Part1 : Exploration de données avec Pandas

In [1]:
```
import pandas as pd
import ydata_profiling as pp

#Lire fichier covid_19
Covid_df = pd.read_csv('covid_19_data.csv')

#Avoir les informations
Covid_df.info()

#Voir l'en-tête de la dataset
Covid_df.head()

#Valeur manquant du datset
Covid_df.isnull()

#Description des information sur la data set
Covid_df.describe()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 306429 entries, 0 to 306428
Data columns (total 8 columns):
 #   Column           Non-Null Count   Dtype
---  ------           --------------   -----
 0   SNo              306429 non-null  int64
 1   ObservationDate  306429 non-null  object
 2   Province/State   228329 non-null  object
 3   Country/Region   306429 non-null  object
 4   Last Update      306429 non-null  object
 5   Confirmed        306429 non-null  float64
 6   Deaths           306429 non-null  float64
 7   Recovered        306429 non-null  float64
dtypes: float64(3), int64(1), object(4)
memory usage: 18.7+ MB
```

Out[1]:

	SNo	Confirmed	Deaths	Recovered
count	306429.000000	3.064290e+05	306429.000000	3.064290e+05
mean	153215.000000	8.567091e+04	2036.403268	5.042029e+04
std	88458.577156	2.775516e+05	6410.938048	2.015124e+05
min	1.000000	-3.028440e+05	-178.000000	-8.544050e+05
25%	76608.000000	1.042000e+03	13.000000	1.100000e+01
50%	153215.000000	1.037500e+04	192.000000	1.751000e+03
75%	229822.000000	5.075200e+04	1322.000000	2.027000e+04
max	306429.000000	5.863138e+06	112385.000000	6.399531e+06

In [2]:
```
#Avoir les informations
Covid_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 306429 entries, 0 to 306428
Data columns (total 8 columns):
 #   Column           Non-Null Count   Dtype
---  ------           --------------   -----
 0   SNo              306429 non-null  int64
 1   ObservationDate  306429 non-null  object
 2   Province/State   228329 non-null  object
 3   Country/Region   306429 non-null  object
 4   Last Update      306429 non-null  object
 5   Confirmed        306429 non-null  float64
 6   Deaths           306429 non-null  float64
 7   Recovered        306429 non-null  float64
dtypes: float64(3), int64(1), object(4)
memory usage: 18.7+ MB
```

In [3]:
```
#Voir l'en-tête de la dataset
Covid_df.head()
```

Out[3]:

	SNo	ObservationDate	Province/State	Country/Region	Last Update	Confirmed	Deaths	Recovered
0	1	01/22/2020	Anhui	Mainland China	1/22/2020 17:00	1.0	0.0	0.0
1	2	01/22/2020	Beijing	Mainland China	1/22/2020 17:00	14.0	0.0	0.0
2	3	01/22/2020	Chongqing	Mainland China	1/22/2020 17:00	6.0	0.0	0.0
3	4	01/22/2020	Fujian	Mainland China	1/22/2020 17:00	1.0	0.0	0.0
4	5	01/22/2020	Gansu	Mainland China	1/22/2020 17:00	0.0	0.0	0.0

In [4]:
```
#Valeur manquant du datset
Covid_df.isnull()
```

Out[4]:

	SNo	ObservationDate	Province/State	Country/Region	Last Update	Confirmed	Deaths	Recovered
0	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...
306424	False	False	False	False	False	False	False	False
306425	False	False	False	False	False	False	False	False
306426	False	False	False	False	False	False	False	False
306427	False	False	False	False	False	False	False	False
306428	False	False	False	False	False	False	False	False

306429 rows × 8 columns

In [5]:
```
#Description des information sur la data set
Covid_df.describe()
```

Out[5]:

	SNo	Confirmed	Deaths	Recovered
count	306429.000000	3.064290e+05	306429.000000	3.064290e+05
mean	153215.000000	8.567091e+04	2036.403268	5.042029e+04
std	88458.577156	2.775516e+05	6410.938048	2.015124e+05
min	1.000000	-3.028440e+05	-178.000000	-8.544050e+05
25%	76608.000000	1.042000e+03	13.000000	1.100000e+01
50%	153215.000000	1.037500e+04	192.000000	1.751000e+03
75%	229822.000000	5.075200e+04	1322.000000	2.027000e+04
max	306429.000000	5.863138e+06	112385.000000	6.399531e+06

Part2 : Exploration de données avec Pandas profiling

In [6]:
```
import pandas as pd
import ydata_profiling as pp

#Lire fichier covid_19
Covid_df = pd.read_csv('covid_19_data.csv')

#Generer un profiling report
profile = pp.ProfileReport(Covid_df, title='Pandas Profiling Report')

# Display the report
profile.to_notebook_iframe()
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]
Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]
Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]
```

t	108911	5.6%
u	82240	4.2%
Other values (16)	454688	23.3%

Uppercase Letter

Value	Count	Frequency (%)
O	34613	10.4%
C	24261	7.3%
S	23987	7.2%
M	23905	7.2%
A	23871	7.2%
N	18613	5.6%
K	18093	5.4%
P	17362	5.2%
R	15614	4.7%
T	14836	4.5%
Other values (16)	117087	35.2%

Other Punctuation

Value	Count	Frequency (%)

In [7]:
```
#Lire fichier covid_19
Covid_df = pd.read_csv('covid_19_data.csv')
```

In [8]:
```
#Generer un profiling report
profile = pp.ProfileReport(Covid_df, title='Pandas Profiling Report')
```

In [9]:
```
# Display the report
#profile.to_notebook_iframe()
profile.to_file("MyCovid_report.html")
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]
Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]
Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]
Export report to file:   0%|          | 0/1 [00:00<?, ?it/s]
```

Résumé sur le rapport du pandas profiling

A l'issue de l'analyse des données sur le Covid_19 de plusieur pays du monde

1. Nous remarquons que la dataset contient 8 variables avec un total de 306429 observations et un pourcentage de 3,2% de valeur manquante dans la dataset.

2. Au niveaux des observationn On note une evolution croissante de la covid_19 en fonction du temps et atteint un pic constant à partir de septembre 2020 jusqu'en Mai 2021 avec la RUSSIE au classement mondiale des pays les plus touché par le Covid_19. Ci-desssous les 10 premiers pays lkes plus touché par le covid_19 entre Janvier 2020 et Mai 2021:

- Russie
- USA
- Japan
- China

- Mainland
- India
- Colombie
- Mexique
- Brasil
- Ukraine

1. La dataset comporte des valeurs manquante et on observe qu'il y a moins de valeur manquante au niveau de la variable Country/Region qu'au niveau des autres variables.

2. Il y a une forte correlation entre le [nombre de cas confirmé et le nombre de cas guéris],[Nombre de deces et nombre de cas confirmé] et [Nobre de cas guéris et nombre de cas confirmé]. Mais nous ne pouvons par affirmer qu'il existe une relation cause et effet entre ces deux variable car bien n'ayant une forte correlation elle ne sont pas lier.

In [ ]: