Universidad Nacional Autónoma de México
Facultad de Ingeniería

Análisis exploratorio de datos
Práctica 2

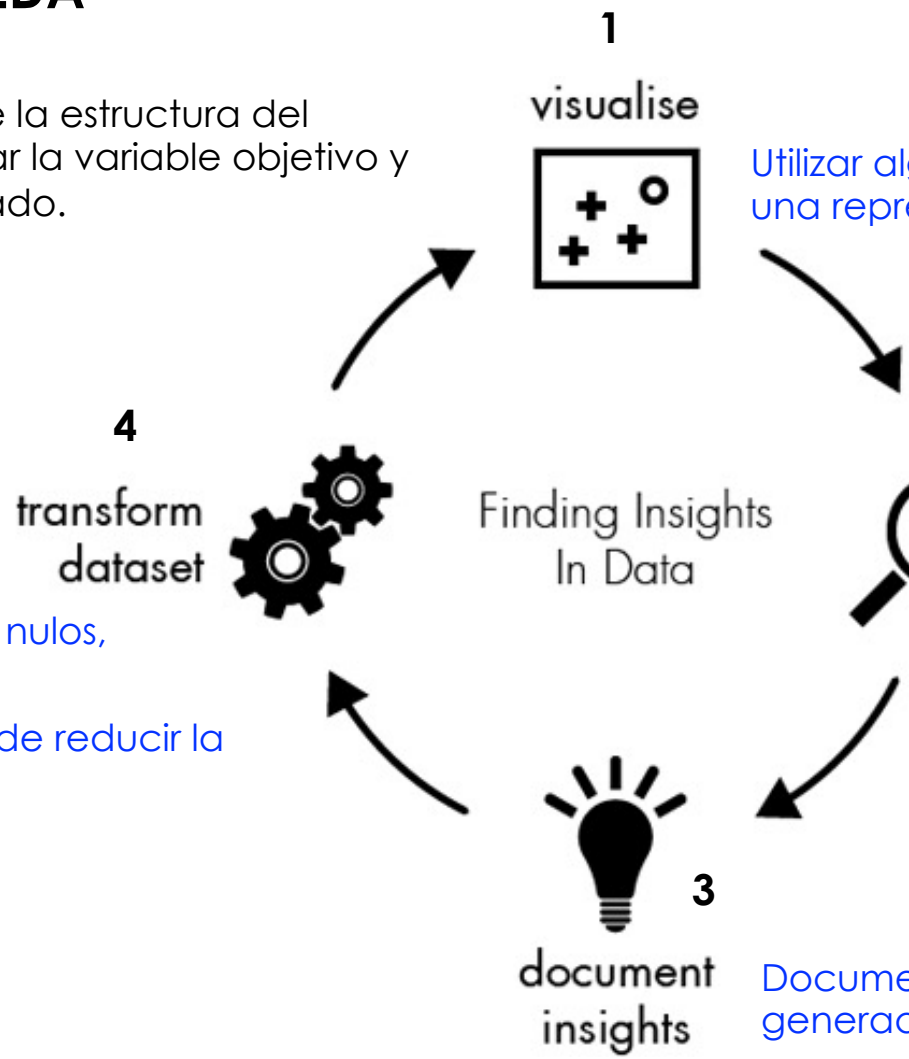
Guillermo Molero-Castillo
guillermo.molero@ingenieria.unam.edu

Septiembre, 2022

Análisis exploratorio de datos

Proceso básico de EDA

Propósito: Tener una idea de la estructura del conjunto de datos, identificar la variable objetivo y posibles técnicas de modelado.



Utilizar alguna herramienta para hacer una representación gráfica de los datos.

Analizar e interpretar los datos a través de diferentes visualizaciones.

Documentar todas las gráficas y estadísticas generadas.

