In [3]:
```
import pandas as pd
```

In [5]:
```
df=pd.read_csv(r"C:\Users\Pooja Shinde\Downloads\covid_19_data.csv")
```

In [7]:
```
df
```

Out[7]:

	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
...	...	...	...	...	...	...	...	...
306424	306425	05/29/2021	Zaporizhia Oblast	Ukraine	2021-05-30 04:20:55	102641.0	2335.0	95289.0
306425	306426	05/29/2021	Zeeland	Netherlands	2021-05-30 04:20:55	29147.0	245.0	0.0
306426	306427	05/29/2021	Zhejiang	Mainland China	2021-05-30 04:20:55	1364.0	1.0	1324.0
306427	306428	05/29/2021	Zhytomyr Oblast	Ukraine	2021-05-30 04:20:55	87550.0	1738.0	83790.0
306428	306429	05/29/2021	Zuid-Holland	Netherlands	2021-05-30 04:20:55	391559.0	4252.0	0.0

306429 rows × 8 columns

In [ ]:
```
### Python Project :
#Dataset : Corona_virus
#Analyse the data and give the answers of below questions :

#1.What is the total number of confirmed cases worldwide?
#2.How many deaths have been reported globally?
#3.What is the total number of recovered cases worldwide?
#4.How many countries/regions are represented in the dataset?
#5.What is the trend of confirmed cases over time globally?
#6.Which province/state has reported the highest number of confirmed cases?
#7.Which country/region has the highest number of deaths?
#8.How does the number of confirmed cases vary across different provinces/states?
#9.What is the trend of deaths over time globally?
#10.Which country/region has the highest number of recovered cases?
#11.How does the number of recovered cases vary across different countries/regions?
#12.What is the distribution of confirmed cases by country/region?
#13.Is there a correlation between the number of confirmed cases and deaths?
#14.Is there a correlation between the number of confirmed cases and recovered cases?
#15.How does the mortality rate vary across different countries/regions?
#16.How does the recovery rate vary across different countries/regions?
#17.What is the trend of new confirmed cases over time globally?
#18.How does the fatality rate vary across different provinces/states?
#19.How does the recovery rate vary across different provinces/states?
#20.What is the trend of active cases over time globally?
```

In [15]:
```
#1.What is the total number of confirmed cases worldwide?
a=df['Confirmed'].sum()
print('the total no. of confirmed cases worldwide are:',a)
```

```
the total no. of confirmed cases worldwide are: 26252051758.0
```

In [7]:
```
#2.How many deaths have been reported globally?
a=df['Deaths'].sum()
print('the globally deaths reported are:',a)
```

```
the globally deaths reported are: 624013017.0
```

In [19]:
```
#3.What is the total number of recovered cases worldwide?
a=df['Recovered'].sum()
print('the total no. of recovered cases worldwide are:',a)
```

```
the total no. of recovered cases worldwide are: 15450237912.0
```

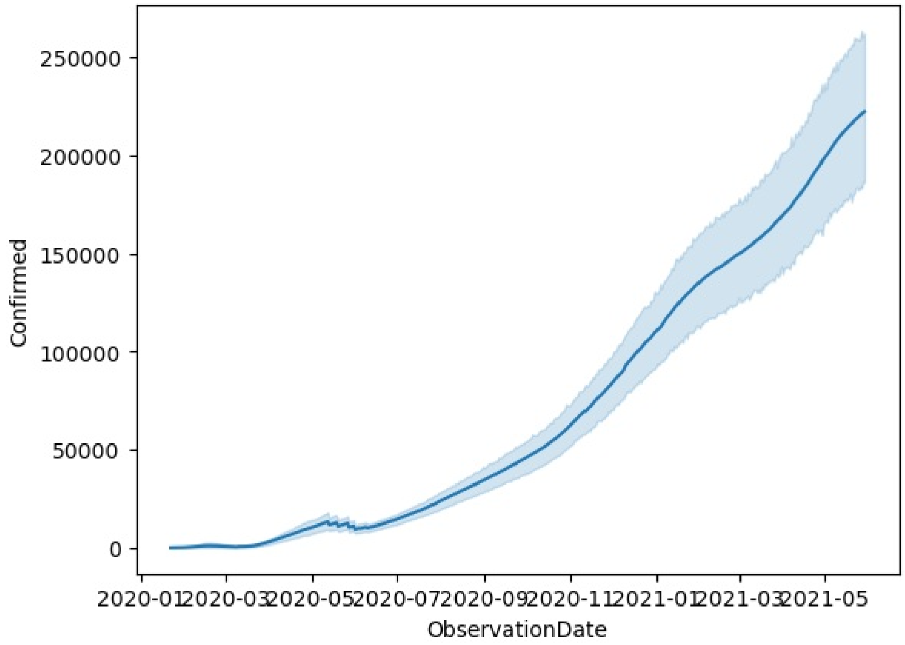
In [9]:
```
#4.How many countries/regions are represented in the dataset?
x=df['Country/Region'].nunique()
print('the countries/regions are:',x)
```

```
the countries/regions are: 229
```

