Covid19India

April 11, 2020

1 Covid19India - EDA

Data Description The dataset consists of the information about Covid19India cases taken from Covid19India API.

Below is a table showing names of all the columns and their description.

Attributes	Dtype
agebracket	object
backupnotes	object
contractedfromwhichpatientsuspected	object
currentstatus	object
dateannounced	object
detectedcity	object
detecteddistrict	object
detectedstate	object
estimatedonsetdate	object
gender	object
nationality	object
notes	object
patientnumber	object
source1	object
source2	object
source3	object
statecode	object
statepatientnumber	object
statuschangedate	object
typeoftransmission	object

1.1 Import Libraries

[1]:

```
import os
from requests import request
import urllib.request
import json
from pandas.io.json import json_normalize
```

```
import numpy as np
import pandas as pd
import pandas_profiling
import seaborn as sns
import matplotlib.pyplot as plt
import plotly
import plotly.graph_objects as go
import plotly.express as px

%matplotlib inline
```

2 Read Data from Covid19India API

[2]:

```
response=request(url='https://api.covid19india.org/raw_data.json', method='get')
elevations = response.json()
rec = elevations['raw_data']
```

[3]:

```
df = json_normalize(rec)
```

[4]:

```
df.head()
```

[4]:

```
  agebracket                                 backupnotes  \
0        20                          Student from Wuhan
1                                    Student from Wuhan
2                                    Student from Wuhan
3        45         Travel history to Italy and Austria
4        24  Travel history to Dubai, Singapore contact

  contractedfromwhichpatientsuspected currentstatus dateannounced  \
0                                          Recovered    30/01/2020
1                                          Recovered    02/02/2020
2                                          Recovered    03/02/2020
3                                          Recovered    02/03/2020
4                                          Recovered    02/03/2020

               detectedcity detecteddistrict detectedstate estimatedonsetdate  \
0                  Thrissur         Thrissur        Kerala
1                 Alappuzha        Alappuzha        Kerala
2                 Kasaragod        Kasaragod        Kerala
3  East Delhi (Mayur Vihar)       East Delhi         Delhi
4                 Hyderabad        Hyderabad     Telangana

  gender nationality                                notes  \
0      F       India                Travelled from Wuhan
1              India                Travelled from Wuhan
2              India                Travelled from Wuhan
```

```
3      M       India                     Travelled from Austria, Italy
4      M       India  Travelled from Dubai to Bangalore on 20th Feb,...

  patientnumber                                            source1  \
0             1  https://twitter.com/vijayanpinarayi/status/122...
1             2  https://www.indiatoday.in/india/story/kerala-r...
2             3  https://www.indiatoday.in/india/story/kerala-n...
3             4  https://www.indiatoday.in/india/story/not-a-ja...
4             5  https://www.deccanherald.com/national/south/qu...

                                             source2  \
0  https://weather.com/en-IN/india/news/news/2020...
1  https://weather.com/en-IN/india/news/news/2020...
2  https://twitter.com/ANI/status/122422148580539...
3  https://economictimes.indiatimes.com/news/poli...
4  https://www.indiatoday.in/india/story/coronavi...

                                             source3 statecode  \
0                                                           KL
1                                                           KL
2  https://weather.com/en-IN/india/news/news/2020...        KL
3                                                           DL
4  https://www.thehindu.com/news/national/coronav...        TG

  statepatientnumber statuschangedate typeoftransmission
0           KL-TS-P1       14/02/2020           Imported
1           KL-AL-P1       14/02/2020           Imported
2           KL-KS-P1       14/02/2020           Imported
3              DL-P1       15/03/2020           Imported
4              TS-P1       02/03/2020           Imported
```

[5]:
```
df.columns
```

[5]:
```
Index(['agebracket', 'backupnotes', 'contractedfromwhichpatientsuspected',
       'currentstatus', 'dateannounced', 'detectedcity', 'detecteddistrict',
       'detectedstate', 'estimatedonsetdate', 'gender', 'nationality', 'notes',
       'patientnumber', 'source1', 'source2', 'source3', 'statecode',
       'statepatientnumber', 'statuschangedate', 'typeoftransmission'],
      dtype='object')
```

