ESCUELA POLITÉCNICA NACIONAL
ESCUELA DE FORMACIÓN DE TECNÓLOGOS
DESARROLLO DE SOFTWARE

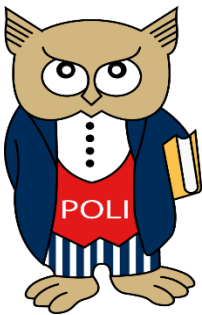
INFORME DE ANÁLISIS DE DATOS

Limpieza, Visualización e Interpretación de Datos. Sobre el Índice de Desarrollo Humano (IDH)

Elaborado por: Jimy Calvo M.
Martin Rosero B.

Fecha de Entrega: 12-Sept-2022

Periodo Académico2022-A.



Resumen

En este informe se presentará una de las diversas formas de extracción, manejo y limpieza de datos obtenido desde un documento (en el caso presentado en este informe, será desde un documento en formato Excel). Siendo el caso de estudio el "Índice de Desarrollo Humano conocido también como (IDH)" donde se extraerá, limpiará, analizará los datos obtenidos sobre este. Cuya finalidad es aplicar los conocimientos de sobre la extracción y almacenamiento de una base de datos.

Para el proceso de extracción y limpieza de los datos se lo realizara mediante el entorno de trabajo interactivo conocido como "Jupyter", entorno donde se emplear el lenguaje de programación "Python". Además, del uso de librerías que gestiona conjuntos de datos como lo son Pandas, Numpy y otras librerías para la realización de gráficos. Donde dichos gráficos serán de utilidad para la interpretación de los datos y su posterior análisis.

"Sin análisis de grandes volúmenes de datos, las empresas son ciegos y sordos, vagando hacia fuera sobre la web como ciervos en una autopista."

– Geoffrey Moore

Tabla de contenido

Tabla de Ilustraciones

1. Introducción

El proyecto consiste en el empleo de los conocimientos sobre los procesos de extracción, limpieza, graficación e identificación de datos obtenidos desde un documento digital o sitio web, empleando los conocimientos adquiridos en la asignatura de Análisis de Datos de la carrera Técnica de Desarrollo de Software. Dichos procesos se lo aplicarán a datos reales, obtenidos desde un documento digital, siendo el caso de estudio de este informe, los datos obtenidos sobre el Índice de Desarrollo Humano [1] . A estos datos se aplicará los procesos anteriormente mencionados, para su posterior análisis de estos. Durante estos procedimientos se emplear herramientas (software como Jupyter) y librerías de Python de gestión y vista de datos, con la finalidad de facilitar el estudio de los datos recopilados y llegar a convertir los datos en ideas y posteriormente en conocimiento.

1.1. Objetivos

1.1.1. General:

- Analizar la información disponible del archivo sobre el Índice de Desarrollo Humano a su elección, para la limpieza, visualización e interpretación de datos.

1.1.2. Específicos:

- Aplicar los conceptos de Análisis de datos en un caso real con la finalidad de llegar a la interpretación.
- Identificar las funciones y librería a utilizar para conversión de datos en conocimiento.
- Interpretar la información obtenida una vez realizada la limpieza, preparación y visualización de datos

1.2. Planteamiento del problema

Se busca la extracción de datos mediante el uso de herramientas de análisis de

datos empleado el lenguaje de programan de alto nivel Python junto a las librerías y gestión y graficacion de datos, y todo esto dentro del entono de Júpiter. Y para esto se ha planteado la extracción de los datos del IDH el cual obtener desde el sitio web de Human Develoment Reports. Acto seguido, empleando Jupiter se extraerá los datos de la Tabla sobre Human Development Index and its components. A dichos datos se creará el código correspondiente parala limpieza, y visualización de datos empleado las respectivas librerías de Python. Y para finalizar, se llevará a cabo la interpretación los datos

2. Marco Teórico

2.1. IDH – Índice de desarrollo Humano

Según el sitio web de Human Development Reports es una media que resume el crecimiento de una nación bajo ciertos criterios como son el PIB, nivel de educación, esperanza de vida, entre otros parámetros cuya finalidad es estimular al debate sobre las necesidades de cada país, en el ámbito de fomentar el desarrollo de este. A demás de servir como medidor sobre qué tan desarrollado se encuentra un país con respecto a los demás [2].

2.2. Python

Python es un lenguaje de programación que es utilizado en varios sectores, este lenguaje se lo puede encontrar en aplicaciones web, en el desarrollo de software, la ciencia de datos (análisis de Datos) y el machine learning (ML) [3]. Python también es un lenguaje de alto nivel orientado a objetos es decir que es un lenguaje ideal para el desarrollo de aplicaciones, así como scripting. Es el leguaje que emplea Jupyter dentro de su entorno por defecto. Este lenguaje incluye librerías que permite la gestión de datos, Dicho lenguaje será empleado en el proceso de extracción de aquí su importancia de conocer sobre este.

2.3. Scripting

Es un lenguaje de programación que permite ejecutar funciones.

Relaciona los objetos de resultados y la interfaz de usuario, además de poder ejecutar sintaxis de comandos. En Python es empleado para personalizar una tabla dinámica [4].

2.4. Jupyter

Es una base de un entorno informático interactivo donde se puede procesar datos como en el caso de la computación científica, la ciencia de datos y el análisis el cual emplea la tecnología de los navegadores web para ejecutar. Este entorno presenta la función de cuaderno el cual permite la creación de documentos que combinan código en vivo con texto narrativo, ecuaciones y visualizaciones [5].

Librerías:

2.5. Pandas

Es una herramienta de análisis y manipulación de datos de código especializada en el manejo y análisis de estructuras de datos. El cual permite leer y escribir fácilmente archivos en distintos formatos como son CSV, Excel, y bases de datos en SQL [6].

2.6. Numpy

Es una librería que permite el cálculo numérico y análisis de datos el cual es ideal para manejar datos de gran columna ya que permite la manipulación de datos mediante arreglos los cuales son colecciones de datos de un mismo tipo lo que permite la rapidez y eficiencia de la manipulación de datos [7].

2.7. Matplotlib

Es una herramienta que permite la creación de varios tipos de gráficos como también la personalización de estos entre los distintos tipos de graficas tenemos:

- Diagramas de áreas
- Diagramas de barras
- Diagramas de caja y bigotes
- Diagramas de contorno
- Diagramas de dispersión o puntos
- Diagramas de líneas
- Diagramas de sectores

- Diagramas de violín
- Histograma
- Mapas de color

Al utilizar esta librería, creará objetos Figure y Axes los cuales pueden llamará a sus métodos para agregar contenido y modificar la apariencia [8].

2.7.1. Pylab

Es un modulo que incluye algunas funciones que permite manipular la interfaz vasado en los distintos estados de la figura generada por matplotlib.

2.8. Seaborn

Librería que permite la visualización de datos mediante gráficos estadísticos de un conjunto de datos, el cual está basado en matplotlib con una estética mejora al momento de generar las gráficas, que a su vez proporciona más información [9].

