Programming Assignment 2

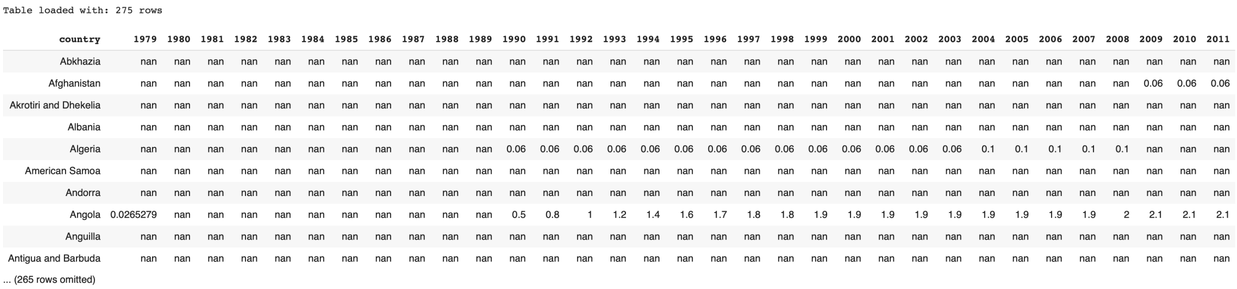
INFO-5502 (Section 002):
Analytic Tools, Techniques and Methods

Ramandeep Harjai

February 13, 2021

```
!pip install datascience

from datascience import Table
import matplotlib.pyplot as plt
import math
```

```
# load HIV data
data_url = 'https://raw.githubusercontent.com/rdharjai/info5502/main/hw2/hiv.csv'
hiv_full_tbl = Table.read_table(data_url)

print('\nTable loaded with: {} rows\n'.format(hiv_full_tbl.num_rows))
hiv_full_tbl
```



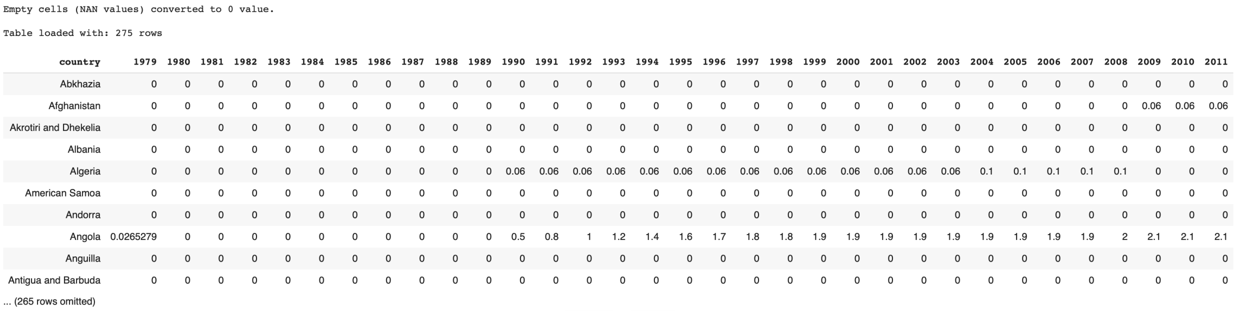
```
Table loaded with: 275 rows
```

country	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Abkhazia	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
Afghanistan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	0.06	0.06	0.06
Akrotiri and Dhekelia	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
Albania	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
Algeria	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.1	0.1	0.1	0.1	0.1	nan	nan	nan
American Samoa	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
Andorra	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
Angola	0.0265279	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	0.5	0.8	1	1.2	1.4	1.6	1.7	1.8	1.8	1.9	1.9	1.9	1.9	1.9	1.9	1.9	1.9	1.9	2	2.1	2.1	2.1
Anguilla	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
Antigua and Barbuda	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan

... (265 rows omitted)

Figure 1: HIV data - loaded raw data

```
# convert empty or NAN values to 0
cols = list(hiv_full_tbl.labels)
cols.remove('country')
for _ in range(len(cols)):
  hiv_full_tbl = hiv_full_tbl.with_column(cols[_],
            hiv_full_tbl.apply(lambda x: 0 if (math.isnan(x)) else x, cols[_]))

print('\nEmpty cells (NAN values) converted to 0 value.')
print('\nTable loaded with: {} rows\n'.format(hiv_full_tbl.num_rows))
hiv_full_tbl
```



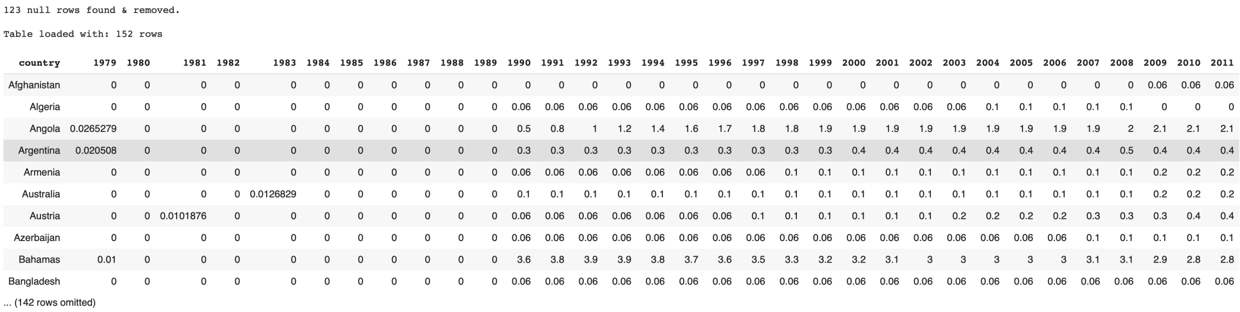
```
Empty cells (NAN values) converted to 0 value.

Table loaded with: 275 rows
```

country	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Abkhazia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Afghanistan	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.06	0.06	0.06
Akrotiri and Dhekelia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Albania	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Algeria	0	0	0	0	0	0	0	0	0	0	0	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.1	0.1	0.1	0.1	0.1	0	0	0
American Samoa	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Andorra	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Angola	0.0265279	0	0	0	0	0	0	0	0	0	0	0.5	0.8	1	1.2	1.4	1.6	1.7	1.8	1.8	1.9	1.9	1.9	1.9	1.9	1.9	1.9	1.9	1.9	2	2.1	2.1	2.1
Anguilla	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Antigua and Barbuda	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

... (265 rows omitted)

Figure 2: HIV data - Converted NAN values to 0

```
# remove rows with no data
# i.e. rows with all 0 values
# hiv_tbl2 = hiv_tbl.copy(shallow=False)
empty_row = (0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
             0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
             0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)
empty_row_idxs = []
row_idx = 0
for row in hiv_full_tbl.rows:
  if tuple(row[1:-1]) == empty_row:
    empty_row_idxs.append(row_idx)
  row_idx += 1

print('\n{} null rows found & removed.'.format(len(empty_row_idxs)))
hiv_tbl = hiv_full_tbl.exclude(empty_row_idxs)

print('\nTable loaded with: {} rows\n'.format(hiv_tbl.num_rows))
```



```
123 null rows found & removed.

Table loaded with: 152 rows
```

