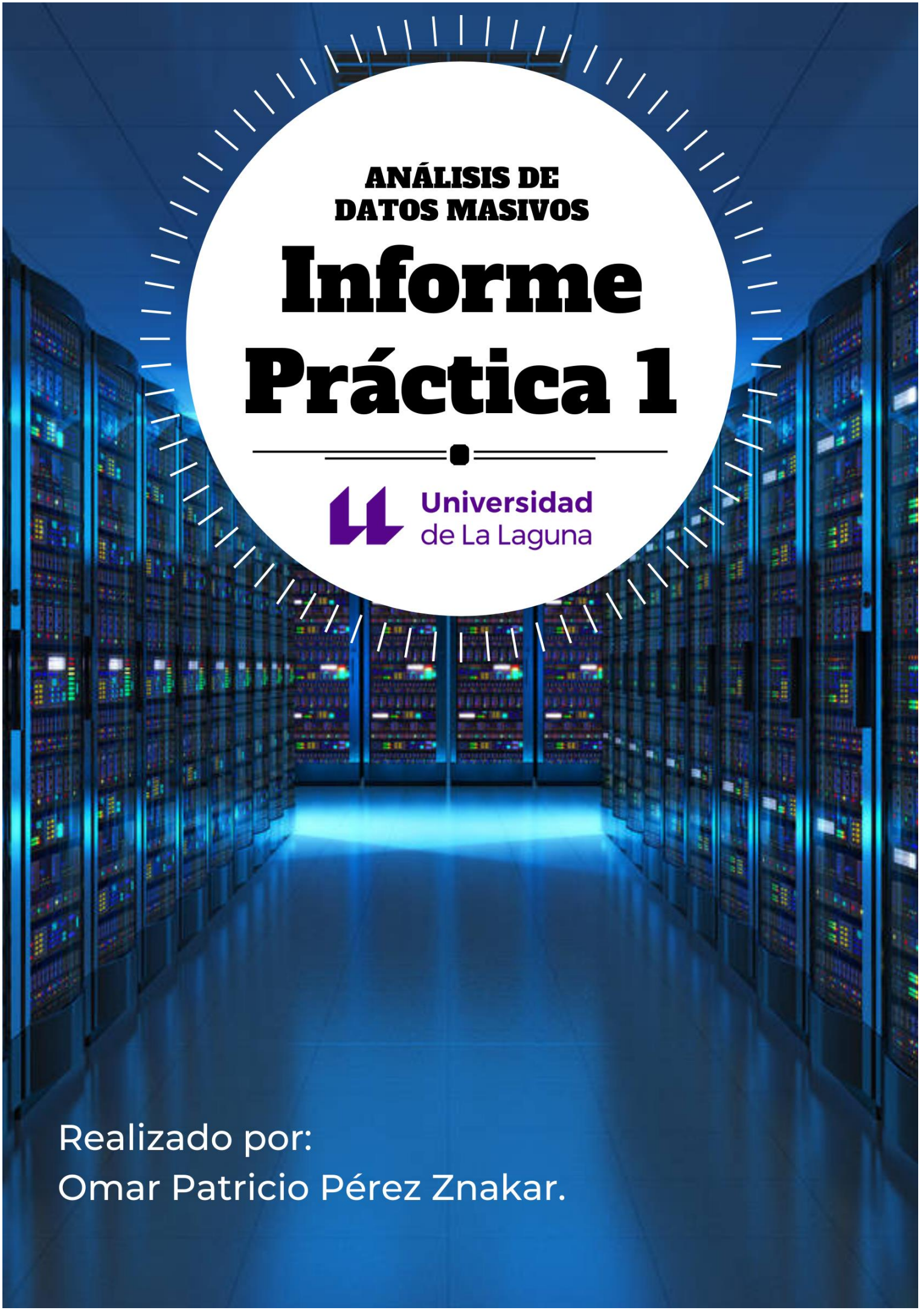


ANÁLISIS DE DATOS MASIVOS

Informe Práctica 1

Universidad de La Laguna

Realizado por:
Omar Patricio Pérez Znakar.

1. Lee el documento "Why scientists need to be better at data visualizationURL" hasta la sección "Ruinous rainbows", esta última sección no incluida. Haz un resumen de no más de una página con las conclusiones que obtienes.