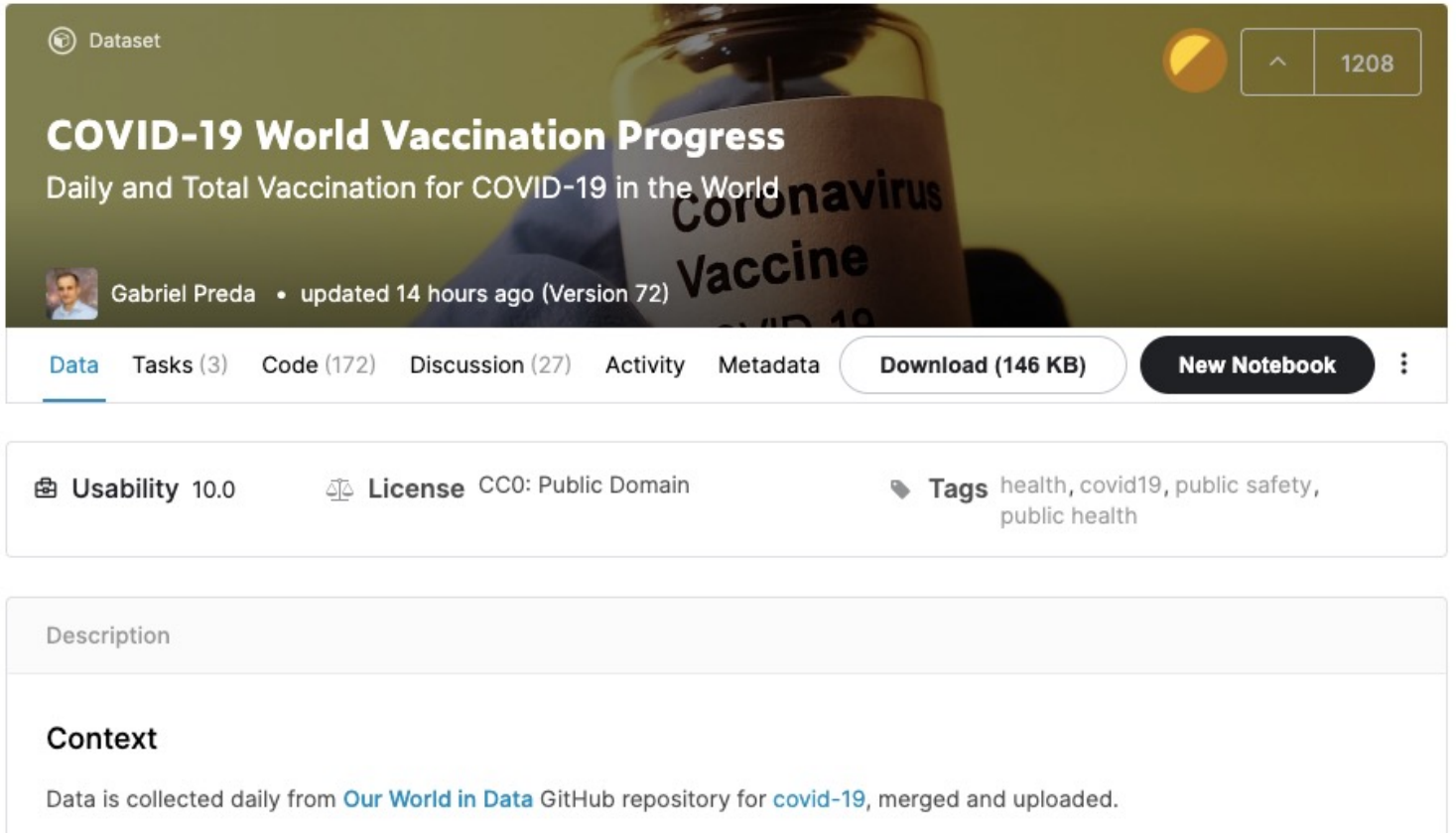
Qué hacer frente a valores nulos, faltantes o datos atípicos.

Establecer si hay necesiad de reducir la dimensionalidad de datos.

Análisis exploratorio de datos

Contexto

- Datos recopilados de manera diaria a nivel país sobre la vacunación contra COVID-19.
- **Objetivo:** Hacer un análisis exploratorio de datos sobre el progreso mundial de vacunación contra COVID-19.