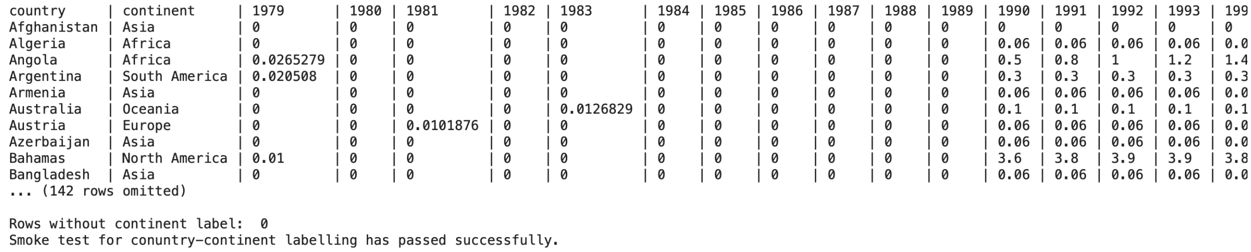
country	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Afghanistan	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.06	0.06	0.06
Algeria	0	0	0	0	0	0	0	0	0	0	0	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.1	0.1	0.1	0.1	0.1	0	0	0
Angola	0.0265279	0	0	0	0	0	0	0	0	0	0	0.5	0.8	1	1.2	1.4	1.6	1.7	1.8	1.8	1.9	1.9	1.9	1.9	1.9	1.9	1.9	1.9	1.9	2	2.1	2.1	2.1
Argentina	0.020508	0	0	0	0	0	0	0	0	0	0	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.5	0.4	0.4	0.4
Armenia	0	0	0	0	0	0	0	0	0	0	0	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2
Australia	0	0	0	0	0.0126829	0	0	0	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2
Austria	0	0	0.0101876	0	0	0	0	0	0	0	0	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.3	0.3	0.3	0.4	0.4
Azerbaijan	0	0	0	0	0	0	0	0	0	0	0	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.1	0.1	0.1	0.1	0.1
Bahamas	0.01	0	0	0	0	0	0	0	0	0	0	3.6	3.8	3.9	3.9	3.8	3.7	3.6	3.5	3.3	3.2	3.2	3.1	3	3	3	3	3	3.1	3.1	2.9	2.8	2.8
Bangladesh	0	0	0	0	0	0	0	0	0	0	0	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06

... (142 rows omitted)

Figure 3: HIV data - removed null rows

Task #1 Add one column as "continent" in the dataset and label each country/region in the dataset to an appropriate continent such as "Europe", "Asia", "Africa", "North America", "South America", "Australia", or "Antarctica". Explain how do validated the correctness of your labelling. Output the updated dataset as a new CSV file. (1 point). (Note: You must write a Python program to complete the labelling, manually labelling will not get any credit).

```python
# load country-continent data
country_data_url = 'https://raw.githubusercontent.com/rdharjai/info5502/' \
'main/hw2/country.csv'
country_tbl = Table.read_table(country_data_url)

# join HIV table with Country table
hiv_tbl = hiv_tbl.join('country', country_tbl, 'country')

# move the country & continent column
# as the first 2 columns in table
hiv_tbl.move_to_start('continent')
hiv_tbl.move_to_start('country')

# print table to validate country-continent names
print(hiv_tbl)
print("\nRows without continent label: ", hiv_tbl.where('continent','').num_rows)

# export data-table as a CSV file
hiv_tbl.to_csv('hiv_tbl.csv')

# smoke test to validate correctness of the
# continent labelling
assert(hiv_tbl.where('country','India').column('continent')=='Asia')
assert(hiv_tbl.where('country','Japan').column('continent')=='Asia')
assert(hiv_tbl.where('country','Canada').column('continent')=='North America')
assert(hiv_tbl.where('country','United States').column('continent')=='North America')
assert(hiv_tbl.where('country','Argentina').column('continent')=='South America')
assert(hiv_tbl.where('country','Brazil').column('continent')=='South America')
assert(hiv_tbl.where('country','Australia').column('continent')=='Oceania')
assert(hiv_tbl.where('country','New Zealand').column('continent')=='Oceania')
print("Smoke test for conuntry-continent labelling has passed successfully.")
```

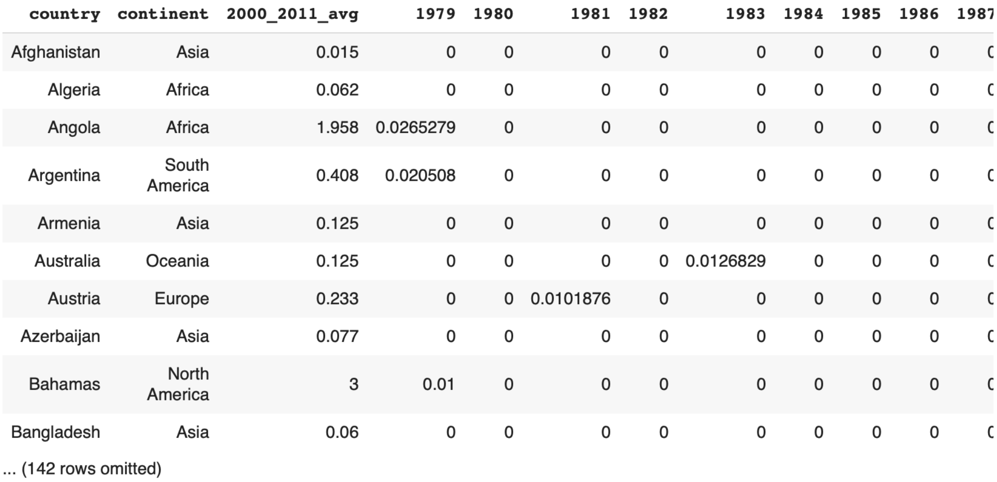


```
country     | continent     | 1979      | 1980 | 1981      | 1982 | 1983      | 1984 | 1985 | 1986 | 1987 | 1988 | 1989 | 1990 | 1991 | 1992 | 1993 | 199
Afghanistan | Asia          | 0         | 0    | 0         | 0    | 0         | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0
Algeria     | Africa        | 0         | 0    | 0         | 0    | 0         | 0    | 0    | 0    | 0    | 0    | 0    | 0.06 | 0.06 | 0.06 | 0.06 | 0.0
Angola      | Africa        | 0.0265279 | 0    | 0         | 0    | 0         | 0    | 0    | 0    | 0    | 0    | 0    | 0.5  | 0.8  | 1    | 1.2  | 1.4
Argentina   | South America | 0.020508  | 0    | 0         | 0    | 0         | 0    | 0    | 0    | 0    | 0    | 0    | 0.3  | 0.3  | 0.3  | 0.3  | 0.3
Armenia     | Asia          | 0         | 0    | 0         | 0    | 0         | 0    | 0    | 0    | 0    | 0    | 0    | 0.06 | 0.06 | 0.06 | 0.06 | 0.0
Australia   | Oceania       | 0         | 0    | 0         | 0    | 0.0126829 | 0    | 0    | 0    | 0    | 0    | 0    | 0.1  | 0.1  | 0.1  | 0.1  | 0.1
Austria     | Europe        | 0         | 0    | 0.0101876 | 0    | 0         | 0    | 0    | 0    | 0    | 0    | 0    | 0.06 | 0.06 | 0.06 | 0.06 | 0.0
Azerbaijan  | Asia          | 0         | 0    | 0         | 0    | 0         | 0    | 0    | 0    | 0    | 0    | 0    | 0.06 | 0.06 | 0.06 | 0.06 | 0.0
Bahamas     | North America | 0.01      | 0    | 0         | 0    | 0         | 0    | 0    | 0    | 0    | 0    | 0    | 3.6  | 3.8  | 3.9  | 3.9  | 3.8
Bangladesh  | Asia          | 0         | 0    | 0         | 0    | 0         | 0    | 0    | 0    | 0    | 0    | 0    | 0.06 | 0.06 | 0.06 | 0.06 | 0.0
... (142 rows omitted)

Rows without continent label:  0
Smoke test for conuntry-continent labelling has passed successfully.
```