El ser humano necesita de una representación gráfica de los datos para poder procesar mejor la información. Para ello, podremos utilizar diferentes tipos de representaciones desde imágenes hasta gráficos. Sin embargo, no siempre se utiliza bien estos recursos, de hecho, hay estudios que validan que no procesamos todas las ayudas visuales de la misma manera, por lo que si utilizamos ciertos gráficos específicos de manera inapropiada el resultado sería contrario al que se quiere alcanzar. Esto no quiere decir que determinados gráficos sean malos o buenos, solo que deben de estar adaptados a la información que se desea transmitir. Por ejemplo, en el caso de los gráficos circulares permiten entender que cada porción genera un todo, no obstante, a la hora de comparar información no resultan del todo útiles, una alternativa correcta sería un diagrama de barras (aunque este limita la visualización del todo).

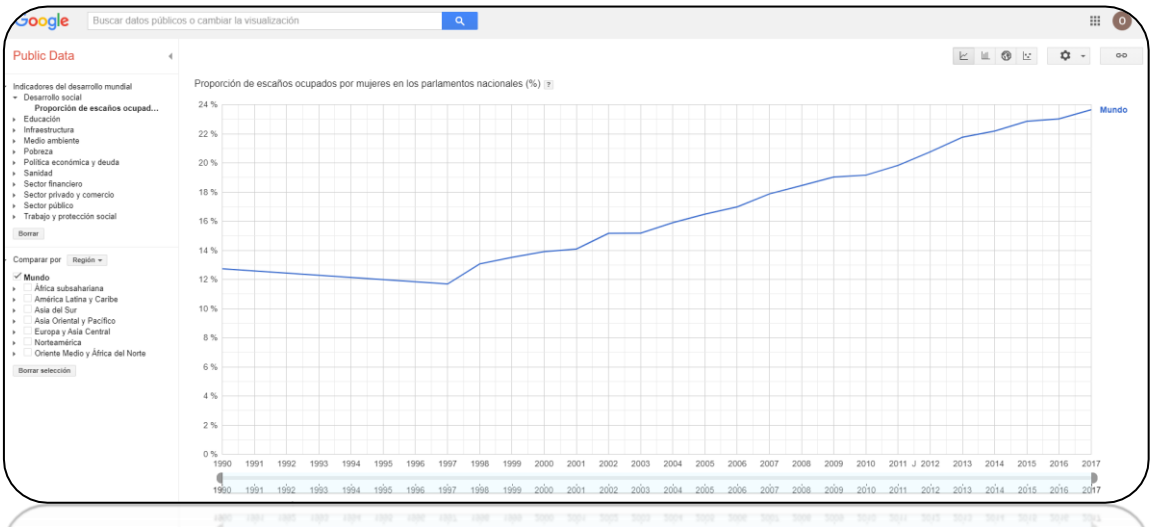
Lo comentado anteriormente no tiene mucha fuerza en el ámbito científico en la actualidad, ya que a pesar de su relevancia la ciencia se mantiene reacia a considerar la importancia de estos, tanto así, que son muy pocos los científicos que actualmente utilizan las representaciones gráficas de forma adecuada. De igual forma, solo existen dos conferencias internacionales dedicadas a este ámbito.

Por todo lo comentado anteriormente, considero que toda la información debe estar consecuentemente apoyada por una representación visual que, ayude a la facilitación en el proceso de adquirir la información que es lo que el autor nos quiere transmitir. No obstante, bajo mi punto de vista, no hay que centrarse ni en la representación gráfica ni en los datos, sino que, debe haber un equilibrio de ambas partes.

2. En la sesión de teoría hemos analizado el portal Gapminder que proporciona información visual sobre muchas temáticas. Realiza una búsqueda en internet intentando localizar portales o frameworks de características similares, en los que grandes volúmenes de datos se muestren de manera gráfica/visual y con una perspectiva genérica.