2.9. Statsmodels

Es un módulo de Python que permite crear modelos estadísticos de diferentes formas y lo realiza mediante la proporción de clase y funciones que admite datos desde una lista de datos las cuales pueden ser de tipo DataFrames. Los resultados se prueban con los paquetes estadísticos existentes para garantizar que sean correctos [10].

2.9.1. `Graphics.Mosaicplot

Permite la creación de un gráfico de tipo mosaico a partir de una tabla de contingencia. Esta a su vez permite la visualizar datos en diversas categorías multivariados de forma rigurosa e informativa.

2.10. Py Mosaic

Es un modulo de datos que permite representar los datos mediante una simulación molecular, el cual permite también la realización de cálculos,

y este se puede combinar con códigos, datos, y documentación [11].

Comandos de las librerías empleadas

Tabla 1. Módulos Empleado Para extracción, limpieza y grafica de datos

Funciones Usadas	
Comando	**Funciones**
pd.read_excel	Función para leer un archivo.xlsx
df.info()	Función para mostrar la información de un DataFrame
df.head()	Función que imprime los primeros registros de un DataFrame
df.tail()	Función que imprime los últimos registros de un DataFrame
df.isna().sum().sort_values()	Función conjunta que imprime la cantidad total de valores nulos
sns.heatmap(df.isna())	Función que imprime los valores nulos de manera grafica
df.dropna()	Función que elimina los registros que contengan algún valor nulo
df.rename(columns={})	Función que renombra las columnas seleccionadas por medio de un diccionario
df.reset_index(drop=True, inplace=True)	Función que crea un Índex nuevo reemplazando al anterior
df.set_index()	Función que te permite establecer una columna existente como Índex
mosaic(df,[])	Función para graficas de Mosaico
df.plot(kind='barch')	Función para graficas de filas
df.plot.area()	Función para graficas de líneas con área

Resumen de los comandos empleados de cada Liberia de Python y su funcionamiento, el cual será empleado para la extracción de datos limpieza y su posterior visualización

3. DESARROLLO

3.1. Instalación e Importación de Comandos

Para la extracción, limpieza y visualización de los datos Jupyter requiere que se instale e importe los módulos y librerías par ello se lo realiza mediante las siguientes líneas de comando

```
#INSTALACI0ON DE LIBRERIOAS
```

```
%pip install matplotlib
%pip install seaborn
%pip install modelo
%pip install pyMosaic
%pip install statsmodels
```

```
#IMPORTACION DE LIBRERIAS
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from statsmodels.graphics.mosaicplot import Mosaic
```

3.2 Obtención de los Datos

Para el proceso de obtención de datos, sea empleado la librería panda junto al comando read el cual permite leer el archivo obtenido de una ruta.

```
#FUNCION PARA LEER ARCHIVO EXCEL
def leer_datos(ruta):
      df = pd.read_excel(ruta)
      return df
```

Dichos datos serna almacenados en una variable df que permite la extracción de los datos.

```
EXTRACCION DE DATOS E IMPRESION DE LA INFORMACION DE LOS MISMOS

df_Original = leer_datos("2020_SAT.xlsx")
print("--------------Información de la data-------------------")
print("-------------------------------------------------------")
print(df_Original.info())
print("-------------Encabezado de los data--------------------")
print("-------------------------------------------------------")
print(df_Original.head(3))
```

```
Output exceeds the size limit. Open the full output data in a text editor
--------------------------Informacion de la data--------------------------
--------------------------------------------------------------------------
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 266 entries, 0 to 265
Data columns (total 9 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   Unnamed: 0                              190 non-null    object
 1   Unnamed: 1                              255 non-null    object
 2   Human Development Index (HDI)           212 non-null    object
 3   Life expectancy at birth                212 non-null    object
 4   Expected years of schooling             212 non-null    object
 5   Mean years of schooling                 212 non-null    object
 6   Gross national income (GNI) per capita  212 non-null    object
 7   GNI per capita rank minus HDI rank      211 non-null    object
 8   HDI rank                                211 non-null    object
dtypes: object(9)
memory usage: 18.8+ KB
None
--------------------------Encabezado de los data--------------------------
--------------------------------------------------------------------------
  Unnamed: 0                  Unnamed: 1 Human Development Index (HDI)   \
0   HDI rank                     Country                          Value
1        NaN                         NaN                           2019
2        NaN  VERY HIGH HUMAN DEVELOPMENT                            NaN
...
  GNI per capita rank minus HDI rank HDI rank
0                                NaN      NaN
1                               2019     2018
2                                NaN      NaN
```

***Fig. 1**. Captura de la ejecución de los comandos de extracción de datos.*
(Fuente: Propia)

Para el proceso de limpieza se realiza la búsqueda de valores nulos para ello se ejecuta los siguientes comandos.

```
    #CONTAVILIZAMOS LOS VALORES NULOS DEL DATAFRAME

    df_Original.isna().sum().sort_values()

    #Esto nos denota que existen dos columnas sin identificar
#Ademas de que existe datos nulos constantes en todas las demás columnas
```

```
Unnamed: 1                                11
Human Development Index (HDI)             54
Life expectancy at birth                  54
Expected years of schooling               54
Mean years of schooling                   54
Gross national income (GNI) per capita    54
GNI per capita rank minus HDI rank        55
HDI rank                                  55
Unnamed: 0                                76
dtype: int64
```

***Fig. 2**. Captura de la ejecución de los comandos de extracción de datos.*
(Fuente: Propia)

Con el fin de representar dichos valores se emplea la librería **Seaborn** el cual permitirá representar dichos valores en una grafica en este caso de tipo heatmap.

```
#REPRESENTACION GRAFICA DE DICHOS DATOS NULOS
sns.heatmap(df_Original.isna())
```

