

Diplomado en Big Data

Gestión de Datos Masivos

LABORATORIO 2 ANACONDA

Docente Evaluador: Ing. Alexander Ramírez Camargo

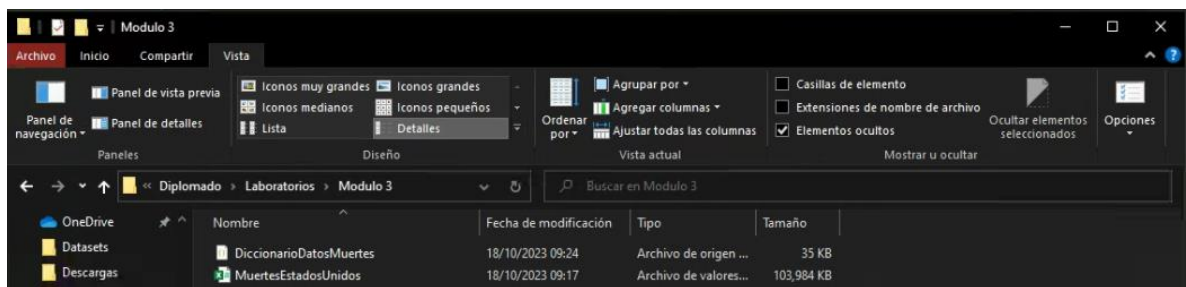
Presenta:

Ing. Jorge Alberto Gómez López

Ing. Guillermo Alexander Cornejo Argueta

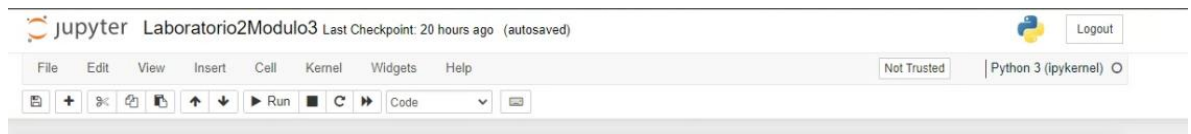
1. Seleccionar un conjunto de datos que tenga por lo menos 6 columnas de datos.

Los datos seleccionados que se utilizarán en el presente laboratorio son los que serán utilizados en el proyecto del diplomado y son tomados del sitio web Kaggle que muestra las causas de muerte de los estados unidos de América, datos que son recolectados por la CDC de dicho país, contando este dataset con más de 1millon de registros.

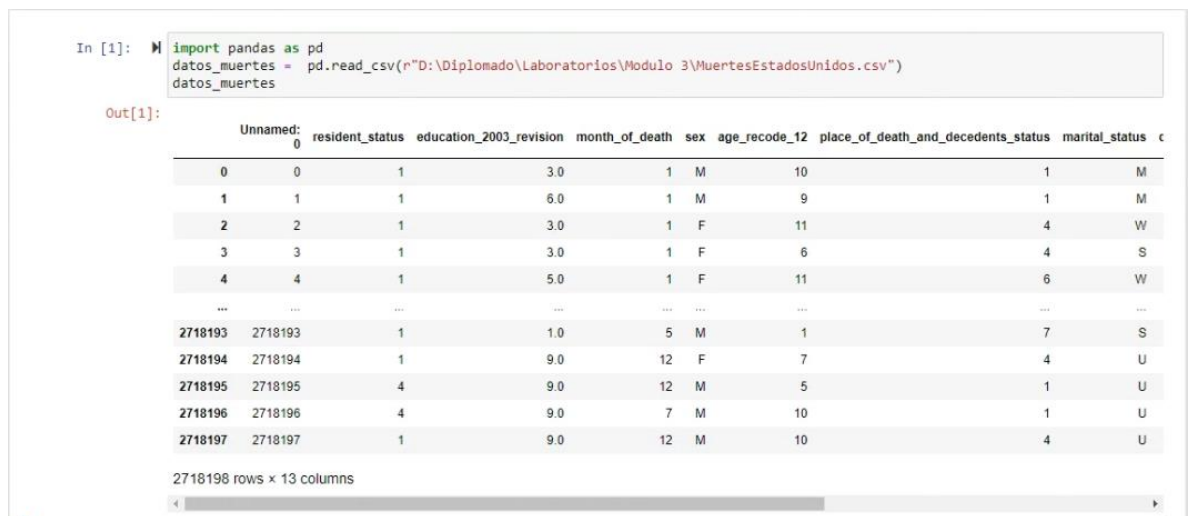


2. Evidenciar poder acceder a los datos ubicando el archivo en una ruta del sistema operativo.

- Primero se crea un nuevo libro llamado Laboratorio1Modulo3 en Jupyter.



