World Development Indicators

Francisco Clavero, Vicente Oyanedel, Fabián Souto

14 de junio de 2017

1. Goal

La meta de este proyecto es obtener los indicadores más presentes en los países y años; esto último (presencia en años) es que se tenga consistencia en el tiempo de cada indicador por país. Con los indicadores seleccionados se desea generar visualizaciones de nuestro país, Chile, y países escogidos para su comparación de distintas partes del mundo; tales como EEUU, China, Portugal y Perú. El primero y el segundo por ser potencias mundiales pero de continentes distintos, Portugal para comparar con un país similar a Chile pero Europeo y Perú para comparar con un país vecino.

A partir de los resultados se hará un análisis de éstos y poder ver correlaciones entre hechos históricos, y los países entre sí.

2. Data

El dataset escogido corresponde a la base da datos World Development Indicators de *Kaggle*. Esta contiene los valores de distintos indicadores económicos de más de 243 países del mundo, tales como su PIB, distintas mediciones demográficas y de emiciones de CO2. La base de datos está disponible tanto en formato *sqlite* como en una serie de archivos CSV. Para poder manejarlos en el sistema distribuido se trabajó con el archivo *Indicators.csv*, que contiene los valores de los indicadores y tiene un tamaño de 560 MB.

3. Methods

En primer lugar, se implementa una serie de programas en **Pig** para la extracción y analisis de los datos.
Luego, se implementa un script en **Python** para generar las visualizaciones de los datos generados en Pig.

Pig: Se implementaron los siguientes archivos, que pueden ser encontrados en el repositorio de código:

1. year_count.pig: Para obtener la cantidad de años que hay en los datos. Esto se utilizó en un comienzo para encontrar los paises que tuvieran todos los indicadores.

2. country_count.pig: Para obtener la cantidad de países que hay en los datos. Se utilizó, como el anterior, en la busqueda de paises con todos los indicadores.

3. nice_indicators.pig: Para obtener una lista de indicadores para los paises, tal que tuvieran una muestra suficientemente grande. Éste resultado se logró sumando las cuentas de los programas anteriores para así obtener una noción de la cantidad de indicadores por pais por año, y luego se fija un numero levemente menor a éste.

4. general_script.pig: Filtrar los datos fuente (Kaggle) por los países que se eligieron para hacer el analisis.

Se llega a esta solución despues de intentar obtener los países que tuvieran todos los indicadores que Chile, dentro de un periodo de tiempo. Solución que se descarta después de intentar de manera ardua programar un nested foreach de divsersas maneras, para finalmente descubrir que el lenguaje no lo permitía.

Otra dificultad ocurre al cargar los datos en formato csv, el cual posee el carácter separador en el texto. Para lo cual descargó la extensión Piggybank.pig, la cual posee un parser de CSV que funciona perfectamente.

Python: Obtenidos la lista de indicadores suficientes y los datos filtrados por los países; se prosiguió a importar los datos en Python y hacer visualizaciones (con pyplot) para cada indicador en función del tiempo. Esto se hace en el script que se encuentra en el repositorio de código.

4. Results

Al comenzar a explorar y analisar los datos mediante Pig, se observa:

1. No hay ningún país con todos los indicadores.

2. Con más de 13000 habían muy pocos.

3. Los indicadores que más tenían países eran muy pocos, y algunos no tan relevantes.

En base a ésto se eligió los atributos más relevantes para nosotros. Luego se graficaron mediante Python en las visualizaciones que se presentan en el anexo.

Del análisis que vá más allá de lo que se deja a ver a simple vista, se concluye:

Población total vs Tiempo ref 12
Se nota que China crece más rapido y gana en tamaño al resto.

Al hacer Zoom se puede dilusidar la baja tasa de fertilidad en paises europes. Por otro lado, Perú tiene tasa de fertilidad grande.
Se puede apoyar en el siguiente grafico: fertilidad

Producto interno bruto vs Tiempo ref 6. Estados Unidos el mayor. Se observa un gran incremento de china. La diferencia entre economías grandes y pequeñas se deja ver claramente.

Además, se pueden ver los efectos de las crisis economicas ocurridas en el periodo.

Al hacer Zoom (7), se deja ver que Chile se acerca a Portugal. Por otro lado, desde mediados de los 80' Chile gana repunte en su economía.

Población urbana vs Tiempo ref 1.
De éste indicador se obtuvo un margen muy pequeño, que al graficarlo deja ver resultados bastante variantes; especialmente los Chinos que perdieron mucho entre el 65.
En contraste, portugal y china persiven un inmenso crecimiento entre el 73 y el 75.

