ASSIGNMENT 1 REPORT

APPLIED DATA SCIENCE I

Kavitha Subramaniyam

St.ID 22097222

10.11.2023

INTRODUCTION

This report is to visualize selected data in three plot graphs such as line plot, pie chart and bar chart which are created using panda and matplotlib with data taken from web resources.

DATA

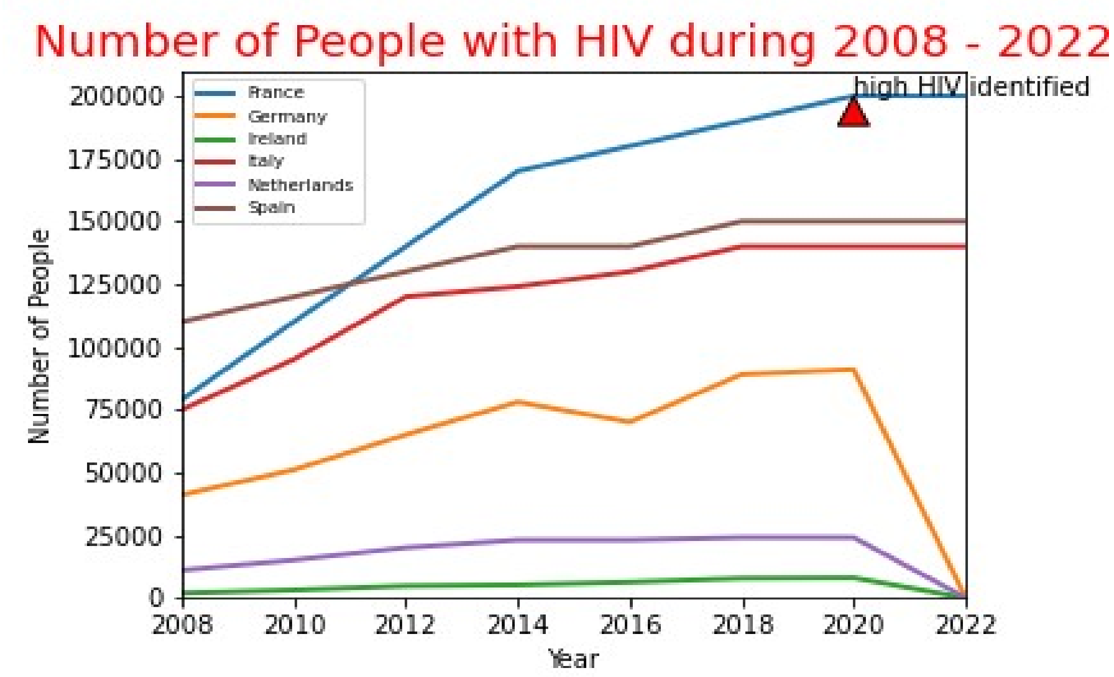
We can see what data is taken for each plot and how it is manipulated in this table.

GRAPH TYPES	DATA	DESCRIPTION
Line plot	Estimated(total) number of people(all ages) living with HIV [1]	Data filtered with below condition: - During year from 2008 to 2022 - Selected(wealthy) 6 countries from WHO region
Pie chart	Estimated(total) number of people(all ages) living with HIV [1]	Data filtered with - From year 2000 to 2022 Data manipulated to get sum of population throughout all years by grouping country
Bar chart	Proportion of population with primary reliance on fuels and technology [2]	Data filtered with below condition: - Year of 2020 - South-East Asia region countries

RESULTS

Visualization I : Line plot

Total number of people(all age group) living with HIV during 2008 to 2022 in wealthiest countries from WHO region



This line plot shows the evolution over time of the HIV population. Data collected from 2008 to 2022 for wealthiest countries in WHO region.

According to this line plot, France has high growth throughout the years and reached high HIV risk in 2020 and 2022. We can notice Ireland has zero HIV affected in 2008 and slightly increased to some number but yet no considerable risk. Also a good observation is that the Netherlands and Germany had a considerable number of people affected in 2008 and got reduced to zero in 2022.