Figure 4: HIV data - Joined country-continent data

Task #2 Write a Python program to find the country/region in each continent that has the highest average HIV estimated prevalence of people ages from 15 to 49 of from year 2000 to 2011. Find the country/region in each continent that has the lowest average HIV estimated prevalence of people ages from 15 to 49 of from year 2000 to 2011. Create a bar chart to show the highest average HIV estimated prevalence of people ages from 15 to 49 of from year 2000 to 2011 in each continent (1 point). Create a bar chart to show the lowest average HIV estimated prevalence of people ages from 15 to 49 of from year 2000 to 2011 in each continent (1 point). Create an overlaid bar chart to show the highest and lowest average HIV estimated prevalence of people ages from 15 to 49 of from year 2000 to 2011 in each continent (1 point). Select a country/region that is different from the average highest or lowest HIV estimated prevalence of people ages from 15 to 49 from year 2000 to 2011 from each continent, then create an overlaid line chart for the selected country/region, the average highest and lowest HIV estimated prevalence of people ages from 15 to 49 from year 2000 to 2011 for each continent (1 point).

```
# add a new column: 2000_2011_avg
# it is the average of values of all columns
# between 2000 (index: 23) and 2011 (index: 34)

avg_lst = []
for row in hiv_tbl.rows:
  avg_lst.append(round(sum(list(row)[23:35]) / len(list(row)[23:35]),3))

try:
  hiv_tbl.drop('2000_2011_avg')
finally:
  hiv_tbl = hiv_tbl.with_column('2000_2011_avg', avg_lst)

hiv_tbl
```

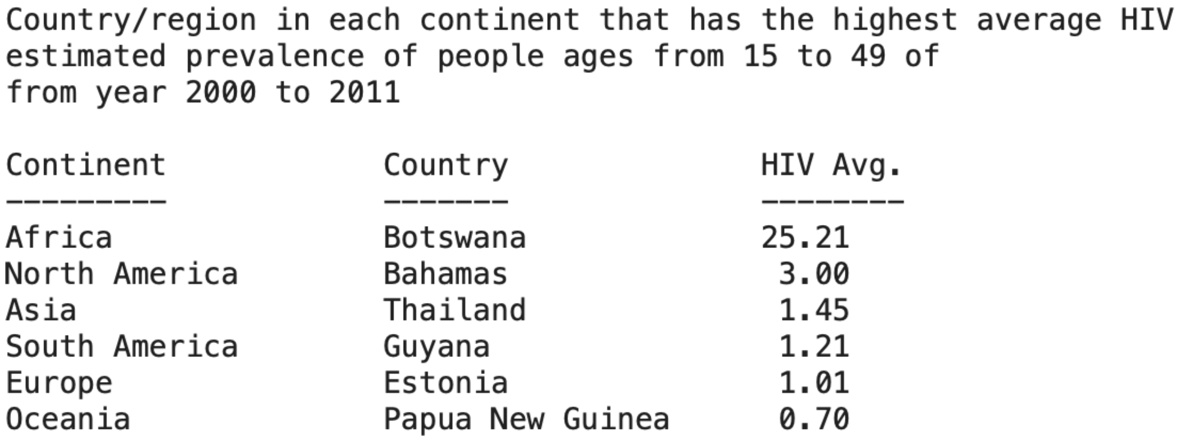


country	continent	2000_2011_avg	1979	1980	1981	1982	1983	1984	1985	1986	1987
Afghanistan	Asia	0.015	0	0	0	0	0	0	0	0	0
Algeria	Africa	0.062	0	0	0	0	0	0	0	0	0
Angola	Africa	1.958	0.0265279	0	0	0	0	0	0	0	0
Argentina	South America	0.408	0.020508	0	0	0	0	0	0	0	0
Armenia	Asia	0.125	0	0	0	0	0	0	0	0	0
Australia	Oceania	0.125	0	0	0	0	0.0126829	0	0	0	0
Austria	Europe	0.233	0	0	0.0101876	0	0	0	0	0	0
Azerbaijan	Asia	0.077	0	0	0	0	0	0	0	0	0
Bahamas	North America	3	0.01	0	0	0	0	0	0	0	0
Bangladesh	Asia	0.06	0	0	0	0	0	0	0	0	0

... (142 rows omitted)

Figure 5: HIV data - Added column for avg. of 2000-2011

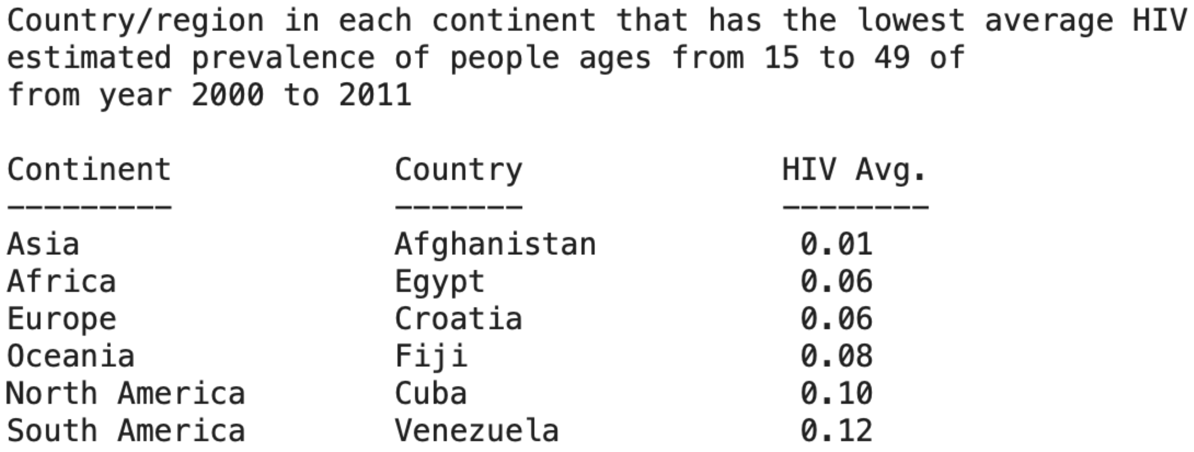
```
lst_2000_2011_avg_high = []
for row in hiv_tbl.group('continent', max).select(0,35).sort('2000_2011_avg max', descending = True).r
  lst_2000_2011_avg_high.append(tuple(hiv_tbl
                                    .where('continent',row[0])
                                    .where('2000_2011_avg',row[1])
                                    .select(1,0,35)
                                    .row(0)))

print('\nCountry/region in each continent that has the highest average HIV ',
      '\nestimated prevalence of people ages from 15 to 49 of ',
      '\nfrom year 2000 to 2011\n')

print('{:<20} {:<20} {:>5}'.format('Continent', 'Country', 'HIV Avg.'))
print('{:<20} {:<20} {:>5}'.format('---------', '-------', '--------'))
for item in lst_2000_2011_avg_high:
  print('{:<20} {:<20} {:>5.2f}'.format(item[0],item[1],item[2]))
```



```
Country/region in each continent that has the highest average HIV
estimated prevalence of people ages from 15 to 49 of
from year 2000 to 2011

Continent            Country              HIV Avg.
---------            -------              --------
Africa               Botswana             25.21
North America        Bahamas               3.00
Asia                 Thailand              1.45
South America        Guyana                1.21
Europe               Estonia               1.01
Oceania              Papua New Guinea      0.70
```