In [15]:
```
import seaborn as sns
#5.What is the trend of confirmed cases over time globally?
df['ObservationDate']=pd.to_datetime(df['ObservationDate'])
global_trend=df.groupby('ObservationDate')['Confirmed'].sum()
```

```
sns.lineplot(data=df,x='ObservationDate',y='Confirmed')
```

```
Out[15]: <Axes: xlabel='ObservationDate', ylabel='Confirmed'>
```



```
In [29]: #6.Which province/state has reported the highest number of confirmed cases?
         b=df.groupby('Province/State')['Confirmed'].sum().reset_index()
         highest=b.loc[b['Confirmed'].idxmax()]
         highest
```

```
Out[29]: Province/State     California
         Confirmed         696898013.0
         Name: 88, dtype: object
```

```
In [37]: #7.Which country/region has the highest number of deaths?
         b=df.groupby('Country/Region')['Deaths'].sum().reset_index()
         highest_deaths=b.loc[b['Deaths'].idxmax()]
         highest_deaths
```

```
Out[37]: Country/Region              US
         Deaths            123303762.0
         Name: 214, dtype: object
```

```
In [61]: #8.How does the number of confirmed cases vary across different provinces/states?
         statewise_cases=df.groupby('Province/State')['Confirmed'].sum().reset_index()
         statewise_cases=statewise_cases.sort_values(by='Confirmed',ascending=False)
         statewise_cases
```

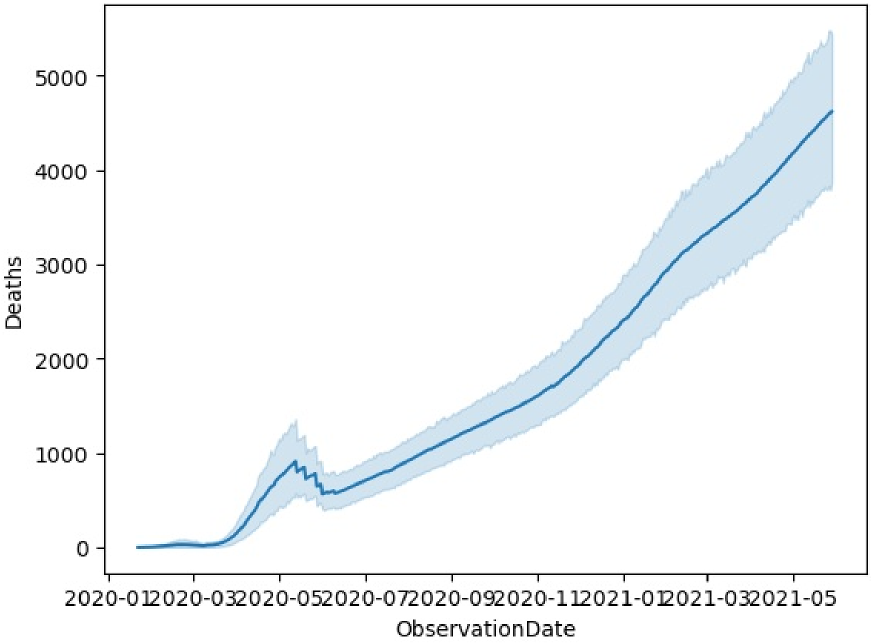
Out[61]:

	Province/State	Confirmed
88	California	696898013.0
365	Maharashtra	681186928.0
171	England	666227518.0
630	Texas	552039886.0
570	Sao Paulo	521308945.0
...	...	...
17	American Samoa	0.0
278	Jervis Bay Territory	0.0
526	Recovered	0.0
404	Montgomery County, TX	0.0
173	External territories	0.0