Fuente: https://www.kaggle.com/gpreda/covid-world-vaccination-progress

Análisis exploratorio de datos

Diccionario de datos

1. **País:** Nombre del país.
2. **Código ISO:** Código ISO del país.
3. **Fecha:** Fecha de registro.
4. **Total de vacunaciones:** Número total de vacunaciones en el país.
5. **Total de personas vacunadas:** Una persona, según el esquema de inmunización, recibirá una o más vacunas (normalmente 2).
6. **Total de personas completamente vacunadas:** Número de personas que recibieron el esquema completo de vacunación.
7. **Vacunas diarias (crudos):** Número de vacunaciones para esa fecha/país.
8. **Vacunas diarias:** Número de vacunaciones para esa fecha/país.
9. **Total de vacunaciones por cien:** Relación (en porcentaje) entre el número de vacunaciones y la población total hasta la fecha.
10. **Total de personas vacunadas por cien:** Relación (en porcentaje) entre la población inmunizada y la población total hasta la fecha.
11. **Total de personas totalmente vacunadas por cien:** Relación (en porcentaje) entre la población totalmente inmunizada y la población total hasta la fecha en el país.
12. **Vacunas diarias por millón:** Relación (en ppm) entre el número de vacunaciones y la población total para la fecha actual en el país.
13. **Vacunas utilizadas en el país:** Nombre de las vacunas utilizadas en el país (hasta la fecha).
14. **Nombre de la fuente:** Fuente de la información (autoridad nacional, organización internacional, organización local, entre otros).
15. **Sitio web de origen:** Fuente de información.

Análisis exploratorio de datos

¿Qué se puede analizar?

Se puede usar este conjunto de datos para obtener información sobre la dinámica de la pandemia, como se refleja en la cantidad de pruebas realizadas, las tasas de infección después de la vacuna y las campañas de vacunación que han seguido determinados países.

- ¿Qué país está usando qué vacuna?
- ¿En qué país está más avanzado el programa de vacunación?
- ¿Dónde se vacunan más personas por día y en qué porcentaje de toda la población?

Análisis exploratorio de datos

Importar las bibliotecas y datos

```
import pandas as pd                          # Para la manipulación y análisis de datos
import numpy as np                           # Para crear vectores y matrices n dimensionales
import matplotlib.pyplot as plt              # Para la generación de gráficas a partir de los datos
import seaborn as sns                        # Para la visualización de datos basado en matplotlib
%matplotlib inline
# Para generar imágenes dentro del cuaderno
```

```
from google.colab import files
files.upload()
```

Análisis exploratorio de datos

Importar las bibliotecas y datos

```
DatosVacunacion = pd.read_csv("country_vaccinations.csv")
DatosVacunacion
```

	country	iso_code	date	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily_vaccinations_raw
0	Afghanistan	AFG	2021-02-22	0.0	0.0	NaN	NaN
1	Afghanistan	AFG	2021-02-23	NaN	NaN	NaN	NaN
2	Afghanistan	AFG	2021-02-24	NaN	NaN	NaN	NaN
3	Afghanistan	AFG	2021-02-25	NaN	NaN	NaN	NaN
4	Afghanistan	AFG	2021-02-26	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...

45257 rows × 15 columns

Análisis exploratorio con respecto a México

- Paso 1: Descripción de la estructura de los datos.
- Paso 2: Identificación de datos faltantes.
- Paso 3: Detección de valores atípicos.
- Paso 4: Identificación de relaciones entre pares variables.

Paso 1. Descripción de la estructura de los datos

1) Forma (dimensiones) del DataFrame

El atributo shape de Pandas proporciona una estructura general de los datos. Devuelve la cantidad de filas y columnas que tiene el conjunto de datos.

```
DatosVacunacion.shape
```

```
(45257, 15)
```

Análisis exploratorio de datos

Paso 1. Descripción de la estructura de los datos

```
DatosVacunacion.dtypes
```

```
country                                object
iso_code                               object
date                                   object
total_vaccinations                    float64
people_vaccinated                     float64
people_fully_vaccinated               float64
daily_vaccinations_raw                float64
daily_vaccinations                    float64
total_vaccinations_per_hundred        float64
people_vaccinated_per_hundred         float64
people_fully_vaccinated_per_hundred   float64
daily_vaccinations_per_million        float64
vaccines                               object
source_name                            object
source_website                         object
dtype: object
```

2) Tipos de datos

- El atributo dtypes muestra los tipos de datos de las variables.
- Se observa que el conjunto de datos tiene una combinación de variables categóricas (objeto) y numéricas (flotante).

