COVID-19 Mexican Analysis

Introduction

The COVID-19 Dataset is a Mexican database that is published daily. This file contains lot of information such as gender, age, pregnancy, location, other diseases that may affect the COVID infected patients among other fields. Moreover, the information is collected from all health institutions such as hospitals and clinics.

However, this file has lot of typos, misspelling fields as well as missing information. The Mexican government was made an effort to consolidate all the information and fill the missing values with an identifier and create some catalogues that are stored separately from the current database. This action allows to download the information from the source and process it by anyone.

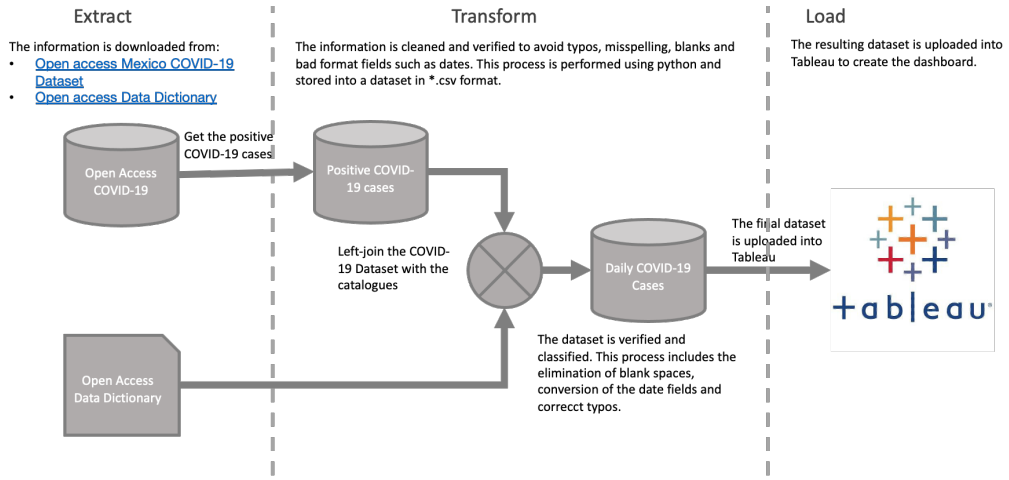
Objectives

- Show the data outliers without issues
- Clean the dataset to be used into Tableau
- Eliminate unnecessary columns
- Validate numerical and date fields

Extract, Transform and Load Process

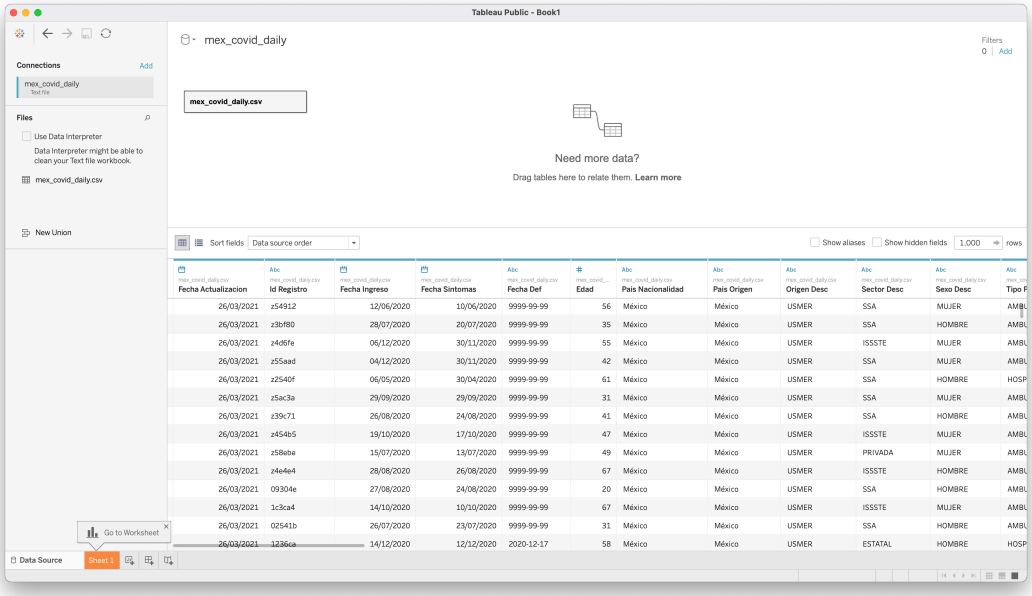
The extract, transform and load (ETL) is a computing procedure, commonly used by the Data Scientists and Data Analysts, to copy information from different sources, process the information to clean and convert into a proper format or structure and load the target database into an operational data store, data mart, data lake or data warehouse. In our particular case, we are going to load the database into tableau.

This procedure was performed into Python because it helps to create special methods to extract and create a statistical analysis to the information. It also helps to save into many different applications as well as databases.



