INFO5502 Assignment 2

The included Excel file lists HIV estimated prevalence of people ages from 15 to 49 in the world from 1979 to 2011. Use the dataset to complete following tasks:

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. (1 point). (Note: You must write a Python program to complete the labelling, manully labelling won't get any credit).

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).

3. Write a Python progrma 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

1979 to 2011 (1 point).

4. 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.

5. Write a report to explain how each question is implemented and its output graphs (2 point).