IBM NAAN MUDHALVAN

PHASE4

DEVELOPMENT PHASE

COVID 19 VACCINE ANALYSIS

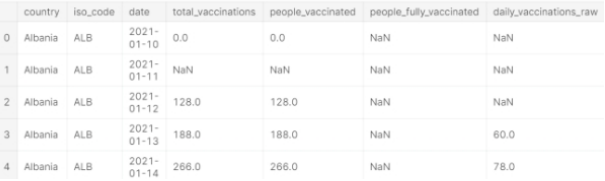
PROBLEM STATEMENT

Forecasting of time taken for completing 100% total vaccinations of particular region over the time period.By this, vaccine manufacturing companies get to know the prior requirements of vaccine which helps to produce the vaccines in large scale and complete the vaccination drive with in calculated time.

1.Data Understanding

importing data and libraries

```
import pandas as pd

import seaborn as sns

import numpy as np

import matplotlib.pyplot as plt

import plotly.express as px

import io

import requests

import warnings

warnings.filterwarnings('ignore')

url = "https://raw.githubusercontent.com/owid/covid-19-data/master/public/data/owid-covid-data.csv"

read_data = requests.get(url).content

address = pd.read_csv(io.StringIO(read_data.decode('utf-8')))

address.head()
```



	country	iso_code	date	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily_vaccinations_raw
0	Albania	ALB	2021-01-10	0.0	0.0	NaN	NaN
1	Albania	ALB	2021-01-11	NaN	NaN	NaN	NaN
2	Albania	ALB	2021-01-12	128.0	128.0	NaN	NaN
3	Albania	ALB	2021-01-13	188.0	188.0	NaN	60.0
4	Albania	ALB	2021-01-14	266.0	266.0	NaN	78.0

```
url2= "https://raw.githubusercontent.com/owid/covid-19-
data/master/public/data/vaccinations/vaccinations-by-manufacturer.csv"

read_data = requests.get(url2).content

vaccine=pd.read_csv(io.StringIO(read_data.decode('utf-8')))

data=address

data.columns
```