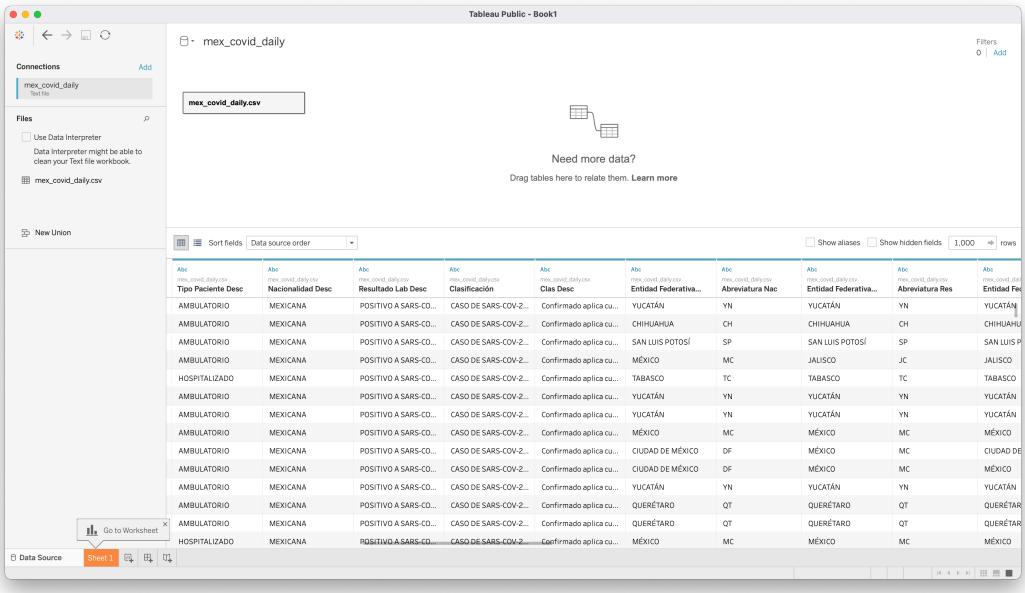
ETL Architectural diagram

Screenshots of the Data Outliers



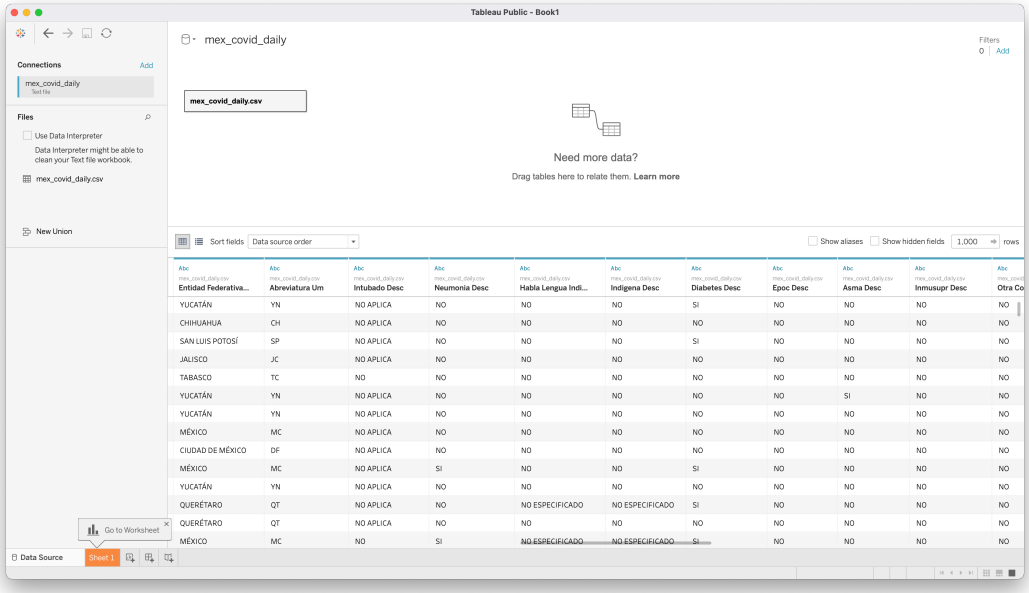
Fecha Actualizacion	Id Registro	Fecha Ingreso	Fecha Sintomas	Fecha Def	Edad	Pais Nacionalidad	Pais Origen	Origen Desc	Sector Desc	Sexo Desc	Tipo P
26/03/2021	z54912	12/06/2020	10/06/2020	9999-99-99	56	México	México	USMER	SSA	MUJER	AMBU
26/03/2021	z3bf80	28/07/2020	20/07/2020	9999-99-99	35	México	México	USMER	SSA	HOMBRE	AMBU
26/03/2021	z4d6fe	06/12/2020	30/11/2020	9999-99-99	55	México	México	USMER	ISSSTE	MUJER	AMBU
26/03/2021	z55aad	04/12/2020	30/11/2020	9999-99-99	42	México	México	USMER	SSA	MUJER	AMBU
26/03/2021	z2540f	06/05/2020	30/04/2020	9999-99-99	61	México	México	USMER	SSA	HOMBRE	HOSP
26/03/2021	z5ac3a	29/09/2020	29/09/2020	9999-99-99	31	México	México	USMER	SSA	MUJER	AMBU
26/03/2021	z39c71	26/08/2020	24/08/2020	9999-99-99	41	México	México	USMER	SSA	HOMBRE	AMBU
26/03/2021	z454b5	19/10/2020	17/10/2020	9999-99-99	47	México	México	USMER	ISSSTE	MUJER	AMBU
26/03/2021	z58ebe	15/07/2020	13/07/2020	9999-99-99	49	México	México	USMER	PRIVADA	MUJER	AMBU
26/03/2021	z4e4e4	28/08/2020	26/08/2020	9999-99-99	67	México	México	USMER	ISSSTE	HOMBRE	AMBU
26/03/2021	09304e	27/08/2020	24/08/2020	9999-99-99	20	México	México	USMER	SSA	HOMBRE	AMBU
26/03/2021	1c3ca4	14/10/2020	10/10/2020	9999-99-99	67	México	México	USMER	ISSSTE	MUJER	AMBU
26/03/2021	02541b	26/07/2020	23/07/2020	9999-99-99	31	México	México	USMER	SSA	HOMBRE	AMBU
26/03/2021	1236ca	14/12/2020	12/12/2020	2020-12-17	58	México	México	USMER	ESTATAL	HOMBRE	HOSP