[6]:
```
df.shape
```

[6]:
```
(8067, 20)
```

[7]:
```
data=df.copy()
data.head()
```

[7]:
```
  agebracket                                 backupnotes  \
0         20                          Student from Wuhan
1                                     Student from Wuhan
```

```
2                                     Student from Wuhan
3         45         Travel history to Italy and Austria
4         24  Travel history to Dubai, Singapore contact

  contractedfromwhichpatientsuspected currentstatus dateannounced  \
0                                         Recovered    30/01/2020
1                                         Recovered    02/02/2020
2                                         Recovered    03/02/2020
3                                         Recovered    02/03/2020
4                                         Recovered    02/03/2020

               detectedcity detecteddistrict detectedstate estimatedonsetdate  \
0                  Thrissur         Thrissur        Kerala
1                 Alappuzha        Alappuzha        Kerala
2                 Kasaragod        Kasaragod        Kerala
3  East Delhi (Mayur Vihar)       East Delhi         Delhi
4                 Hyderabad        Hyderabad     Telangana

  gender nationality                                              notes  \
0      F       India                               Travelled from Wuhan
1              India                               Travelled from Wuhan
2              India                               Travelled from Wuhan
3      M       India                      Travelled from Austria, Italy
4      M       India  Travelled from Dubai to Bangalore on 20th Feb,...

  patientnumber                                            source1  \
0             1  https://twitter.com/vijayanpinarayi/status/122...
1             2  https://www.indiatoday.in/india/story/kerala-r...
2             3  https://www.indiatoday.in/india/story/kerala-n...
3             4  https://www.indiatoday.in/india/story/not-a-ja...
4             5  https://www.deccanherald.com/national/south/qu...

                                             source2  \
0  https://weather.com/en-IN/india/news/news/2020...
1  https://weather.com/en-IN/india/news/news/2020...
2  https://twitter.com/ANI/status/122422148580539...
3  https://economictimes.indiatimes.com/news/poli...
4  https://www.indiatoday.in/india/story/coronavi...

                                             source3 statecode  \
0                                                           KL
1                                                           KL
2  https://weather.com/en-IN/india/news/news/2020...        KL
3                                                           DL
4  https://www.thehindu.com/news/national/coronav...        TG

  statepatientnumber statuschangedate typeoftransmission
```

```
0           KL-TS-P1       14/02/2020           Imported
1           KL-AL-P1       14/02/2020           Imported
2           KL-KS-P1       14/02/2020           Imported
3              DL-P1       15/03/2020           Imported
4              TS-P1       02/03/2020           Imported
```

[8]:

```python
profile = pandas_profiling.ProfileReport(df)
profile.to_file(output_file="covid19_data_before_preprocessing.html")
```

Observations - `agebracket` has a high cardinality: 86 distinct values - `backupnotes` has a high cardinality: 223 distinct values - `contractedfromwhichpatientsuspected` has a high cardinality: 144 distinct values - `detectedcity` has a high cardinality: 313 distinct values
- `detecteddistrict` has a high cardinality: 349 distinct values
- `estimatedonsetdate` has constant value as NULL NEEDS TO BE Rejected - `notes` has a high cardinality: 709 distinct values
- `source1` has a high cardinality: 785 distinct values - `source2` has a high cardinality: 338 distinct values - `source3` has a high cardinality: 102 distinct values - `statepatientnumber` has a high cardinality: 1463 distinct values

[9]:

```python
print("Data Shape : Rows = {} , Columns = {}".format(df.shape[0],df.shape[1]))
```

```
Data Shape : Rows = 8067 , Columns = 20
```

[10]:

```python
print("Column Names are : \n", df.columns)
```

```
Column Names are :
 Index(['agebracket', 'backupnotes', 'contractedfromwhichpatientsuspected',
       'currentstatus', 'dateannounced', 'detectedcity', 'detecteddistrict',
       'detectedstate', 'estimatedonsetdate', 'gender', 'nationality', 'notes',
       'patientnumber', 'source1', 'source2', 'source3', 'statecode',
       'statepatientnumber', 'statuschangedate', 'typeoftransmission'],
      dtype='object')
```

[11]:

```python
df.drop(['estimatedonsetdate', 'notes', 'contractedfromwhichpatientsuspected', 'source1', 'source2', 'source3', 'backupnotes' ], axis = 1, inplace = True)
df.sample(10)
```