736 rows × 2 columns

```
In [39]: #9.What is the trend of deaths over time globally?
         df['ObservationDate']=pd.to_datetime(df['ObservationDate'])
         global_trend=df.groupby('ObservationDate')['Deaths'].sum()
         sns.lineplot(data=df,x='ObservationDate',y='Deaths')
```

```
Out[39]: <Axes: xlabel='ObservationDate', ylabel='Deaths'>
```



In [45]:

```
#10.Which country/region has the highest number of recovered cases?
b=df.groupby(by='Country/Region')['Recovered'].sum().reset_index()
highest_recovered=b.loc[b['Recovered'].idxmax()]
highest_recovered
```

Out[45]:

```
Country/Region            India
Recovered          2900589824.0
Name: 96, dtype: object
```

In [7]:

```
#11.How does the number of recovered cases vary across different countries/regions?
country_recoveries = df.groupby('Country/Region')['Recovered'].sum().reset_index()
country_recoveries.sort_values(by='Recovered', ascending=False)
```

Out[7]:

	Country/Region	Recovered
96	India	2.900590e+09
27	Brazil	2.313677e+09
172	Russia	7.907057e+08
212	Turkey	5.641706e+08
214	US	5.033710e+08
...	...	...
166	Puerto Rico	0.000000e+00
168	Republic of Ireland	0.000000e+00
169	Republic of the Congo	0.000000e+00
170	Reunion	0.000000e+00
228	occupied Palestinian territory	0.000000e+00

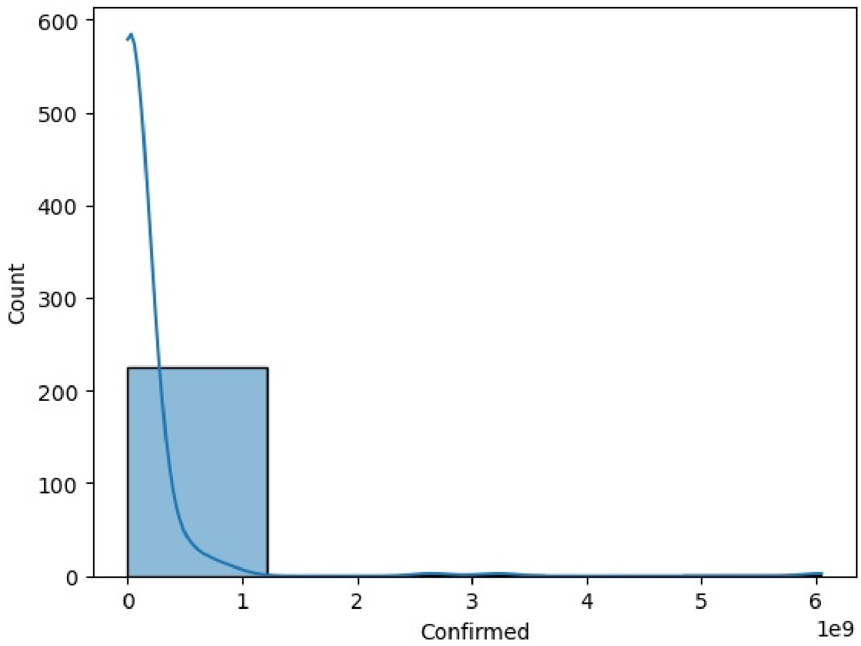
229 rows × 2 columns

In [51]:

```
#12.What is the distribution of confirmed cases by country/region?
country_cases=df.groupby('Country/Region')['Confirmed'].sum().reset_index()
sns.histplot(data=country_cases,x='Confirmed',bins=5,kde=True)
```