Tableau Data Outliers - part 1



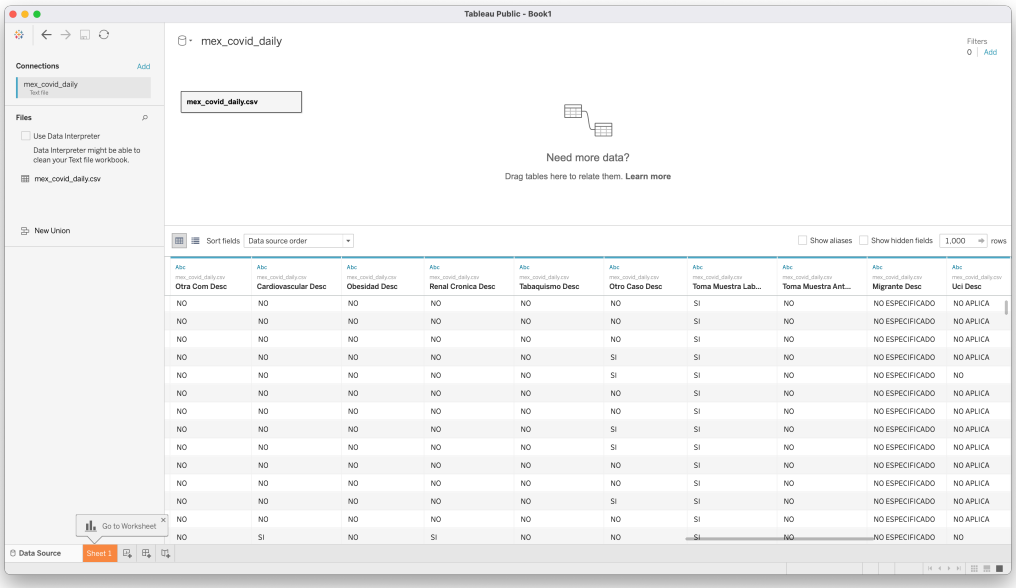
Tipo Paciente Desc	Nacionalidad Desc	Resultado Lab Desc	Clasificación	Clas Desc	Entidad Federativa...	Abreviatura Nac	Entidad Federativa...	Abreviatura Res	Entidad Fe
AMBULATORIO	MEXICANA	POSITIVO A SARS-CO...	CASO DE SARS-COV-2...	Confirmado aplica cu...	YUCATÁN	YN	YUCATÁN	YN	YUCATÁN
AMBULATORIO	MEXICANA	POSITIVO A SARS-CO...	CASO DE SARS-COV-2...	Confirmado aplica cu...	CHIHUAHUA	CH	CHIHUAHUA	CH	CHIHUAHU
AMBULATORIO	MEXICANA	POSITIVO A SARS-CO...	CASO DE SARS-COV-2...	Confirmado aplica cu...	SAN LUIS POTOSÍ	SP	SAN LUIS POTOSÍ	SP	SAN LUIS P
AMBULATORIO	MEXICANA	POSITIVO A SARS-CO...	CASO DE SARS-COV-2...	Confirmado aplica cu...	MÉXICO	MC	JALISCO	JC	JALISCO
HOSPITALIZADO	MEXICANA	POSITIVO A SARS-CO...	CASO DE SARS-COV-2...	Confirmado aplica cu...	TABASCO	TC	TABASCO	TC	TABASCO
AMBULATORIO	MEXICANA	POSITIVO A SARS-CO...	CASO DE SARS-COV-2...	Confirmado aplica cu...	YUCATÁN	YN	YUCATÁN	YN	YUCATÁN
AMBULATORIO	MEXICANA	POSITIVO A SARS-CO...	CASO DE SARS-COV-2...	Confirmado aplica cu...	YUCATÁN	YN	YUCATÁN	YN	YUCATÁN
AMBULATORIO	MEXICANA	POSITIVO A SARS-CO...	CASO DE SARS-COV-2...	Confirmado aplica cu...	MÉXICO	MC	MÉXICO	MC	MÉXICO
AMBULATORIO	MEXICANA	POSITIVO A SARS-CO...	CASO DE SARS-COV-2...	Confirmado aplica cu...	CIUDAD DE MÉXICO	DF	MÉXICO	MC	CIUDAD DE
AMBULATORIO	MEXICANA	POSITIVO A SARS-CO...	CASO DE SARS-COV-2...	Confirmado aplica cu...	CIUDAD DE MÉXICO	DF	MÉXICO	MC	MÉXICO
AMBULATORIO	MEXICANA	POSITIVO A SARS-CO...	CASO DE SARS-COV-2...	Confirmado aplica cu...	YUCATÁN	YN	YUCATÁN	YN	YUCATÁN
AMBULATORIO	MEXICANA	POSITIVO A SARS-CO...	CASO DE SARS-COV-2...	Confirmado aplica cu...	QUERÉTARO	QT	QUERÉTARO	QT	QUERÉTAR
AMBULATORIO	MEXICANA	POSITIVO A SARS-CO...	CASO DE SARS-COV-2...	Confirmado aplica cu...	QUERÉTARO	QT	QUERÉTARO	QT	QUERÉTAR
HOSPITALIZADO	MEXICANA	POSITIVO A SARS-CO...	CASO DE SARS-COV-2...	Confirmado aplica cu...	MÉXICO	MC	MÉXICO	MC	MÉXICO