- Importamos la librería de pandas y ingresamos la ruta en donde se encuentra el csv que contiene los datos y mostramos en pantalla.



3. Evidenciar poder abrir el archivo con las funciones strip() y Split() para contar filas y columnas.

- Primero abrimos el archivo en la máquina y lo recorremos y verificamos que lea la primera fila de datos.

```
In [4]: #CONATANDO COLUMNAS Y FILAS
datos_muertes = open("D:\\Diplomado\\Laboratorios\\Modulo 3\\MuertesEstadosUnidos.csv",'r')
cols = datos_muertes.readline()
cols

Out[4]: ',resident_status,education_2003_revision,month_of_death,sex,age_recode_12,place_of_death_and_decedents_status,marital_status,day_of_week_of_death,injury_at_work,358_cause_recoding,race,hispanic_origin'
```

- Iniciamos la rutina que cuenta las filas y columnas de datos.

Como primer paso leemos la primera línea para obtener los nombres de las columnas:

```
In [47]: #RUTINA QUE CUENTA LAS FILAS Y COLUMNA DE DATOS
#Leemos la primera línea para obtener los nombres de las columnas
cols = datos_muertes.readline().strip().split(",")
cols

Out[47]: ['',
'resident_status',
'education_2003_revision',
'month_of_death',
'sex',
'age_recode_12',
'place_of_death_and_decedents_status',
'marital_status',
'day_of_week_of_death',
'injury_at_work',
'358_cause_recoding',
'race',
'hispanic_origin']
```

Ahora el puntero del archivo está al comienzo de la segunda línea y así se contará correctamente las filas que contiene el archivo sin tener en cuenta los nombres de las columnas y dando como resultado el número de filas y columnas del archivo.

```
In [55]: # Ahora el puntero del archivo está al comienzo de la segunda línea, por lo que puedes contar las filas correctamente
num_filas = 0
for lineas in datos_muertes:
    num_filas += 1

# Imprime el número de filas y columnas
print("El archivo contiene un número {} filas y {} columnas".format(num_filas, len(cols)))

El archivo contiene un número 2718198 filas y 13 columnas
```

Dando como resultado:

Filas = 2,718,198

Columnas = 13

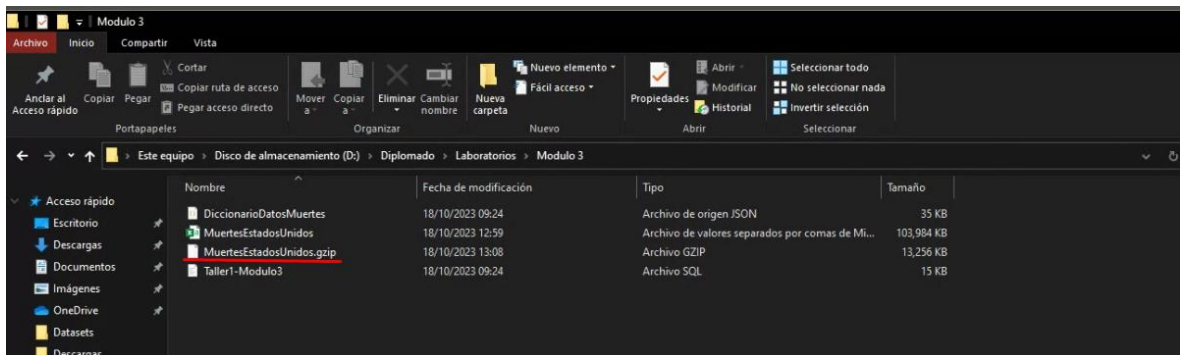
4. Migrar el conjunto de datos en al menos dos formatos de archivos diferentes al original

Ya que el archivo original está en formato csv, se procederá a transformar a formato parquet y el archivo original.