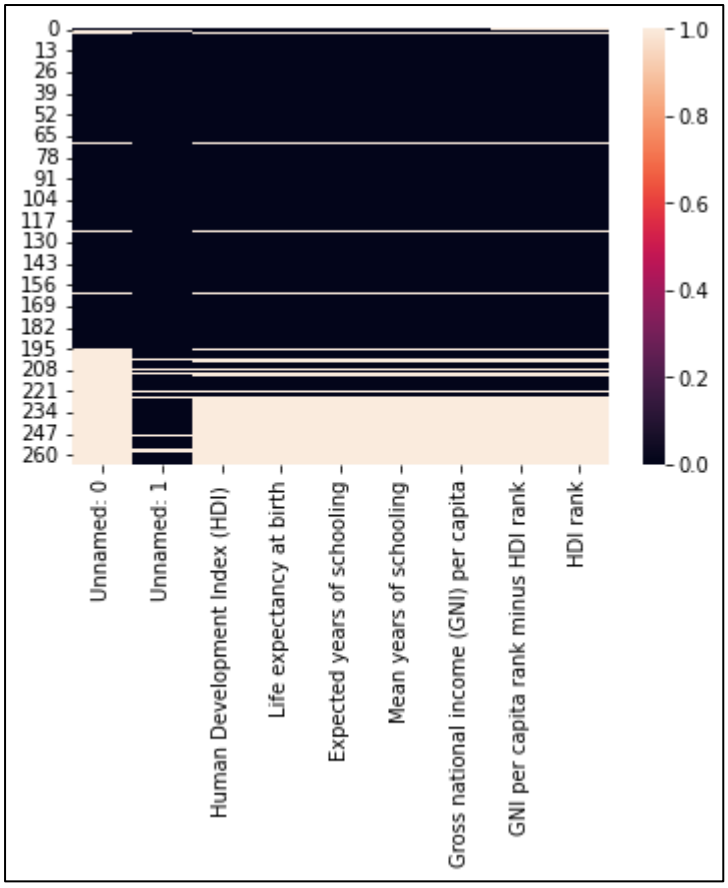


***Fig. 3**. Grafica de un mapa de calor de los valores vacíos por columna*
(Fuente: Propia)

Para una mayor visualización de los datos se imprime los DataFrame que contiene la variable.

```
#IMPRIMIREMOS EL DATAFRAME PARA LOGRAR IDENTIFICAR LA
 UBICACIÓN DELOS DATOS NULOS

df_Original

#Podemos observar cómo los datos NaN se ubican especialmente
 en la última sección del DATAFRAME
#Se denota que existen además datos no correspondientes en
 varias columnas
#Recordaremos los años en los que se hicieron estos datos ya
 que es importante para su análisis
```

	Unnamed: 0	Unnamed: 1	Human Development Index (HDI)	Life expectancy at birth	Expected years of schooling	Mean years of schooling	Gross national income (GNI) per capita	GNI per capita rank minus HDI rank	HDI rank
0	HDI rank	Country	Value	(years)	(years)	(years)	(2017 PPP $)	NaN	NaN
1	NaN	NaN	2019	2019	2019	2019	2019	2019	2018
2	NaN	VERY HIGH HUMAN DEVELOPMENT	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	1	Norway	0.957	82.4	18.06615	12.89775	66494.25217	7	1
4	2	Ireland	0.955	82.31	18.70529	12.666331	68370.58737	4	3
...	...	...	...	...	...	...	...	...	...
61	NaN	Column 2: UNDESA (2019a),	NaN	NaN	NaN	NaN	NaN	NaN	NaN
62	NaN	Column 3: UNESCO Institute for Statistics (202...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
63	NaN	Column 4: UNESCO Institute for Statistics (202...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
64	NaN	Column 5: World Bank (2020a), IMF (2020) and U...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
65	NaN	Column 6: Calculated based on data in columns ...	NaN	NaN	NaN	NaN	NaN	NaN	NaN

6 rows × 9 columns

Fig. 4*. Captura de la ejecución de los comandos de la variable que contiene los datos.*
(Fuente: Propia)

3.2.Integridad de Datos

Para este proceso primero se verificara los datos obtenidos para ello se ejecutara el comando head de la Liberia panda el cual permite visualizar los datos de un determinado número de filas.

	Unnamed: 0	Unnamed: 1	Human Development Index (HDI)	Life expectancy at birth	Expected years of schooling	Mean years of schooling	Gross national income (GNI) per capita	GNI per capita rank minus HDI rank	HDI rank
0	HDI rank	Country	Value	(years)	(years)	(years)	(2017 PPP $)	NaN	NaN
1	NaN	NaN	2019	2019	2019	2019	2019	2019	2018
2	NaN	VERY HIGH HUMAN DEVELOPMENT	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	1	Norway	0.957	82.4	18.06615	12.89775	66494.25217	7	1
4	2	Ireland	0.955	82.31	18.70529	12.666331	68370.58737	4	3
5	2	Switzerland	0.955	83.78	16.32844	13.380812	69393.52076	3	2
6	4	Hong Kong, China (SAR)	0.949	84.86	16.92947	12.27996	62984.76553	7	4
7	4	Iceland	0.949	82.99	19.08309	12.772787	54682.38057	14	4
8	6	Germany	0.947	81.33	16.97719	14.15168	55314.35355	11	4
9	7	Sweden	0.945	82.8	19.48234	12.54847	54507.80504	12	7
10	8	Australia	0.944	83.44	21.95433	12.724691	48084.84207	15	7
11	8	Netherlands	0.944	82.28	18.48513	12.4148	57707.06867	6	9
12	10	Denmark	0.94	80.9	18.89342	12.613803	58661.87084	2	10
13	11	Finland	0.938	81.91	19.39712	12.82413	48511.3978	11	11
14	11	Singapore	0.938	83.62	16.44088	11.62461	88155.21431	-8	12
15	13	United Kingdom	0.932	81.32	17.49843	13.24287	46070.64481	13	14
16	14	Belgium	0.931	81.63	19.77438	12.05027	52084.59209	6	13
17	14	New Zealand	0.931	82.29	18.83857	12.782077	40798.7196	18	14
18	16	Canada	0.929	82.43	16.15789	13.366105	48527.03573	5	14
19	17	United States	0.926	78.86	16.31039	13.41344	63825.65748	-7	17
20	18	Austria	0.922	81.54	16.09207	12.546144	56196.89869	-3	18
21	19	Israel	0.919	82.97	16.16283	13.034854	40186.84583	14	21
22	19	Japan	0.919	84.63	15.23057	12.85	42931.69575	9	20
23	19	Liechtenstein	0.919	80.67	14.90233	12.53854	131031.5898	-18	19
24	22	Slovenia	0.917	81.32	17.57206	12.65874	38079.53399	15	24
25	23	Korea (Republic of)	0.916	83.03	16.48165	12.213765	43043.71057	4	22
26	23	Luxembourg	0.916	82.25	14.25582	12.31145	72711.67412	-19	23
27	25	Spain	0.904	83.57	17.61613	10.25178	40974.52408	6	25
28	26	France	0.901	82.66	15.64217	11.4795	47172.53748	-1	26
29	27	Czechia	0.9	79.38	16.78511	12.722382	38108.56662	9	26
30	28	Malta	0.895	82.53	16.10407	11.32613	39554.52078	6	28
31	29	Estonia	0.892	78.75	15.989	13.143061	36019.26692	9	30
32	29	Italy	0.892	83.51	16.0889	10.37955	42776.35692	0	29
33	31	United Arab Emirates	0.89	77.97	14.3441	12.11122	67462.0953	-24	30
34	32	Greece	0.888	82.24	17.90792	10.55207	30154.63438	14	33
35	33	Cyprus	0.887	80.98	15.16887	12.17123	38206.84049	2	32
36	34	Lithuania	0.882	75.93	16.63731	13.07721	35798.67041	5	35
37	35	Poland	0.88	78.73	16.301171	12.47232	31622.55302	8	34
38	36	Andorra	0.868	81.91	13.300239	10.50176	56000.30336	-20	36
39	37	Latvia	0.866	75.29	16.16769	13.03085	30282.39353	8	37
40	38	Portugal	0.864	82.05	16.53579	9.26347	33966.82756	2	38
41	39	Slovakia	0.86	77.54	14.48704	12.693157	32113.04569	3	39
42	40	Hungary	0.854	76.88	15.20418	11.96329	31328.80599	4	42
43	40	Saudi Arabia	0.854	75.13	16.13575	10.22517	47495.42697	-16	40
44	42	Bahrain	0.852	77.29	16.25482	9.5192	42521.70451	-12	41
45	43	Chile	0.851	80.18	16.44755	10.58295	23261.29726	16	43
46	43	Croatia	0.851	78.49	15.22672	11.44932	28069.84837	6	44
47	45	Qatar	0.848	80.23	12.03595	9.73144	92418.23036	-43	45
48	46	Argentina	0.845	76.67	17.6542	10.940601	21190.17661	16	46
49	47	Brunei Darussalam	0.838	75.86	14.31391	9.14	63965.09279	-38	47