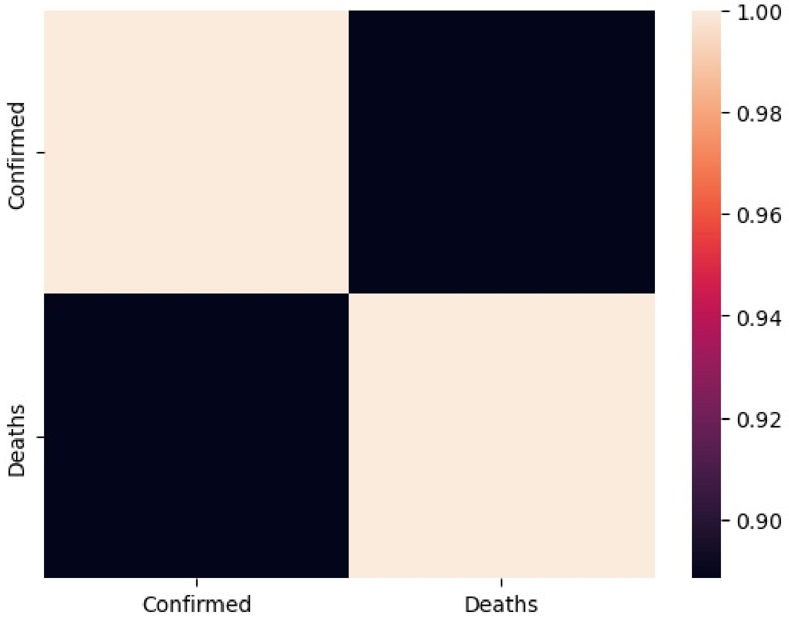
Out[51]:

```
<Axes: xlabel='Confirmed', ylabel='Count'>
```



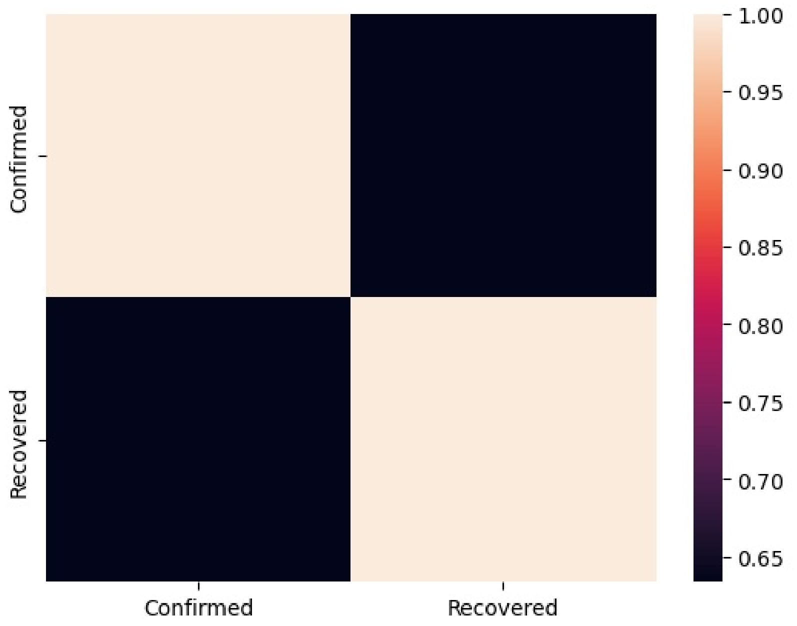
```
In [63]: #13.Is there a correlation between the number of confirmed cases and deaths?
         df[['Confirmed','Deaths']].corr()
         sns.heatmap(df[['Confirmed','Deaths']].corr())
```

```
Out[63]: <Axes: >
```



```
In [65]: #14.Is there a correlation between the number of confirmed cases and recovered cases?
         df[['Confirmed','Recovered']].corr()
         sns.heatmap(df[['Confirmed','Recovered']].corr())
```

```
Out[65]: <Axes: >
```



In [9]:

```
#15.How does the mortality rate vary across different countries/regions?
country_data=df.groupby('Country/Region')[['Confirmed','Deaths']].sum().reset_index()
country_data=country_data[country_data['Confirmed']>0]
country_data['Mortality Rate(%)']=(country_data['Deaths']/country_data['Confirmed'])*100
country_data_sorted=country_data.sort_values(by='Mortality Rate(%)',ascending=False)
country_data_sorted.head(100)
```

Out[9]:

	Country/Region	Confirmed	Deaths	Mortality Rate(%)
225	Yemen	962066.0	237613.0	24.698202
123	MS Zaandam	3824.0	848.0	22.175732
220	Vanuatu	406.0	39.0	9.605911
137	Mexico	460463678.0	43005509.0	9.339609
197	Sudan	7632455.0	488709.0	6.403038
...	...	...	...	...
143	Morocco	104557135.0	1823724.0	1.744237
35	Cameroon	11346589.0	197906.0	1.744189
109	Kenya	27728648.0	482736.0	1.740929
12	Austria	97965875.0	1678309.0	1.713157
93	Hong Kong	2655935.0	45325.0	1.706555