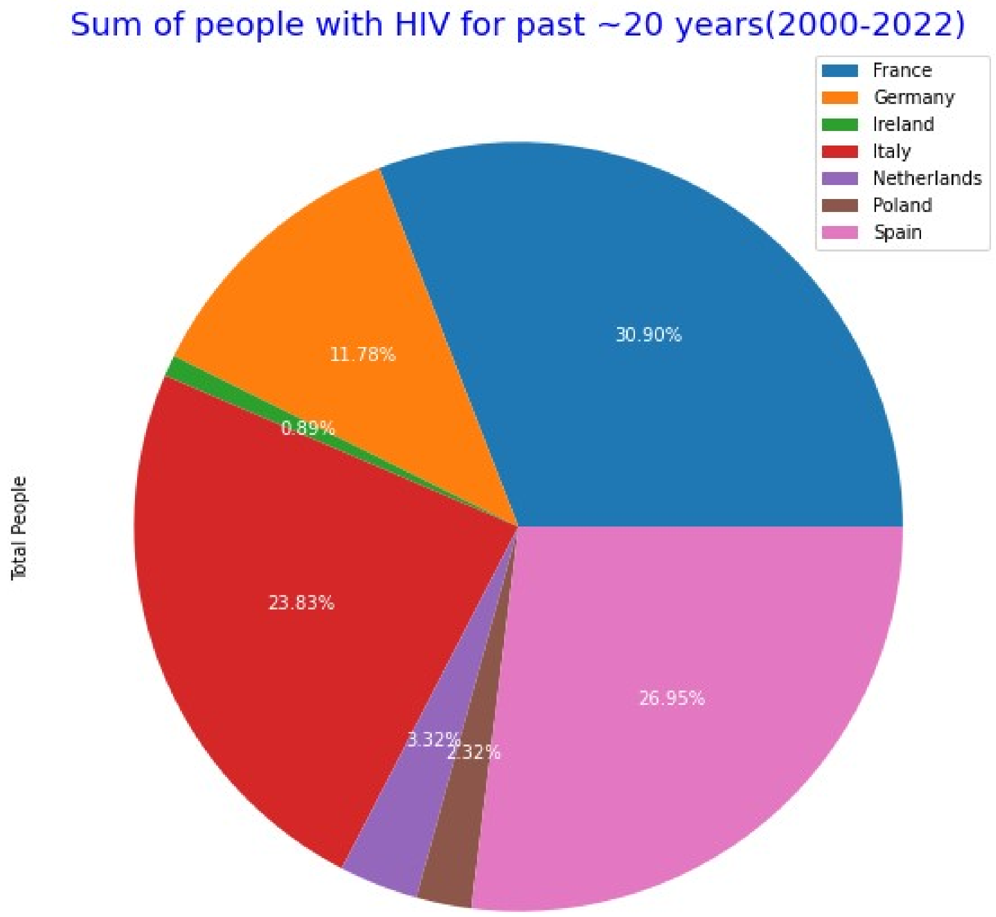
Script:

```python
# -*- coding: utf-8 -*-
"""
Created on Sun Nov  5 18:37:31 2023

@author: kthat
"""

import matplotlib.pyplot as plt
import pandas as pd


def draw_lineplot(df, countries):
    """
        This method is used to create a line plot. Arguments:
        A dataframe with a column 'Period' (used for years) taken as x and
        other columns taken as y. List 'country' containing column to plot.
    """
    plt.figure()

    for country in countries:
        plt.plot(df["Period"], df[country], label=country, linewidth=2.0)

    plt.annotate("high HIV identified", xy=(2020, 200000),
                 arrowprops=dict(facecolor='red', shrink=0.05))

    # Add label
    plt.xlabel("Year")
    plt.ylabel("Number of People")

    # Add title and other adjustments
    plt.title("Number of People with HIV during 2008 - 2022", color='red',
              fontsize=18)
    plt.legend(loc="center right", fontsize='small')

    # Remove white space from left and right and yaxis
    plt.xlim(min(df["Period"]))
    plt.ylim(ymin=0)

    # Save as image
    plt.savefig("line_plot.png")

    plt.show()
    return


# Read data file
data = pd.read_csv(
    "C:\\Users\\kthat\\OneDrive\\data\\number_people_with_HIV_sheet2.csv")

df = pd.DataFrame(data)

# Replace nan values with zeros
df = df.fillna(0)

# Get column names by excluding first index(Period) column
countries = df.columns[1:]

# Call function
draw_lineplot(df, countries)
```

Source code : https://github.com/kavithasub/uh-study-modules/blob/main/applied-data-science/assignment/assignment-1/lineplot.py

Visualization II : Pie chart

Sum of people with HIV in wealthiest WHO region countries for past ~20 years



This pie chart describes the sum of the number of people affected by HIV in wealthiest countries from the WHO region.

According to the pie chart, France and Spain are at highest risk of HIV impact as they have 30.9% & 26.95% total HIV population in order. Comparatively Ireland and Poland have very less HIV impact whereas Italy and Germany have a growing number of HIV people.