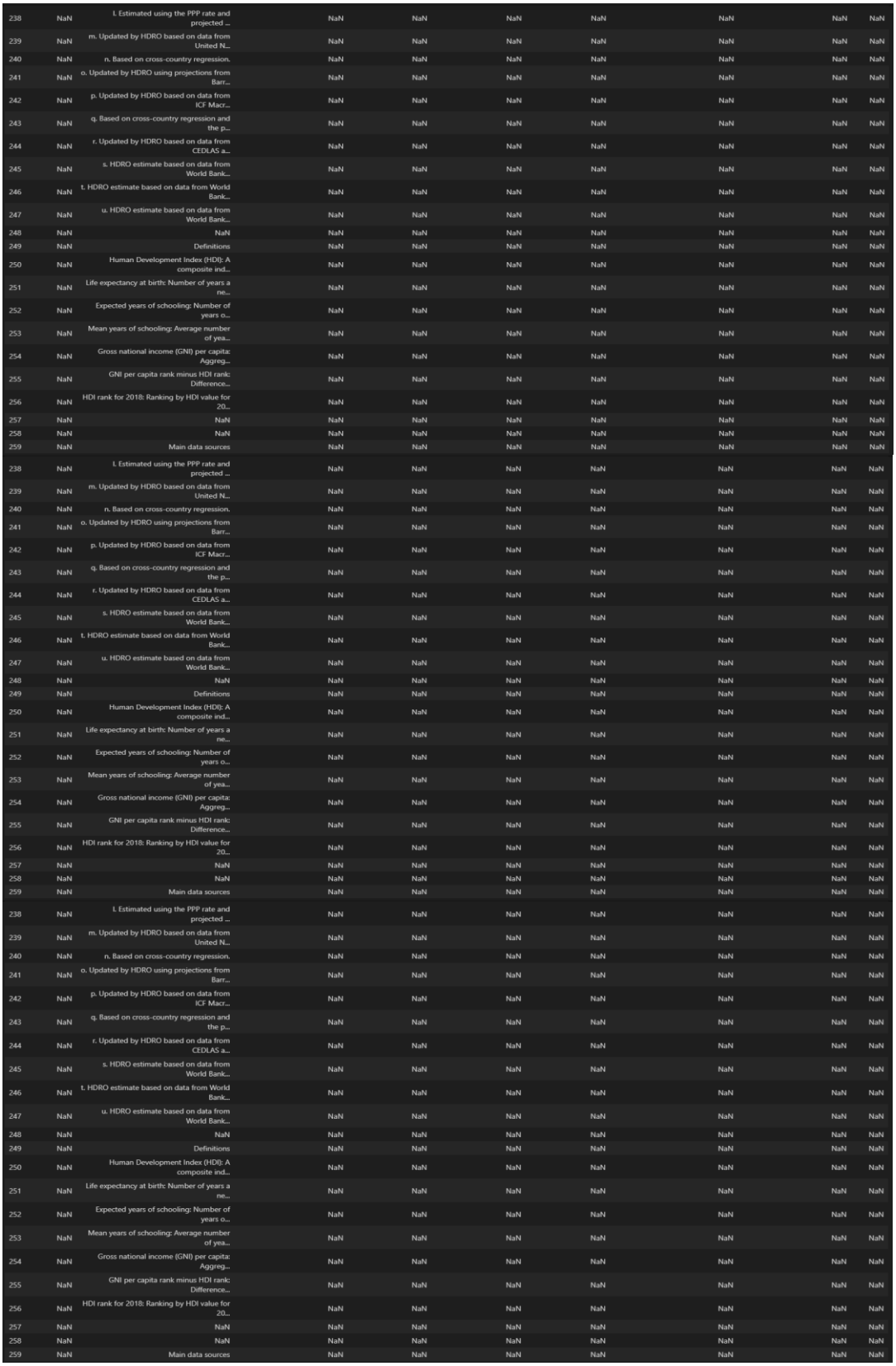
***Fig. 5**. Captura de la ejecución de los comandos de la variable que contiene los datos.*
(Fuente: Propia)

De la misma forma se imprime los datos de las ultimas filas para verificar si los datos a utilizar son de utilidad para ello recursos el comando del módulo de panda Tail.

```
#IMPRIMIMOS LOS ULTIMOS 50 DATOS PARA VERIFICAR SI SON UTILES O NO
   LAS SECCIONES CON DATOS NULOS

   df_Original.tail(50)

   #Existe una gran cantidad de datos adicionales provenientes del
    archivo original que no sirve
```