Análisis exploratorio de datos

Paso 2. Identificación de datos faltantes

Una función útil de Pandas es isnull().sum() que regresa la suma de todos los valores nulos en cada variable.

```
DatosVacunacion.isnull().sum()
```

```
country                                0
iso_code                               0
date                                   0
total_vaccinations                 20550
people_vaccinated                  21676
people_fully_vaccinated            24567
daily_vaccinations_raw             24999
daily_vaccinations                   304
total_vaccinations_per_hundred     20550
people_vaccinated_per_hundred      21676
people_fully_vaccinated_per_hundred 24567
daily_vaccinations_per_million       304
vaccines                               0
source_name                            0
source_website                         0
dtype: int64
```

Paso 2. Identificación de datos faltantes

También se puede usar info() para obtener el tipo de datos y la suma de valores nulos.

```
DatosVacunacion.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45257 entries, 0 to 45256
Data columns (total 15 columns):
 #   Column                              Non-Null Count  Dtype  
---  ------                              --------------  -----  
 0   country                             45257 non-null  object 
 1   iso_code                            45257 non-null  object 
 2   date                                45257 non-null  object 
 3   total_vaccinations                  24707 non-null  float64
 4   people_vaccinated                   23581 non-null  float64
 5   people_fully_vaccinated             20690 non-null  float64
 6   daily_vaccinations_raw              20258 non-null  float64
 7   daily_vaccinations                  44953 non-null  float64
 8   total_vaccinations_per_hundred      24707 non-null  float64
 9   people_vaccinated_per_hundred       23581 non-null  float64
 10  people_fully_vaccinated_per_hundred 20690 non-null  float64
 11  daily_vaccinations_per_million      44953 non-null  float64
 12  vaccines                            45257 non-null  object 
 13  source_name                         45257 non-null  object 
 14  source_website                      45257 non-null  object 
dtypes: float64(9), object(6)
memory usage: 5.2+ MB
```

Paso 3. Detección de valores atípicos

- Se pueden utilizar gráficos para tener una idea general de las distribuciones de los datos, y se sacan estadísticas para resumir los datos. Estas dos estrategias son recomendables y se complementan.

- La distribución se refiere a cómo se distribuyen los valores en una variable o con qué frecuencia ocurren.

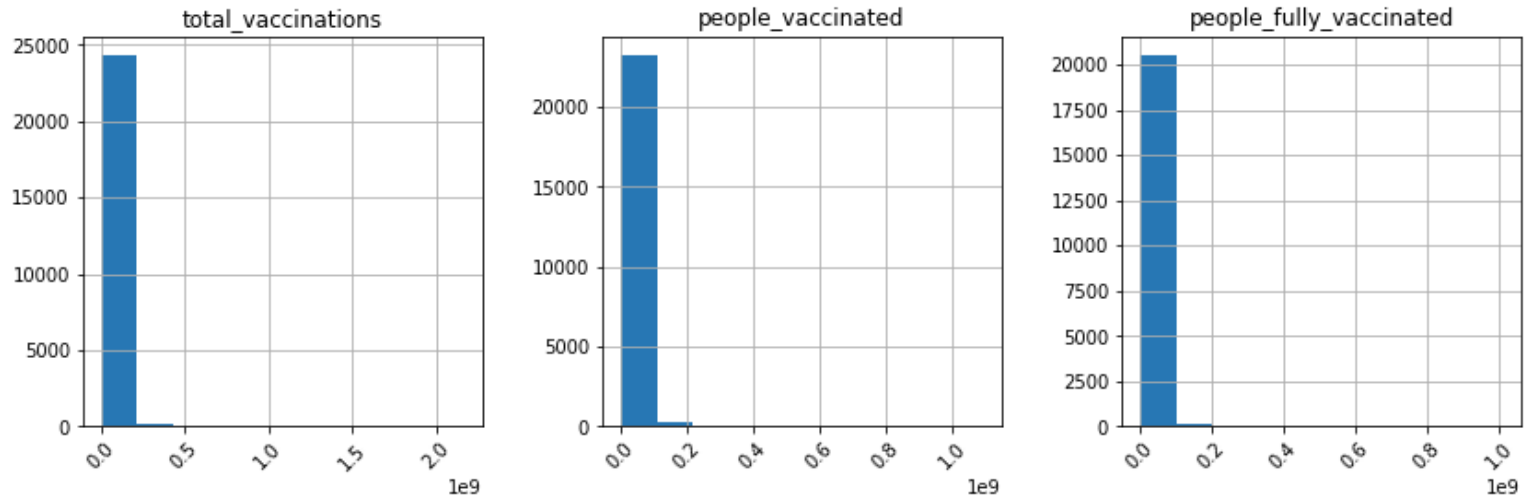
- Para las **variables numéricas**, se observa cuántas veces aparecen grupos de números en una columna. Mientras que para las **variables categóricas**, son las clases de cada columna y su frecuencia.