Script:

```
# -*- coding: utf-8 -*-
"""
Created on Mon Nov  6 14:50:22 2023
```

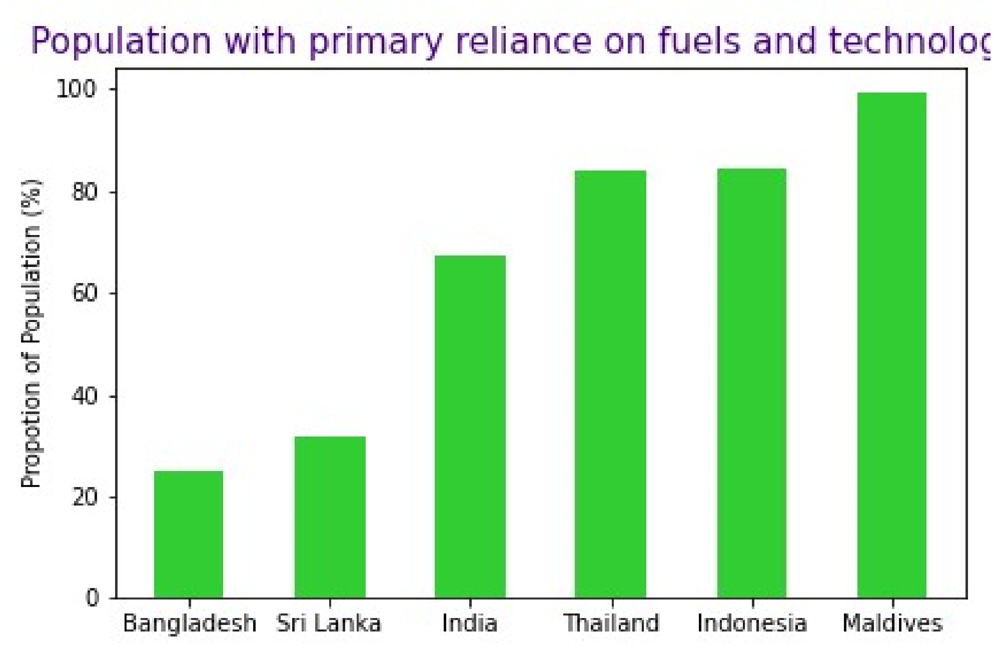
```
@author: kthat
"""
import matplotlib.pyplot as plt
import pandas as pd


def draw_pieplot(df):
    """
        This method is used to create a pie plot. Arguments:
        A dataframe with total population with HIV from each country taken
as y
    """

    plt.figure()
    plot = df.plot.pie(y="Number_Of_People", figsize=(
        8, 8), autopct='%1.2f%%', textprops=dict(color='w'))

    # Add label and title
    plot.set_ylabel("Total People")
    plt.title("Sum of people with HIV for past ~20 years(2000-2022)",
fontsize=18, color="blue")
    plt.legend(loc="upper right")

    # Save as png image
    plt.savefig("pie_plot.png")
    plt.show()

    return


# Read data file
data = pd.read_csv(
    "C:\\Users\\kthat\\OneDrive\\data\\number_people_with_HIV.csv")

# Create data frame
df = pd.DataFrame(data)
print(df)

# Replace NAN values from number_of_people column with zeros
df["Number_Of_People"] = df["Number_Of_People"].fillna(0)
```

```
# Get total number of people for each country through out given years
df_manipulate = df.groupby(["Country"]).sum(["Number_Of_People"])
print(df_manipulate)
# Call function
draw_pieplot(df_manipulate)
```

Source code : https://github.com/kavithasub/uh-study-modules/blob/main/applied-data-science/assignment/assignment-1/pieplot.py

Visualization III : Bar chart

Proportion of population (%) with primary reliance on fuels and technology



This bar chart shows population(%) primarily rely on fuels and technology as a proportion of total population in South East region countries. Having more fuels and technology usage for cooking etc will be an indicator for air pollution.

According to the bar chart, the high usage is shown in Maldives and it is nearly reaching

99% of total population. Overall, most of the countries from the SouthEast region are highly relying on fuels and technology use for primary sources except Sri lanka and Bangladesh where they have below 50% of population.

Script:

```
# -*- coding: utf-8 -*-
"""
Created on Tue Nov  7 13:36:27 2023

@author: kthat
"""

import pandas as pd
import matplotlib.pyplot as plt


def draw_barplot(df_m):
    """
    This method is used to create a bar plot. Arguments:
    A dataframe with a column 'Period' (used for years) taken as x and
    other columns taken as y. List 'country' containing column to plot.
    """

    fig, ax = plt.subplots()
    ax.bar(df_m["Country"], df_m["ValueAccurate"],
           width=0.5, color="limegreen")

    # Add label and title
    ax.set_ylabel('Propotion of Population (%)')
    ax.set_title('Population with primary reliance on fuels and
technologies',
                 color="indigo", fontsize=15)

    # Save as png image
    plt.savefig("bar_plot.png")
    plt.show()

    return


# Read data file
data = pd.read_csv(
"C:\\Users\\kthat\\OneDrive\\data\\popuation-cleanfuels-technology2.csv")

# Create data frame
df = pd.DataFrame(data)

# Remove column 'ParentLocation' which is not useful
df = df.drop(columns=["ParentLocation", "IsLatestYear", "CountryCode"])

# Rename column 'FactValueNumeric' to 'ValueAccurate' for meaningful
df = df.rename(columns={"FactValueNumeric": "ValueAccurate"})

# Call function
draw_barplot(df)
```

Source code : https://github.com/kavithasub/uh-study-modules/blob/main/applied-data-science/assignment/assignment-1/barplot.py

REFERENCES

[1] Dataset 1 :

https://www.who.int/data/gho/data/indicators/indicator-details/GHO/estimated-number-of-people--living-with-hiv

[2] Dataset 2 :

https://www.who.int/data/gho/data/themes/air-pollution/household-air-pollution

[3] Plot styles :
https://matplotlib.org/stable/gallery/lines_bars_and_markers/index.html

[4] Dataframe manipulation :

https://pandas.pydata.org/docs/reference/frame.html