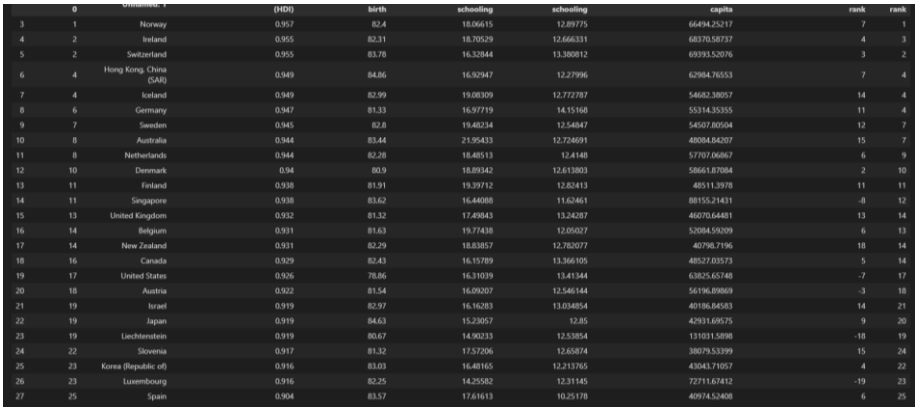
238	NaN	l. Estimated using the PPP rate and projected ...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
239	NaN	m. Updated by HDRO based on data from United N...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
240	NaN	n. Based on cross-country regression.	NaN	NaN	NaN	NaN	NaN	NaN	NaN
241	NaN	o. Updated by HDRO using projections from Barr...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
242	NaN	p. Updated by HDRO based on data from ICF Macr...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
243	NaN	q. Based on cross-country regression and the p...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
244	NaN	r. Updated by HDRO based on data from CEDLAS a...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
245	NaN	s. HDRO estimate based on data from World Bank...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
246	NaN	t. HDRO estimate based on data from World Bank...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
247	NaN	u. HDRO estimate based on data from World Bank...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
248	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
249	NaN	Definitions	NaN	NaN	NaN	NaN	NaN	NaN	NaN
250	NaN	Human Development Index (HDI): A composite ind...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
251	NaN	Life expectancy at birth: Number of years a ne...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
252	NaN	Expected years of schooling: Number of years o...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
253	NaN	Mean years of schooling: Average number of yea...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
254	NaN	Gross national income (GNI) per capita: Aggreg...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
255	NaN	GNI per capita rank minus HDI rank: Difference...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
256	NaN	HDI rank for 2018: Ranking by HDI value for 20...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
257	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
258	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
259	NaN	Main data sources	NaN	NaN	NaN	NaN	NaN	NaN	NaN
238	NaN	l. Estimated using the PPP rate and projected ...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
239	NaN	m. Updated by HDRO based on data from United N...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
240	NaN	n. Based on cross-country regression.	NaN	NaN	NaN	NaN	NaN	NaN	NaN
241	NaN	o. Updated by HDRO using projections from Barr...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
242	NaN	p. Updated by HDRO based on data from ICF Macr...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
243	NaN	q. Based on cross-country regression and the p...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
244	NaN	r. Updated by HDRO based on data from CEDLAS a...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
245	NaN	s. HDRO estimate based on data from World Bank...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
246	NaN	t. HDRO estimate based on data from World Bank...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
247	NaN	u. HDRO estimate based on data from World Bank...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
248	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
249	NaN	Definitions	NaN	NaN	NaN	NaN	NaN	NaN	NaN
250	NaN	Human Development Index (HDI): A composite ind...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
251	NaN	Life expectancy at birth: Number of years a ne...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
252	NaN	Expected years of schooling: Number of years o...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
253	NaN	Mean years of schooling: Average number of yea...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
254	NaN	Gross national income (GNI) per capita: Aggreg...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
255	NaN	GNI per capita rank minus HDI rank: Difference...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
256	NaN	HDI rank for 2018: Ranking by HDI value for 20...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
257	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
258	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
259	NaN	Main data sources	NaN	NaN	NaN	NaN	NaN	NaN	NaN
238	NaN	l. Estimated using the PPP rate and projected ...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
239	NaN	m. Updated by HDRO based on data from United N...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
240	NaN	n. Based on cross-country regression.	NaN	NaN	NaN	NaN	NaN	NaN	NaN
241	NaN	o. Updated by HDRO using projections from Barr...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
242	NaN	p. Updated by HDRO based on data from ICF Macr...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
243	NaN	q. Based on cross-country regression and the p...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
244	NaN	r. Updated by HDRO based on data from CEDLAS a...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
245	NaN	s. HDRO estimate based on data from World Bank...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
246	NaN	t. HDRO estimate based on data from World Bank...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
247	NaN	u. HDRO estimate based on data from World Bank...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
248	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
249	NaN	Definitions	NaN	NaN	NaN	NaN	NaN	NaN	NaN
250	NaN	Human Development Index (HDI): A composite ind...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
251	NaN	Life expectancy at birth: Number of years a ne...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
252	NaN	Expected years of schooling: Number of years o...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
253	NaN	Mean years of schooling: Average number of yea...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
254	NaN	Gross national income (GNI) per capita: Aggreg...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
255	NaN	GNI per capita rank minus HDI rank: Difference...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
256	NaN	HDI rank for 2018: Ranking by HDI value for 20...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
257	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
258	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
259	NaN	Main data sources	NaN	NaN	NaN	NaN	NaN	NaN	NaN

***Fig. 6**. Captura de la ejecución de los comandos de la variable que contiene los datos.*
(Fuente: Propia)

3.3.Limpieza

Para este procedimiento se empleara la librería de panda y se invocara la función .dropna el cual eliminara estos espacios , y dicho resultado será enviado una nueva variable pero sin las filas que no contienen datos.

```
#CREACION DE UN DATAFRAME EXTRA PARA NO ALTERAR EL ORIGINAL,
EL CUAL CONTIENE LOS DATOS YA FILTRANDO LOS 'NaN'
df = df_Original.dropna()
df.head(25)
```



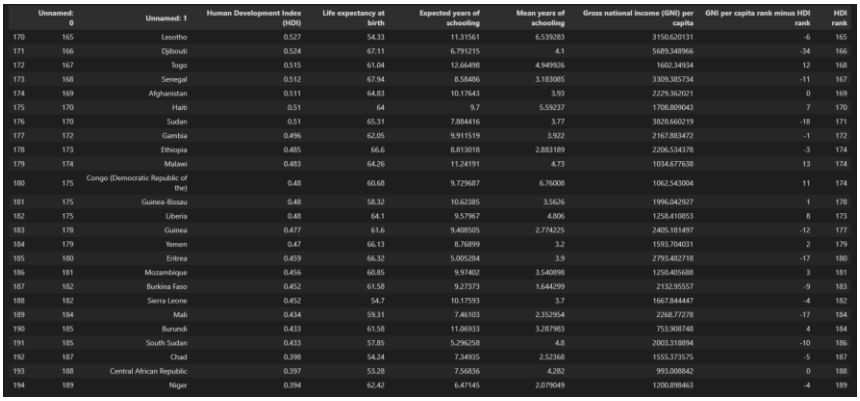
	0	Unnamed: 1	(HDI)	birth	schooling	schooling	capita	rank	rank
3	1	Norway	0.957	82.4	18.06615	12.89775	66494.25217	7	1
4	2	Ireland	0.955	82.31	18.70529	12.666331	68370.58737	4	3
5	2	Switzerland	0.955	83.78	16.32844	13.380812	69393.52076	3	2
6	4	Hong Kong, China (SAR)	0.949	84.86	16.92947	12.27996	62984.76553	7	4
7	4	Iceland	0.949	82.99	19.08309	12.772787	54682.38057	14	4
8	6	Germany	0.947	81.33	16.97719	14.15168	55314.35355	11	4
9	7	Sweden	0.945	82.8	19.48234	12.54847	54507.80504	12	7
10	8	Australia	0.944	83.44	21.95433	12.724691	48084.84207	15	7
11	8	Netherlands	0.944	82.28	18.48513	12.4148	57707.06867	6	9
12	10	Denmark	0.94	80.9	18.89342	12.613803	58661.87084	2	10
13	11	Finland	0.938	81.91	19.39712	12.82413	48511.3978	11	11
14	11	Singapore	0.938	83.62	16.44088	11.62461	88155.21431	-8	12
15	13	United Kingdom	0.932	81.32	17.49843	13.24287	46070.64481	13	14
16	14	Belgium	0.931	81.63	19.77438	12.05027	52084.59209	6	13
17	14	New Zealand	0.931	82.29	18.83857	12.782077	40798.7196	18	14
18	16	Canada	0.929	82.43	16.15789	13.366105	48527.03573	5	14
19	17	United States	0.926	78.86	16.31039	13.41344	63825.65748	-7	17
20	18	Austria	0.922	81.54	16.09207	12.546144	56196.89869	-3	18
21	19	Israel	0.919	82.97	16.16283	13.034854	40186.84583	14	21
22	19	Japan	0.919	84.63	15.23057	12.85	42931.69575	9	20
23	19	Liechtenstein	0.919	80.67	14.90233	12.53854	131031.5898	-18	19
24	22	Slovenia	0.917	81.32	17.57206	12.65874	38079.53399	15	24
25	23	Korea (Republic of)	0.916	83.03	16.48165	12.213765	43043.71057	4	22
26	23	Luxembourg	0.916	82.25	14.25582	12.31145	72711.67412	-19	23
27	25	Spain	0.904	83.57	17.61613	10.25178	40974.52408	6	25