Esperanza de vida vs Tiempo: ref 2.
Sórpresivamente, Chile muestra la mayor espectativa de vida.
China presiente un boom en los 70s.
Por su lado, Peru y chile son iguales muy similares. Teniendo Chile un crecimiento muy uniforme.

Hospital: ref 5.
Una conclusión interesante es que a medida que crece la poblacion, la cantidad de camas no aumenta de la misma manera y se ve disminuido.
Todos excepto estados unidos crecen más que este, dado que siempre dismnuye.

Exports: ref 8.
Se ven claramente las crisis. En los 80 china no hacía nada; montó las factorias y se disparó su crecimiento.

Se noto más la crisis de los 80 en comparacion al otro grafico.

Nuestra teoría es que China logró suplir las pérdidas mediante el consumo interno, por ello no se ve una baja en el producto interno bruto durante el mismo período.

Renovable: ref 9.
Malas noticias para el ambiente dado que se ve una clara tendencia a la baja. En especial Estados Unidos y China que demuestran muy bajo consumo de energía limpia. Aunque China muestra un leve crecimiento ultimamente; esperamos que siga así.

5. Conclusion

La generación de las visualizaciones fue bastante sencilla una vez que se tenían los datos a procesar. El manejo en python jugó de nuestro lado.
En un comienzo fue difícil realizar las consultas en *PIG*, puesto que eran bastante estrictas y complejas. Luego notamos que para obtener los resultados deseados se podían generar de forma distintas -relajando las restricciones a los indicadores- y logramos sobrellevar nuestras dificultados con *PIG*.
Para mejorar se podrían haber manejado los datos con *MapReduce*, que da más libertades y así crear consultas más complejas.

Appendix

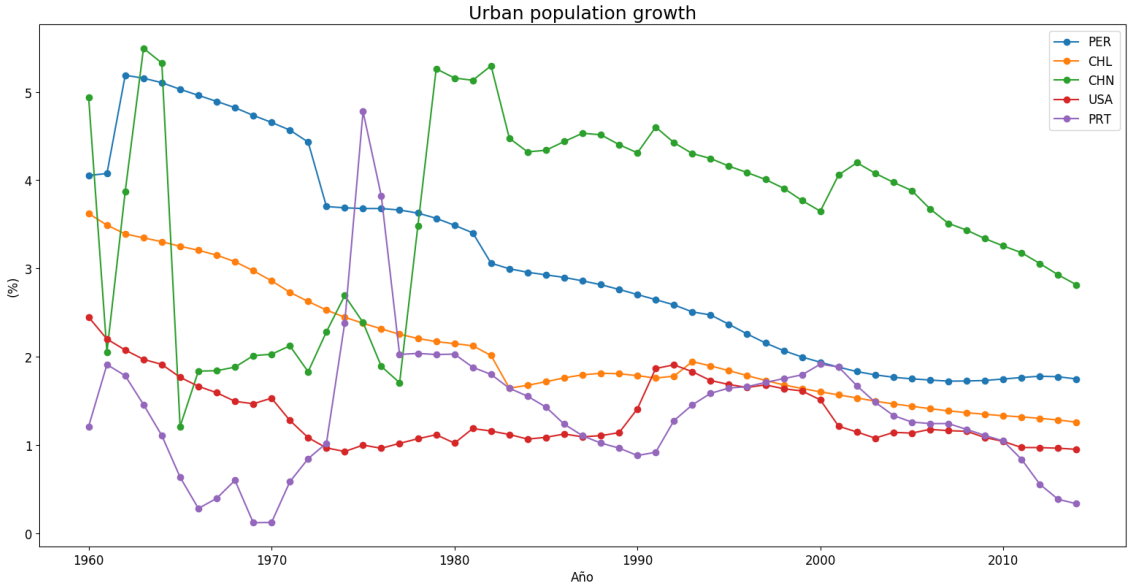


Figura 1: Crecimiento de población urbana.

