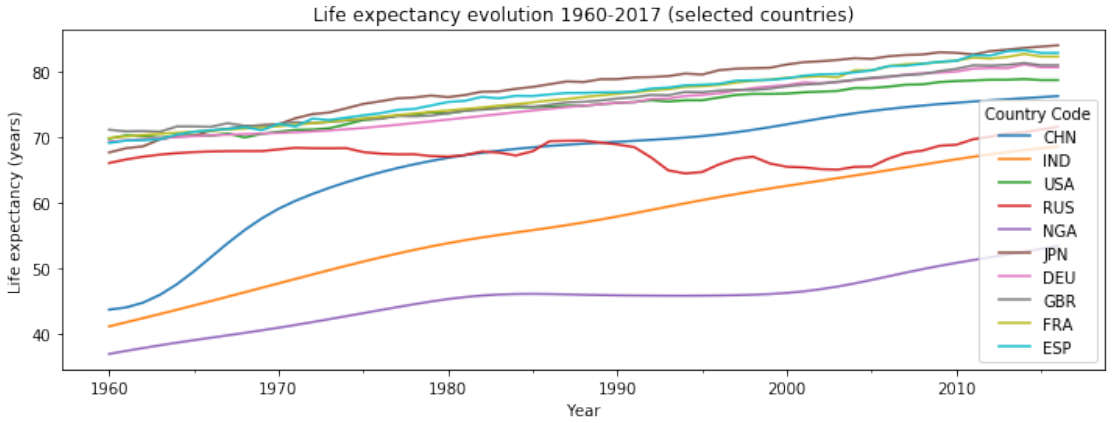
Year 1960 1970 1980 1990 2000 2010 2015 Country Code CHN 141.0 104.0 83.0 86.0 78.0 65.0 66.0 IND 157.0 149.0 147.0 150.0 145.0 143.0 142.0 USA 18.0 25.0 22.0 27.0 35.0 39.0 43.0 RUS 41.0 52.0 80.0 91.0 134.0 131.0 124.0 NGA 171.0 173.0 178.0 187.0 197.0 195.0 195.0 JPN 38.0 11.0 2.0 1.0 1.0 2.0 2.0 DEU 21.0 27.0 28.0 26.0 21.0 29.0 33.0 GBR 9.0 10.0 20.0 20.0 25.0 24.0 30.0 FRA 14.0 13.0 13.0 14.0 10.0 10.0 13.0 ESP 24.0 9.0 7.0 12.0 11.0 11.0 5.0

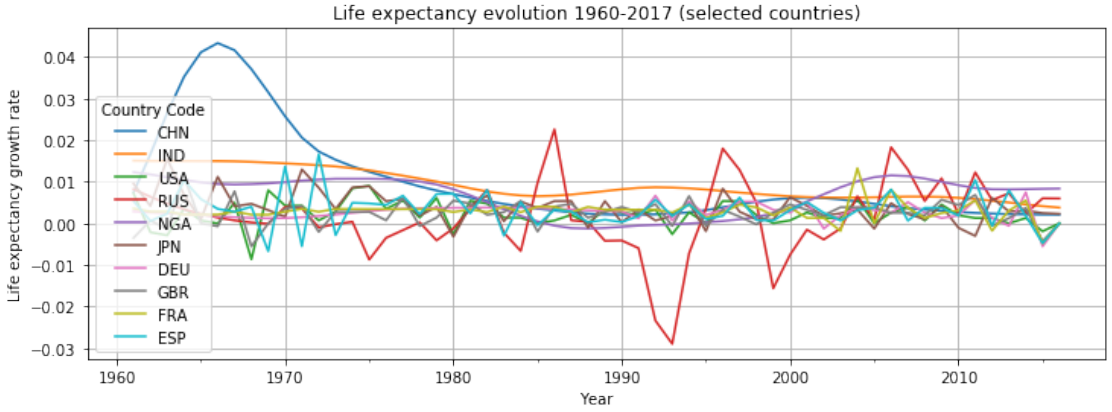

Boxplot
2015
2010
2000
1990
1980
1970
1960
30
40
50
60
70
80
Life expectancy (years)

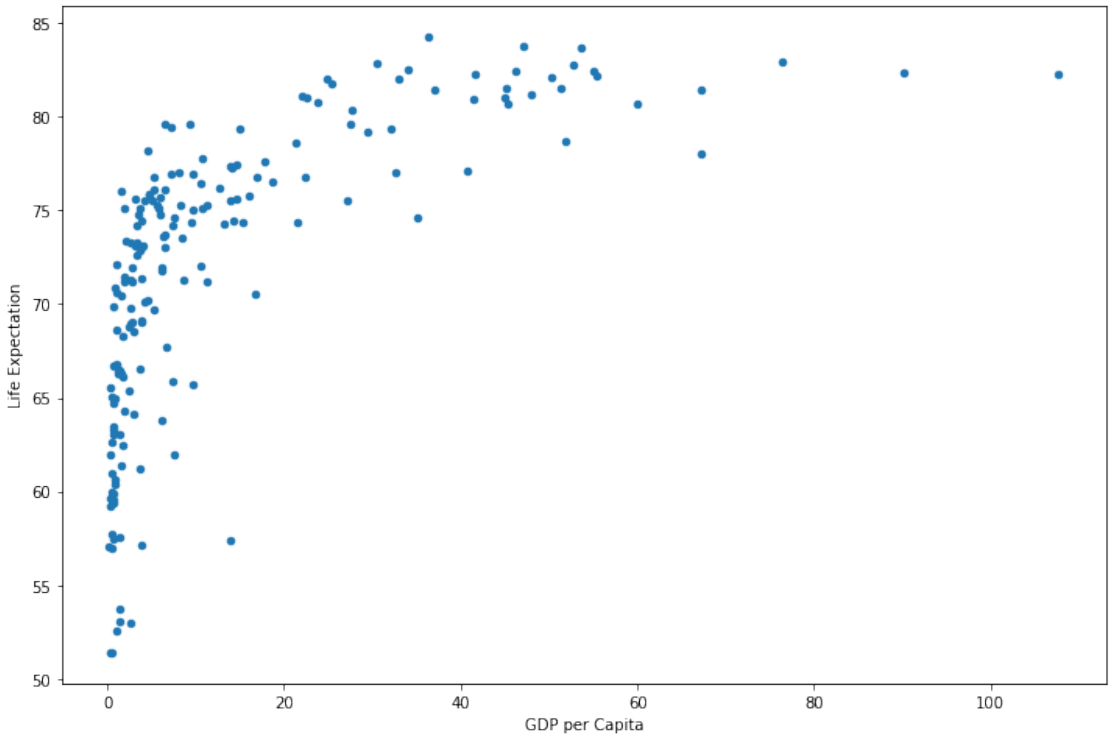

Histogram: 1960
Frequency
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
30
40
50
60
70
Life expectancy (years)


Histogram: 2015
Frequency
25
20
15
10
5
0
50
55
60
65
70
75
80
85
Life expectancy (years)

1960 1970 1980 1990 2000 2010 2015 mean 53.9 58.2 61.9 64.9 67.1 70.5 72.0 median 54.7 60.2 64.6 68.1 70.3 72.8 73.9 skew -0.2 -0.4 -0.7 -0.8 -0.8 -0.7 -0.6 kurt -1.2 -1.1 -0.4 -0.2 -0.4 -0.4 -0.4







Life Expectation GDP per Capita Life Expectation 1.0 0.6 GDP per Capita 0.6 1.0