Tableau Data Outliers - part 2



Entidad Federativa...	Abreviatura Um	Intubado Desc	Neumonia Desc	Habla Lengua Indi...	Indigena Desc	Diabetes Desc	Epoc Desc	Asma Desc	Inmusupr Desc	Otra Co
YUCATÁN	YN	NO APLICA	NO	NO	NO	SI	NO	NO	NO	NO
CHIHUAHUA	CH	NO APLICA	NO	NO	NO	NO	NO	NO	NO	NO
SAN LUIS POTOSÍ	SP	NO APLICA	NO	NO	NO	SI	NO	NO	NO	NO
JALISCO	JC	NO APLICA	NO	NO	NO	NO	NO	NO	NO	NO
TABASCO	TC	NO	NO	NO	NO	NO	NO	NO	NO	NO
YUCATÁN	YN	NO APLICA	NO	NO	NO	NO	NO	SI	NO	NO
YUCATÁN	YN	NO APLICA	NO	NO	NO	NO	NO	NO	NO	NO
MÉXICO	MC	NO APLICA	NO	NO	NO	NO	NO	NO	NO	NO
CIUDAD DE MÉXICO	DF	NO APLICA	NO	NO	NO	NO	NO	NO	NO	NO
MÉXICO	MC	NO APLICA	SI	NO	NO	SI	NO	NO	NO	NO
YUCATÁN	YN	NO APLICA	NO	NO	NO	NO	NO	NO	NO	NO
QUERÉTARO	QT	NO APLICA	NO	NO ESPECIFICADO	NO ESPECIFICADO	SI	NO	NO	NO	NO
QUERÉTARO	QT	NO APLICA	NO	NO	NO	NO	NO	NO	NO	NO
MÉXICO	MC	NO	SI	NO ESPECIFICADO	NO ESPECIFICADO	SI	NO	NO	NO	NO