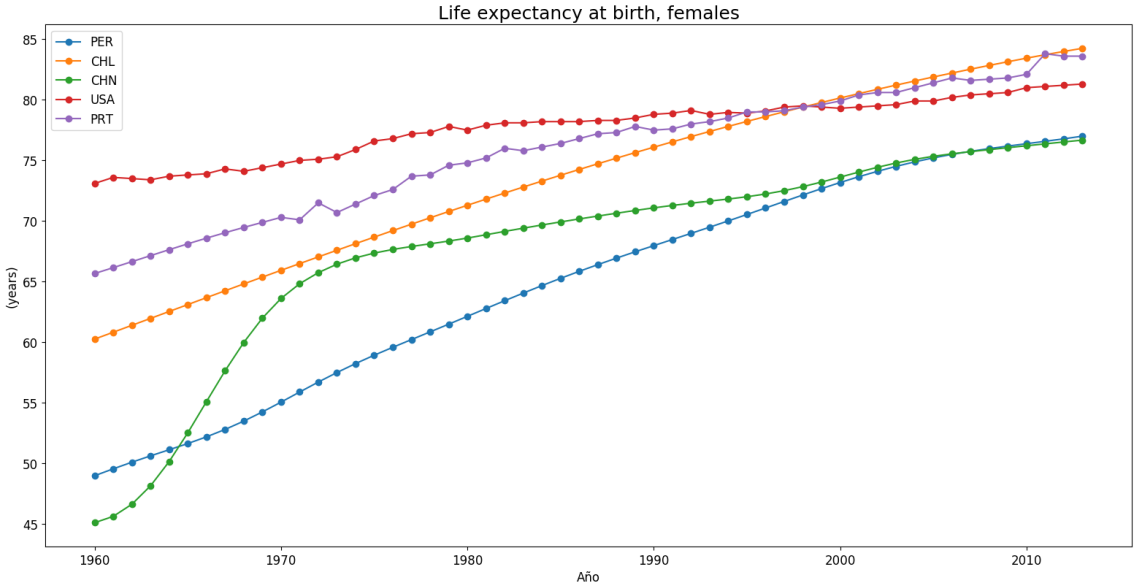


Figura 2: Esperanza de vida al nacer, mujeres.

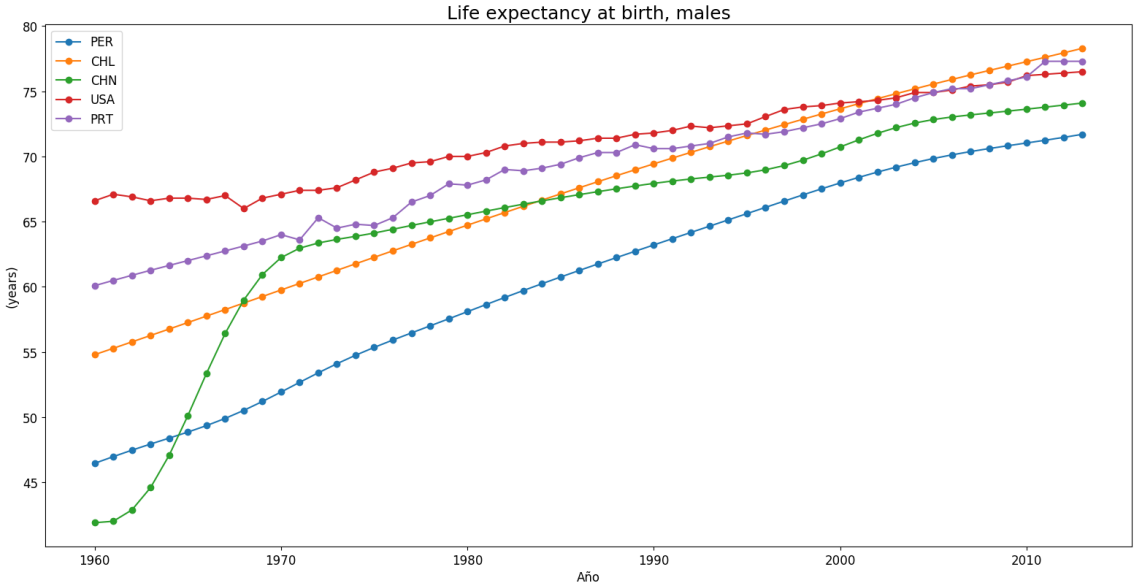


Figura 3: Esperanza de vida al nacer, hombres.