Figure 6: HIV data - highest average data for 2000-2011 by continents

```
lst_2000_2011_avg_low = []
for row in hiv_tbl.group('continent', min).select(0,35).sort('2000_2011_avg min').rows:
  lst_2000_2011_avg_low.append(tuple(hiv_tbl
                                    .where('continent',row[0])
                                    .where('2000_2011_avg',row[1])
                                    .select(1,0,35)
                                    .row(0)))

print('\nCountry/region in each continent that has the lowest average HIV ',
      '\nestimated prevalence of people ages from 15 to 49 of ',
      '\nfrom year 2000 to 2011\n')

print('{:<20} {:<20} {:>5}'.format('Continent', 'Country', 'HIV Avg.'))
print('{:<20} {:<20} {:>5}'.format('---------', '-------', '--------'))
for item in lst_2000_2011_avg_low:
```



```
Country/region in each continent that has the lowest average HIV
estimated prevalence of people ages from 15 to 49 of
from year 2000 to 2011

Continent            Country              HIV Avg.
---------            -------              --------
Asia                 Afghanistan           0.01
Africa               Egypt                 0.06
Europe               Croatia               0.06
Oceania              Fiji                  0.08
North America        Cuba                  0.10
South America        Venezuela             0.12
```

Figure 7: HIV data - lowest average data for 2000-2011 by continents

```
x_val = [x[0] for x in lst_2000_2011_avg_high]
y_high_val = [x[2] for x in lst_2000_2011_avg_high]

fig, (ax1) = plt.subplots(1,1,constrained_layout=True)
fig.set_size_inches(8, 6)
fig.set_dpi(100)
fig.suptitle('highest average HIV estimated prevalence of people ages from \
15 to 49 \n from year 2000 to 2011 in each continent \n (using log scale)')

# Bar Plot
ax1.set_yscale('log')
ax1.set(xlabel='Continents',
        ylabel='Average HIV estimated prevalence \n (in log scale)')
ax1.bar(x_val,
        y_high_val,
        label="Highest avg. HIV estimated prevalence")
ax1.legend()
```

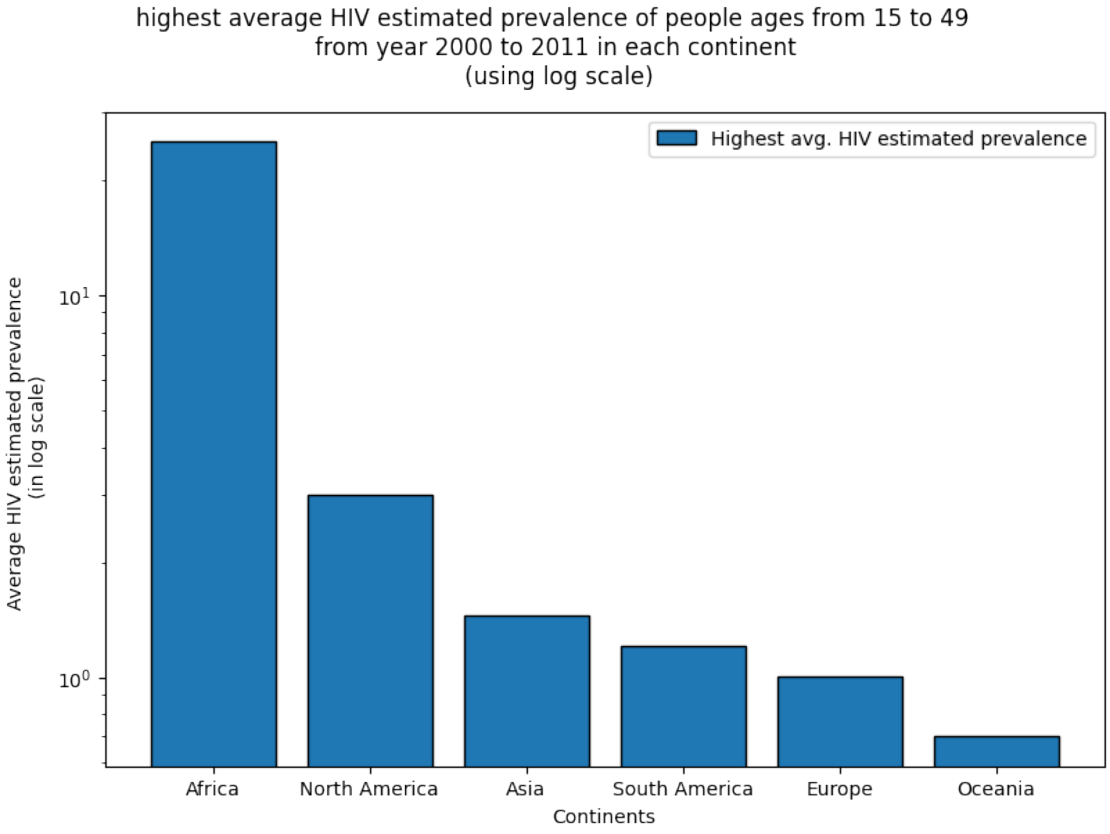


Figure 8: HIV data - bar chart for highest average data for 2000-2011 by continents

```
x_val = [x[0] for x in lst_2000_2011_avg_low]
y_low_val = [x[2] for x in lst_2000_2011_avg_low]

fig, (ax1) = plt.subplots(1,1,constrained_layout=True)
fig.set_size_inches(8, 6)
fig.set_dpi(100)
fig.suptitle('lowest average HIV estimated prevalence of people ages from \
15 to 49 \n from year 2000 to 2011 in each continent \n (using log scale)')

# Bar Plot
ax1.set_yscale('log')
ax1.set(xlabel='Continents',
        ylabel='Average HIV estimated prevalence \n (in log scale)')
ax1.bar(x_val,
        y_low_val,
        label="Lowest avg. HIV estimated prevalence")

ax1.legend()
```