Tableau Data Outliers - part 3



Otra Com Desc	Cardiovascular Desc	Obesidad Desc	Renal Cronica Desc	Tabaquismo Desc	Otro Caso Desc	Toma Muestra Lab...	Toma Muestra Ant...	Migrante Desc	Uci Desc
NO	NO	NO	NO	NO	NO	SI	NO	NO ESPECIFICADO	NO APLICA
NO	NO	NO	NO	NO	NO	SI	NO	NO ESPECIFICADO	NO APLICA
NO	NO	NO	NO	NO	NO	SI	NO	NO ESPECIFICADO	NO APLICA
NO	NO	NO	NO	NO	SI	SI	NO	NO ESPECIFICADO	NO APLICA
NO	NO	NO	NO	NO	SI	SI	NO	NO ESPECIFICADO	NO
NO	NO	NO	NO	NO	NO	SI	NO	NO ESPECIFICADO	NO APLICA
NO	NO	NO	NO	NO	NO	SI	NO	NO ESPECIFICADO	NO APLICA
NO	NO	NO	NO	NO	SI	SI	NO	NO ESPECIFICADO	NO APLICA
NO	NO	NO	NO	NO	SI	SI	NO	NO ESPECIFICADO	NO APLICA
NO	NO	NO	NO	NO	NO	SI	NO	NO ESPECIFICADO	NO APLICA
NO	NO	NO	NO	NO	NO	SI	NO	NO ESPECIFICADO	NO APLICA
NO	NO	NO	NO	NO	SI	SI	NO	NO ESPECIFICADO	NO APLICA
NO	NO	NO	NO	NO	NO	SI	NO	NO ESPECIFICADO	NO APLICA
NO	SI	NO	SI	NO	NO	SI	NO	NO ESPECIFICADO	NO

Tableau Data Outliers - part 4

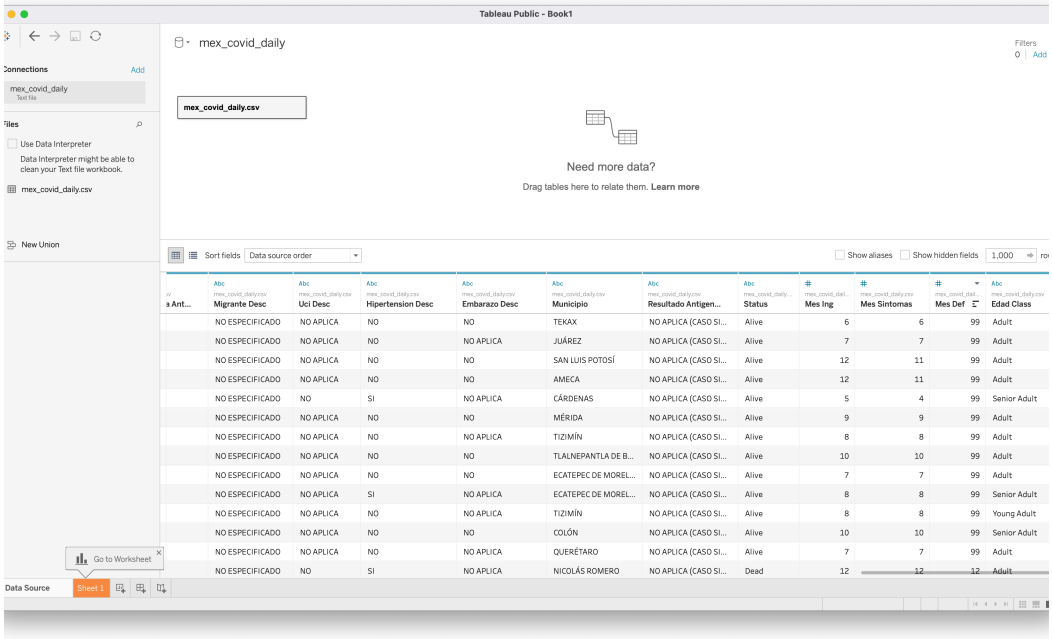


Tableau Data Outliers - part 5

Extract

The information is downloaded from the source: Open access Mexico COVID-19 Dataset. This one is presented into *.CSV format. The information needs to be downloaded daily and process continuously. Also, it is necessary to download the dictionaries to complete the database. These dictionaries are available in the source: Open access Data Dictionary.

If the Database doens't exist, then it download from the source

The following code downloads the latest dataset from the openacess database, this file is updated daily.

```
path = info.get('paths').get('Databases')
dataset = info.get('paths').get('dataset_url')
dictionary = info.get('paths').get('dictionary_url')
files = glob.glob(os.path.join(path,'*'))

[flag,fileName] = checkFiles(info.get('files').get('dataset_zip'),files)
if flag == 1:
    os.system('rm -r '+os.path.join(path,fileName))
    wget.download(dataset,path)
else:
    wget.download(dataset,path)

[flag,fileName] = checkFiles(info.get('files').get('dictionary_zip'),files)
if flag == 1:
    os.system('rm -r '+os.path.join(path,fileName))
    wget.download(dictionary,path)
else:
    wget.download(dictionary,path)

[flag,fileName] = checkFiles(info.get('files').get('open_data'),files)
if flag == 1:
    os.system('rm -r '+os.path.join(path,fileName))
[flag,fileName] = checkFiles(info.get('files').get('dictionary'),files)
if flag == 1:
    os.system('rm -r '+os.path.join(path,fileName))
```