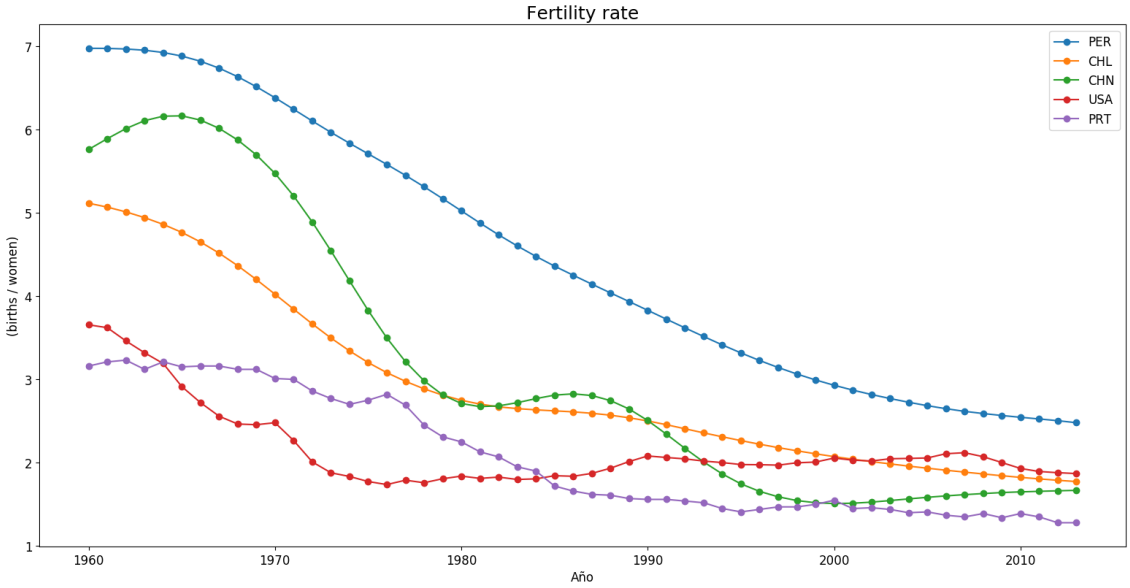


Figura 4: Fertilidad

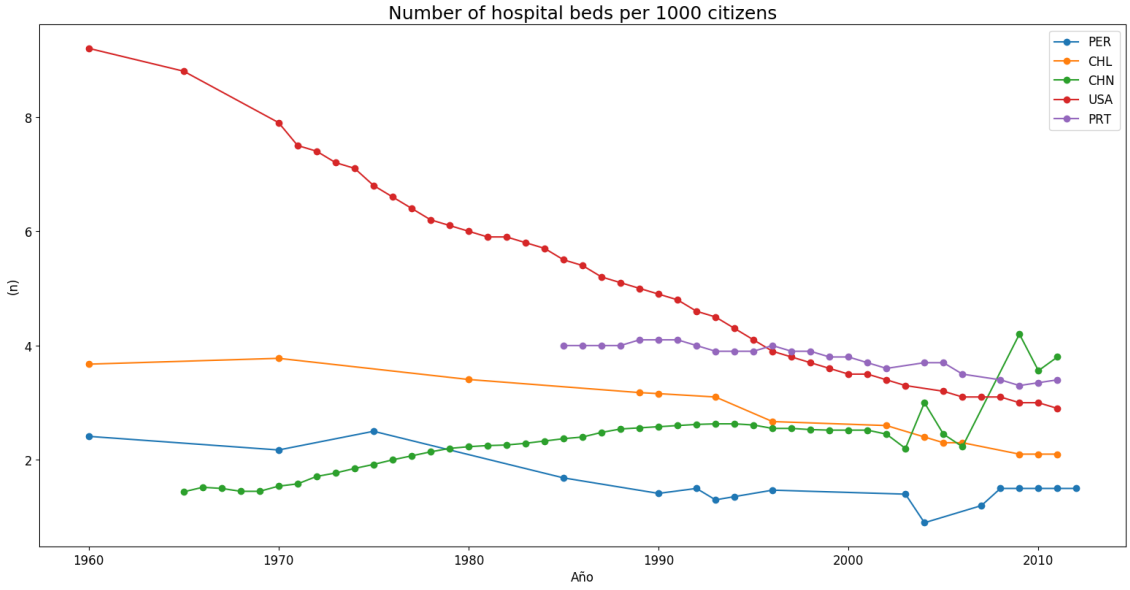


Figura 5: Numero de camas en hospitales por cada 1000 habitantes.