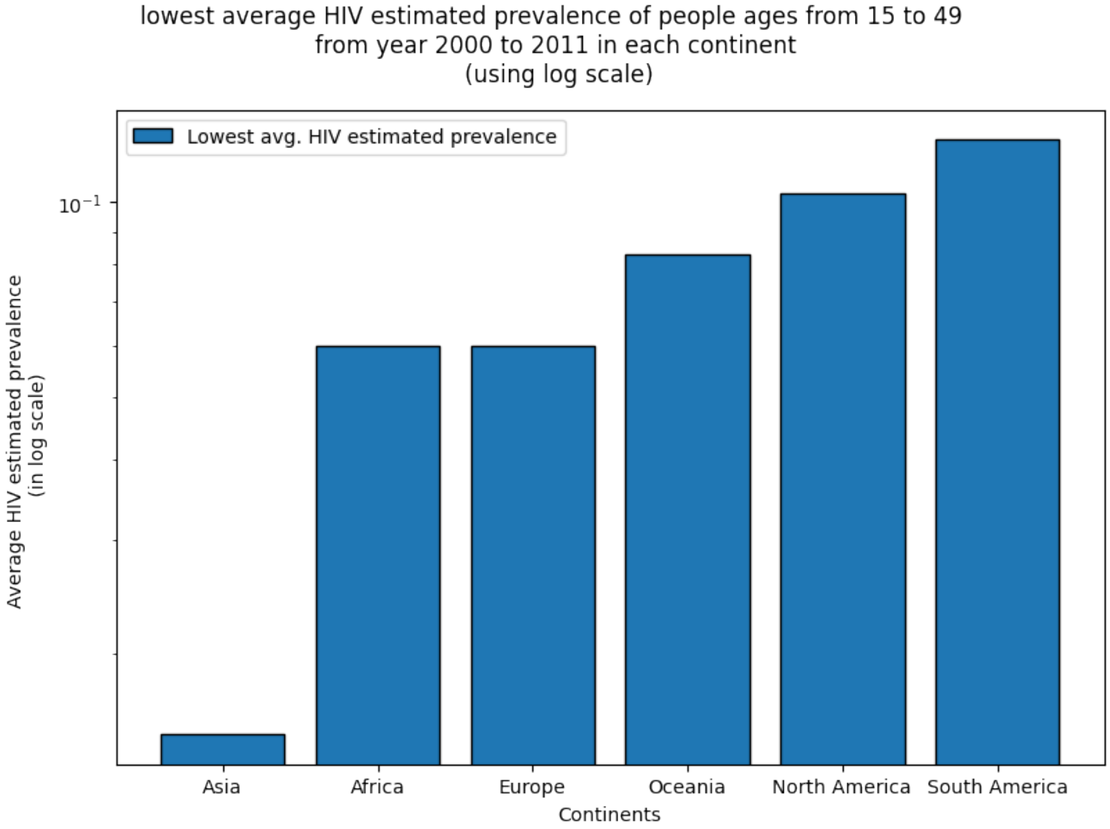


Figure 9: HIV data - bar chart for lowest average data for 2000-2011 by continents

```
fig, (ax1) = plt.subplots(1,1,constrained_layout=True)
fig.set_size_inches(8, 6)
fig.set_dpi(100)
fig.suptitle('highest & lowest average HIV estimated prevalence of people ages from \
15 to 49 \n from year 2000 to 2011 in each continent \n (using log scale)')

# Bar Plot
ax1.set_yscale('log')
ax1.set(xlabel='Continents',
        ylabel='Average HIV estimated prevalence \n (in log scale)')

ax1.bar(x_val,
        y_high_val,
        width=0.9,
        label="Highest avg. HIV estimated prevalence")

ax1.bar(x_val,
        y_low_val,
        width=0.7,
        label="Lowesr avg. HIV estimated prevalence")

ax1.legend()
```

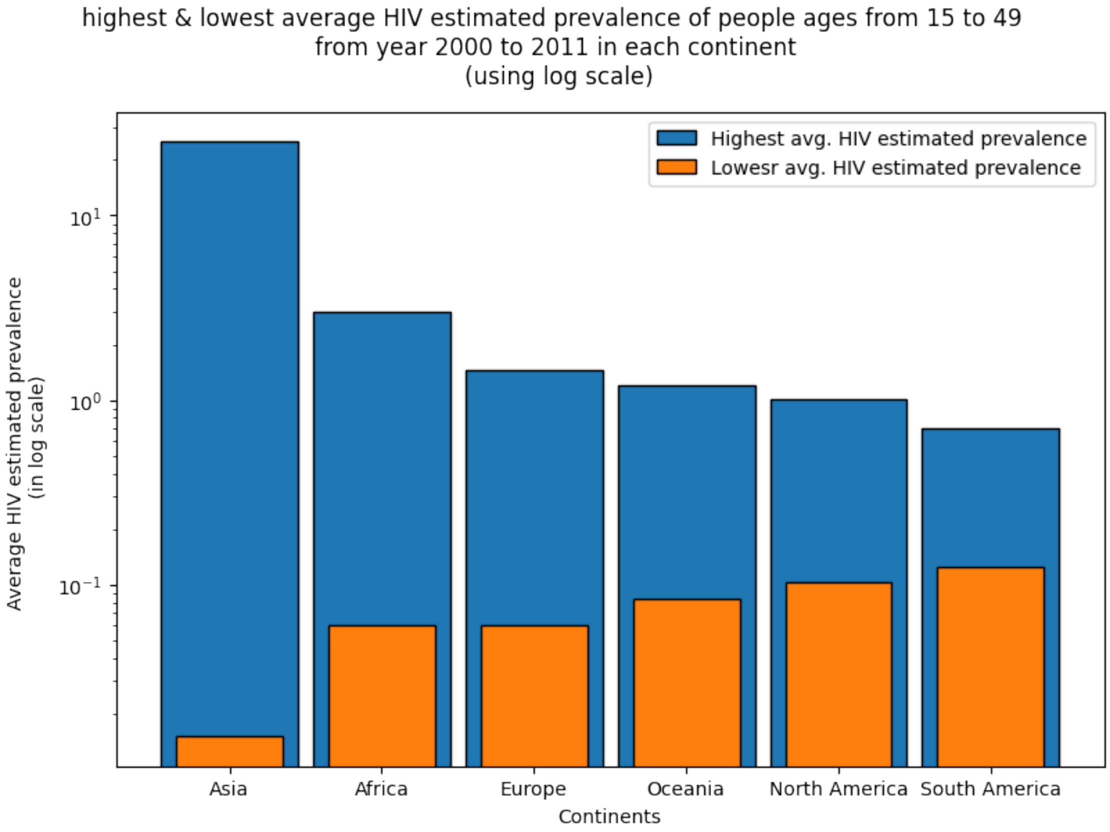
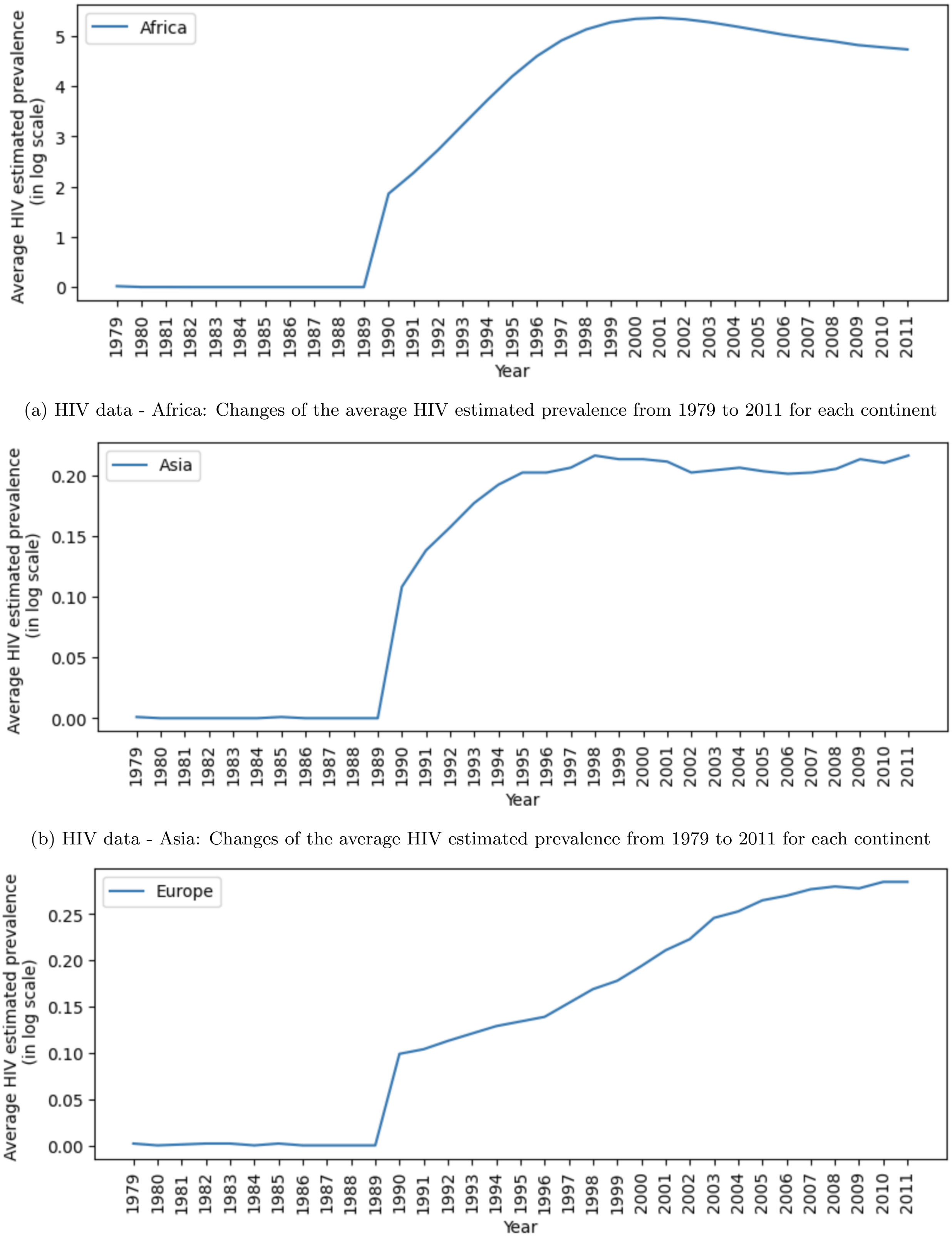