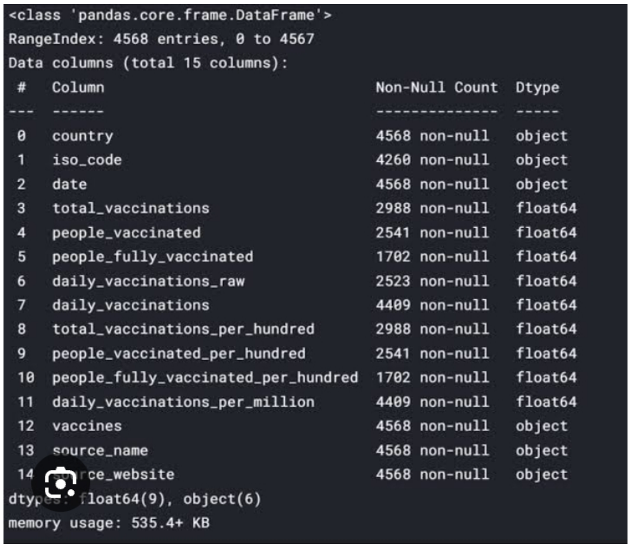
```
Index(['iso_code', 'continent', 'location', 'date', 'total_cases', 'new_cases',

       'new_cases_smoothed', 'total_deaths', 'new_deaths',

       'new_deaths_smoothed', 'total_cases_per_million',

       'new_cases_per_million', 'new_cases_smoothed_per_million',

       'total_deaths_per_million', 'new_deaths_per_million',

       'new_deaths_smoothed_per_million', 'reproduction_rate', 'icu_patients',

       'icu_patients_per_million', 'hosp_patients',

       'hosp_patients_per_million', 'weekly_icu_admissions',

       'weekly_icu_admissions_per_million', 'weekly_hosp_admissions',

       'weekly_hosp_admissions_per_million', 'total_tests', 'new_tests',

       'total_tests_per_thousand', 'new_tests_per_thousand',

       'new_tests_smoothed', 'new_tests_smoothed_per_thousand',

       'positive_rate', 'tests_per_case', 'tests_units', 'total_vaccinations',

       'people_vaccinated', 'people_fully_vaccinated', 'total_boosters',

       'new_vaccinations', 'new_vaccinations_smoothed',
       'total_vaccinations_per_hundred', 'people_vaccinated_per_hundred',
       'people_fully_vaccinated_per_hundred', 'total_boosters_per_hundred',
       'new_vaccinations_smoothed_per_million',
       'new_people_vaccinated_smoothed',
       'new_people_vaccinated_smoothed_per_hundred', 'stringency_index',
       'population', 'population_density', 'median_age', 'aged_65_older',
       'aged_70_older', 'gdp_per_capita', 'extreme_poverty',
       'cardiovasc_death_rate', 'diabetes_prevalence', 'female_smokers',
       'male_smokers', 'handwashing_facilities', 'hospital_beds_per_thousand',
       'life_expectancy', 'human_development_index',
       'excess_mortality_cumulative_absolute', 'excess_mortality_cumulative',
       'excess_mortality', 'excess_mortality_cumulative_per_million'],
      dtype='object')
data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 191376 entries, 0 to 191375
Data columns (total 67 columns):
 #   Column                                      Non-Null Count   Dtype
---  ------                                      --------------   -----
 0   iso_code                                    191376 non-null  object
 1   continent                                   180250 non-null  object
 2   location                                    191376 non-null  object
 3   date                                        191376 non-null  object
 4   total_cases                                 183834 non-null  float64
 5   new_cases                                   183621 non-null  float64
```

```
6   new_cases_smoothed                     182447 non-null  float64
7   total_deaths                           165368 non-null  float64
8   new_deaths                             165361 non-null  float64
9   new_deaths_smoothed                    164198 non-null  float64
10  total_cases_per_million                182986 non-null  float64
11  new_cases_per_million                  182773 non-null  float64
12  new_cases_smoothed_per_million         181604 non-null  float64
13  total_deaths_per_million               164533 non-null  float64
14  new_deaths_per_million                 164526 non-null  float64
15  new_deaths_smoothed_per_million        163368 non-null  float64
16  reproduction_rate                      140710 non-null  float64
17  icu_patients                           25496 non-null   float64
18  icu_patients_per_million               25496 non-null   float64
19  hosp_patients                          26747 non-null   float64
20  hosp_patients_per_million              26747 non-null   float64
21  weekly_icu_admissions                  6222 non-null    float64
22  weekly_icu_admissions_per_million      6222 non-null    float64
23  weekly_hosp_admissions                 12397 non-null   float64
24  weekly_hosp_admissions_per_million     12397 non-null   float64
25  total_tests                            77683 non-null   float64
26  new_tests                              74008 non-null   float64
27  total_tests_per_thousand               77683 non-null   float64
28  new_tests_per_thousand                 74008 non-null   float64
29  new_tests_smoothed                     101315 non-null  float64
30  new_tests_smoothed_per_thousand        101315 non-null  float64
```

```
31  positive_rate                        93441 non-null   float64
32  tests_per_case                       91681 non-null   float64
33  tests_units                          104079 non-null  object
34  total_vaccinations                   52388 non-null   float64
35  people_vaccinated                     49909 non-null   float64
36  people_fully_vaccinated              47375 non-null   float64
37  total_boosters                        24452 non-null   float64
38  new_vaccinations                      42912 non-null   float64
39  new_vaccinations_smoothed             103578 non-null  float64
40  total_vaccinations_per_hundred        52388 non-null   float64
41  people_vaccinated_per_hundred         49909 non-null   float64
42  people_fully_vaccinated_per_hundred   47375 non-null   float64
43  total_boosters_per_hundred            24452 non-null   float64
44  new_vaccinations_smoothed_per_million  103578 non-null  float64
45  new_people_vaccinated_smoothed        102491 non-null  float64
46  new_people_vaccinated_smoothed_per_hundred  102491 non-null  float64
47  stringency_index                      148621 non-null  float64
48  population                            190211 non-null  float64
49  population_density                    170524 non-null  float64
50  median_age                            158052 non-null  float64
51  aged_65_older                         156377 non-null  float64
52  aged_70_older                         157223 non-null  float64
53  gdp_per_capita                        157205 non-null  float64
54  extreme_poverty                       102625 non-null  float64
55  cardiovasc_death_rate                 157692 non-null  float64
```

```
 56  diabetes_prevalence                       165401 non-null  float64
 57  female_smokers                            119268 non-null  float64
 58  male_smokers                              117633 non-null  float64
 59  handwashing_facilities                     77477 non-null   float64
 60  hospital_beds_per_thousand                139914 non-null  float64
 61  life_expectancy                           178964 non-null  float64
 62  human_development_index                   153621 non-null  float64
 63  excess_mortality_cumulative_absolute       6553 non-null    float64
 64  excess_mortality_cumulative                6553 non-null    float64
 65  excess_mortality                           6553 non-null    float64
 66  excess_mortality_cumulative_per_million    6553 non-null    float64
dtypes: float64(62), object(5)
memory usage: 97.8+ MB
```

```
data.describe(include='all')
```