[11]:

```
     agebracket currentstatus dateannounced detectedcity detecteddistrict  \
5221             Hospitalized    07/04/2020
4026             Hospitalized    05/04/2020                        Mumbai
5127             Hospitalized    07/04/2020                        Mumbai
5746         57  Hospitalized    08/04/2020                      Vadodara
2523             Hospitalized    02/04/2020                         Thane
1560             Hospitalized    31/03/2020                       Chennai
5588             Hospitalized    08/04/2020
6826         57  Hospitalized    10/04/2020                     Bengaluru
6700             Hospitalized    09/04/2020                         Akola
4246         62  Hospitalized    05/04/2020                       Kachchh
```

```
     detectedstate gender nationality patientnumber statecode  \
5221         Delhi                             5222        DL
4026   Maharashtra                             4027        MH
5127   Maharashtra                             5128        MH
5746       Gujarat      M                      5747        GJ
2523   Maharashtra                             2524        MH
1560    Tamil Nadu                             1561        TN
5588     Telangana                             5589        TG
6826     Karnataka      M                      6827        KA
6700   Maharashtra                             6701        MH
4246       Gujarat      M                      4247        GJ

     statepatientnumber statuschangedate typeoftransmission
5221                          07/04/2020
4026                          05/04/2020
5127                          07/04/2020
5746                          08/04/2020
2523                          02/04/2020
1560            TN-P121       31/03/2020                TBD
5588                          08/04/2020
6826            KA-P199       10/04/2020
6700                          09/04/2020
4246                          05/04/2020
```

[12]:
```
#df['agebracket'] = pd.to_numeric(df['agebracket'], errors='coerce')
df['agebracket'] = df['agebracket'].astype('str')
df['patientnumber'] = df['patientnumber'].astype('float')
```

[13]:
```
df['statuschangedate'] = pd.to_datetime(df['statuschangedate'])
df['dateannounced'] = pd.to_datetime(df['dateannounced'])

df['durationOfAnyStatus'] = df['statuschangedate'] - df['dateannounced']
df['durationOfAnyStatus'] = df['durationOfAnyStatus'].dt.days

df['statuschangedate'] = df['statuschangedate'].dt.strftime('%Y-%m-%d')
df['dateannounced'] = df['dateannounced'].dt.strftime('%Y-%m-%d')
```

[14]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8067 entries, 0 to 8066
Data columns (total 14 columns):
agebracket             8067 non-null object
currentstatus          8067 non-null object
dateannounced          8067 non-null object
detectedcity           8067 non-null object
detecteddistrict       8067 non-null object
detectedstate          8067 non-null object
gender                 8067 non-null object
```

```
nationality            8067 non-null object
patientnumber          8067 non-null float64
statecode              8067 non-null object
statepatientnumber     8067 non-null object
statuschangedate       8067 non-null object
typeoftransmission     8067 non-null object
durationOfAnyStatus    7766 non-null float64
dtypes: float64(2), object(12)
memory usage: 882.4+ KB
```

[15]:

```
df.sample(10)
```

[15]:

```
     agebracket currentstatus dateannounced detectedcity detecteddistrict  \
5544         32  Hospitalized    2020-08-04                        Kannur
362          58  Hospitalized    2020-03-22                     Kasaragod
5131             Hospitalized    2020-07-04                        Mumbai
3599             Hospitalized    2020-04-04                     Osmanabad
3891             Hospitalized    2020-05-04                  The Nilgiris
3728             Hospitalized    2020-05-04                         Dausa
2290             Hospitalized    2020-02-04                  Thoothukkudi
6094             Hospitalized    2020-09-04                        Mumbai
1904             Hospitalized    2020-03-31
650          18  Hospitalized    2020-03-25      Chennai          Chennai

     detectedstate gender nationality  patientnumber statecode  \
5544        Kerala              India         5545.0        KL
362         Kerala      M       India          363.0        KL
5131   Maharashtra                            5132.0        MH
3599   Maharashtra                            3600.0        MH
3891    Tamil Nadu      M                     3892.0        TN
3728     Rajasthan                            3729.0        RJ
2290    Tamil Nadu                            2291.0        TN
6094   Maharashtra                            6095.0        MH
1904   West Bengal                            1905.0        WB
650     Tamil Nadu      M       India          651.0        TN

     statepatientnumber statuschangedate typeoftransmission  \
5544                          2020-08-04
362                           2020-03-22           Imported
5131                          2020-07-04
3599                                 NaT
3891            TN-P541       2020-05-04              Local
3728                          2020-05-04
2290            TN-P296       2020-02-04              Local
6094                          2020-09-04
1904                          2020-03-31
650              TN-P24       2020-03-25              Local
```