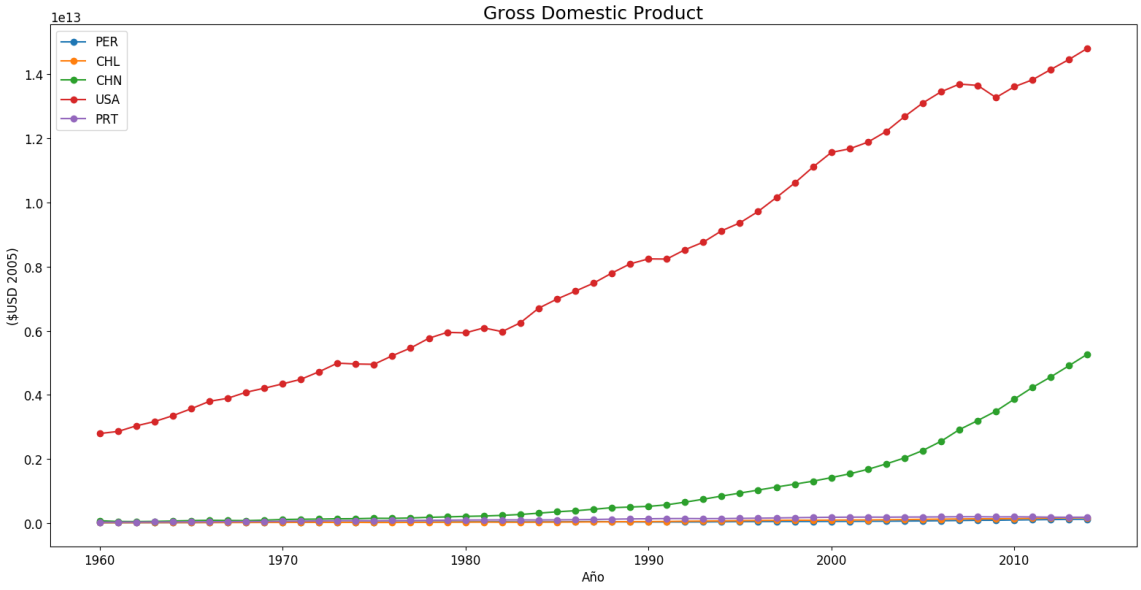


Figura 6: Producto interno bruto.

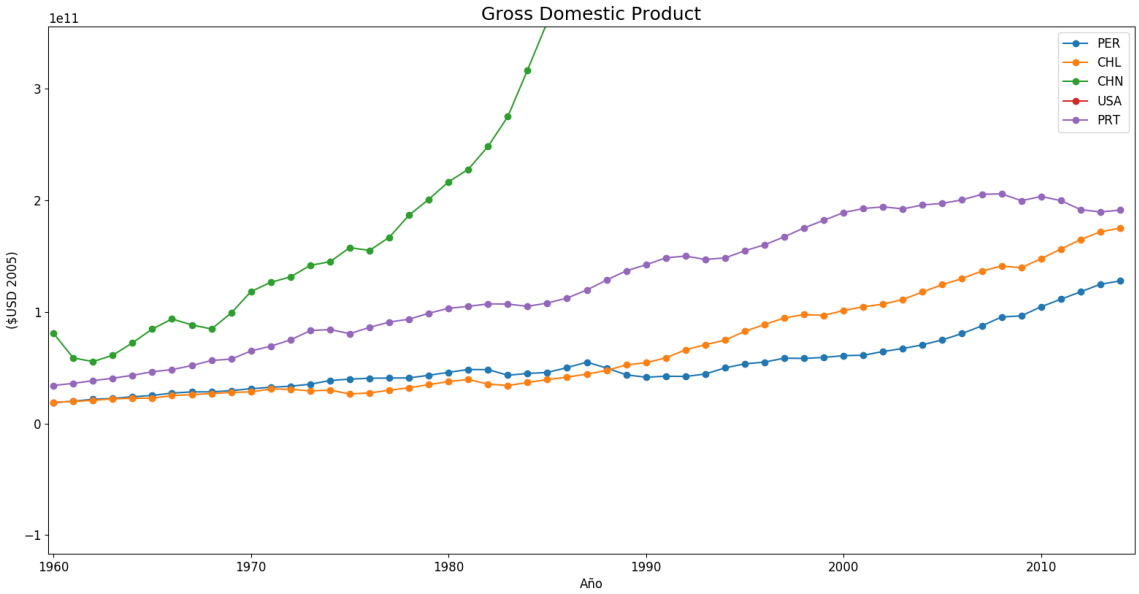


Figura 7: Producto interno bruto: Zoom.