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4568 entries, 0 to 4567
Data columns (total 15 columns):
 #   Column                               Non-Null Count  Dtype
---  ------                               --------------  -----
 0   country                              4568 non-null   object
 1   iso_code                             4260 non-null   object
 2   date                                 4568 non-null   object
 3   total_vaccinations                   2988 non-null   float64
 4   people_vaccinated                    2541 non-null   float64
 5   people_fully_vaccinated              1702 non-null   float64
 6   daily_vaccinations_raw               2523 non-null   float64
 7   daily_vaccinations                   4409 non-null   float64
 8   total_vaccinations_per_hundred       2988 non-null   float64
 9   people_vaccinated_per_hundred        2541 non-null   float64
 10  people_fully_vaccinated_per_hundred  1702 non-null   float64
 11  daily_vaccinations_per_million       4409 non-null   float64
 12  vaccines                             4568 non-null   object
 13  source_name                          4568 non-null   object
 14  source_website                       4568 non-null   object
dtypes: float64(9), object(6)
memory usage: 535.4+ KB
```

```
vaccine.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 42395 entries, 0 to 42394
Data columns (total 4 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   location            42395 non-null  object
 1   date                42395 non-null  object
 2   vaccine             42395 non-null  object
 3   total_vaccinations  42395 non-null  int64
dtypes: int64(1), object(3)
```

```
vaccine.describe()
```

```
total_vaccinations
count   4.239500e+04
mean    1.782378e+07
std     5.733925e+07
min     0.000000e+00
25%     1.075450e+05
50%     1.528400e+06
75%     9.792642e+06
max     6.141617e+08
```

2.data preprocessing

```
data.isnull().sum()
```

```
iso_code                           0
continent                      11126
```

```
location                                  0
date                                      0
total_cases                            7542
                                        ...
human_development_index               37755
excess_mortality_cumulative_absolute 184823
excess_mortality_cumulative          184823
excess_mortality                     184823
excess_mortality_cumulative_per_million 184823
Length: 67, dtype: int64
```

```
data['date']=pd.to_datetime(data['date'])
vaccine['date']=pd.to_datetime(data['date'])
data.drop([ 'new_cases_smoothed','new_deaths_smoothed', 'new_cases_smoothed_per_million',
       'new_deaths_smoothed_per_million', 'reproduction_rate', 'icu_patients',
       'new_tests_smoothed', 'new_tests_smoothed_per_thousand',
       'new_vaccinations_smoothed',
       'new_vaccinations_smoothed_per_million',
       'new_people_vaccinated_smoothed',
       'new_people_vaccinated_smoothed_per_hundred'], axis=1, inplace=True)
data.drop(['icu_patients_per_million','hosp_patients','hosp_patients_per_million','weekly_icu_admissions',

'weekly_icu_admissions_per_million','weekly_hosp_admissions','weekly_hosp_admissions_per_million',

'new_tests_per_thousand','excess_mortality_cumulative_absolute','excess_mortality_cumulative',
'excess_mortality','excess_mortality_cumulative_per_million','stringency_index','life_expectancy','
human_development_index','extreme_poverty',

'cardiovasc_death_rate',

'diabetes_prevalence',

'female_smokers',

'male_smokers',

'handwashing_facilities',

'hospital_beds_per_thousand'],axis= 1,inplace=True)
```

checking for the null values