***Fig. 7**. Captura de la ejecución de los comandos .dropna() de la variable que contiene los datos.*
(Fuente: Propia)

También se comprará los valores de las filas finales empleando los comandos anteriores

```
#COMPROBAMOS LA INTEGRIDAD DE LA DATA EN SUS ULTIMAS SECCIONES
df = df_Original.dropna()
df.tail(25)
```

‘



	Unnamed: 0	Unnamed: 1	Human Development Index (HDI)	Life expectancy at birth	Expected years of schooling	Mean years of schooling	Gross national income (GNI) per capita	GNI per capita rank minus HDI rank	HDI rank
170	165	Lesotho	0.527	54.33	11.31561	6.539283	3150.620131	-6	165
171	166	Djibouti	0.524	67.11	6.791215	4.1	5689.348966	-34	166
172	167	Togo	0.515	61.04	12.66498	4.949926	1602.34934	12	168
173	168	Senegal	0.512	67.94	8.58486	3.183085	3309.385734	-11	167
174	169	Afghanistan	0.511	64.83	10.17643	3.93	2229.362021	0	169
175	170	Haiti	0.51	64	9.7	5.59237	1708.809043	7	170
176	170	Sudan	0.51	65.31	7.884416	3.77	3828.660219	-18	171
177	172	Gambia	0.496	62.05	9.911519	3.922	2167.883472	-1	172
178	173	Ethiopia	0.485	66.6	8.813018	2.883189	2206.534378	-3	174
179	174	Malawi	0.483	64.26	11.24191	4.73	1034.677638	13	174
180	175	Congo (Democratic Republic of the)	0.48	60.68	9.729687	6.76008	1062.543004	11	174
181	175	Guinea-Bissau	0.48	58.32	10.62385	3.5626	1996.042927	1	178
182	175	Liberia	0.48	64.1	9.57967	4.806	1258.410853	8	173
183	178	Guinea	0.477	61.6	9.408505	2.774225	2405.181497	-12	177
184	179	Yemen	0.47	66.13	8.76899	3.2	1593.704031	2	179
185	180	Eritrea	0.459	66.32	5.005284	3.9	2793.482718	-17	180
186	181	Mozambique	0.456	60.85	9.97402	3.540898	1250.405688	3	181
187	182	Burkina Faso	0.452	61.58	9.27373	1.644299	2132.95557	-9	183
188	182	Sierra Leone	0.452	54.7	10.17593	3.7	1667.844447	-4	182
189	184	Mali	0.434	59.31	7.46103	2.352954	2268.77278	-17	184
190	185	Burundi	0.433	61.58	11.06933	3.287983	753.908748	4	184
191	185	South Sudan	0.433	57.85	5.296258	4.8	2003.318894	-10	186
192	187	Chad	0.398	54.24	7.34935	2.52368	1555.373575	-5	187
193	188	Central African Republic	0.397	53.28	7.56836	4.282	993.008842	0	188
194	189	Niger	0.394	62.42	6.47145	2.079049	1200.898463	-4	189

***Fig. 8**. Captura de la ejecución de los comandos .dropna() de la variable que contiene los datos.*
(Fuente: Propia)

Y por último se comprobara de forma general mediante la generación de un informe

```
#COMPROBAMOS QUE YA NO EXISTEN VALORES NaN EN EL DATAFRAME

df.isna().sum().sort_values()
```

```
Unnamed: 0                                0
Unnamed: 1                                0
Human Development Index (HDI)             0
Life expectancy at birth                  0
Expected years of schooling               0
Mean years of schooling                   0
Gross national income (GNI) per capita    0
GNI per capita rank minus HDI rank        0
HDI rank                                  0
dtype: int64
```

***Fig. 9**. Captura de la ejecución de .sort_values() de la variable que contiene los datos.*
(Fuente: Propia)

Una vez verificado que toda las filas se realizar el mismo proceso a las columnas pero ante de eso se visualizar los datos dentro de una grafica de mapa de calor

```
#VERIFICAMOS LOS DATOS COMPLETADOS

print()
sns.heatmap(df.isna())

#Aun podemos ver que existen 2 columnas sin nombre del cual una es de los nombres de los
#paises y otra que representa al ranking de los países según vemos inferimos de los datos
```

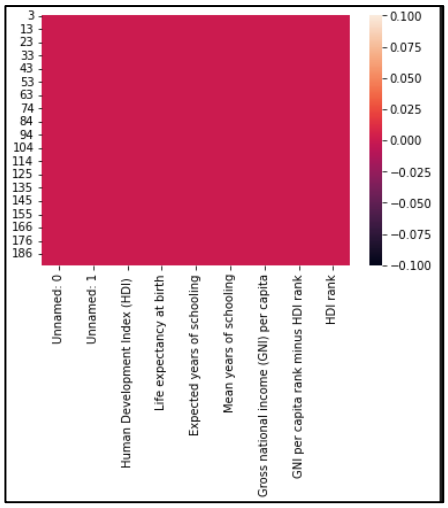


***Fig. 10**. Grafica de un mapa de calor de los valores de las columnas con datos*
(Fuente: Propia)

Finalmente se cambiará los valores de las columnas con los valores correspondiente a cada campo aplicando el siguiente comando.

```
#COLOCAREMOS LOS NOMBRES A LAS COLUMNAS QUE NO LO TIENEN Y LE
AGREGAMOS EL AÑO (DEL DATAFRAME ORIGINAL) A LAS COLUMNAS
 NECESARIAS

df = df.rename(columns={'Unnamed: 0':'RANK (2019)','Unnamed:
      1':'CONTRY','HDI rank':'HDI rank (2018)'})
```

Y finalmente verificamos los resultados

```
#COMPROBAMOS QUE EL CAMBIO EN SU INDEX SE EFECTUO
df
```

