**FUNCIONAMIENTO CÓDIGO PANDA\_DATANALYSIS**

**PRESENTADO POR:**

**JOHAN SEBASTIAN FUENTES ORTEGA**

**COD: 230172002**

**UNIDAD CENTRAL DEL VALLE DEL CAUCA**

**INTELIGENCIA ARTIFICIAL**

**INGENIERÍA DE SISTEMAS**

**2021**

**PANDA DATANALYSIS**

Los datos a continuación consisten en la población total y el número de muertes por tuberculosis en cada uno de los países de todo el mundo en el año 2013.

Como primera medida llamamos la librería Warning para desactivarla, debido que en algún proceso más adelante nos podría generar mensajes de advertencia o nos detiene el programa. Por ende se usa el código:

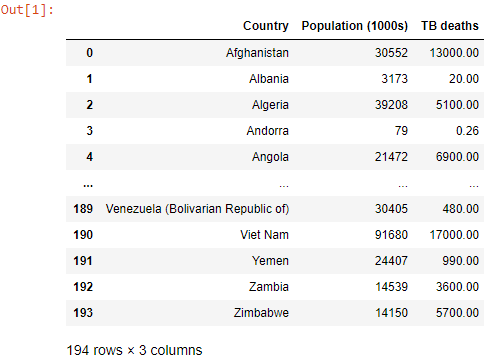


Posteriormente importamos las librerías Pandas con el fin de llevar a cabo códigos como datasets.



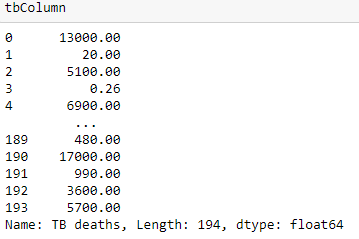
Ahora bien, el programa comienza creando una variable de tipo objeto para almacenar los datos que obtiene del archivo WHO POP TB all.xls, un archivo de excel que se encuentra en la misma carpeta donde se encuentra el programa.

Cuando ejecutamos solamente el comando data, jupyter lo lee como si fuera un printf, permitiéndonos visualizar los datos que están en ese objeto llamado data.

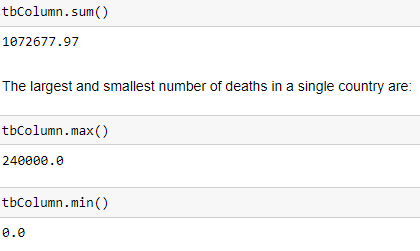


Como podemos visualizar para obtener el rango del problema definimos un dataset de nombre tbColumn que almacenará los datos que se encuentran en la columna TB deaths que están en el objeto data.

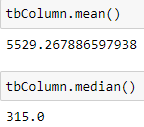




Posteriormente se procede a calcular el número de muertes que hubieron en el año 2013 por tuberculosis, el máximo número de muertes y el mínimo número de muertes. Esto se hace con las funciones para dataset que brinda panda (sum(), max() y min())

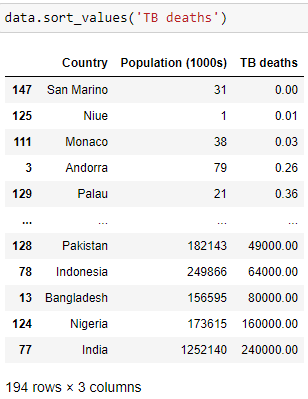


Ahora bien, como se explica en jupyter, de 0 a casi un cuarto de millón de muertes es un rango enorme. El número promedio de muertes puede dar una mejor idea de la gravedad del problema en cada país. Para ello calculan la media de los datos con la función mean() o utilizan la mediana con la función median().



El hecho de que la mediana sea más bajo que la media implica que en algunos países hubo un número muy grande de muertes en el 2013 por tuberculosis, razón por la que la media es bastante alta.

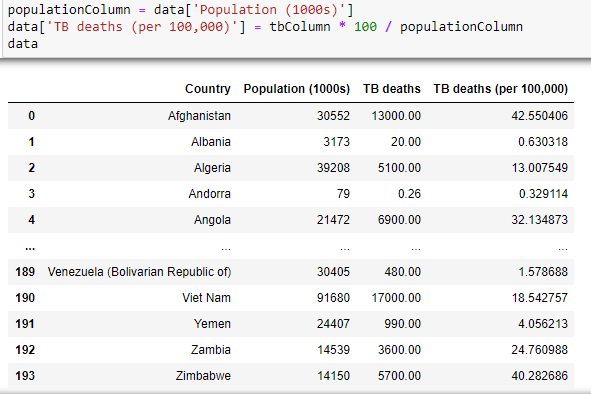
Como se planea visualizar los países más afectados se debe hacer un orden, por ende se aplica la función sort() (en el código no funciona debido a que está mal escrito o hubo una actualización de la función en panda, por ende lo reemplacé como sort\_values()) este método va a ordenar los datos de acuerdo a TB deaths, como se pasa solamente el parámetro nombre de la columna, implica que el orden se va a hacer de menor a mayor, si se quiere que sea de mayor a menor pasamos el otro parámetro que sería como ‘ascending=False’.



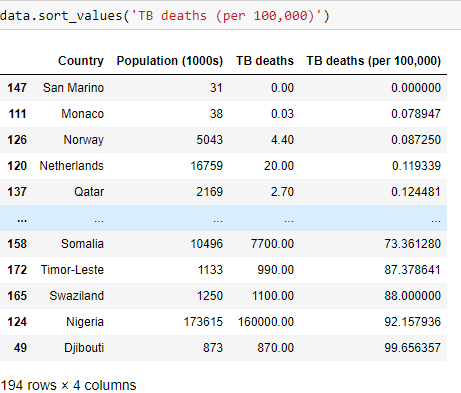
Posteriormente como dice el archivo el cuadro plantea que el número de muertes es proporcional al número de personas, proponen calcular la tasa de mortalidad por cada 100000 habitantes. Para ello entonces se hará un nuevo objeto llamado populationColumn que almacenará los datos de Population (1000s) que tomó el dataset data del archivo de excel.

Ahora bien, al dataset data le creamos una nueva columna que denominaremos ‘TB deaths (per 100,000)’ el cual almacenará los resultados de la multiplicación de tbColumn por 100 dividido entre populationColumn.

Posteriormente lo mostramos en pantalla llamando la variable data (esto funciona porque como dije antes es una función que jupyter nos permite efectuar).



Por último ejecutamos el comando data.sort\_values(TB deaths (‘per 100,000’) para ordenar los datos de acuerdo al número de muertes por cada cien mil habitantes.



Ya hecho esto, en el código se realiza la conclusión del estudio.