100 rows × 4 columns

In [11]:

```
#16.How does the recovery rate vary across different countries/regions?
country_data=df.groupby('Country/Region')[['Confirmed','Recovered']].sum().reset_index()
country_data=country_data[country_data['Confirmed']>0]
country_data['Recovery Rate(%)']=(country_data['Recovered']/country_data['Confirmed'])*100
country_data_sorted=country_data.sort_values(by='Recovery Rate(%)',ascending=False)
country_data_sorted.head(100)
```

Out[11]:

	Country/Region	Confirmed	Recovered	Recovery Rate(%)
219	Uzbekistan	22207571.0	21035683.0	94.723025
78	Ghana	20784664.0	19586296.0	94.234364
55	Diamond Princess	306872.0	288580.0	94.039209
138	Micronesia	129.0	121.0	93.798450
95	Iceland	1729527.0	1621682.0	93.764480
...	...	...	...	...
97	Indonesia	265186050.0	226416174.0	85.380122
151	Nigeria	33407947.0	28514090.0	85.351219
150	Niger	1047041.0	892393.0	85.229996
172	Russia	930548849.0	790705716.0	84.971973
165	Portugal	141962632.0	120619045.0	84.965348

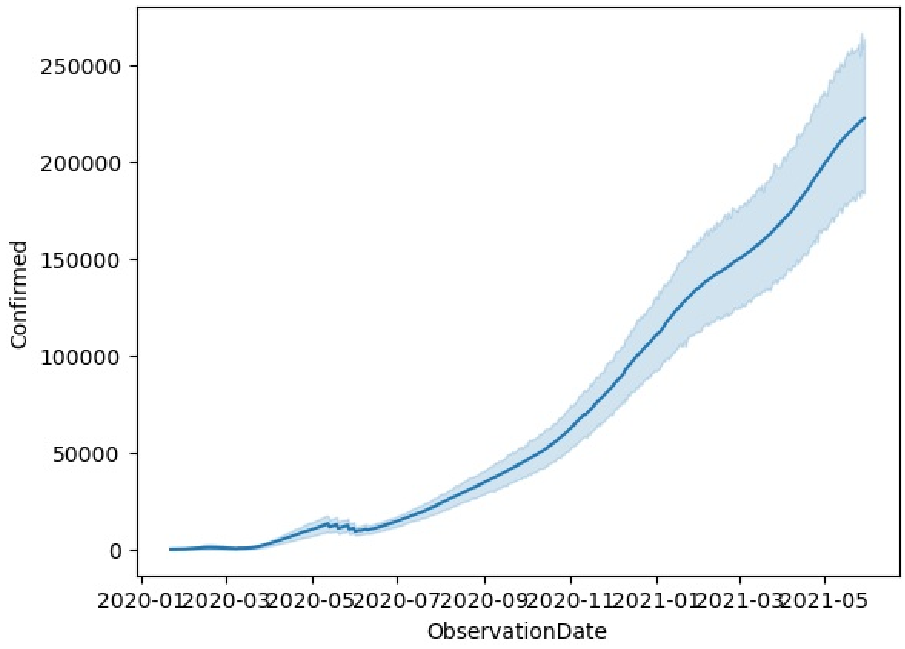
100 rows × 4 columns

In [18]:

```
#17.What is the trend of new confirmed cases over time globally?
df['ObservationDate']=pd.to_datetime(df['ObservationDate'])
global_trend=df.groupby('ObservationDate')['Confirmed'].sum().reset_index()
global_trend.sort_values('ObservationDate')
sns.lineplot(data=df,x='ObservationDate',y='Confirmed')
```

Out[18]:

```
<Axes: xlabel='ObservationDate', ylabel='Confirmed'>
```



In [28]:

```
#18.How does the fatality rate vary across different provinces/states?
country_data=df.groupby('Province/State')[['Confirmed','Deaths']].sum().reset_index()
country_data=country_data[country_data['Confirmed']>0]
country_data['Mortality Rate(%)']=(country_data['Deaths']/country_data['Confirmed'])*100
country_data_sorted=country_data.sort_values(by='Mortality Rate(%)',ascending=False)
country_data_sorted.head(100)
```