✓ 0.1s — Python

	RANK (2019)	CONTRY	Human Development Index (HDI)	Life expectancy at birth	Expected years of schooling	Mean years of schooling	Gross national income (GNI) per capita	GNI per capita rank minus HDI rank	HDI rank (2018)
0	1	Norway	0.957	82.4	18.06615	12.89775	66494.25217	7	1
1	2	Ireland	0.955	82.31	18.70529	12.666331	68370.58737	4	3
2	2	Switzerland	0.955	83.78	16.32844	13.380812	69393.52076	3	2
3	4	Hong Kong, China (SAR)	0.949	84.86	16.92947	12.27996	62984.76553	7	4
4	4	Iceland	0.949	82.99	19.08309	12.772787	54682.38057	14	4
...	...	...	...	...	...	...	...	...	...
184	185	Burundi	0.433	61.58	11.06933	3.287983	753.908748	4	184
185	185	South Sudan	0.433	57.85	5.296258	4.8	2003.318894	-10	186
186	187	Chad	0.398	54.24	7.34935	2.52368	1555.373575	-5	187
187	188	Central African Republic	0.397	53.28	7.56836	4.282	993.008842	0	188
188	189	Niger	0.394	62.42	6.47145	2.079049	1200.898463	-4	189

189 rows × 9 columns

***Fig. 11**. Captura de la ejecución del dataframe que contiene los datos.*
(Fuente: Propia)

Y para finalizar el proceso de limpieza y ajuste de la tabla de datos se actulizara los índices y se comprobara

```
#REAPRACION DEL INDEX YA QUE FALTAN LOS QUE SE ELIMINARON
df.reset_index(drop=True, inplace=True)
#COMPROBAMOS QUE EL CAMBIO EN SU INDEX SE EFECTUO
df
```

#COMPROBAMOS QUE EL CAMBIO EN SU INDEX SE EFECTUO
df

✓ 0.1s — Python

	RANK (2019)	CONTRY	Human Development Index (HDI)	Life expectancy at birth	Expected years of schooling	Mean years of schooling	Gross national income (GNI) per capita	GNI per capita rank minus HDI rank	HDI rank (2018)
0	1	Norway	0.957	82.4	18.06615	12.89775	66494.25217	7	1
1	2	Ireland	0.955	82.31	18.70529	12.666331	68370.58737	4	3
2	2	Switzerland	0.955	83.78	16.32844	13.380812	69393.52076	3	2
3	4	Hong Kong, China (SAR)	0.949	84.86	16.92947	12.27996	62984.76553	7	4
4	4	Iceland	0.949	82.99	19.08309	12.772787	54682.38057	14	4
...	...	...	...	...	...	...	...	...	...
184	185	Burundi	0.433	61.58	11.06933	3.287983	753.908748	4	184
185	185	South Sudan	0.433	57.85	5.296258	4.8	2003.318894	-10	186
186	187	Chad	0.398	54.24	7.34935	2.52368	1555.373575	-5	187
187	188	Central African Republic	0.397	53.28	7.56836	4.282	993.008842	0	188
188	189	Niger	0.394	62.42	6.47145	2.079049	1200.898463	-4	189

189 rows × 9 columns

***Fig. 12**. Captura de la ejecución del dataframe que contiene los datos.*
(Fuente: Propia)

3.4. Visualización de los datos

Para este pardo se empleara los datos de top ten delos países que ms desarrollados entre los anos del 2019 y 2018.

```
#APARTIR DEL DATAFRAME YA COMPLETO CREAMOS OTROS PARA EL ANALISIS DEL IDH
 EN LOS 10 RANKIS SUPERIORES E INFERIORES en el 2019 y 2018

df_Superior = df.head(10).drop(['Human Development Index (HDI) ', 'Life
expectancy at birth',   'Expected years of schooling',  'Mean years of
schooling',  'Gross national income (GNI) per capita',   'GNI per capita
rank minus HDI rank'], axis=1)
df_Superior = df_Superior.set_index('CONTRY')
df_Superior

df_Inferior = df.tail(10).drop(['Human Development Index (HDI) ', 'Life
expectancy at birth',   'Expected years of schooling',  'Mean years of
schooling',  'Gross national income (GNI) per capita',   'GNI per capita
rank minus HDI rank'], axis=1)
df_Inferior = df_Inferior.set_index('CONTRY')
df_Inferior
```

```
#USAREMOS UN GRAFICO DE MOSAICO PARA REALIZAR LA COMPARACION DEL RANKIN
 DE LOS 10 PRIEMROS PAISES EN 2019 Y 2018
mosaic(df_Superior, ['RANK (2019)', 'HDI rank (2018)'])
```

Graficas en Mosaico

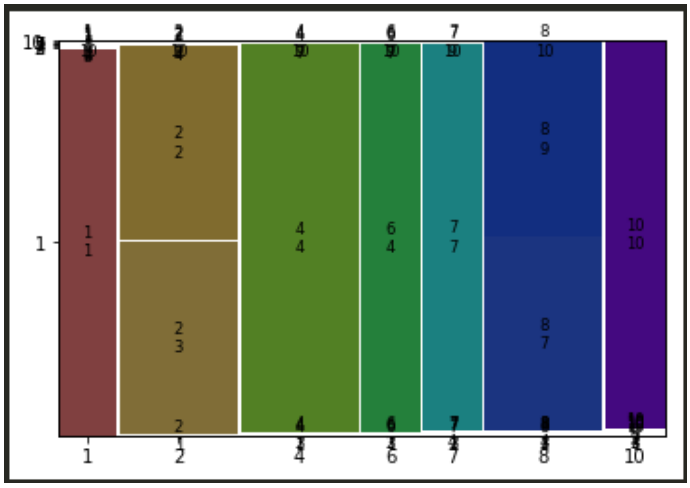


***Fig. 13**. Grafica tipo mosaico del top 10 del rankig del IDH*
(Fuente: Propia)

En este gráfico de mosaico permite visualizar los datos dentro de un diagrama de barras apiladas que muestra los porcentajes de datos en distintos grupos en este caso seria los valores de los países que ocupan el top 10 de los datos de año 2018-2019 está grafica nos permite compara entre distintos grupos.

Graficas en Barras

```
#CON LA AYUDA DE UNA GRAFICO DOBLE BARRA HORIZONTAL DENOTAREMOS LA
VARIACION DEL RANKING DE LOS 10 PRIMEROS PAISES

df_Superior.plot(kind = 'barh')
```

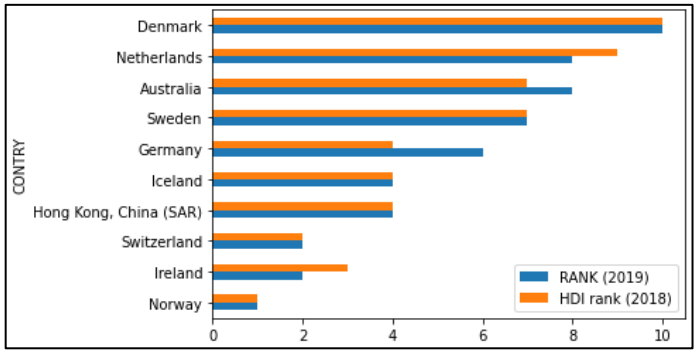


***Fig. 14**. Grafica tipo barra del top 10 del rankig de países y el IDH*
(Fuente: Propia)

Gráfico en barras el cual indica el incremento o deceso del índice de desarrollo humano de cada país con respeto a los anos 2018 y 2019 donde la mayoría de países se han mantenido pero los existe casos como el de Alemania ha incrementado pero también existe caso donde el IDH de países como Australia, Noruega e Irlanda han decrecido.