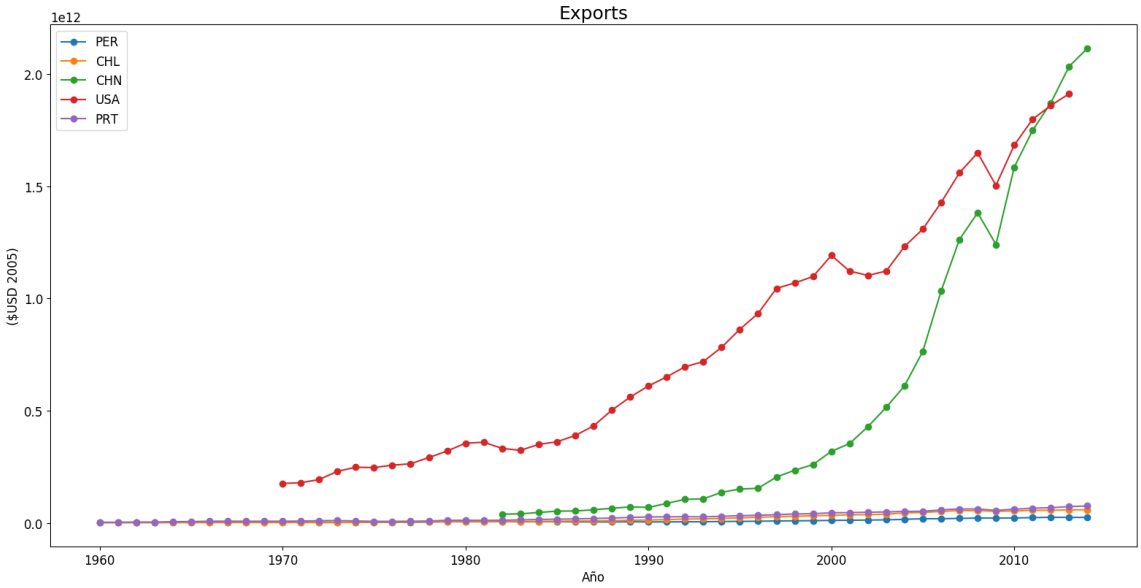


Figura 8: Exportaciones.

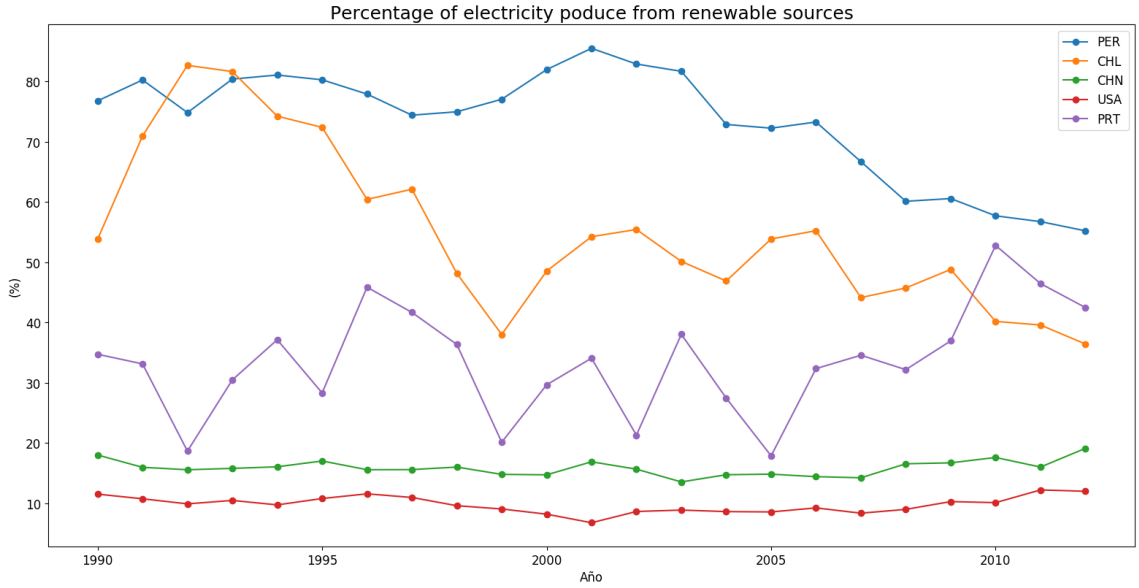


Figura 9: Porcentaje de electricidad producida por fuentes renovables.

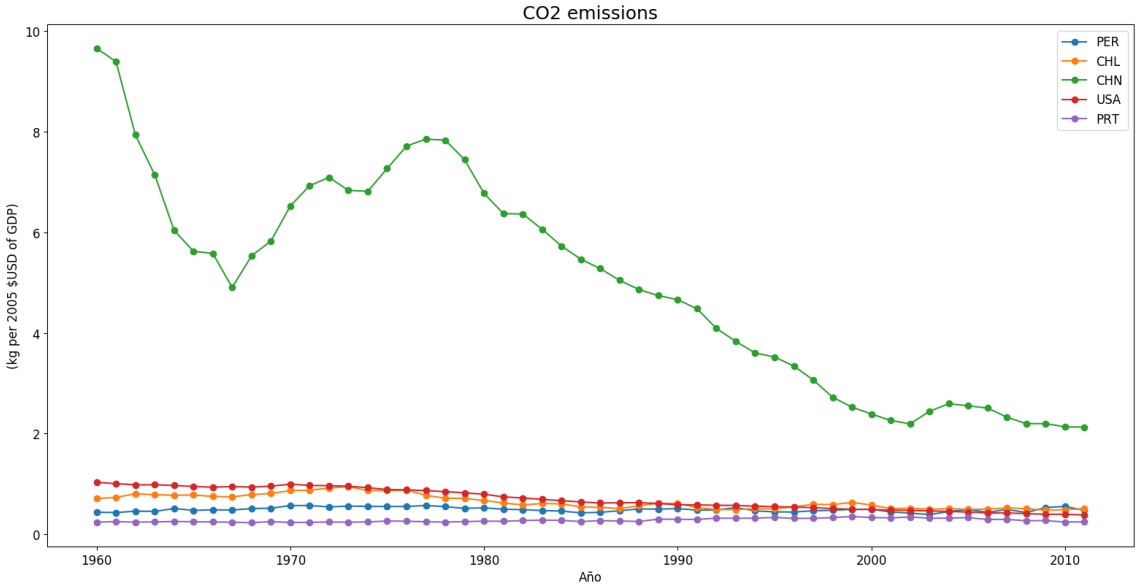
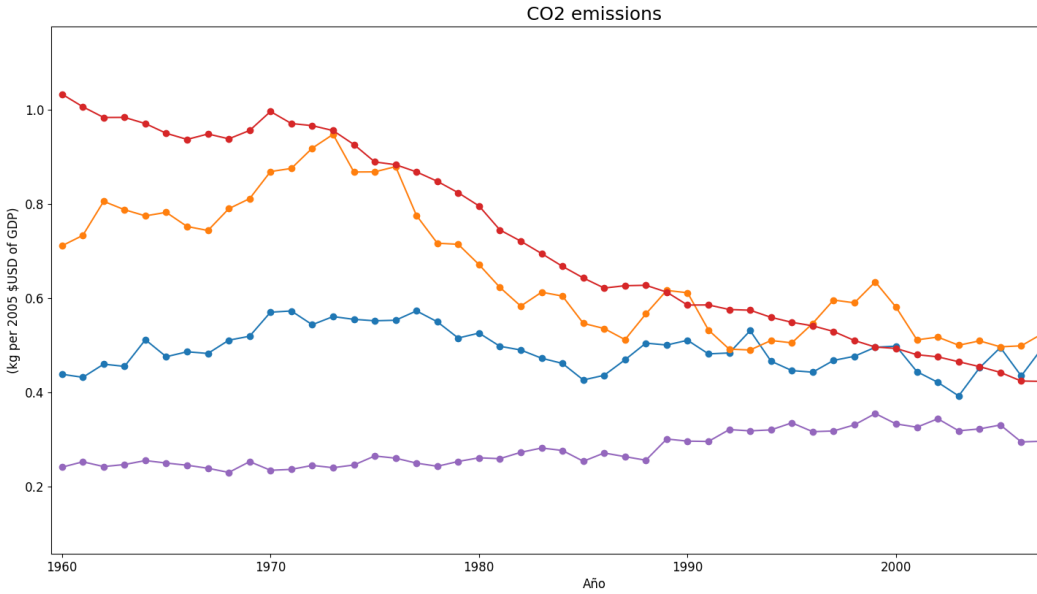


Figura 10: Emisiones de CO2.



- co2.png

Figura 11: Emisiones de CO2: Zoom.