Paso 3. Detección de valores atípicos

1) Distribución de variables numéricas

- Se utilizan histogramas que agrupan los números en rangos.
- La altura de una barra muestra cuántos números caen en ese rango.
- Se emplea hist() para trazar el histograma de las variables numéricas. También se pueden usar los parámetros: figsize y xrot para aumentar el tamaño de la cuadrícula y rotar el eje x 45 grados.

```
DatosVacunacion.hist(figsize=(14,14), xrot=45)
plt.show()
```



Análisis exploratorio de datos

Paso 3. Detección de valores atípicos

1) Distribución de variables numéricas (México)

```
DatosVacunacion[DatosVacunacion.country == 'Mexico']
```

	country	iso_code	date	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily_vaccinations_raw
25968	Mexico	MEX	2020-12-24	2924.0	2924.0	NaN	NaN
25969	Mexico	MEX	2020-12-25	NaN	NaN	NaN	NaN
25970	Mexico	MEX	2020-12-26	NaN	NaN	NaN	NaN

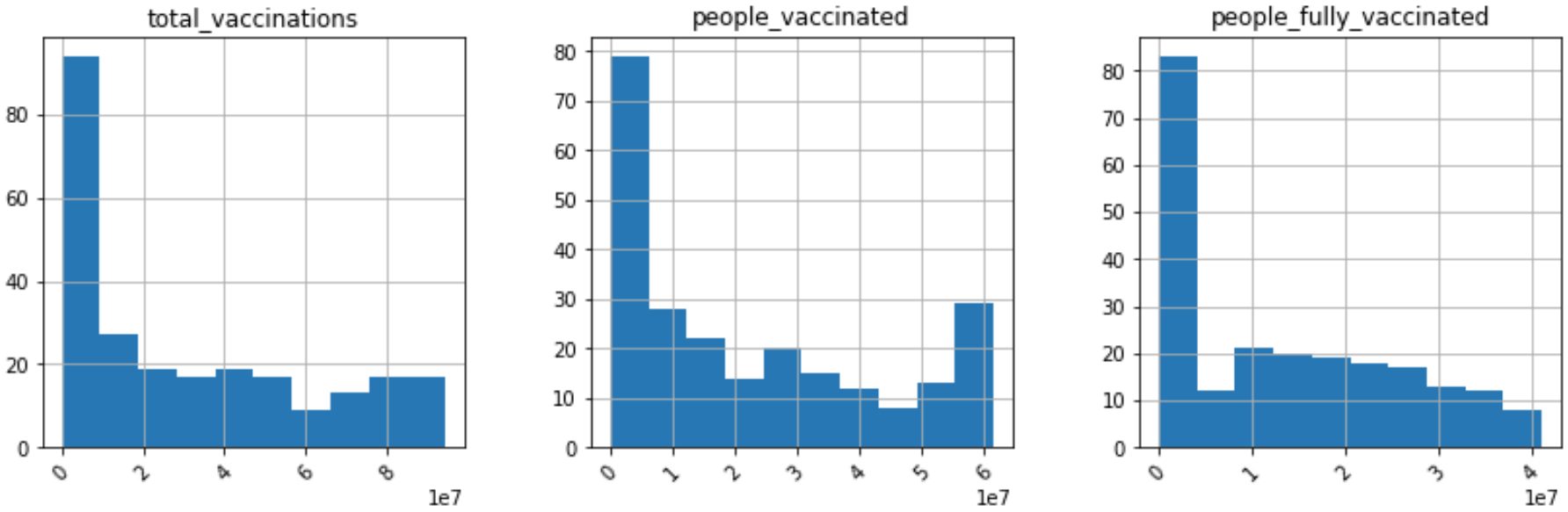
267 rows × 15 columns

Análisis exploratorio de datos