```
      durationOfAnyStatus
5544                  0.0
362                   0.0
5131                  0.0
3599                  NaN
3891                  0.0
3728                  0.0
2290                  0.0
6094                  0.0
1904                  0.0
650                   0.0
```

[16]:
```
profile = pandas_profiling.ProfileReport(df)
profile.to_file(output_file="covid19_data_after_preprocessing.html")
```

Observations

- Dataset info

Data	Info
Number of variables	14
Number of observations	8067
Missing cells	301 (0.3%)
Duplicate rows	0 (0.0%)
Total size in memory	882.4 KiB

- Variables types

Varibale	Count
Numeric	2
Categorical	12

- `agebracket` has a high cardinality: 86 distinct values
- `detectedcity` has a high cardinality: 314 distinct values
- `detecteddistrict` has a high cardinality: 349 distinct values
- `durationOfAnyStatus` has 7579 (94.0%) zeros
- `durationOfAnyStatus` has 301 (3.7%) missing values
- `statepatientnumber` has a high cardinality: 1463 distinct values
- `currentstatus` distribution

Value	Count	Frequency (%)
Hospitalized	7706	95.5%
Unknown	192	2.4%
Recovered	137	1.7%
Deceased	31	0.4%
Migrated	1	< 0.1%

- `typeoftransmission` distribution

Value	Count	Frequency (%)
Unknown	5233	64.9%
Local	1606	19.9%
TBD	630	7.8%
Imported	596	7.4%

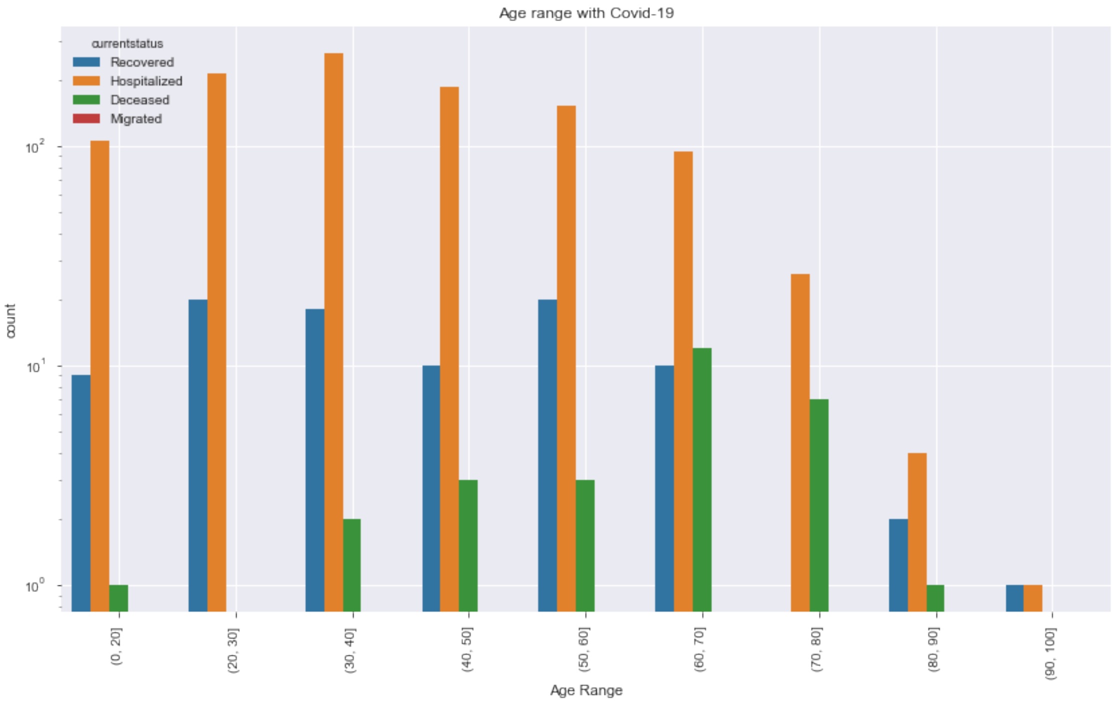
[17]:

```
df['agebracket'] = pd.to_numeric(df['agebracket'], errors='coerce')
```

2.1 Age range distribution with Covid-19

[18]:

```
age = df['agebracket']
status = df['currentstatus']
age_bins = [0,20,30,40,50,60,70,80,90,100]
plt.figure(figsize=(14,8))
sns.countplot(x=pd.cut(age, age_bins), hue=status)
plt.xticks(rotation=90)
plt.xlabel("Age Range")
plt.yscale('log')
plt.title("Age range with Covid-19")
plt.grid(True)
plt.show()
```



2.2 Covid-19 Cases Distribution across States

[19]:
```
state = df.groupby('detectedstate').count()
fig = px.pie(state, values='currentstatus', names=state.index
             ,color_discrete_sequence=px.colors.sequential.
 ↪Plasma_r,title='Covid19 cases based on State')
fig.update_traces(textposition='outside', textinfo='value+label')
fig.show()
```

2.3 Covid-19 cases distribution based on Nationality

[20]:
```
nationality = df.groupby('nationality').count()
fig = px.pie(nationality, values='currentstatus', names=nationality.index
             ,color_discrete_sequence=px.colors.qualitative.G10,title='Covid19␣
 ↪cases based on Nationality in India')
fig.update_traces(textposition='outside', textinfo='value+label')
fig.show()
```

2.4 No. of foreign citizens affected by Covid-19 in India

[21]:
```
temp = df.groupby('nationality')['patientnumber'].count().reset_index()
temp = temp.sort_values('patientnumber')
temp = temp[temp['nationality']!='']
temp = temp[temp['nationality']!='India']
fig = px.bar(temp, x='patientnumber', y='nationality', orientation='h',␣
 ↪text='patientnumber', width=600,
       color_discrete_sequence = ['#35495e'], title='No. of foreign citizens')
fig.update_xaxes(title='')
fig.update_yaxes(title='')
fig.show()
```

2.5 Covid-19 distribution based on Type of Transmission

[22]:
```
temp = pd.DataFrame(df[['typeoftransmission']].
 ↪groupby('typeoftransmission')['typeoftransmission'].count())
temp = temp.dropna()
temp.columns = ['count']
temp = temp.reset_index().sort_values(by='count')

fig = px.bar(temp, x='count', y='typeoftransmission', orientation='h',␣
 ↪text='count', width=600, height=300,
       color_discrete_sequence = ['#35495e'], title='Type of transmission')
fig.update_xaxes(title='')
fig.update_yaxes(title='')
fig.show()
```

2.6 Covid-19 cases Vs Age Brackets along with current status

[23]:
```
fig = plotly.subplots.make_subplots(
    rows=1, cols=2, column_widths=[0.8, 0.2],
    subplot_titles = ['Cases vs Age', ''],
    specs=[[{"type": "histogram"}, {"type": "pie"}]]
)

temp = df[['agebracket', 'currentstatus']].dropna()
print('Total no. of values :', df.shape[0], '\nNo. of missing values :', df.
 ↪shape[0]-temp.shape[0], '\nNo. of available values :', df.shape[0]-(df.
 ↪shape[0]-temp.shape[0]))
gen_grp = temp.groupby('currentstatus').count()

fig.add_trace(go.Pie(values=gen_grp.values.reshape(-1).tolist(),␣
 ↪labels=['Deceased', 'Hospitalized', 'Recovered'],
                     marker_colors = ['#fd0054', '#393e46', '#40a798'], hole=.
 ↪3),1, 2)
```

```
fig.add_trace(go.
 ↪Histogram(x=temp[temp['currentstatus']=='Deceased']['agebracket'],␣
 ↪nbinsx=50, name='Deceased', marker_color='#fd0054'), 1, 1)
fig.add_trace(go.
 ↪Histogram(x=temp[temp['currentstatus']=='Recovered']['agebracket'],␣
 ↪nbinsx=50, name='Recovered', marker_color='#40a798'), 1, 1)
fig.add_trace(go.
 ↪Histogram(x=temp[temp['currentstatus']=='Hospitalized']['agebracket'],␣
 ↪nbinsx=50, name='Hospitalized', marker_color='#393e46'), 1, 1)

fig.update_layout(showlegend=False)
fig.update_layout(barmode='stack')
fig.data[0].textinfo = 'label+text+value+percent'

fig.show()
```

```
Total no. of values : 8067
No. of missing values : 6901
No. of available values : 1166
```

2.7 Covid-19 cases Gender Vs Age Brackets along with gender distribution

[24]:

```
fig = plotly.subplots.make_subplots(
    rows=1, cols=2, column_widths=[0.8, 0.2],
    subplot_titles = ['Gender vs Age', ''],
    specs=[[{"type": "histogram"}, {"type": "pie"}]]
)

temp = df[['agebracket', 'gender']].dropna()
print('Total no. of values :', df.shape[0], '\nNo. of missing values :', df.
 ↪shape[0]-temp.shape[0], '\nNo. of available values :', df.shape[0]-(df.
 ↪shape[0]-temp.shape[0]))
gen_grp = temp.groupby('gender').count()

fig.add_trace(go.Histogram(x=temp[temp['gender']=='F']['agebracket'],␣
 ↪nbinsx=50, name='Female', marker_color='#6a0572'), 1, 1)
fig.add_trace(go.Histogram(x=temp[temp['gender']=='M']['agebracket'],␣
 ↪nbinsx=50, name='Male', marker_color='#39065a'), 1, 1)

fig.add_trace(go.Pie(values=gen_grp.values.reshape(-1).tolist(),␣
 ↪labels=['Female', 'Male'], marker_colors = ['#6a0572', '#39065a']),1, 2)

fig.update_layout(showlegend=False)
fig.update_layout(barmode='stack')
fig.data[2].textinfo = 'label+text+value+percent'
fig.show()
```

```
Total no. of values : 8067
No. of missing values : 6901
No. of available values : 1166
```

2.8 Covid-19 cases Age distribution of confirmed patients

[25]:
```
print('Total no. of values :', df.shape[0], '\nNo. of missing values :', df.
 ↪shape[0]-df[['agebracket']].dropna().shape[0],
      '\nNo. of available values :', df.shape[0]-(df.
 ↪shape[0]-df[['agebracket']].dropna().shape[0]))
px.histogram(df, x='agebracket', color_discrete_sequence = ['#35495e'],␣
 ↪nbins=50,
             title='Distribution of ages of confirmed patients')
```

```
Total no. of values : 8067
No. of missing values : 6901
No. of available values : 1166
```

2.9 Covid-19 cases distribution across states

[26]:
```
dist = df.groupby(['detectedstate', 'detecteddistrict'])['patientnumber'].
 ↪count().reset_index()
dist.head()
fig = px.treemap(dist, path=["detectedstate", "detecteddistrict"],␣
 ↪values="patientnumber", height=700,
           title='Number of Confirmed Cases', color_discrete_sequence = px.
 ↪colors.qualitative.Prism)
fig.data[0].textinfo = 'label+text+value'
fig.show()
```

[27]:
```
df['statuschangedate'] = pd.to_datetime(df['statuschangedate'])
df['dateannounced'] = pd.to_datetime(df['dateannounced'])
```

[28]:
```
temp = df[['dateannounced', 'statuschangedate', 'currentstatus']].dropna()
temp = temp[temp['statuschangedate']!=temp['dateannounced']]
temp['no_of_days'] = temp['statuschangedate'] - temp['dateannounced']
temp['no_of_days'] = temp['no_of_days'].dt.days
temp = temp[temp['no_of_days']>0]
temp.head()
```

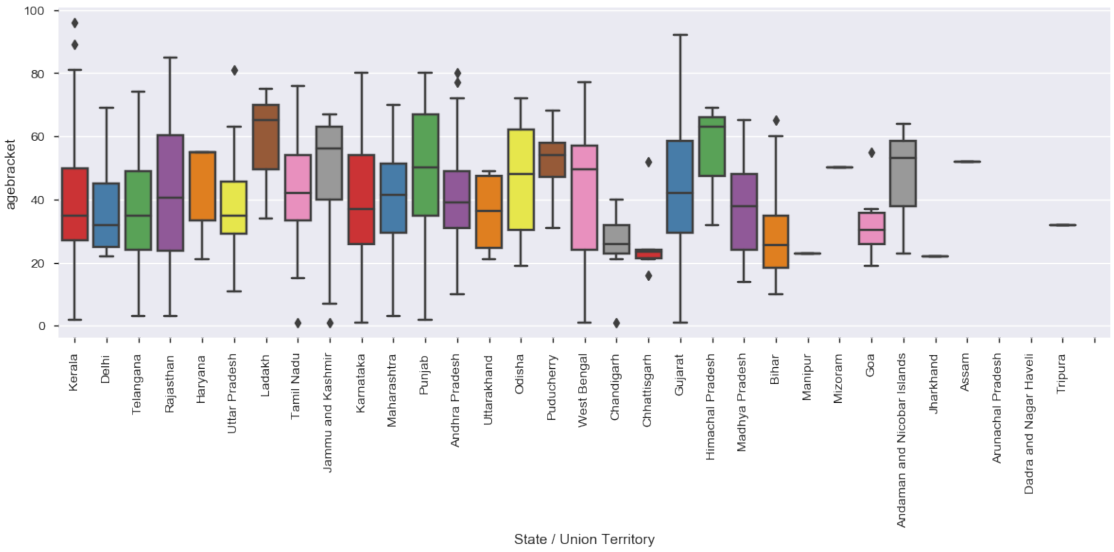
[28]:
```
   dateannounced statuschangedate currentstatus  no_of_days
0     2020-01-30       2020-02-14     Recovered          15
1     2020-02-02       2020-02-14     Recovered          12
3     2020-02-03       2020-03-15     Recovered          41
77    2020-11-03       2020-12-03   Hospitalized         30
```