Python extraction method

The python code above download automatically the dataset and data dictionary from the source. This automation also store the information in the system. In order to use it in python it is needed to unzip the files.

```
import zipfile
files = glob.glob(os.path.join(path,'*.zip'))
for file in files:
    if file == os.path.join(path,info.get('files').get('dictionary_zip')):
        with zipfile.ZipFile(file,"r") as zip_ref:
            zip_ref.extractall(os.path.join(path,dictionary.split('/')[len(dictionary.split('/'))-1].split('.')[0],'.'))
    else:
        with zipfile.ZipFile(file,"r") as zip_ref:
            zip_ref.extractall(os.path.join(path,'.'))
```

The length of the dataset is 6,006,949 registers that contains information from infected, non-infected patients as well as false-positive and true-negative cases. The information needed for the analysis is the infected patients information.

TRANSFORM

The downloaded information contains all the reported COVID-19 cases in the country. This information represents the positive, negative and false detections. For this particular study, only the positive cases will take into consideration. The following code extract the confirmed cases:

Transforming the data

The database are very large and it increases daily. However, the confirm cases are used for this analysis. In this database, the confirm cases also have other relations s
as smoking, pneumonia, COPD, asthma, hyperthension, diabetes, among other complications. In this case, we are going to extract only the confirma cases by means
laboratory analysis.

```
# Get positive COVID 19 infected people
data = db.loc[(db['RESULTADO_LAB']==1)]
start_number = len(data)
print('Total registers: {:,.0f}'.format(len(data)))
data.head()
```

```
Total registers: 1,621,273
```

	FECHA_ACTUALIZACION	ID_REGISTRO	ORIGEN	SECTOR	ENTIDAD_UM	SEXO	ENTIDAD_NAC	ENTIDAD_RES	MUNICIPIO_RES	TIPO_PACIENTE
8	2021-03-26	z33a15	2	12	12	2	12	12	29	1
14	2021-03-26	z4f06b	1	6	22	2	15	22	14	1
15	2021-03-26	z2770b	1	12	9	1	9	9	6	1
16	2021-03-26	z166d5	1	12	1	1	1	1	1	2
26	2021-03-26	zz8e57	1	12	31	1	31	31	96	1

Python Code, extract confirmed COVID-19 cases

This actions eliminate almost 4 million rows that has false detections as well as negative cases. Also, this action let the computer works faster, and clean. The next part is to join the catalogues to the main database. There are 11 catalogues that are needed to understand the dataset. These catalogues can download directly from the source In the following link Open access Data Dictionary.

In this part, the catalogue is open and merge to the database\ ORIGIN Catalogue It describes the origin description.

Code	Descriptio
1	USMER
2	Out of USMER
99	Not specify

Sentinel surveillance is carried out through the system of respiratory disease monitoring units (USMER). The USMER includes medical units of the first, second or third level of care, and third level units also participate as USMERs, which due to their characteristics contribute to expanding the epidemiological information panorama, including those with a specialty in pulmonology, infectology or pediatrics.

In [12]:
```
origen = catalogues.parse('Catálogo ORIGEN')
data = data.join(origen.set_index(['CLAVE']),on=['ORIGEN'])
data = data.rename(columns={'DESCRIPCIÓN':'ORIGEN_DESC'})
print('Total registers: {:,.0f}'.format(len(data)),'\nStarting register number: {:,.0f}'.format(start_number))
```

```
Total registers: 1,621,273
Starting register number: 1,621,273
```

Example: python left-join of the dataset and Origen catalogue.

The following code do the left join of the catalogue with the database. As it was mentioned before, there are 11 catalogues but there are almost 20 columns that used these catalogues. The reader can refer to COVID-19 Mexican Analysis github to check the full code.
Another important task to perform is to validate the date stamp fields. There are 3 columns in the database that must be checked, FECHA_INGRESO that denotes when the patient goes to the hospital or clinic; FECHA_Syntomas that are when the patient presents the first COVID-19 symptoms and finally, FECHA_DEFUNCION, this field mentioned when the patient died due to the disease. This columns are used to create a special field of patient status.