Figure 10: HIV data - bar chart for highest & lowest average data for 2000-2011 by continents

Task #3 Write a Python program to calculate the average HIV estimated prevalence of people ages from 15 to 49 for each year in the dataset for each continent (you only need simply add the estimate prevalence number of all countries/regions and divided by the number of the countries/regions in the continent). Based on the calculation, create a line chart for each continent to show the changes of the average HIV estimated prevalence from 1979 to 2011 (1 point). Create an overlaid line chart for all continents to show their changes of the average HIV estimated prevalence from 1 1979 to 2011 (1 point).

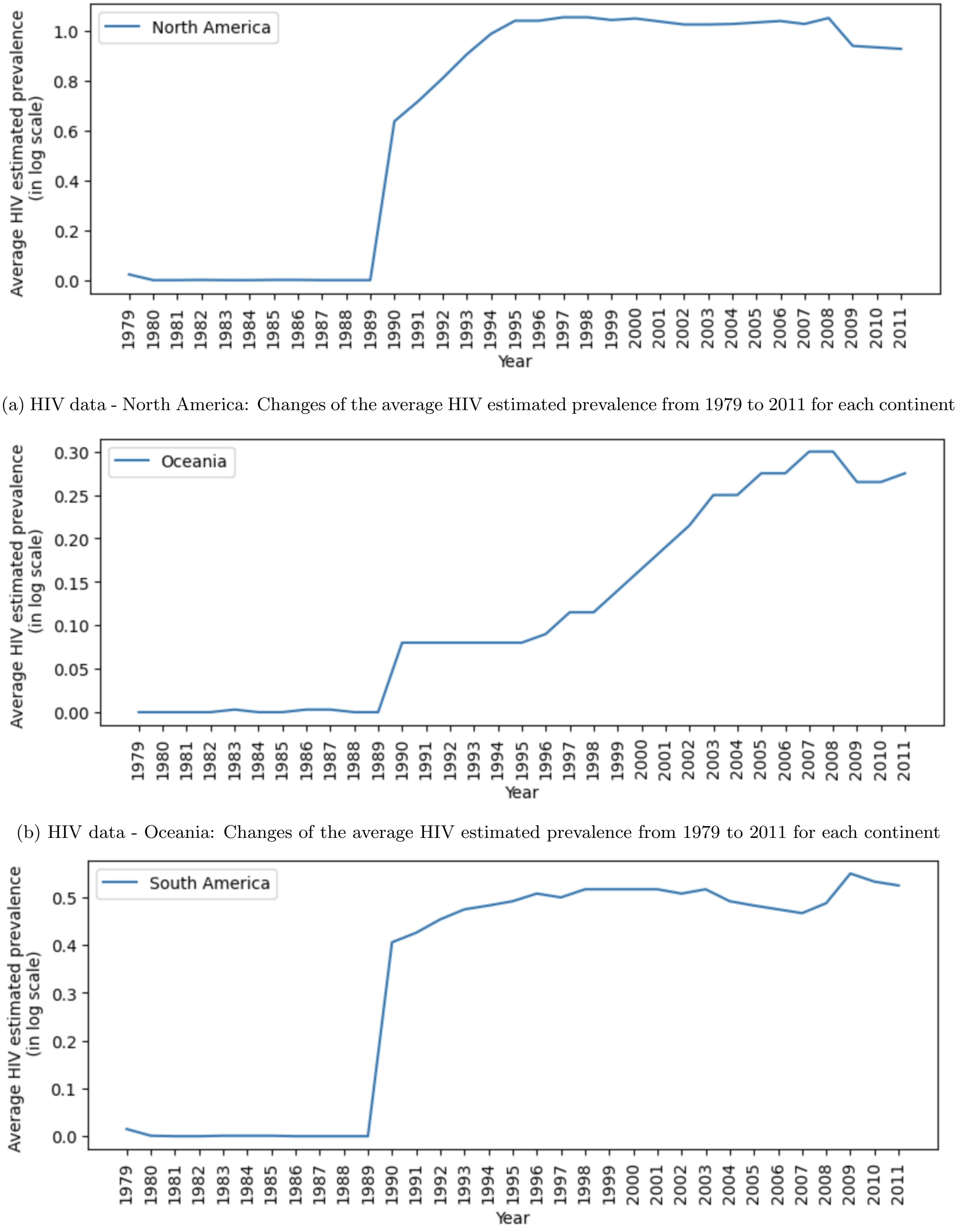
```
lst_continents = []
for row in hiv_tbl.group('continent').rows:
  lst_continents.append(tuple(row))

lst_year_avg = []
for continent in lst_continents:
  lst = []
  for col in range(2, hiv_tbl.num_columns-1):
    lst.append((hiv_tbl.labels[col],
                round(hiv_tbl.where('continent',continent[0]).select(col)
                .mean()
                .row(0)[0],3)))
  lst_year_avg.append([continent[0],lst])


fig, axs = plt.subplots(len(lst_year_avg),1,constrained_layout=True)
fig.set_size_inches(8, 20)
fig.set_dpi(100)
fig.suptitle('Changes of the average HIV estimated prevalence \
\n from 1979 to 2011 for each continent')

idx = 0
for ax in axs:
  x_data = [x[0] for x in lst_year_avg[idx][1]]
  y_data = [x[1] for x in lst_year_avg[idx][1]]
  ax.set(xlabel='Year',
        ylabel='Average HIV estimated prevalence \n (in log scale)')
  ax.set_xticklabels(x_data, rotation=90)
  ax.plot(x_data, y_data, label=lst_year_avg[idx][0])
  ax.legend()
  idx += 1
```



