AVANCE PRACTICA 1 EN EQUIPO

Objetivo

En esta actividad nos encargaremos de limpiar los datos de nuestra Base de Datos, que como ya lo hemos comentado anteriormente es de una BD que tiene informacion de Suicidios y contamos con algunas columnas las cuales las mas importantes para nosotros son la generacion y el lugar donde han ocurrido estos mismos, ya que nuestro objetivo principal es agruparlos y ver la clasificacion que se tiene en cuestion de generaciones para que de esta manera nos ayude a nosotros a poder saber cual de estas Generaciones ha sido más propensa a esta triste situacion y poder tomar accion, creando conciencia a futuras generaciones y hacer todo lo que este en nuestras manos para tomarlo como "foco rojo" a esta misma.

Integrantes:

Mata Martínez Missael 1672902

Celestino Tovar Angel Gabriel 1668653

López Sánchez Maribel 1672709

Limpieza de datos

El primer paso sera importar librerias y utilizamos la libreria pandas para poder cargar archivos csv.

In [1]:

```
import matplotlib as plt
import seaborn as sn
import numpy as np
import pandas as pd
import json
df = pd.read_csv("master.csv")
```

Verificamos las dimensiones de la tabla

In [2]:

```
df.shape
```

Out[2]:

```
(27820, 12)
```

Ahora visualizaremos la tabla con solamente 12 de sus filas de las 27820 que tiene

In [4]:

```
df.head(12)
```

Out[4]:

	country	year	sex	age	suicides_no	population	suicides/100k pop	country-year	HDI for year	gdp_for_year ($)	gdp_per_capita ($)	generation
0	Albania	1987	male	15-24 years	21	312900	6.71	Albania1987	NaN	2,156,624,900	796	Generation X
1	Albania	1987	male	35-54 years	16	308000	5.19	Albania1987	NaN	2,156,624,900	796	Silent

	country	year	sex	age	suicides_no	population	suicides/100k pop	country-year	HDI for year	gdp_for_year ($)	gdp_per_capita ($)	generation
2	Albania	1987	female	15-24 years	14	289700	4.83	Albania1987	NaN	2,156,624,900	796	Generation X
3	Albania	1987	male	75+ years	1	21800	4.59	Albania1987	NaN	2,156,624,900	796	G.I. Generation
4	Albania	1987	male	25-34 years	9	274300	3.28	Albania1987	NaN	2,156,624,900	796	Boomers
5	Albania	1987	female	75+ years	1	35600	2.81	Albania1987	NaN	2,156,624,900	796	G.I. Generation
6	Albania	1987	female	35-54 years	6	278800	2.15	Albania1987	NaN	2,156,624,900	796	Silent
7	Albania	1987	female	25-34 years	4	257200	1.56	Albania1987	NaN	2,156,624,900	796	Boomers
8	Albania	1987	male	55-74 years	1	137500	0.73	Albania1987	NaN	2,156,624,900	796	G.I. Generation
9	Albania	1987	female	5-14 years	0	311000	0.00	Albania1987	NaN	2,156,624,900	796	Generation X
10	Albania	1987	female	55-74 years	0	144600	0.00	Albania1987	NaN	2,156,624,900	796	G.I. Generation
11	Albania	1987	male	5-14 years	0	338200	0.00	Albania1987	NaN	2,156,624,900	796	Generation X

A continuacion mostraremos los nombres y tipos de datos que tiene nuestra BD

In [5]:

```
df.dtypes
```

Out[5]:

```
country                object
year                    int64
sex                    object
age                    object
suicides_no             int64
population              int64
suicides/100k pop     float64
country-year           object
HDI for year          float64
 gdp_for_year ($)      object
gdp_per_capita ($)      int64
generation             object
dtype: object
```

De estas columnas solamente necesitaremos country, sex, suicides_no y generation. Ahora cambiaremos este nombre de las columnas al español

In [13]:

```
df = df.rename(columns = {'country':'Pais', 'sex':'Sexo', 'suicides_no':'NumeroDeSuicidios', 'gener
ation':'Generacion'})
df.columns
```

Out[13]:

```
Index(['Pais', 'year', 'Sexo', 'edad', 'NumeroDeSuicidios', 'population',
       'suicides/100k pop', 'country-year', 'HDI for year',
       ' gdp_for_year ($) ', 'gdp_per_capita ($)', 'Generacion'],
      dtype='object')
```

Ahora, eliminaremos las filas que contengan datos nulos para quitar lo que no necesitamos. Nosotros lo hemos echo asi porque sabemos que las columnas de interes (PAIS, SEXO, NUMERO DE SUICIDIOS Y GENERACION) no se encuentran con ninguna celda vacia.