- Formato Parquet:

```
In [10]: #Transformacion de datos parquet y
datos_muertes = pd.read_csv("D:\Diplomado\Laboratorios\Modulo 3\MuertesEstadosUnidos.csv")
datos_muertes.to_parquet("D:\Diplomado\Laboratorios\Modulo 3\MuertesEstadosUnidos.gzip", compression='gzip')
```

Vemos que se ha creado el Gzip en la carpeta contenedora:



Consultamos el archivo con extensión Gzip:

```
In [26]: pd.read_parquet("D:\Diplomado\Laboratorios\Modulo 3\MuertesEstadosUnidos.gzip")
```

Out[26]:

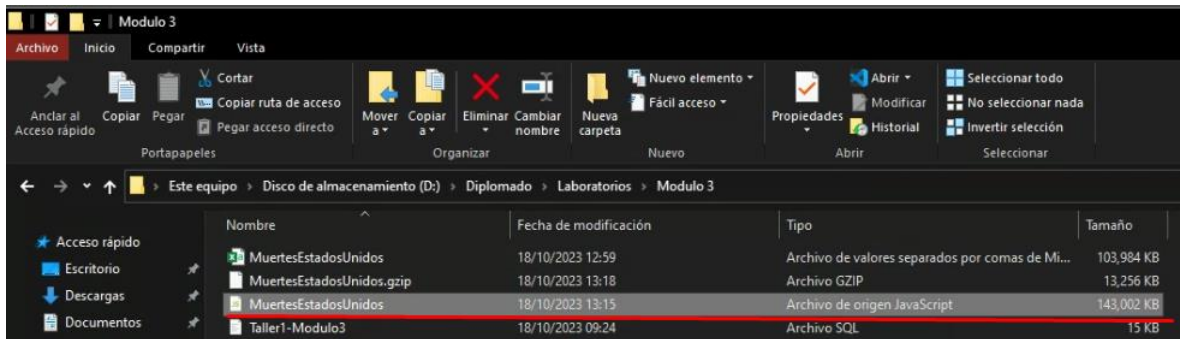
	number	resident_status	education_2003_revision	month_of_death	sex	age_recode_12	place_of_death_and_decedents_status	marital_status	day
0	0	1	3.0	1	M	10	1	M	
1	1	1	6.0	1	M	9	1	M	
2	2	1	3.0	1	F	11	4	W	
3	3	1	3.0	1	F	6	4	S	
4	4	1	5.0	1	F	11	6	W	
...
2718193	2718193	1	1.0	5	M	1	7	S	
2718194	2718194	1	9.0	12	F	7	4	U	
2718195	2718195	4	9.0	12	M	5	1	U	
2718196	2718196	4	9.0	7	M	10	1	U	
2718197	2718197	1	9.0	12	M	10	4	U	

2718198 rows x 13 columns

- Formato Json:

```
In [16]: #Transformacion de datos parquet y
datos_muertes = pd.read_csv(r"D:\Diplomado\Laboratorios\Modulo 3\MuertesEstadosUnidos.csv")
datos_muertes.to_parquet("D:\Diplomado\Laboratorios\Modulo 3\MuertesEstadosUnidos.gzip", compression='gzip')
datos_muertes.to_json("D:\Diplomado\Laboratorios\Modulo 3\MuertesEstadosUnidos.js",orient="split")
```

Vemos que se ha creado el Js en la carpeta contenedora:



Consultamos el archivo js:

```
In [27]: pd.read_json("D:\Diplomado\Laboratorios\Modulo 3\MuertesEstadosUnidos.js",orient="split")
```

```
Out[27]:
```

	number	resident_status	education_2003_revision	month_of_death	sex	age_recode_12	place_of_death_and_decendants_status	marital_status	day
0	0	1	3.0	1	M	10	1	M	
1	1	1	6.0	1	M	9	1	M	
2	2	1	3.0	1	F	11	4	W	
3	3	1	3.0	1	F	6	4	S	
4	4	1	5.0	1	F	11	6	W	
...
2718193	2718193	1	1.0	5	M	1	7	S	
2718194	2718194	1	9.0	12	F	7	4	U	
2718195	2718195	4	9.0	12	M	5	1	U	
2718196	2718196	4	9.0	7	M	10	1	U	
2718197	2718197	1	9.0	12	M	10	4	U	

2718198 rows x 13 columns

5. Consultas sobre el dataset

- Consulta 1: Conocer la cantidad de personas que eran residentes y murieron, dado el dataset de muertes.

```
#Consulta 1: Conociendo la cantidad de personas que eran residentes y murieron, dado el dataset de muertes
len(datos_muertes[datos_muertes["resident_status"] == 1])
```

[5] ✓ 0.2s Python

... 2197091

La consulta muestra que, a partir del dataset de 2.7 millones de datos, aproximadamente 2.19 de los registros de personas que murieron, eran residentes de Estados Unidos.