```
x=data.isnull().sum()*100/len(data)

x
```

```
iso_code                        0.000000

continent                       5.813686

location                       0.000000

date                          0.000000

total_cases                      3.940933

new_cases                         4.052232

total_deaths                     13.590001

new_deaths                        13.593659

total_cases_per_million             4.384040

new_cases_per_million                4.495339

total_deaths_per_million           14.026315

new_deaths_per_million              14.029972

total_tests                     59.408181

new_tests                        61.328484

total_tests_per_thousand            59.408181
```

```
positive_rate                      51.174128
tests_per_case                     52.093784
tests_units                        45.615438
total_vaccinations                 72.625617
people_vaccinated                  73.920972
people_fully_vaccinated            75.245067
total_boosters                     87.223058
new_vaccinations                   77.577126
total_vaccinations_per_hundred     72.625617
people_vaccinated_per_hundred      73.920972
people_fully_vaccinated_per_hundred    75.245067
total_boosters_per_hundred         87.223058
population                          0.608749
population_density                 10.895828
median_age                         17.412842
aged_65_older                      18.288082
aged_70_older                      17.846020
gdp_per_capita                     17.855426
dtype: float64
```

checking for duplicate values

```
duplicate = data[data.duplicated()]
duplicate
```

```
iso_code	continent	location	date	total_cases	new_cases	total_deaths	new_deaths	total_cases_per_million	new_cases_per_million	...	total_vaccinations_per_hundred	people_vaccinated_per_hundred	people_fully_vaccinated_per_hundred	total_boosters_per_hundred	population	population_density	median_age	aged_65_older	aged_70_older	gdp_per_capita
```

```
0 rows × 33 columns
```

```
print(data.isnull().values.any())
```

```
True
```

```
data['total_deaths'].mean()
```

```
64774.858037830774
```

```
data['total_deaths'].median()
```

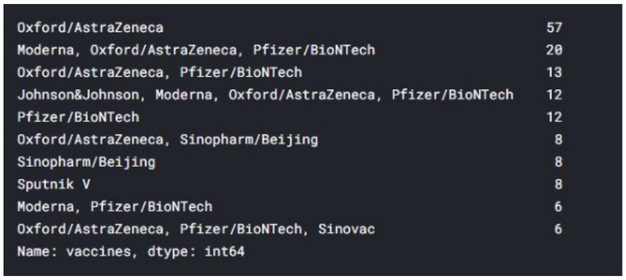
```
917.0
```

```
data['total_deaths'].replace(np.nan,data['total_deaths'].median()).head(10)
```

```
0    917.0
1    917.0
2    917.0
3    917.0
4    917.0
5    917.0
6    917.0
7    917.0
8    917.0
9    917.0
Name: total_deaths, dtype: float64
```

using bfill method to fill nan cells

```
data.fillna(method="bfill")
```



```
Oxford/AstraZeneca                                         57
Moderna, Oxford/AstraZeneca, Pfizer/BioNTech               20
Oxford/AstraZeneca, Pfizer/BioNTech                        13
Johnson&Johnson, Moderna, Oxford/AstraZeneca, Pfizer/BioNTech    12
Pfizer/BioNTech                                            12
Oxford/AstraZeneca, Sinopharm/Beijing                       8
Sinopharm/Beijing                                           8
Sputnik V                                                   8
Moderna, Pfizer/BioNTech                                    6
Oxford/AstraZeneca, Pfizer/BioNTech, Sinovac                6
Name: vaccines, dtype: int64
```

```
data.isnull().values.any() #Checking fo nan values in whole dataframe
```

```
True
```

```
data.head()
```

```
iso_code	continent	location	date	total_cases	new_cases	total_deaths
new_deaths	total_cases_per_million	new_cases_per_million	...
total_vaccinations_per_hundred	people_vaccinated_per_hundred
people_fully_vaccinated_per_hundred	total_boosters_per_hundred	population
population_density	median_age	aged_65_older	aged_70_older	gdp_per_capita

0	AFG	Asia	Afghanistan	2020-02-24	5.0	5.0	NaN	NaN	0.126	0.126
	...	NaN	NaN	NaN	NaN	39835428.0	54.422	18.6	2.581	1.337
1803.987