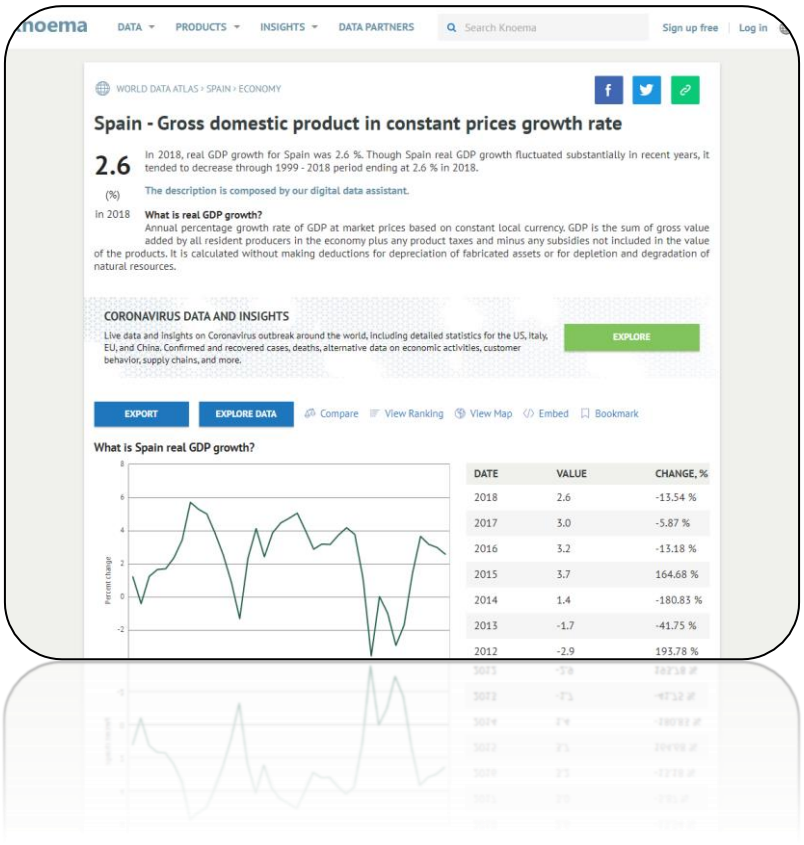
En la actualidad existen diversas herramientas similares a las descritas en el enunciado de este ejercicio. Algunas de ellas las podemos ver a continuación:

Google Public Data Explorer

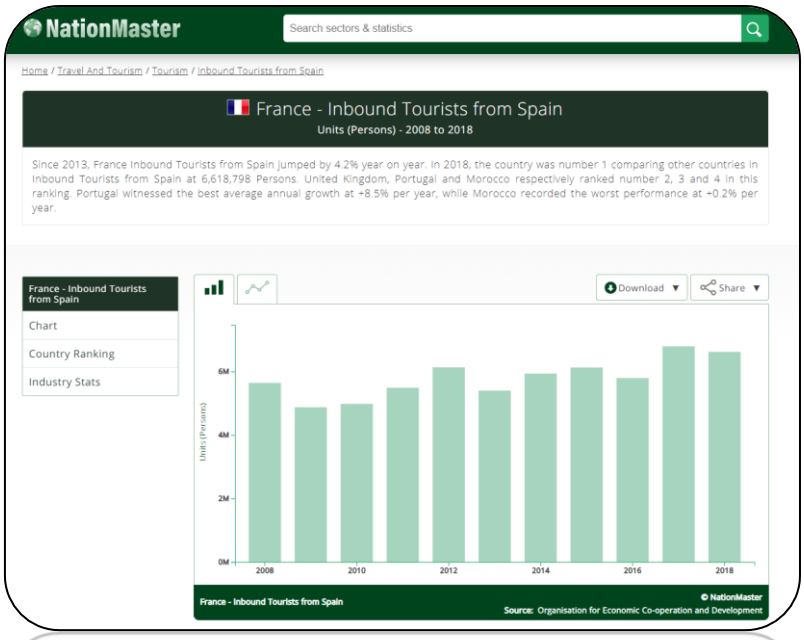


Statista

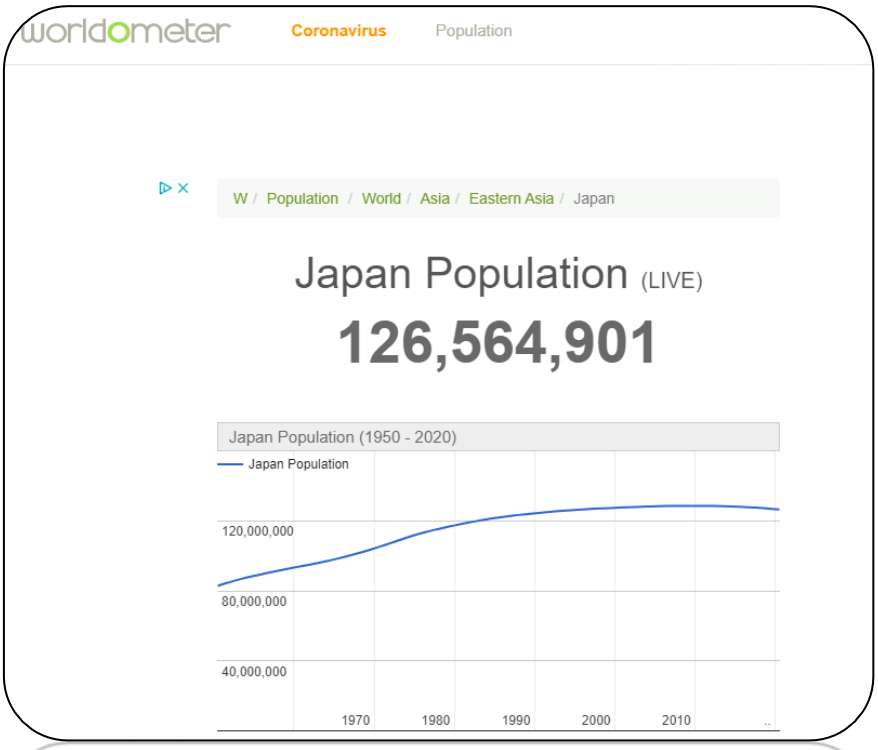
Knoema



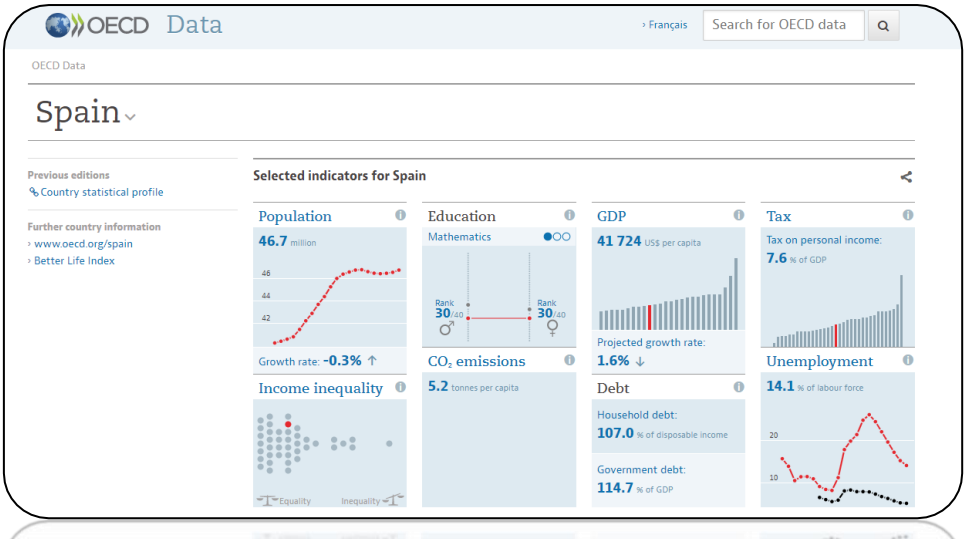
NationMaster



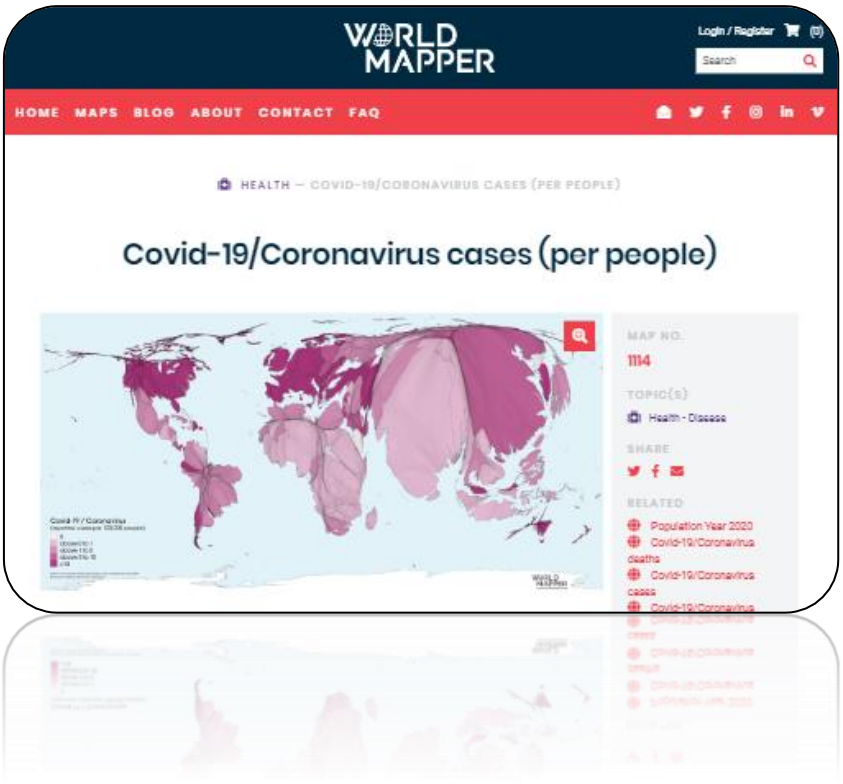
Worldometers



OECD



Worldmapper

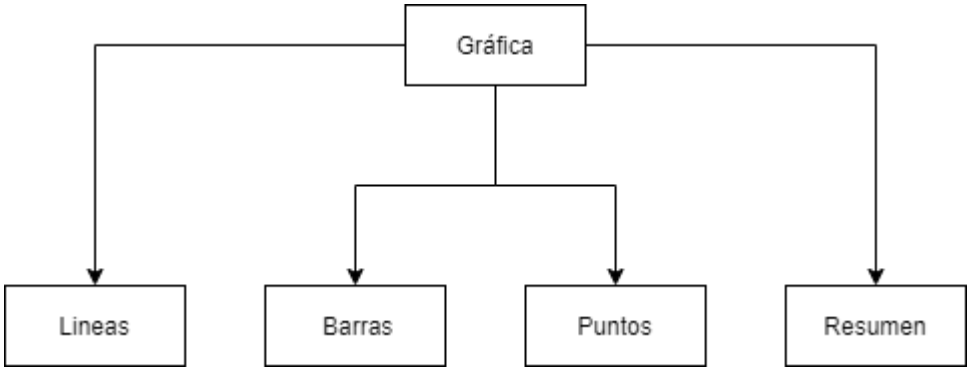


3. Desarrolla un pequeño entorno de visualización. El entorno debe separar las rutinas de representación de los datos, de las fuentes de las que se obtienen datos mediante una capa de abstracción en la que se estandaricen las interfaces mediante dataframes unificados que permitirán realizar representaciones gráficas para determinados patrones de datos. Las fuentes de datos podrían ser diversas, por tanto, asociado a cada una de ellas deberíamos disponer de un analizador/transformador que convierta el conjunto de datos desde la fuente de origen al dataframe. El patrón de diseño "Estrategia" [3, 4] puede serte útil en este nivel del desarrollo. Para una determinada fuente de datos deberías generar al menos dos representaciones, diferentes, por ejemplo, líneas y diagramas de barras. Describe en una página el diagrama de clases del framework y la tecnología utilizada para su desarrollo. Considera al menos dos fuentes de datos para ilustrar su uso:

a) Datos de los coronavirus procedentes del repositorio de la Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE), [1].