- Consulta 2: Conocer las mujeres entre 25 a 44 años que murieron (dado el dataset de muertes en USA)

```
#Consulta 2: Conociendo las mujeres entre 25 a 44 años que murieron para esta dataset
muertes_25_44 = datos_muertes[datos_muertes["age_recode_12"].between(5,6)]
muertes_25_44[muertes_25_44["sex"] == "F"]
```

[29] ✓ 0.1s Python

Unnamed: 0	resident_status	education_2003_revision	month_of_death	sex	age_recode_12	place_of_death_and_decedents_status	n
3	3	1	3.0	1	F	6	4
29	29	1	4.0	1	F	6	1
55	55	1	4.0	1	F	5	4
170	170	1	5.0	1	F	5	7
180	180	1	6.0	1	F	6	7
...
2718075	2718075	1	3.0	12	F	5	1
2718125	2718125	3	4.0	12	F	6	2
2718134	2718134	2	7.0	12	F	6	1
2718143	2718143	2	4.0	12	F	6	1
2718190	2718190	1	9.0	11	F	6	4

43314 rows x 13 columns

Esta consulta muestra que 43,314 mujeres se encontraban entre los 25 y 44 años al momento de su muerte.

- Consulta 3: Conocer todas las personas que murieron a causa de una herida en el lugar de trabajo y obtener el porcentaje respecto a todo el dataset

```
# Consulta 3: Conociendo todas las personas que murieron a causa de una herida en el lugar de trabajo
# Y obtener el porcentaje respecto a todo el dataset
heridas_lugar_trabajo = datos_muertes[datos_muertes["injury_at_work"] == "Y"]
(len(heridas_lugar_trabajo) / len(datos_muertes))*100.0
```

[33] ✓ 0.0s Python

... 0.15624321701362445

Esta consulta muestra que, solo el 0.15% de todas las muertes registradas en el dataset fueron por una herida al momento de estar realizando su trabajo.

- Consulta 4: Obtener el mes donde se registran más muertes en el dataset

```
# Consulta 4: obtener el mes donde se registran mas muertes en el dataset
meses_muertes = datos_muertes["month_of_death"].unique()
diccionario_muertes = {}
for mes in meses_muertes:
    cantidad = len(datos_muertes[datos_muertes["month_of_death"] == mes])
    diccionario_muertes[mes] = cantidad

diccionario_muertes = sorted(diccionario_muertes.items(), key=lambda x:x[1])
diccionario_muertes
```

[36] ✓ 0.3s Python

... [(9, 210347),
(6, 211631),
(8, 214917),
(7, 217451),
(11, 220234),
(10, 223987),
(5, 224035),
(4, 224841),
(2, 227471),
(12, 234228),
(3, 243174),
(1, 265882)]

El resultado se encuentra organizado de manera ascendente, por lo que, se concluye que el mes de enero, es donde se ha presentado más muertes registradas en los Estados Unidos (con 265 mil muertes aproximadamente, seguido por el mes de marzo con 243 mil muertes aproximadamente).

- Consulta 5: Conocer las personas cuyo lugar de muerte es desconocido

```
#Consulta 5: Obtener la lista de personas cuyo lugar de muerte es desconocido
datos_muertes[datos_muertes["place_of_death_and_decedents_status"] == 9]
```

[37] ✓ 0.0s Python

Unnamed: 0	resident_status	education_2003_revision	month_of_death	sex	age_recode_12	place_of_death_and_decedents_status	n
5892	5892	1	NaN	1	M	11	9
6741	6741	2	NaN	1	M	8	9
7017	7017	1	NaN	1	F	8	9
8322	8322	1	NaN	1	M	9	9
8500	8500	2	NaN	1	M	8	9
...
2638162	2638162	2	NaN	1	M	7	9
2638325	2638325	2	NaN	1	M	9	9
2638384	2638384	2	NaN	1	F	9	9
2640004	2640004	1	NaN	2	F	8	9
2640019	2640019	1	NaN	2	M	5	9

1290 rows x 13 columns

Del dataset de 2.7 millones de datos aproximadamente, se presenta que solo 1290 personas se desconoce el lugar donde se registró la muerte.