1	AFG	Asia	Afghanistan	2020-02-25	5.0	0.0	NaN	NaN	0.126	0.000
	...	NaN	NaN	NaN	NaN	39835428.0	54.422	18.6	2.581	1.337
1803.987

2	AFG	Asia	Afghanistan	2020-02-26	5.0	0.0	NaN	NaN	0.126	0.000
	...	NaN	NaN	NaN	NaN	39835428.0	54.422	18.6	2.581	1.337
1803.987

3	AFG	Asia	Afghanistan	2020-02-27	5.0	0.0	NaN	NaN	0.126	0.000
	...	NaN	NaN	NaN	NaN	39835428.0	54.422	18.6	2.581	1.337
1803.987

4	AFG	Asia	Afghanistan	2020-02-28	5.0	0.0	NaN	NaN	0.126	0.000
	...	NaN	NaN	NaN	NaN	39835428.0	54.422	18.6	2.581	1.337
1803.987

5 rows × 33 columns
```

```
data.info(

)
```

```
<class 'pandas.core.frame.DataFrame'>

RangeIndex: 191376 entries, 0 to 191375

Data columns (total 33 columns):

 #   Column                              Non-Null Count   Dtype

---  ------                              --------------   -----

 0   iso_code                            191376 non-null  object

 1   continent                           180250 non-null  object

 2   location                            191376 non-null  object

 3   date                                191376 non-null  datetime64[ns]

 4   total_cases                         183834 non-null  float64

 5   new_cases                           183621 non-null  float64

 6   total_deaths                        165368 non-null  float64

 7   new_deaths                          165361 non-null  float64

 8   total_cases_per_million             182986 non-null  float64

 9   new_cases_per_million               182773 non-null  float64

 10  total_deaths_per_million            164533 non-null  float64

 11  new_deaths_per_million              164526 non-null  float64

 12  total_tests                         77683 non-null   float64

 13  new_tests                           74008 non-null   float64

 14  total_tests_per_thousand            77683 non-null   float64

 15  positive_rate                       93441 non-null   float64

 16  tests_per_case                      91681 non-null   float64

 17  tests_units                         104079 non-null  object

 18  total_vaccinations                  52388 non-null   float64

 19  people_vaccinated                   49909 non-null   float64
 20  people_fully_vaccinated             47375 non-null   float64
 21  total_boosters                      24452 non-null   float64
 22  new_vaccinations                    42912 non-null   float64
 23  total_vaccinations_per_hundred      52388 non-null   float64
 24  people_vaccinated_per_hundred       49909 non-null   float64
 25  people_fully_vaccinated_per_hundred  47375 non-null   float64
 26  total_boosters_per_hundred          24452 non-null   float64
 27  population                          190211 non-null  float64
 28  population_density                  170524 non-null  float64
 29  median_age                          158052 non-null  float64
 30  aged_65_older                       156377 non-null  float64
 31  aged_70_older                       157223 non-null  float64
 32  gdp_per_capita                      157205 non-null  float64
dtypes: datetime64[ns](1), float64(28), object(4)
memory usage: 48.2+ MB
```

```
data.drop(['tests_units'],axis=1,inplace=True)
null_percentage=data.isna().sum()*100/len(data)
null_percentage.head(38)
```

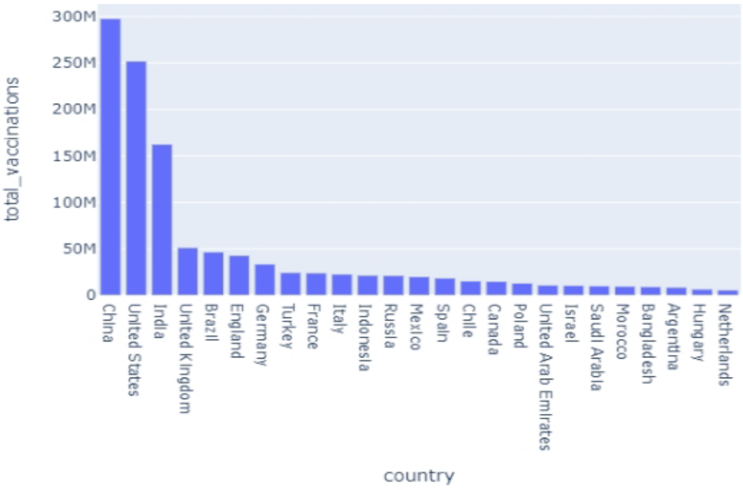
```
iso_code                                0.000000
continent                               5.813686
location                                0.000000
date                                    0.000000
total_cases                             3.940933
new_cases                               4.052232
total_deaths                           13.590001
```

