Aditya Pitchuka  
INFO 5502.002 (14676)  
Spring 2020

**Assignment – 2**

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 you validate the correctness of your labelling. Output the updated dataset as a new CSV file.

**Approach:**

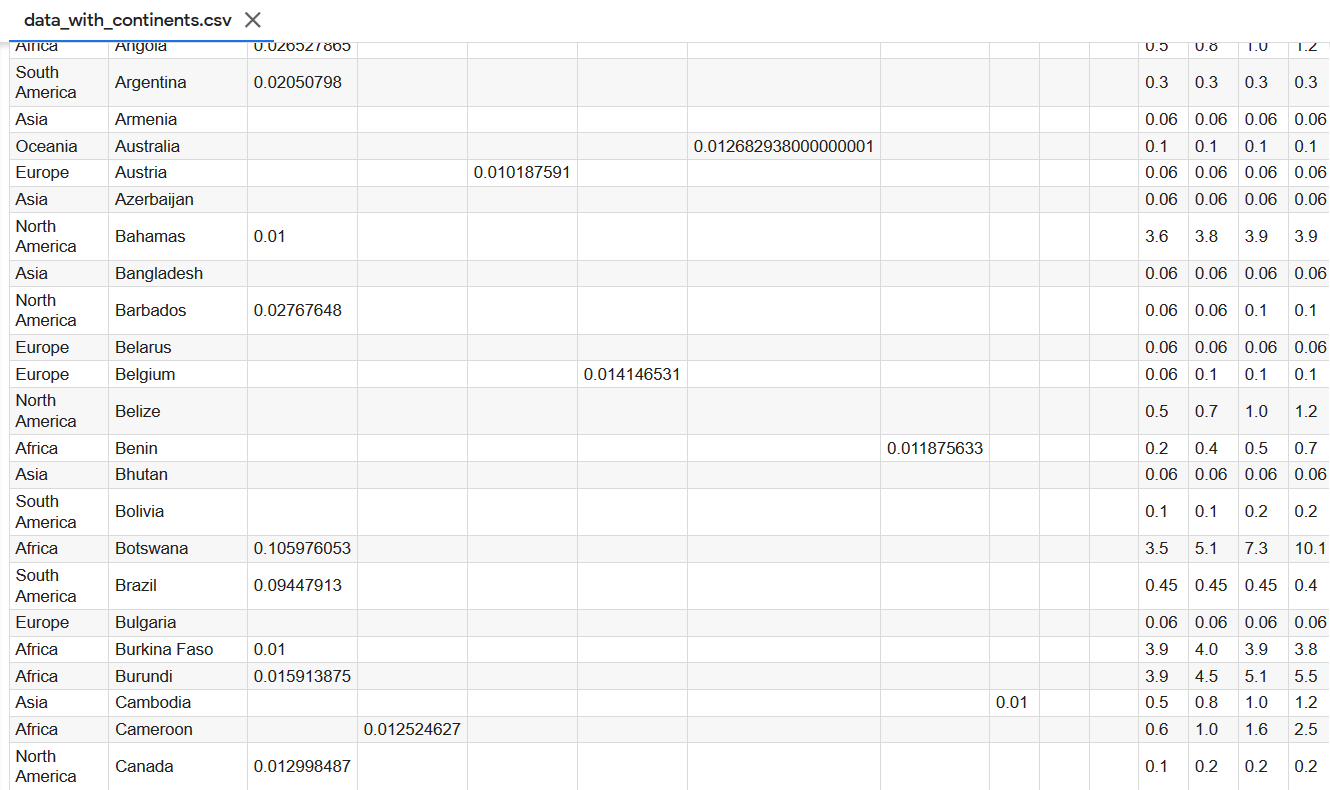
Read the given dataset as a Table using read\_table () method. Converted the table to pandas data frame to perform some data cleaning and clear the empty values in the column. Once it is cleared, again changed it to Table format.

Now using the function country\_to\_continent () present in the module pycountry\_convert I created a new column and applied that function to whole column of Countries in the main Table. Now the Table with new appended column is written to a new csv file.

data\_reader.append\_column('Continents',data\_reader.apply(country\_to\_continent,0))

data\_reader.to\_csv('/content/sample\_data/data\_with\_continents.csv')

**Visualization Result:**

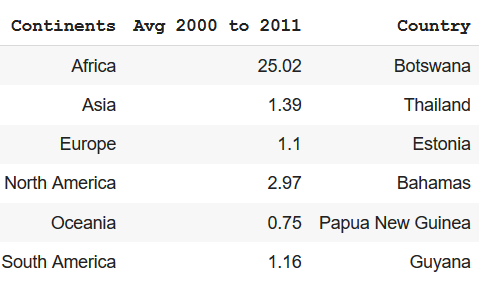


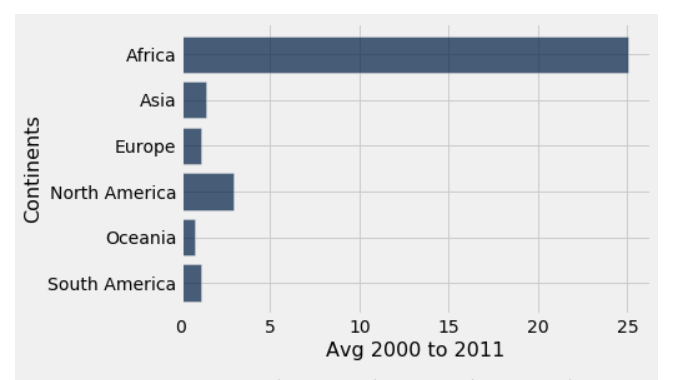
New File with Continents Column added

1. 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. 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. 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.

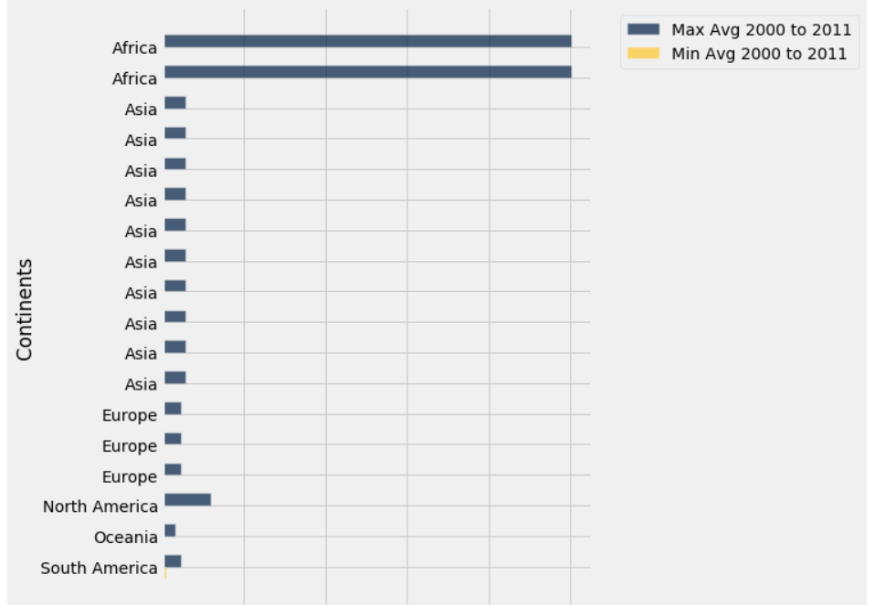
**Approach:**

Read the table from the dataset extract from Q1 and then convert to pandas data frame. Now we create a column for Averages 2000 to 2011 and find average for each country in each continent. Convert the data frame back to Table. Now we sort the averages in descending order for the table obtained. Now two tables are created for finding the maximum & minimum HIV % for country in each continent. We do it by selecting the first 3 columns of main table and group by continents with max and min function respectively. We join the above two columns with the main table. We select the columns with continents & country where averages are equal for each country are equal to the max average value for years 2000 to 2011. We display the table to obtain the country having maximum HIV %. Similarly, we do the same for minimum HIV %

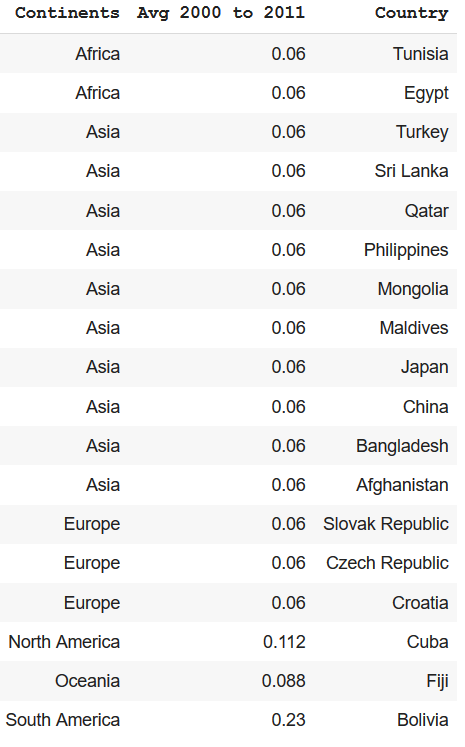
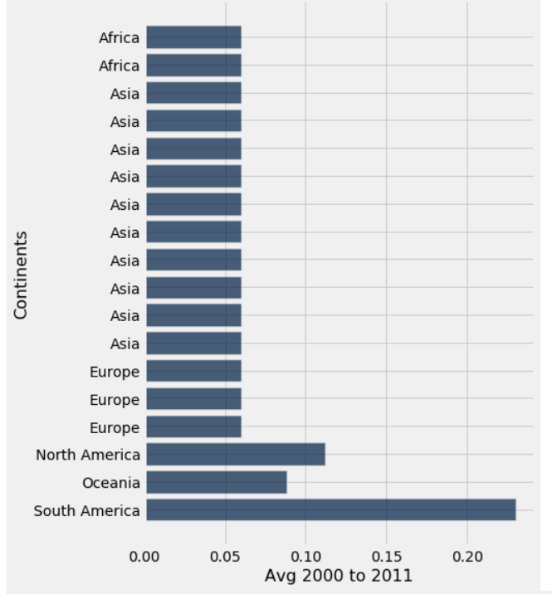




Max Average HIV % for years 2000 to 2011 and its respective country and continent



Overlaid bar chart with highest and lowest HIV% for ages 15-29 for years 2000 to 2011



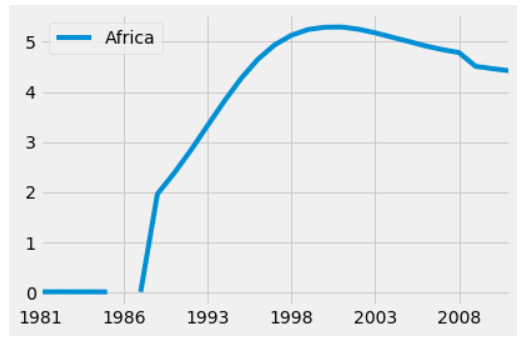
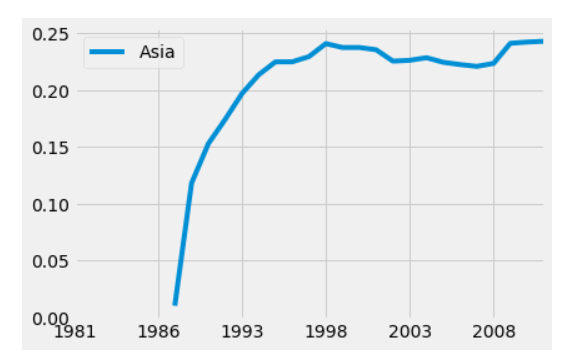
Min Average HIV % for years 2000 to 2011 and its repective country and continent

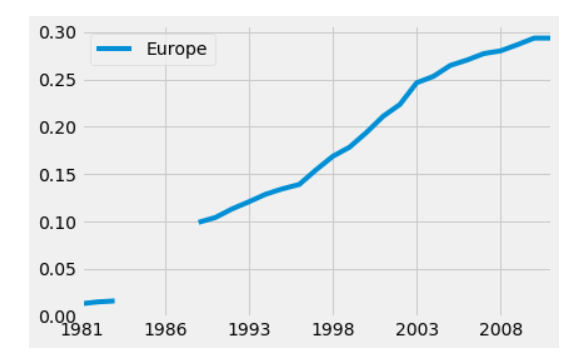
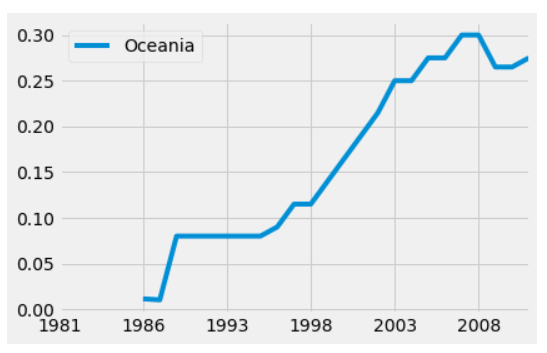
1. Write a Python program to calculate the average HIV estimated prevalence of people ages from 15 to 49 for each year in the dataset for each continent (you only need simply add the estimate prevalence number of all countries/regions and divided by the number of the countries/regions in the continent).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

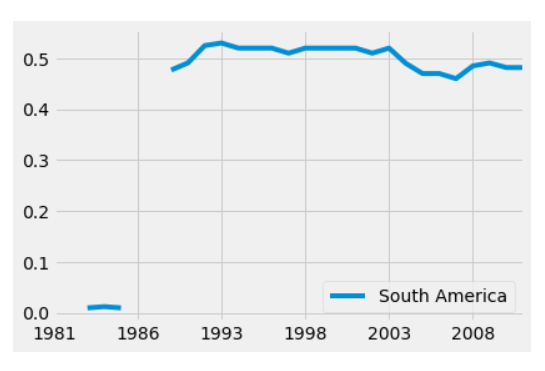
**Approach:**

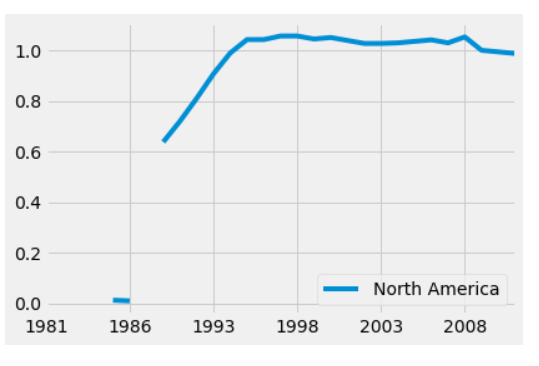
First, we read the table that we got from Q1 and we convert into pandas data frame. Now we use filtering to find out list of data with columns based on each continent. Now the resultant data set is used to find average of each column yearly using the mean () function and plot the graph for each continent. For an overlaid plot, which contains all the continents’ data each of the continent’s average values are considered and a graph is plotted.

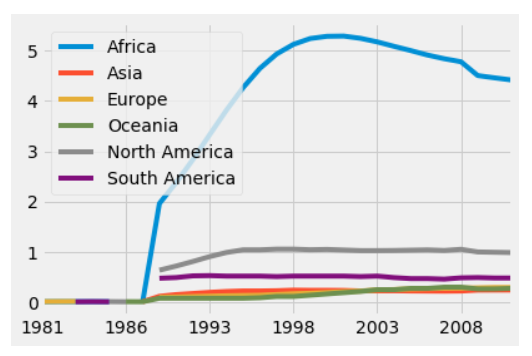
**Visualization Results:**









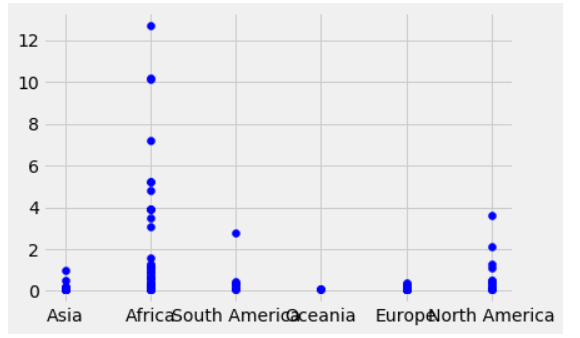


1. Create two scatter plots to show the data (i.e. each country/region) in year 1990 and year 2010, respectively. The vertical axis in the scatter plot is the HIV estimated prevalence, and the horizontal axis is the corresponding year average HIV estimated prevalence in each continent, which you calculated above. Using different color to show data from different continent.

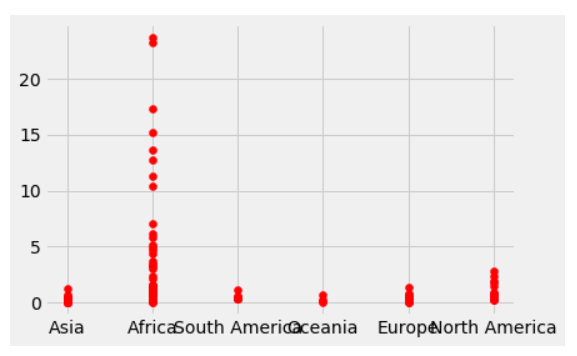
**Approach:**

We read the table and select 2 tables from them, one for year 1990 and the other for year 2010. We create two lists one for the values of Estimated HIV % for year 1990 and 2010. We create another list for the list of continents. Now we create two scatter plots using the lists, list with 1990 and continents for 1990-year plot and for list with 2010 and continents for 2010 year.

**Visualization Results:**



Year 1990



Year 2010