```
 84     2020-03-13        2020-03-24      Recovered           11
```

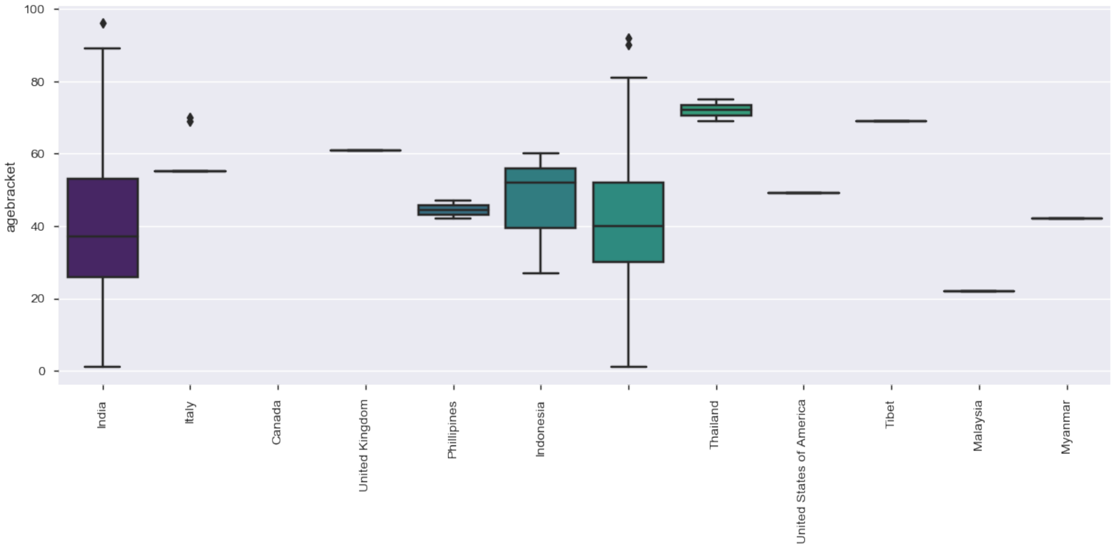
[29]:
```python
print('Total no. of values :', df.shape[0], '\nNo. of missing values :', df.
 ↪shape[0]-temp.shape[0], '\nNo. of available values :', df.shape[0]-(df.
 ↪shape[0]-temp.shape[0]))
px.box(temp, x="currentstatus", y="dateannounced", color='currentstatus')
```

```
Total no. of values : 8067
No. of missing values : 7996
No. of available values : 71
```

[30]:
```python
plt.figure(figsize=(12, 6), dpi = 100)
sns.boxplot(x = 'detectedstate', y = 'agebracket', data = df, palette = 'Set1')
plt.xlabel('State / Union Territory')
plt.ylabel('agebracket')
plt.xticks(rotation = 90)
plt.tight_layout()
plt.show()
```



[31]:
```python
plt.figure(figsize=(12, 6), dpi = 100)
sns.boxplot(x = 'nationality', y = 'agebracket', data = df, palette = 'viridis')
plt.xlabel('')
plt.xticks(rotation=90)
plt.ylabel('agebracket')
plt.tight_layout()
plt.show()
```


100
80
60
40
20
0
agebracket
India
Italy
Canada
United Kingdom
Phillipines
Indonesia
Thailand
United States of America
Tibet
Malaysia
Myanmar