In [15]:

```
df = df.dropna()
```

Ahora mostramos lo que hicimos

In [16]:

```
df
```

Out[16]:

	Pais	year	Sexo	edad	NumeroDeSuicidios	population	suicides/100k pop	country-year	HDI for year	gdp_for_year ($)	gdp_per_c
72	Albania	1995	male	25-34 years	13	232900	5.58	Albania1995	0.619	2,424,499,009	
73	Albania	1995	male	55-74 years	9	178000	5.06	Albania1995	0.619	2,424,499,009	
74	Albania	1995	female	75+ years	2	40800	4.90	Albania1995	0.619	2,424,499,009	
75	Albania	1995	female	15-24 years	13	283500	4.59	Albania1995	0.619	2,424,499,009	
76	Albania	1995	male	15-24 years	11	241200	4.56	Albania1995	0.619	2,424,499,009	
...	...	...	...	...	...	...	...	...	...	...	
27815	Uzbekistan	2014	female	35-54 years	107	3620833	2.96	Uzbekistan2014	0.675	63,067,077,179	
27816	Uzbekistan	2014	female	75+ years	9	348465	2.58	Uzbekistan2014	0.675	63,067,077,179	
27817	Uzbekistan	2014	male	5-14 years	60	2762158	2.17	Uzbekistan2014	0.675	63,067,077,179	
27818	Uzbekistan	2014	female	5-14 years	44	2631600	1.67	Uzbekistan2014	0.675	63,067,077,179	
27819	Uzbekistan	2014	female	55-74 years	21	1438935	1.46	Uzbekistan2014	0.675	63,067,077,179	

8364 rows × 12 columns

Ahora eliminaremos todas las columnas que tengan datos nulos. Igual que en lo anterior, ya confirmamos que en nuestras columnas de interes no habia datos nulos.

In [17]:

```
df
```

Out[17]:

	Pais	year	Sexo	edad	NumeroDeSuicidios	population	suicides/100k pop	country-year	HDI for year	gdp_for_year ($)	gdp_per_c
72	Albania	1995	male	25-34 years	13	232900	5.58	Albania1995	0.619	2,424,499,009	
73	Albania	1995	male	55-74 years	9	178000	5.06	Albania1995	0.619	2,424,499,009	
74	Albania	1995	female	75+ years	2	40800	4.90	Albania1995	0.619	2,424,499,009	
75	Albania	1995	female	15-24	13	283500	4.59	Albania1995	0.619	2,424,499,009	

	Pais	year	Sexo	edad	NumeroDeSuicidios	population	suicides/100k pop	country-year	HDI for year	gdp_for_year ($)	gdp_per_c
76	Albania	1995	male	25-34 years	11	241200	4.56	Albania1995	0.619	2,424,499,009	
...	...	...	...	...	...	...	...	...	...	...	
27815	Uzbekistan	2014	female	35-54 years	107	3620833	2.96	Uzbekistan2014	0.675	63,067,077,179	
27816	Uzbekistan	2014	female	75+ years	9	348465	2.58	Uzbekistan2014	0.675	63,067,077,179	
27817	Uzbekistan	2014	male	5-14 years	60	2762158	2.17	Uzbekistan2014	0.675	63,067,077,179	
27818	Uzbekistan	2014	female	5-14 years	44	2631600	1.67	Uzbekistan2014	0.675	63,067,077,179	
27819	Uzbekistan	2014	female	55-74 years	21	1438935	1.46	Uzbekistan2014	0.675	63,067,077,179	

8364 rows × 12 columns

A continuacion eliminaremos las columnas que no necesitamos. Para solo quedarnos con Pais, Sexo, NumeroDeSuicidios y Generacion

In [18]:

```
df.columns
```

Out[18]:

```
Index(['Pais', 'year', 'Sexo', 'edad', 'NumeroDeSuicidios', 'population',
       'suicides/100k pop', 'country-year', 'HDI for year',
       ' gdp_for_year ($) ', 'gdp_per_capita ($)', 'Generacion'],
      dtype='object')
```

In [20]:

```
df = df.drop(['year', 'edad', 'population', 'suicides/100k pop', 'country-year', 'HDI for year',
       ' gdp_for_year ($) ', 'gdp_per_capita ($)'], axis=1)
```

Ahora pasamos a mostrar la tabla sin esas columnas.

In [21]:

```
df
```

Out[21]:

	Pais	Sexo	NumeroDeSuicidios	Generacion
72	Albania	male	13	Generation X
73	Albania	male	9	Silent
74	Albania	female	2	G.I. Generation
75	Albania	female	13	Generation X
76	Albania	male	11	Generation X
...	...	...	...	...
27815	Uzbekistan	female	107	Generation X
27816	Uzbekistan	female	9	Silent
27817	Uzbekistan	male	60	Generation Z
27818	Uzbekistan	female	44	Generation Z
27819	Uzbekistan	female	21	Boomers

8364 rows × 4 columns

Ahora quisiéramos ver que generaciones son las que más números de suicidios tienen

In [22]:

```
df.groupby('Generacion').sum()
```

Out[22]:

Generacion	NumeroDeSuicidios
Boomers	435081
G.I. Generation	129523
Generation X	529371
Generation Z	7991
Millenials	242303
Silent	379755

Preguntas de interes

Cuantos grupos de hombres y mujeres mayores de 30 años se han suicidado?

Tomando un registro de la BD, se puede saber como afecto el suicidio de la persona?

Conclusion

Si bien, lo anterior aplicado fue la manera en la cual se debe a empezar a desarrollar mas el uso cognitivo sobre la investigacion de nuestra BD, tomando en cuenta la eliminacion de columnas que no utilicemos ya se por funciones desconocidas o simplemente datos que no son relevantes.

Las preguntas fueran replanteadas mas a fondo para dar a entender con mayor facilidad el tema que se esta tratando con la informacion. Al final, fue la manera mas correcta de formularlas.

In [ ]:

```

```