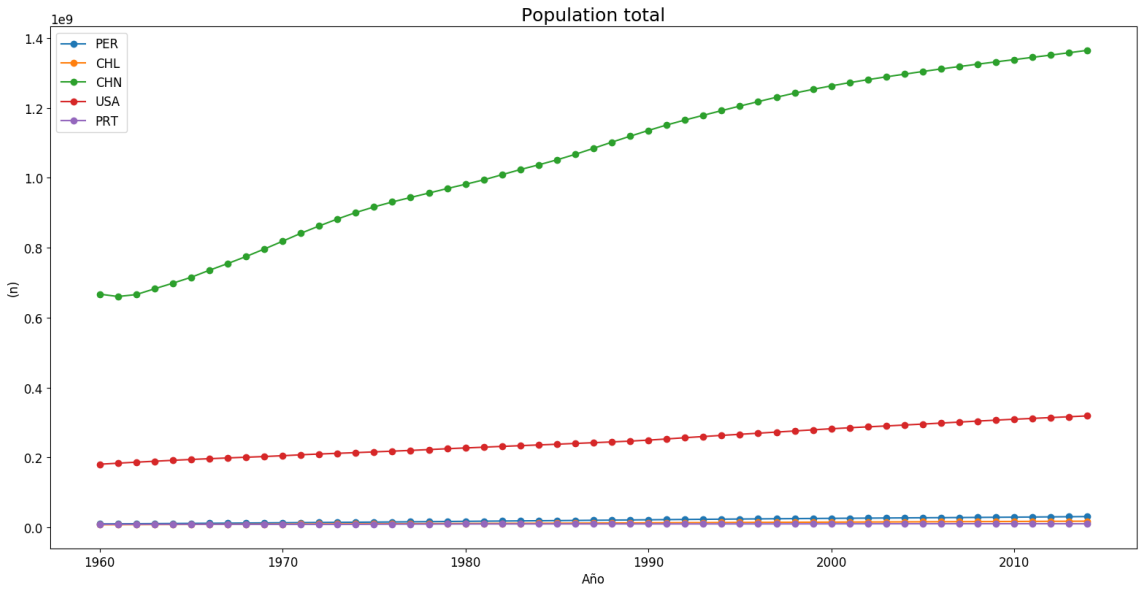


Figura 12: Población total.

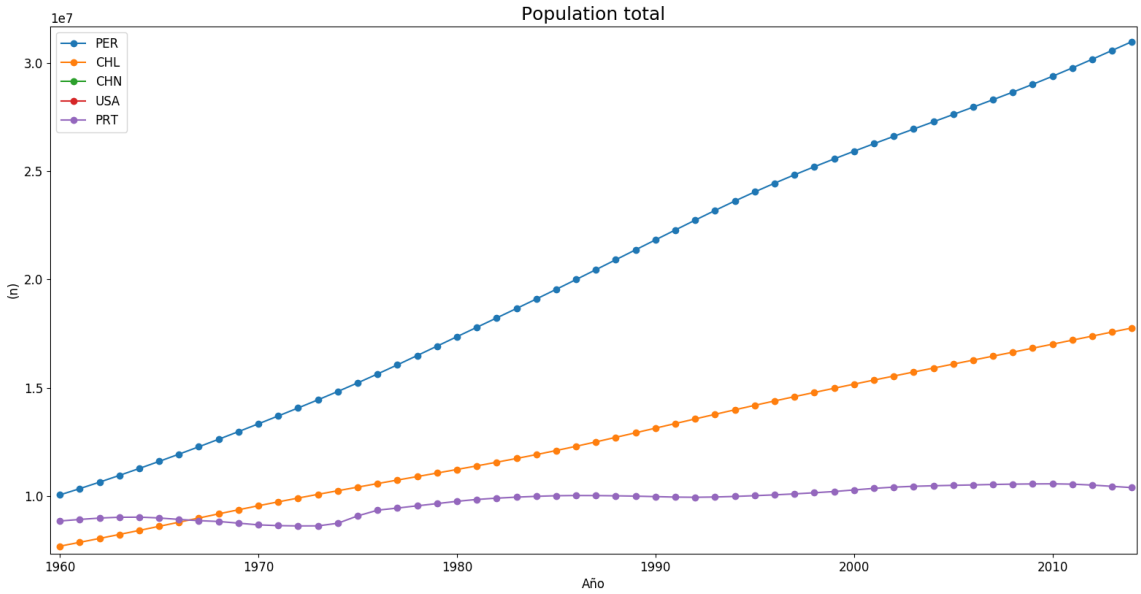


Figura 13: Población total: Zoom.