Paso 3. Detección de valores atípicos

1) Distribución de variables numéricas (México)

```
DatosVacunacion[DatosVacunacion.country == 'Mexico'].hist(figsize=(14,14), xrot=45)
plt.show()
```



Paso 3. Detección de valores atípicos

2) Resumen estadístico de variables numéricas

Se sacan estadísticas usando describe() que muestra un resumen estadístico de las variables numéricas.

```
DatosVacunacion[DatosVacunacion.country == 'Mexico'].describe()
```

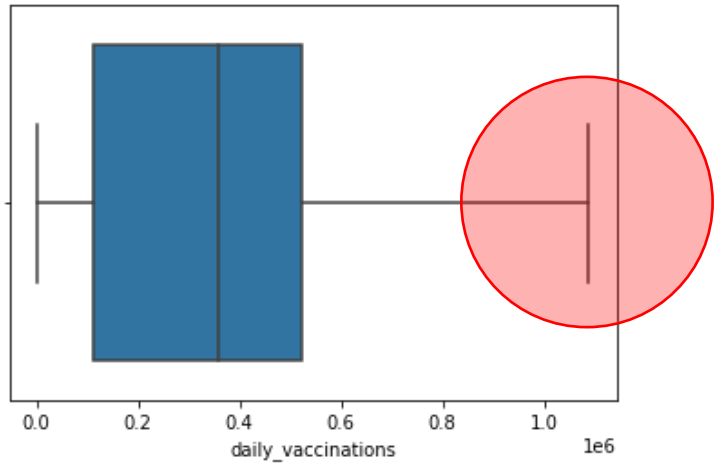
	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily_vaccinations_raw
count	2.490000e+02	2.400000e+02	2.230000e+02	2.350000e+02
mean	3.080362e+07	2.218104e+07	1.341907e+07	3.692836e+05
std	3.012615e+07	2.058768e+07	1.240475e+07	3.192058e+05
min	2.924000e+03	2.924000e+03	1.958000e+03	0.000000e+00
25%	2.676035e+06	2.036169e+06	6.712345e+05	9.965800e+04
50%	2.100862e+07	1.483526e+07	1.179455e+07	3.113180e+05
75%	5.270496e+07	3.782661e+07	2.245211e+07	5.545190e+05
max	9.430053e+07	6.161690e+07	4.111521e+07	1.454578e+06