Change the date 9999-99-99 to Alive

In [46]:
```
import datetime

data['FECHA_INGRESO'] = pd.to_datetime(data['FECHA_INGRESO'])
data.loc[(data['FECHA_DEF'].isin(['9999-99-99'])),'Status'] = 'Alive'
data.loc[(~data['FECHA_DEF'].isin(['9999-99-99'])),'Status'] = 'Dead'

today = datetime.datetime.now()
delta_time = datetime.timedelta(days = 15)
newDate = today - delta_time
data.loc[(data['FECHA_DEF'].isin(['9999-99-99'])) & (data['FECHA_INGRESO']>pd.to_datetime(newDate.date())) & (data['TIPO_PACIE
NTE_DESC'].isin(['AMBULATORIO'])),'Status']='Active'

today = datetime.datetime.now()
delta_time = datetime.timedelta(days = 25)
newDate = today - delta_time
data.loc[(data['FECHA_DEF'].isin(['9999-99-99'])) & (data['FECHA_INGRESO']>pd.to_datetime(newDate.date())) & (data['TIPO_PACIE
NTE_DESC'].isin(['HOSPITALIZADO'])),'Status']='Active'
```

Patient status grouped into Alive, Death and Recovered

Finally, a special group is created based on age range to classified the people by age. This helps to analyze which particular sector is the most affected. This recommendation was done by the medical council to determine and create a vaccination campaign to demise the spread of the diseases.

In [49]:
```
data.loc[(data['EDAD']<5),'EDAD_CLASS'] = 'Baby'
data.loc[(data['EDAD'].between(5,8)),'EDAD_CLASS'] = 'Child'
data.loc[(data['EDAD'].between(9,17)),'EDAD_CLASS'] = 'Teenage'
data.loc[(data['EDAD'].between(18,30)),'EDAD_CLASS'] = 'Young Adult'
data.loc[(data['EDAD'].between(31,60)),'EDAD_CLASS'] = 'Adult'
data.loc[(data['EDAD']>60),'EDAD_CLASS'] = 'Senior Adult'
```

Once all these actions were done, the database is stored into a *.csv file and uploaded into different platforms to be analyzed. In this particular case, the information is uploaded into Tableau to create a dashboard that may help the local government to create a vaccination campaign as well as the lockdown semaphore to stop unnecessary activities in the region.

```
In [52]: os.system('rm -r ../Databases/*')
         data.to_csv(os.path.join(path,'mex_covid_daily.csv'),index=False,encoding='utf-8',line_terminator='\n')

In [53]: os.system('zip ../Databases/db.zip ../Databases/mex_covid_daily.csv')
         os.system('rm -r ../Databases/*.csv')

Out[53]: 0
```

This is the result of the whole analysis in terms of the information:

```
ata.info()

class 'pandas.core.frame.DataFrame'>
nt64Index: 1621273 entries, 8 to 6006524
ata columns (total 49 columns):
#   Column                      Non-Null Count    Dtype
--  ------                      --------------    -----
0   FECHA_ACTUALIZACION         1621273 non-null  object
1   ID_REGISTRO                 1621273 non-null  object
2   FECHA_INGRESO               1621273 non-null  datetime64[ns
3   FECHA_SINTOMAS              1621273 non-null  datetime64[ns
4   FECHA_DEF                   1621273 non-null  object
5   EDAD                        1621273 non-null  int64
6   PAIS_NACIONALIDAD           1621273 non-null  object
7   PAIS_ORIGEN                 1621273 non-null  object
8   ORIGEN_DESC                 1621273 non-null  object
9   SECTOR_DESC                 1621273 non-null  object
10  SEXO_DESC                   1621273 non-null  object
11  TIPO_PACIENTE_DESC          1621273 non-null  object
12  NACIONALIDAD_DESC           1621273 non-null  object
13  RESULTADO_LAB_DESC          1621273 non-null  object
14  CLASIFICACIÓN               1621273 non-null  object
15  CLAS_DESC                   1621273 non-null  object
16  ENTIDAD_FEDERATIVA_NAC      1621273 non-null  object
17  ABREVIATURA_NAC             1621273 non-null  object
18  ENTIDAD_FEDERATIVA_RES      1621273 non-null  object
19  ABREVIATURA_RES             1621273 non-null  object
20  ENTIDAD_FEDERATIVA_UM       1621273 non-null  object
21  ABREVIATURA_UM              1621273 non-null  object
22  INTUBADO_DESC               1621273 non-null  object
23  NEUMONIA_DESC               1621273 non-null  object
24  HABLA_LENGUA_INDIG_DESC     1621273 non-null  object
25  INDIGENA_DESC               1621273 non-null  object
26  DIABETES_DESC               1621273 non-null  object
27  EPOC_DESC                   1621273 non-null  object
28  ASMA_DESC                   1621273 non-null  object
29  INMUSUPR_DESC               1621273 non-null  object
30  OTRA_COM_DESC               1621273 non-null  object
31  CARDIOVASCULAR_DESC         1621273 non-null  object
32  OBESIDAD_DESC               1621273 non-null  object
33  RENAL_CRONICA_DESC          1621273 non-null  object
34  TABAQUISMO_DESC             1621273 non-null  object
35  OTRO_CASO_DESC              1621273 non-null  object
36  TOMA_MUESTRA_LAB_DESC       1621273 non-null  object
37  TOMA_MUESTRA_ANTIGENO_DESC  1621273 non-null  object
38  MIGRANTE_DESC               1621273 non-null  object
39  UCI_DESC                    1621273 non-null  object
40  HIPERTENSION_DESC           1621273 non-null  object
41  EMBARAZO_DESC               1621273 non-null  object
42  MUNICIPIO                   1621267 non-null  object
43  RESULTADO_ANTIGENO_DESC     1621273 non-null  object
44  Status                      1621273 non-null  object
45  MES_ING                     1621273 non-null  int64
46  MES_SINTOMAS                1621273 non-null  int64
47  MES_DEF                     1621273 non-null  int64
48  EDAD_CLASS                  1621273 non-null  object
ypes: datetime64[ns](2), int64(4), object(43)
emory usage: 618.5+ MB
```