(a) HIV data - Africa: Changes of the average HIV estimated prevalence from 1979 to 2011 for each continent

(b) HIV data - Asia: Changes of the average HIV estimated prevalence from 1979 to 2011 for each continent

(c) HIV data - Europe: Changes of the average HIV estimated prevalence from 1979 to 2011 for each continent



(a) HIV data - North America: Changes of the average HIV estimated prevalence from 1979 to 2011 for each continent

(b) HIV data - Oceania: Changes of the average HIV estimated prevalence from 1979 to 2011 for each continent

(c) HIV data - South America: Changes of the average HIV estimated prevalence from 1979 to 2011 for each continent

```
fig, ax = plt.subplots(1,1,constrained_layout=True)
fig.set_size_inches(8, 6)
fig.set_dpi(100)
fig.suptitle('Changes of the average HIV estimated prevalence \
\n from 1979 to 2011 for each continent')

for item in lst_year_avg:
  x_data = [x[0] for x in item[1]]
  y_data = [x[1] for x in item[1]]
  ax.set(xlabel='Year',
        ylabel='Average HIV estimated prevalence \n (in log scale)')
  ax.set_xticklabels(x_data, rotation=90)
  ax.plot(x_data, y_data, label=item[0])
  ax.legend()
```

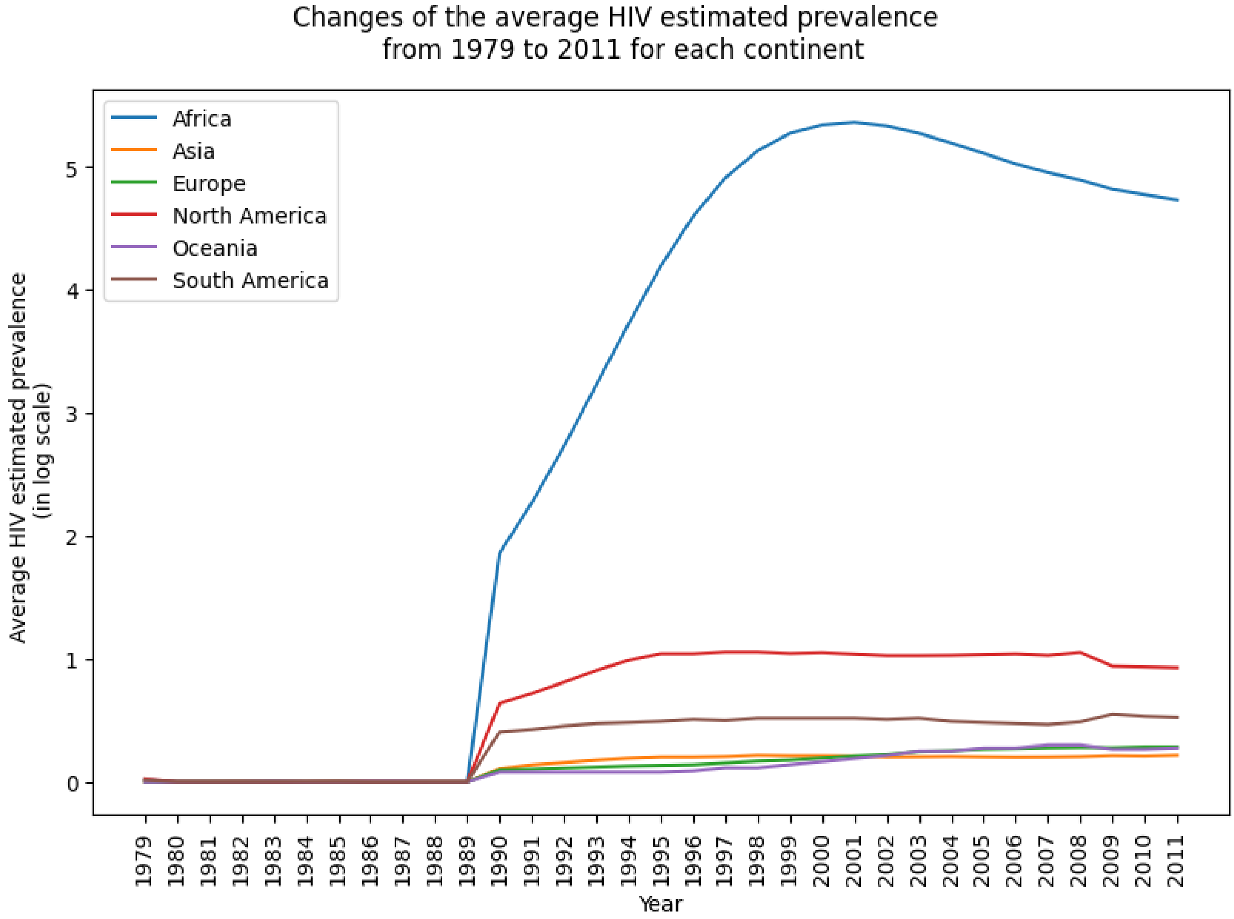
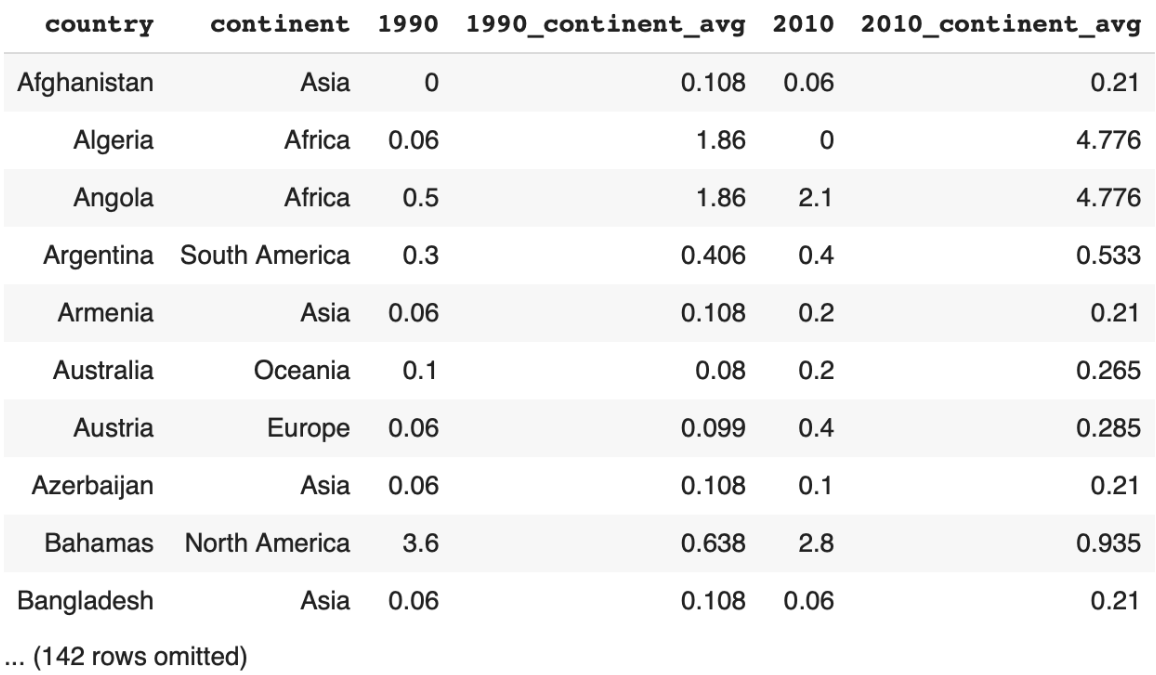