- Consulta 6: Conocer las personas que, se desconoce el lugar donde murieron, y la edad que tenían al momento de morir

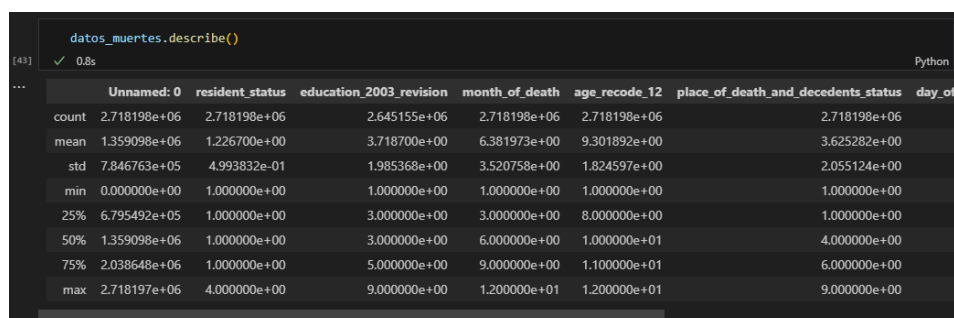
```
# Consulta 6: Conocer las personas que, se desconoce su lugar donde murieron, y su edad a la hora de morir
muertes_desconocidas = datos_muertes[datos_muertes["place_of_death_and_decedents_status"] == 9]
muertes_desconocidas[muertes_desconocidas["age_recode_12"] == 12]
```

[42] ✓ 0.0s Python

Unnamed: 0	resident_status	education_2003_revision	month_of_death	sex	age_recode_12	place_of_death_and_decedents_status	n
1444296	1444296	4	9.0	7	F	12	9
1756657	1756657	2	9.0	9	M	12	9
1762506	1762506	4	9.0	10	M	12	9
1768133	1768133	4	9.0	10	M	12	9

El resultado muestra que, solamente 4 personas de las 2.7 millones se desconoce la edad al momento de su muerte, y el lugar donde se registró la muerte (quedando completamente en el anonimato)

6. Obtener valores estadísticos de tendencia central y hacer al menos dos análisis con respecto a las variables numéricas que se observen.



	Unnamed: 0	resident_status	education_2003_revision	month_of_death	age_recode_12	place_of_death_and_decedents_status	day_of
count	2.718198e+06	2.718198e+06	2.645155e+06	2.718198e+06	2.718198e+06		2.718198e+06
mean	1.359098e+06	1.226700e+00	3.718700e+00	6.381973e+00	9.301892e+00		3.625282e+00
std	7.846763e+05	4.993832e-01	1.985368e+00	3.520758e+00	1.824597e+00		2.055124e+00
min	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00		1.000000e+00
25%	6.795492e+05	1.000000e+00	3.000000e+00	3.000000e+00	8.000000e+00		1.000000e+00
50%	1.359098e+06	1.000000e+00	3.000000e+00	6.000000e+00	1.000000e+01		4.000000e+00
75%	2.038648e+06	1.000000e+00	5.000000e+00	9.000000e+00	1.100000e+01		6.000000e+00
max	2.718197e+06	4.000000e+00	9.000000e+00	1.200000e+01	1.200000e+01		9.000000e+00

El análisis estadístico presenta la siguiente información:

- Para la columna “resident_status” (estatus migratorio de la persona fallecida), por lo menos el 75% de los datos cuentan con el valor 1.0 (que, dado el diccionario de datos, representa que por lo menos el 75% del dataset al momento de morir, eran residentes de USA)
- Para la columna “education_2003_revision”, la mediana tiene un valor de 3.0, que, en base al diccionario de datos de esta columna (que representa el nivel de educación de los fallecidos) que la mediana la mitad de la población de fallecidos, contaba con educación nivel high school o menor.
- Para la columna “month_of_death” (mes que se registró la muerte), por lo menos el 25% de la población del dataset, murió entre el mes de enero y marzo
- Para la columna “age_recode_12” (rango de edades de las personas que murieron), la mediana es de 10.0, lo que significa que aproximadamente el 50% de la población tenía una edad de 74 años o menor al momento de registrada la muerte