In [1]:
```
import pandas as pd
covidData = pd.read_csv("Datasets/covid.csv")
```

In [2]:
```
covidData.head()
```

Out[2]:

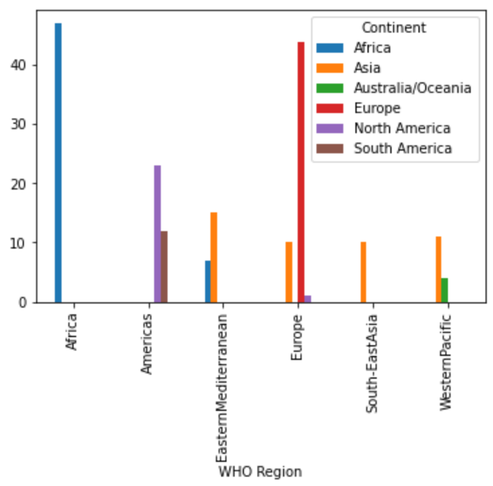
	Country/Region	Continent	Population	TotalCases	NewCases	TotalDeaths	NewDeaths	TotalRecovered	NewRecovered	ActiveCases	Serious,Critical	Tot Cases/1M pop	Deaths/1M pop	TotalTests	Tests/1M pop	WHO Region
0	USA	North America	3.311981e+08	5032179	NaN	162804.0	NaN	2576668.0	NaN	2292707.0	18296.0	15194.0	492.0	63139605.0	190640.0	Americas
1	Brazil	South America	2.127107e+08	2917562	NaN	98644.0	NaN	2047660.0	NaN	771258.0	8318.0	13716.0	464.0	13206188.0	62085.0	Americas
2	India	Asia	1.381345e+09	2025409	NaN	41638.0	NaN	1377384.0	NaN	606387.0	8944.0	1466.0	30.0	22149351.0	16035.0	South-EastAsia
3	Russia	Europe	1.459409e+08	871894	NaN	14606.0	NaN	676357.0	NaN	180931.0	2300.0	5974.0	100.0	29716907.0	203623.0	Europe
4	South Africa	Africa	5.938157e+07	538184	NaN	9604.0	NaN	387316.0	NaN	141264.0	539.0	9063.0	162.0	3149807.0	53044.0	Africa

Expectations

I expect these 2 to be very similar, since theyre almost the same. This dataset didnt have that many non numerical columns, so these were the best i could compare.

In [3]:
```
covidData.groupby(['Continent','WHO Region']).size().unstack('Continent', fill_value=0).plot(kind='bar')
```

Out[3]:
```
<AxesSubplot:xlabel='WHO Region'>
```



In [4]:
```
covidData.groupby(['Continent','WHO Region']).size().unstack('Continent', fill_value=0)
```

Out[4]:

Continent	Africa	Asia	Australia/Oceania	Europe	North America	South America
WHO Region						
Africa	47	0	0	0	0	0
Americas	0	0	0	0	23	12
EasternMediterranean	7	15	0	0	0	0
Europe	0	10	0	44	1	0
South-EastAsia	0	10	0	0	0	0
WesternPacific	0	11	4	0	0	0

In [5]:
```
from scipy.stats import chi2_contingency
```

In [6]:
```
chi2_contingency(covidData.groupby(['Continent','WHO Region']).size().unstack('Continent', fill_value=0))
```

Out[6]:
```
(479.55382395382395,
 1.3152901775915486e-85,
 25,
 array([[13.79347826, 11.75      ,  1.02173913, 11.23913043,  6.13043478,
          3.06521739],
        [10.27173913,  8.75      ,  0.76086957,  8.36956522,  4.56521739,
          2.2826087 ],
        [ 6.45652174,  5.5       ,  0.47826087,  5.26086957,  2.86956522,
          1.43478261],
        [16.14130435, 13.75      ,  1.19565217, 13.15217391,  7.17391304,
          3.58695652],
        [ 2.93478261,  2.5       ,  0.2173913 ,  2.39130435,  1.30434783,
          0.65217391],
        [ 4.40217391,  3.75      ,  0.32608696,  3.58695652,  1.95652174,
          0.97826087]]))
```

There seems to be a 132% chance the columns arent the same. Since this number is higher than 100% I pressume this means

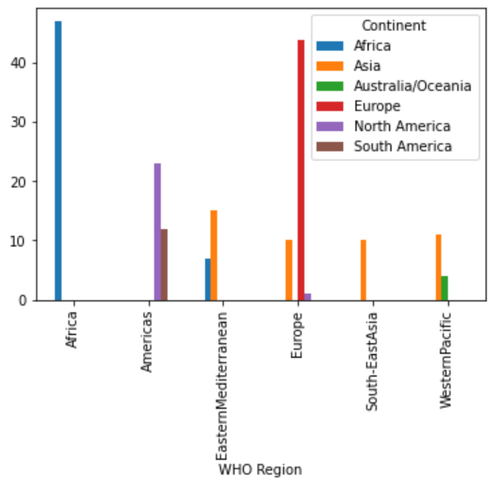
In [8]:
```
contingencyTableCovid = covidData.groupby(['Continent','WHO Region']).size().unstack('Continent', fill_value=0)
contingencyTableCovid
```

Out[8]:

Continent	Africa	Asia	Australia/Oceania	Europe	North America	South America
WHO Region						
Africa	47	0	0	0	0	0
Americas	0	0	0	0	23	12
EasternMediterranean	7	15	0	0	0	0
Europe	0	10	0	44	1	0
South-EastAsia	0	10	0	0	0	0
WesternPacific	0	11	4	0	0	0

In [9]:
```
contingencyTableCovid.plot(kind='bar')
```

Out[9]:
```
<AxesSubplot:xlabel='WHO Region'>
```



I understand how this works. In the example notebooks there was an extra example which showed multiple rolls of a dice. In my dataset this isnt something that has to be done, but this shows you howmany times each continent would have gotten a certain amount.