Python final dataset

LOAD

The dataset is uploaded into Tableau and it is verified to check problems or issues in the data. For example: Checking null content or bad classify information.

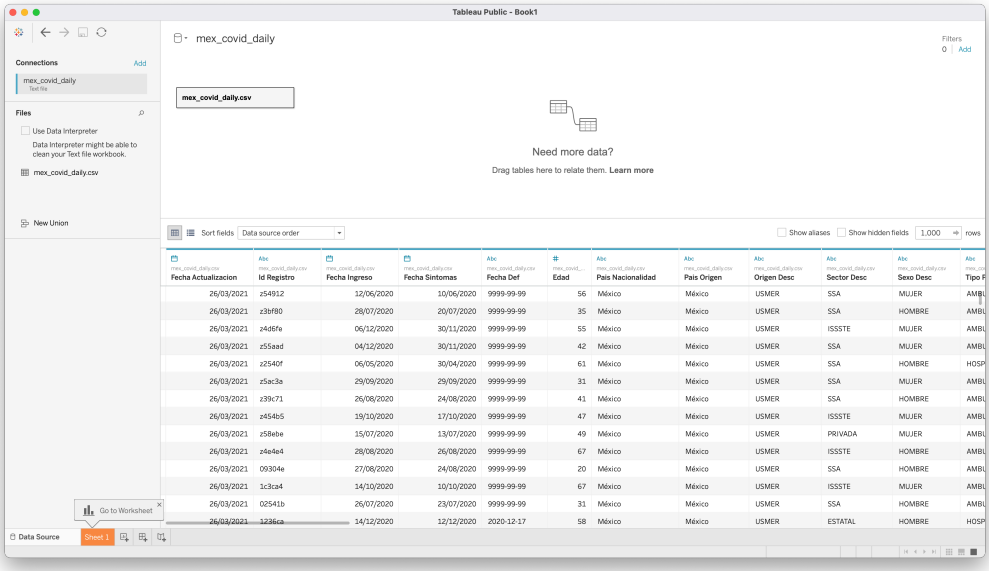
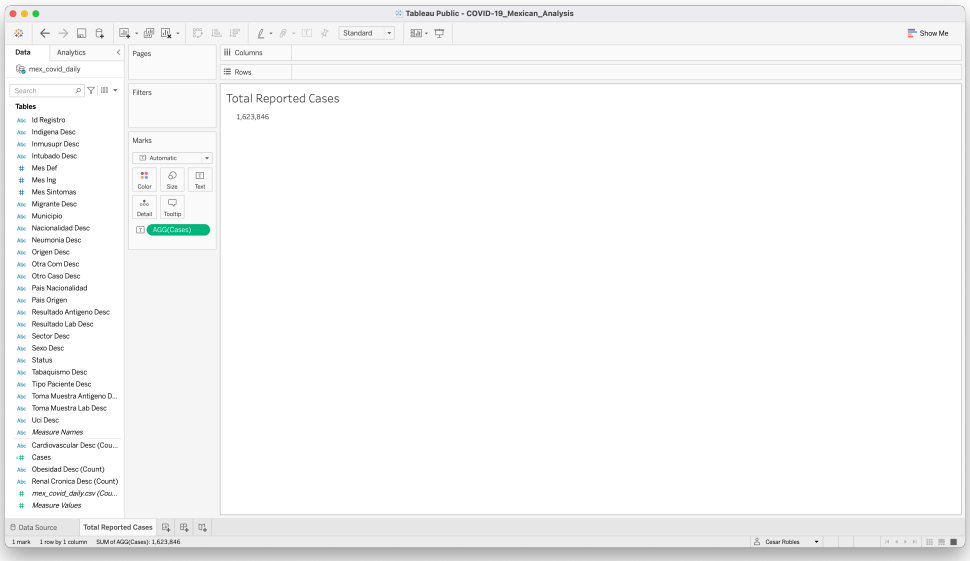


Tableau uploading information

To validate the uploaded information, it is verified the number of uploaded rows against the total rows in the dataset. In this case, python reports 1,623,846 cases that are the same in Tableau.



Total Reported Cases uploaded to Tableau