Out[28]:

	Province/State	Confirmed	Deaths	Mortality Rate(%)
668	Unknown	7804169.0	4247616.0	54.427525
568	Santa Rosa County, FL	5.0	2.0	40.000000
338	Lee County, FL	6.0	2.0	33.333333
314	King County, WA	412.0	91.0	22.087379
499	Placer County, CA	28.0	6.0	21.428571
...	...	...	...	...
492	Perm Krai	9426150.0	341592.0	3.623876
601	Sormland	3171168.0	114559.0	3.612518
171	England	666227518.0	24042130.0	3.608697
95	Caqueta	4000046.0	144258.0	3.606409
287	Junin	10474345.0	372782.0	3.559001

100 rows × 4 columns

In [22]:

```
#19.How does the recovery rate vary across different provinces/states?
country_data=df.groupby('Province/State')[['Confirmed','Recovered']].sum().reset_index()
country_data=country_data[country_data['Confirmed']>0]
country_data['Recovery Rate(%)']=(country_data['Recovered']/country_data['Confirmed'])*100
country_data_sorted=country_data.sort_values(by='Recovery Rate(%)',ascending=False)
country_data_sorted.head(100)
```

Out[22]:

	Province/State	Confirmed	Recovered	Recovery Rate(%)
656	US	4.0	532.0	13300.000000
668	Unknown	7804169.0	619474280.0	7937.735331
527	Repatriated Travellers	2431.0	2431.0	100.000000
549	Saint Helena, Ascension and Tristan da Cunha	882.0	863.0	97.845805
439	Ningxia	35904.0	34976.0	97.415330
...	...	...	...	...
82	Buryatia Republic	6975683.0	6404564.0	91.812716
626	Telangana	84952428.0	77992105.0	91.806799
694	Volgograd Oblast	10840874.0	9952071.0	91.801371
709	West Bengal	154070425.0	141428487.0	91.794702
97	Casanare	2583680.0	2371658.0	91.793798

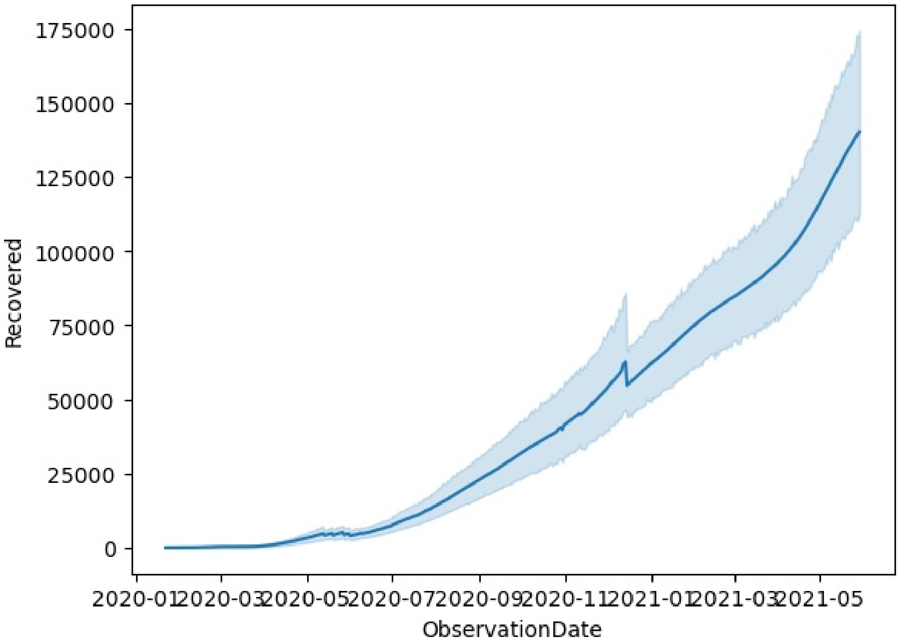
100 rows × 4 columns

In [55]:

```
#20.What is the trend of active cases over time globally?
df['ObservationDate']=pd.to_datetime(df['ObservationDate'])
global_trend=df.groupby('ObservationDate')['Recovered'].sum()
sns.lineplot(data=df,x='ObservationDate',y='Recovered')
```

Out[55]:

```
<Axes: xlabel='ObservationDate', ylabel='Recovered'>
```



```
In [ ]:
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js