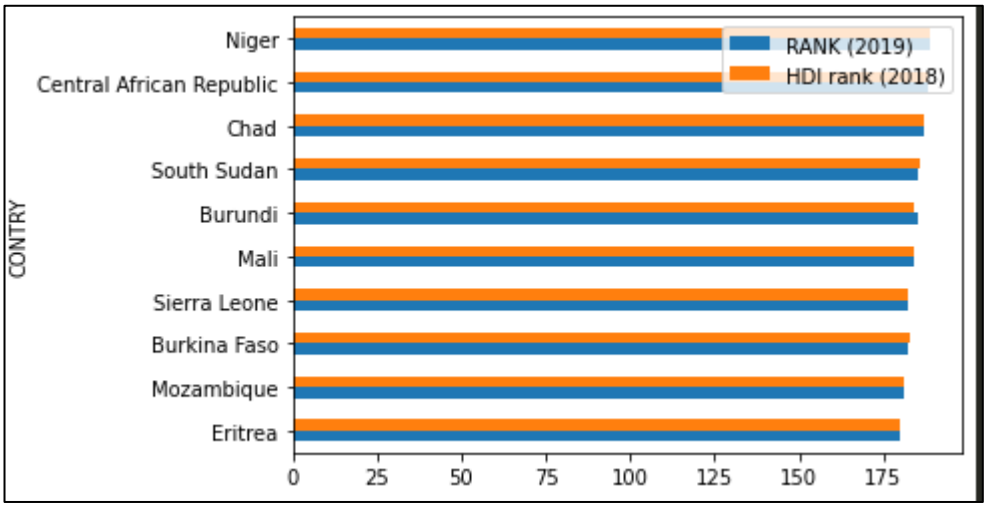


***Fig. 15**. Grafica tipo barra de los últimos 10 paises del rankig del IDH*
(Fuente: Propia)

```
df_Superior.plot.area(alpha=0.4);
```

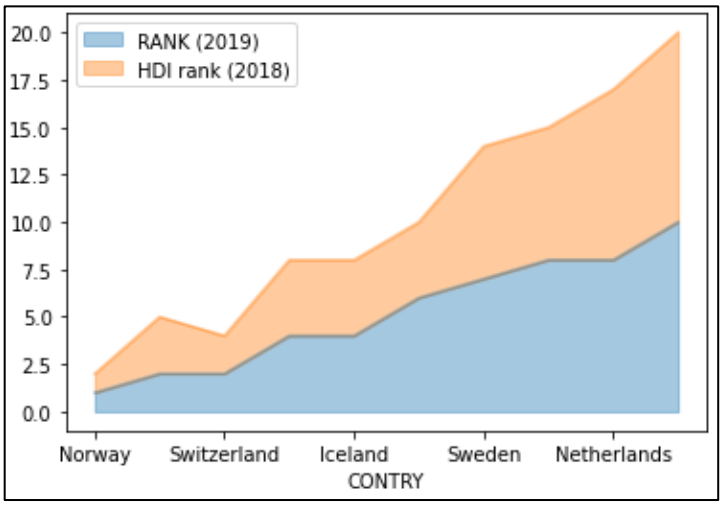


***Fig. 16**. Grafica de área de los últimos 10 países del rankig del IDH*
(Fuente: Propia)

Para este ultimo grafico se aprecia la diferencia entre el año 2018 y el 2019. Siendo este ultimo el cual indica que el crecimiento del año 2018 fue mayor que del año 2019.

4. Conclusiones

El uso de herramientas de informática ayuda a la agilización del manejo, análisis y visualización de datos, los cuales por su gran cantidad serían muy extensos de analizarlos de manera manual.

El procesamiento de datos de archivos digitales a través de software permite la recopilación de datos de manera automática pero dicha información puede venir con datos vacíos. De ahí la importancia de realizar los procesos de limpieza de datos y verificación de la integridad de estos.

Las variaciones de IDH en los países seleccionados a través de un año son pocas, muchos países conservan su ranking, las variaciones más marcadas se encuentran en los últimos 10 países del top

Bibliografía

[1] Z. Yanchun, «Human Development Reports,» Table Human Development Index and its components, 15 07 2022. [En línea]. Available: https://hdr.undp.org/sites/default/files/2021-22_HDR/HDR21-22_Statistical_Annex_HDI_Table.xlsx. [Último acceso: 10 09 2022].

[2] Human Development Reports, «Human development index (HDI),» Human Development Reports, 10 07 2022. [En línea]. Available: https://hdr.undp.org/. [Último acceso: 11 09 2022].

[3] Python.org, «What is python?,» Python.org, [En línea]. Available: https://www.python.org/doc/essays/blurb/. [Último acceso: 11 09 2022].

[4] IBM, «Creación de scripts en lenguaje de programación Python,» IBM - Deutschland | IBM, [En línea]. Available: https://www.ibm.com/docs/es/spss-statistics/25.0.0?topic=SSLVMB_25.0.0/spss/base/scripts_python.html. [Último acceso: 11 9 2022].

[5] Jupyter Team, «Architecture - jupyter documentation,» Jupyter Project Documentation, 30 9 2019. [En línea]. Available: https://docs.jupyter.org/en/latest/projects/architecture/content-architecture.html. [Último acceso: 11 09 2022].

[6] Pandas, «Pandas - python data analysis library,» pandas - Python Data Analysis Library, 31 8 2022. [En línea]. Available: https://pandas.pydata.org/. [Último acceso: 11 09 2022].

[7] Aprende con Alf, «La librería Numpy,» Aprende con Alf, 12 04 2022. [En línea]. Available: https://aprendeconalf.es/docencia/python/manual/numpy/. [Último acceso: 11 9 2022].

[8] J. Hunter, D. Dale, E. Firing y M. Droettboom, «Matplotlib — Visualization with Python,» Matplotlib, 2012. [En línea]. Available: https://matplotlib.org/stable/devel/documenting_mpl.html. [Último acceso: 11 09 2022].

[9] M. Waskom, «Seaborn: Statistical data visualization,» Seabor, 2012. [En línea]. Available: https://seaborn.pydata.org/. [Último acceso: 09 11 2022].

[10] J. Perktold, S. Seabold y J. Taylor, «Introduction Statsmodels,» Statsmodels-developers, 27 8 2022. [En línea]. Available: https://www.statsmodels.org/dev/index.html. [Último acceso: 11 09 2022].

[11] PyPI, «PyMosaic,» PyPI, [En línea]. Available: https://pypi.org/project/pyMosaic/. [Último acceso: 11 09 2022].

[12] Aprende con Alf, «La librería Pandas,» Aprende con Alf, 12 4 2022. [En línea]. Available: https://aprendeconalf.es/docencia/python/manual/pandas/. [Último acceso: 11 09 202].