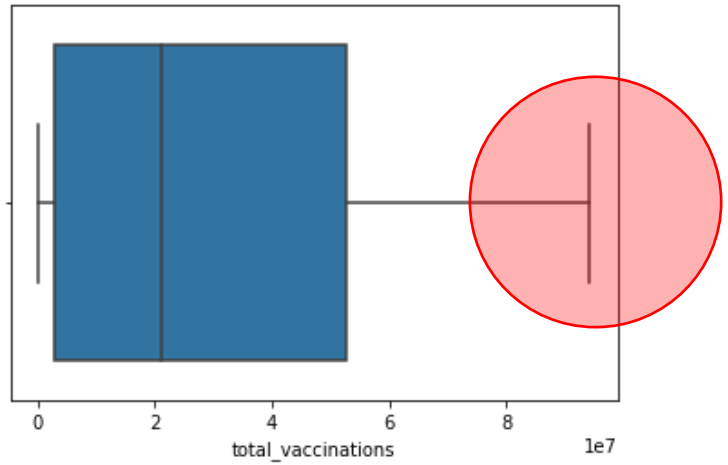
Paso 3. Detección de valores atípicos

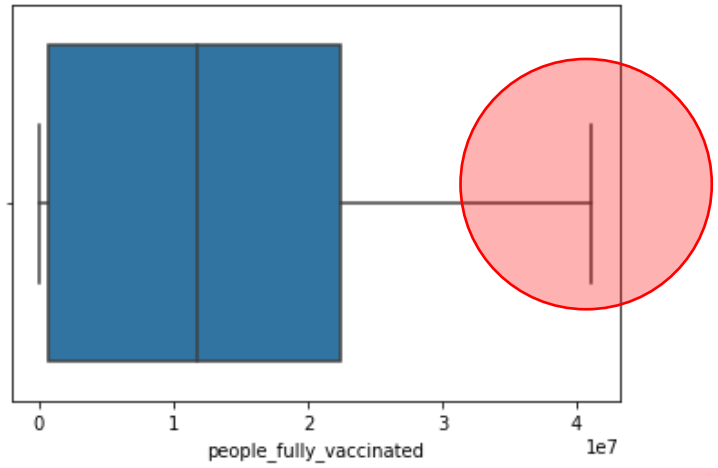
3) Diagramas para detectar posibles valores atípicos

Para este tipo de gráficas se utiliza Seaborn, que permite generar diagramas de cajas.

```
VariablesValoresAtipicos = ['daily_vaccinations', 'total_vaccinations', 'people_fully_vaccinated']
for col in VariablesValoresAtipicos:
  sns.boxplot(col, data=DatosVacunacion[DatosVacunacion.country == 'Mexico'])
  plt.show()
```







Paso 3. Detección de valores atípicos

4) Distribución de variables categóricas

- Se refiere a la observación de las clases de cada columna (variable) y su frecuencia.
- Aquí, las gráficas ayudan para tener una idea general de las distribuciones, mientras que las estadísticas dan números reales.

```
DatosVacunacion[DatosVacunacion.country == 'Mexico'].describe(include='object')
```

	country	iso_code	date	vaccines	source_name	source_website
count	267	267	267	267	267	267
unique	1	1	267	1	1	1
top	Mexico	MEX	2021-01-01	CanSino, Johnson&Johnson, Moderna, Oxford/Astr...	Secretary of Health	http://www.gob.mx/cms/uploads/attachment/file/...
freq	267	267	1	267	267	267

Paso 3. Detección de valores atípicos

4) Distribución de variables categóricas

```
plt.figure(figsize=(20, 5))
plt.plot(DatosVacunacion[DatosVacunacion.country == 'Mexico']['date'],
         DatosVacunacion[DatosVacunacion.country == 'Mexico']['people_vaccinated'], color='green', marker='o')
plt.xlabel('Fecha')
plt.ylabel('Personas vacunadas')
plt.title('Personas vacunadas en México')
plt.grid(True)
plt.show()
```

Paso 4. Identificación de relaciones entre variables

- Una matriz de correlaciones es útil para analizar la relación entre las variables numéricas.
- Se emplea la función corr()

```
DatosVacunacion[DatosVacunacion.country == 'Mexico'].corr()
```

	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily_vaccinations_raw
total_vaccinations	1.000000	0.999259	0.995915	0.643654
people_vaccinated	0.999259	1.000000	0.992050	0.648258
people_fully_vaccinated	0.995915	0.992050	1.000000	0.601532
daily_vaccinations_raw	0.643654	0.648258	0.601532	1.000000
daily_vaccinations	0.817368	0.823410	0.781864	0.792243
total_vaccinations_per_hundred	1.000000	0.999260	0.995915	0.643656
people_vaccinated_per_hundred	0.999259	1.000000	0.992047	0.648273
people_fully_vaccinated_per_hundred	0.995914	0.992049	1.000000	0.601544
daily_vaccinations_per_million	0.817365	0.823407	0.781860	0.792255

Análisis exploratorio de datos

Paso 4. Identificación de relaciones entre pares variables

```
plt.figure(figsize=(14,14))
sns.heatmap(DatosVacunacion[DatosVacunacion.country == 'Mexico'].corr(), cmap='RdBu_r', annot=True)
plt.show()
```