Programming Assignment 2: Data Visualization

**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 you validate the correctness of your labelling (1 point).**

1. The given file containing HIV estimated prevalence data was imported to Jupyter Notebook in Python. The column consisting country names was renamed to “Country” for an easy reference. A new file (countries\_by\_continents.xlsx) that had a list of countries with their respective continents was downloaded from the web and merged with the HIV.xlsx file that had the original data of estimated HIV prevalence from 1979 to 2011 and age 15 to 49. Left inner join was performed to merge with HIV.xlsx data on the left and contries\_by\_continents.xlsx data on the right. The inner join compared values in each row in the ‘Country’ column in the HIV.xlsx file to each row in the ‘Country’ column in the contries\_by\_continents.xlsx file and appended the matching values i.e. the continent names for the county names that matched, to the HIV.xlsx file. The new merged file was named merged\_HIV. This is how each country was validated and respective continent was assigned.

A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

**2. A. 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 year2000 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 (1point). 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).**

1. A location reference was made using .loc function in Python that located columns from year 2000 to 2011. The mean was then calculated for years 2000 to 2011, the highest and lowest values were calculated using .idmax() and idmin() functions, and the results were grouped by each continent using.

A close up of a logo

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

1. Bar chart showing the *highest* average HIV in each continent from 2000 to 2011, ages 15-49.

A screenshot of a cell phone

Description automatically generated

1. Bar chart showing the *lowest* average HIV in each continent from 2000 to 2011, ages 15-49.

A screenshot of a cell phone

Description automatically generated

1. The mean for each country was recalculated and grouped by their respective continent. A new binary variable was created to assign the binary values ‘high’ and ‘low’. This was done to plot the highest and lowest values by continent together in one bar chart in an easier way.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

1. Overlaid bar chart showing highest and lowest averages by continent from 2000 to 2011, ages 15-49.

A screenshot of a social media post

Description automatically generated

**2. B. Select a country/region that is different from the average highest or lowest HIV estimated prevalence of people ages from 15 to 49 from year2000 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).**

1. Eleven countries that did not have the highest or lowest average HIV estimated prevalence in their respective continent were selected and their average HIV prevalence from 2000 to 2011 was calculated.

A screenshot of a cell phone

Description automatically generated

A close up of a piece of paper

Description automatically generated

1. Line chart showing average HIV prevalence by selected countries.

A screenshot of a cell phone

Description automatically generated

1. Line chart showing average HIV prevalence by each continent.

A screenshot of a cell phone

Description automatically generated

**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 1979 to 2011 (1 point).**

1. Average HIV prevalence for each country was calculated from 1979 to 2011 and the results were stored in the new variable “average1979to2011”. Countries were then grouped by their respective continent based on the averages calculated from 1979 to 2011.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

1. A new table was created with columns 'Years', 'Africa', 'Asia', 'Europe', 'North America', 'South America', and 'Oceania' and their values from the data set was imported to this new table to make indexing more simplified for plotting the charts.

A screenshot of a cell phone

Description automatically generated

1. Codes and line chart for each continent showing changes in the average HIV estimated prevalence from 1979 to 2011.

A screenshot of a cell phone

Description automatically generated

A close up of a map

Description automatically generated

A close up of a map

Description automatically generated

1. Codes and overlaid line chart showing estimated HIV prevalence (1979-2011) by continent.

A screenshot of a social media post

Description automatically generated

**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 scatterplot 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.**

1. A new variable was created to store records from the columns “Country”, “Continent”, and “1990” from the merged\_HIV file that had all the original data of all the countries labeled by their respective continent. The values for each country for year 1990 was pulled and stored in each new variable, which were then used for visualization. Same was done for year 2010.

A screenshot of a social media post

Description automatically generated

1. Scatter plot showing HIV prevalence data of each country for year 1990.

A screenshot of a cell phone

Description automatically generated

The scatter plot for year 1990 looks empty because most of the data values were ‘Nan’ or too small to fit in within the axis limits. The axis limits can be customized to fit the data (in progress).

1. Scatter plot showing HIV prevalence data of each country for year 1990.

A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated

The chart for year 2010 clearly shows that the estimated HIV prevalence in Africa is much higher than in any other continent, followed by North America. Other continents have values higher than averages, but they seem to be doing better in controlling the spread of HIV.