```
new_deaths                       13.593659

total_cases_per_million             4.384040

new_cases_per_million               4.495339

total_deaths_per_million           14.026315

new_deaths_per_million              14.029972

total_tests                     59.408181

new_tests                        61.328484

total_tests_per_thousand            59.408181

positive_rate                    51.174128

tests_per_case                     52.093784

total_vaccinations                 72.625617

people_vaccinated                   73.920972

people_fully_vaccinated             75.245067

total_boosters                    87.223058

new_vaccinations                    77.577126

total_vaccinations_per_hundred        72.625617

people_vaccinated_per_hundred         73.920972

people_fully_vaccinated_per_hundred    75.245067

total_boosters_per_hundred           87.223058

population                       0.608749

population_density                  10.895828

median_age                        17.412842

aged_65_older                      18.288082

aged_70_older                      17.846020

gdp_per_capita                     17.855426
```

```
dtype: float64

data=data.fillna(method="bfill")

null_percentage=data.isna().sum()*100/len(data)


null_percentage.head(38)

iso_code                 0.000000

continent                0.000000

location                0.000000

date                   0.000000

total_cases               0.000000

new_cases                  0.000000

total_deaths              0.000000

new_deaths                 0.000000

total_cases_per_million        0.000000

new_cases_per_million           0.000000

total_deaths_per_million       0.000000

new_deaths_per_million          0.000000

total_tests              0.000523

new_tests                 0.007838

total_tests_per_thousand        0.000523

positive_rate             0.000523

tests_per_case              0.000523

total_vaccinations           0.001045

people_vaccinated             0.001045

people_fully_vaccinated        0.001045
```

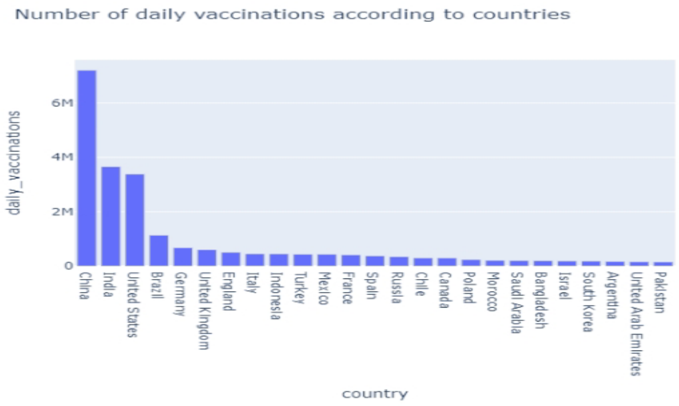
```
total_boosters                      0.001045
new_vaccinations                    0.001045
total_vaccinations_per_hundred      0.001045
people_vaccinated_per_hundred       0.001045
people_fully_vaccinated_per_hundred    0.001045
total_boosters_per_hundred          0.001045
population                          0.000000
population_density                  0.000000
median_age                          0.000000
aged_65_older                       0.000000
aged_70_older                       0.000000
gdp_per_capita                      0.000000
dtype: float64
```

```
data = new_df[['country','total_vaccinations']].nlargest(25,'total_vaccinations')

fig = px.bar(data, x = 'country',y = 'total_vaccinations',title="Number of total vaccinations according to countries",)

fig.show()
```



Covid-19 Vaccination country wise

```
data = new_df[['country','daily_vaccinations']].nlargest(25,'daily_vaccinations')

fig = px.bar(data, x = 'country',y = 'daily_vaccinations',title="Number of daily vaccinations according to countries",)

fig.show()
```


Number of daily vaccinations according to countries
daily_vaccinations
6M
4M
2M
China
India
United States
Brazil
Germany
United Kingdom
England
Italy
Indonesia
Turkey
Mexico
France
Spain
Russia
Chile
Canada
Poland
Morocco
Saudi Arabia
Bangladesh
Israel
South Korea
Argentina
United Arab Emirates
Pakistan
country