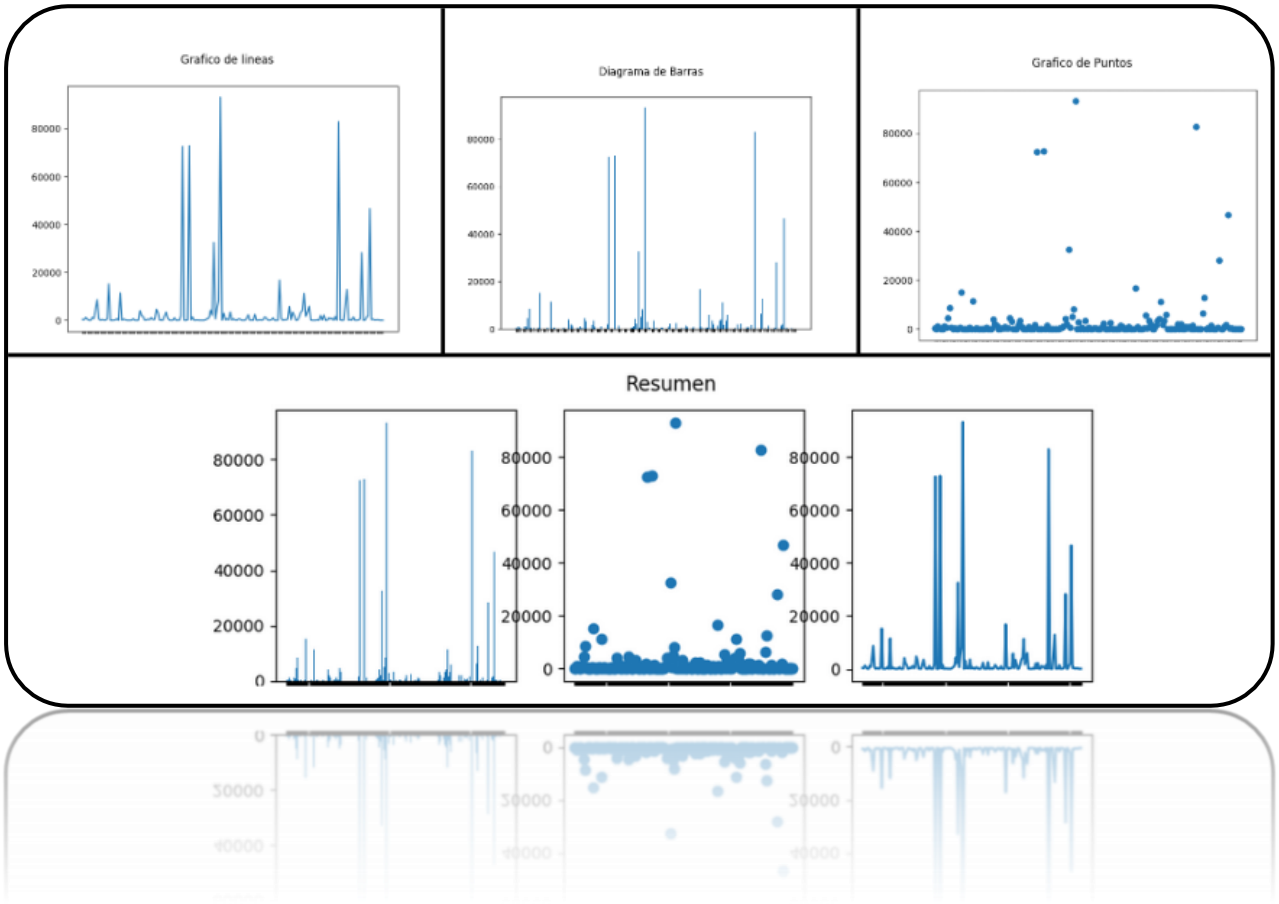
b) Alguna base de datos que esté en abierto y que sea de tu interés, por ejemplo, alguna de las que se encuentran disponibles en [2]. También puede ser el conjunto de datos para que el ya estás haciendo el proyecto en esta asignatura.

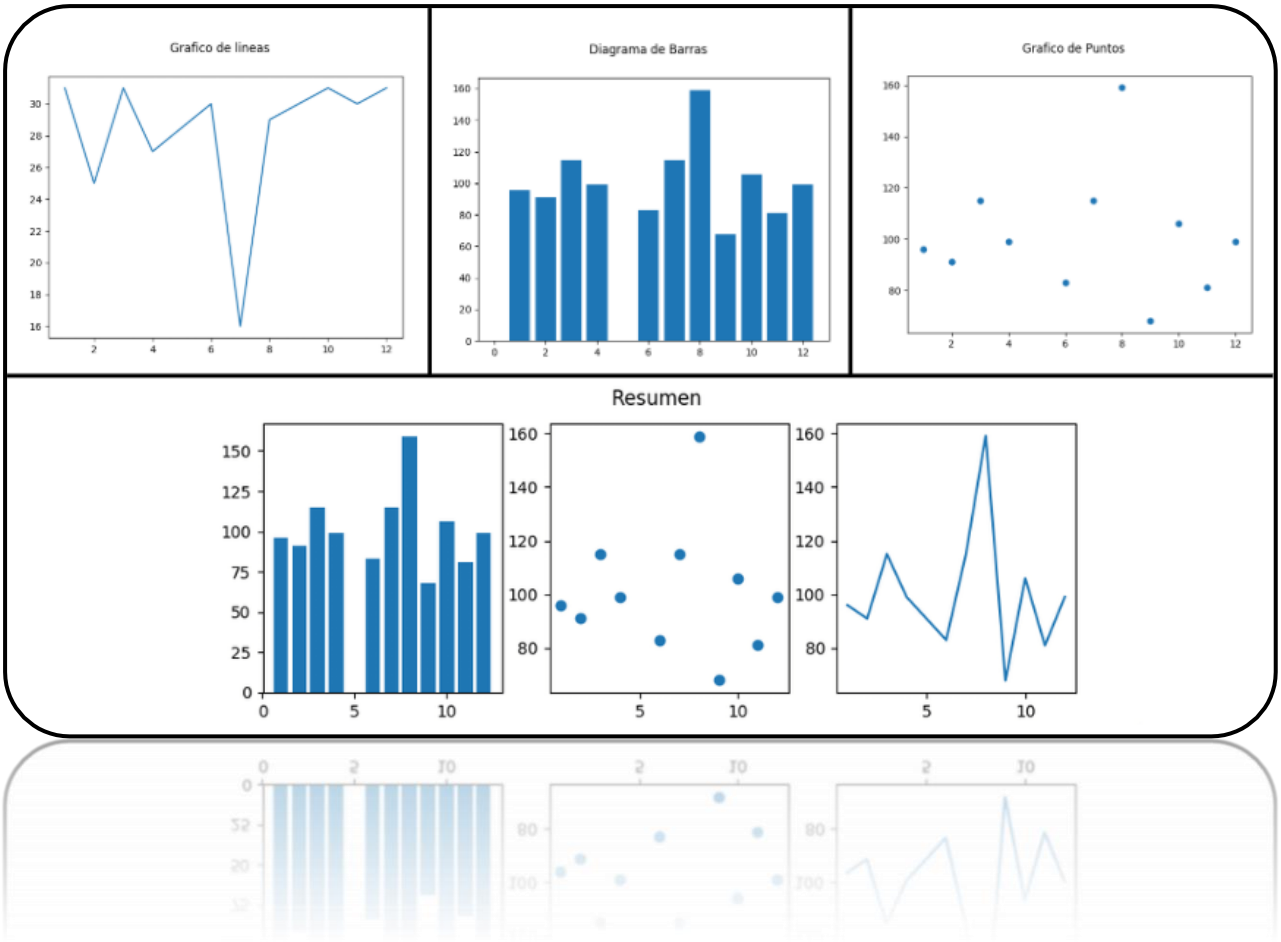
Para este ejercicio se ha usado Python junto junto con las librerías Pandas (tratar los datos) y Matplotlib (representar los datos). A su vez, se ha utilizado el patrón de patrón de diseño "Estrategia", bajo el siguiente esquema:



De igual forma, se ha utilizado el Dataset del Coronavirus propuesto por el profesor junto con un dataset que nos proporciona la calidad del aire en la Zona de TomeCano (recogido de la página del Cabildo de Tenerife). Las salidas obtenidas han sido:

Coronavirus 06-04-2020





A la hora de ejecutar el código (“make runTomeCano” o “make runCoronavirus”) nos pedirá diferentes parámetros. Estos son:

- **Valor eje X:** valor del eje X y por el que se van a agrupar los datos.
- **Valor eje Y:** valor del eje Y.
- **Tipo representación:** en caso de existir agrupaciones con el mismo nombre, este parámetro nos permite saber que hacer en ese caso (sumar, máximo o mínimo).
- **Gráfica:** nos permitirá seleccionar el tipo de gráfico que deseamos (Líneas, Barras, Puntos o un resumen de todos).