Figure 13: HIV data - All Continents: Changes of the average HIV estimated prevalence from 1979 to 2011 for each continent

Task #3 Create two scatter plots to show the data (i.e. each country/region) in year 1990 and year 2010, respectively. The vertical axis in the scatter plot is the HIV estimated prevalence, and the horizontal axis is the corresponding year average HIV estimated prevalence in each continent, which you calculated above. Using different color to show data from different continent (1 point). If you found any interesting result from the charts, explain it.

```
# preparing data
data1990 = {}
data2010 = {}
for item in lst_year_avg:
  for subitem in item:
    for subitem_year in subitem:
      if subitem_year[0] == '1990':
        data1990.update({item[0]: subitem_year[1]})
        # lst_1990data.append((item[0], subitem_year[1]))
        hiv_tbl.where('continent',item[0])
      if subitem_year[0] == '2010':
        # lst_2010data.append((item[0], subitem_year[1]))
        data2010.update({item[0]: subitem_year[1]})

lst1990 = []
lst2010 = []
for row in hiv_tbl.rows:
  lst1990.append(data1990.get(row[1]))
  lst2010.append(data2010.get(row[1]))

try:
  hiv_tbl.drop('1990_continent_avg')
  hiv_tbl.drop('2010_continent_avg')
finally:
  hiv_tbl = hiv_tbl.with_column('1990_continent_avg',lst1990)
  hiv_tbl = hiv_tbl.with_column('2010_continent_avg',lst2010)

hiv_tbl.select('country','continent','1990','1990_continent_avg',
               '2010','2010_continent_avg')
```

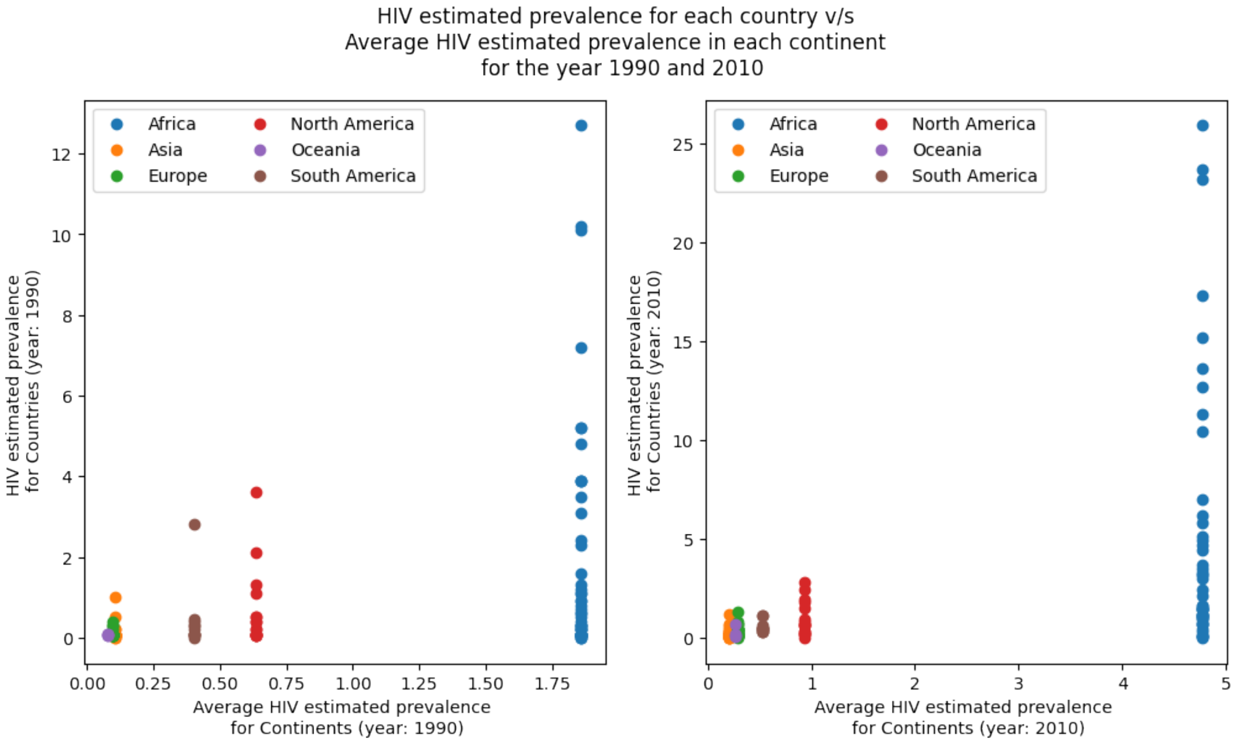


country	continent	1990	1990_continent_avg	2010	2010_continent_avg
Afghanistan	Asia	0	0.108	0.06	0.21
Algeria	Africa	0.06	1.86	0	4.776
Angola	Africa	0.5	1.86	2.1	4.776
Argentina	South America	0.3	0.406	0.4	0.533
Armenia	Asia	0.06	0.108	0.2	0.21
Australia	Oceania	0.1	0.08	0.2	0.265
Austria	Europe	0.06	0.099	0.4	0.285
Azerbaijan	Asia	0.06	0.108	0.1	0.21
Bahamas	North America	3.6	0.638	2.8	0.935
Bangladesh	Asia	0.06	0.108	0.06	0.21

... (142 rows omitted)

(a) HIV data - added columns for continent average for 1990 and 2010

```
fig, (ax1, ax2) = plt.subplots(1,2,constrained_layout=True)
fig.set_size_inches(10, 6)
fig.set_dpi(100)
fig.suptitle('HIV estimated prevalence for each country v/s \
\nAverage HIV estimated prevalence in each continent \
\n for the year 1990 and 2010')

ax1.set(xlabel='Average HIV estimated prevalence \
\n for Continents (year: 1990)',
      ylabel='HIV estimated prevalence \
\n for Countries (year: 1990)')

ax2.set(xlabel='Average HIV estimated prevalence \
\n for Continents (year: 2010)',
      ylabel='HIV estimated prevalence \
\n for Countries (year: 2010)')

for continent in hiv_tbl.group('continent')[0]:
  ax1.scatter(
      hiv_tbl.where('continent',continent).column('1990_continent_avg'),
      hiv_tbl.where('continent',continent).column('1990'),
      label=continent)

  ax2.scatter(
      hiv_tbl.where('continent',continent).column('2010_continent_avg'),
      hiv_tbl.where('continent',continent).column('2010'),
      label=continent)
```



(a) HIV estimated prevalence for each country v/s Average HIV estimated prevalence in each continent for the year 1990 and 2010

```
ax1.legend(ncol=2)
ax2.legend(ncol=2)
```