University
of Economics, Law and
Social Sciences, International
Relations and Computer Science (HSG)

Group Project 02: United States Region Data

Gorup ID: 2329

Alois Blum 19 – 617 – 976

Arne Christes 17-618-737

Zhiqing Pan 20-624-839

Group Project 02

University of St. Gallen

Skills: Programming – Introduction Level

Dr. Mario Silic

27.05.2021

Contents

1 Intro (What does the project do?)

In our group project „Data United States", a series of functions have been developed to extract meaningful information from a dataset with different economic information about the United States.

The given dataset contains economic information of different states in different regions in the United States. For each state, the Population, GDP, Personal Income, Subsidies, Compensation of Employees and the Taxes on Production and Imports information are given.

With our program, users can extract data of a selected region by entering a region name. Data of all states located in that region and the highest and lowest GDP per Capita and Personal Income information of that region will be displayed. Furthermore, users can choose to plot a scatter plot with a regression line by selecting 2 variables.

2 Process

To make it as easy as possible for the user, all the functions that are retrieved in the last function (main function) are defined first.

First, users must enter a valid filename so that the data can be retrieved. If the user enters a wrong filename, he will be asked to enter a correct filename until he has done so.

The data in its original form contains information on population, GDP, personal income, subsidies, compensation of employees and the taxes on production and imports. In addition, we have calculated the GDP per person and per capita personal income and appended them in two separate columns.

Second, users can enter a region name within the United States to extract data. With the input, users will then receive information on the states with the highest and lowest GDP per capita and per capita personal income in the selected region. In addition, the data of all states in the selected region will be clearly displayed.

Finally, users can opt to visualize the data within a graph. For this purpose, we have implemented a scatter plot with a regression line. Users can select 2 variables within his previously selected region and compare between the individual states.

3 Code with Description

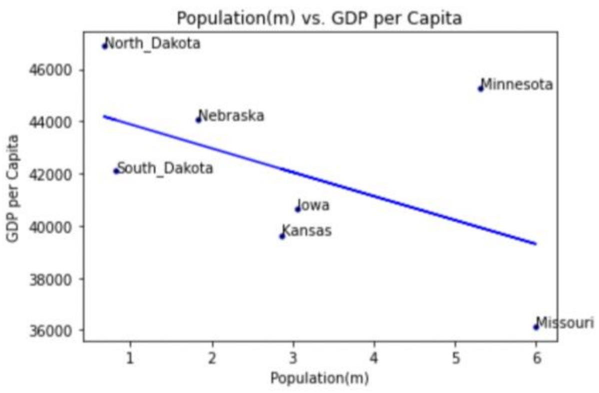
```
import pandas as pd
import matplotlib.pylab as pylab
import matplotlib.pyplot as plt


REGION_LIST = ['Far_West','Great_Lakes', 'Mideast', 'New_England', 'Plains', 'Rocky_Mountain',
'Southeast', 'Southwest','all']
VALUES_LIST = ['Pop', 'GDP', 'PI', 'Sub', 'CE', 'TPI', 'GDPp', 'PIp']



#Create dictionaries for user input
thisdic = {"far_west":"Far_West","great_lakes":"Great_Lakes",
"mideast":"Mideast","new_england":"New_England", "plains":"Plains",
"rocky_mountain":"Rocky_Mountain", "southeast":"Southeast",
"southwest":"Southwest","all":"all"}

thisdic1 = {"Pop":"Population (millions)", "GDP":"GDP (billions)", "PI":"Personal Income
(billions)","Sub":"Subsidies (millions)","CE":"Comp of Emp (billions)","TPI":"Tax on Prod/Imp
(billions)", "GDPp":"GDP per Capita","PIp":"Per capita personal income"}

thisdic2 = {"Pop":"Population(m)","GDP":
"GDP(b)","PI":"Income(b)","Sub":"Subsidies(m)","CE":"Compensation(b)","TPI":"Taxes(b)","GDPp":
"GDP per Capita","PIp":"Per capita personal income"}



#Create function to read file with given filename

def read_file():
    try:
        file_name = input("What is the filename? ")
        df = pd.read_csv(file_name) #read the csv file into dataframe
        return df                   #return dataframe
    except FileNotFoundError:
        print("File not found").    #if filename cannot be found, ask to enter again
        df = read_file()
        return df



#Create function to calculate GDP per Capita and Per capita personla income

def add_GDPp_PIp(df):
  df['GDP per Capita'] = round(df['GDP (billions)'] / df['Population (millions)']*1000, 2)
#calculate and append GDP per Capita to df
  df['Per capita personal income'] = round(df['Personal Income (billions)'] / df['Population
(millions)']*1000, 2) #calculate and append Per capita personal income to df
  return df



#Create function to get region input

def get_region_data(df):

  while True:
    data = input("Specify a region from this list --
far_west,great_lakes,mideast,new_england,plains,rocky_mountain,southeast,southwest,all:
").lower()   #change user input into lowercase
```

```
    if data == "all":      #If user input == "all", return full df and data
          return df, data

    elif data in (x.lower() for x in REGION_LIST): #If data is in region list, return df of
selected region and data
          df = df.loc[df["Region"]==thisdic[data]]
          return df, data


    else:             #If data is not all or in region list, keep prompting
        print("Sorry, your response is invalid.")



# Create function to find out the states with highest and lowest GDP per capita and Per capita
personal income
def get_max_min(df, data):
    state_max1 = df['State'][df['GDP per Capita'] == df['GDP per
Capita'].max()].to_string(index=False).replace(" ", "")
    GDP_max = df['GDP per Capita'][df['GDP per Capita'] == df['GDP per Capita'].max()].apply(
        lambda x: "${:,.2f}".format(x)).to_string(index=False)


    state_min1 = df['State'][df['GDP per Capita'] == df['GDP per
Capita'].min()].to_string(index=False).replace(" ", "")
    GDP_min = df['GDP per Capita'][df['GDP per Capita'] == df['GDP per Capita'].min()].apply(
        lambda x: "${:,.2f}".format(x)).to_string(index=False)


    state_max2 = df['State'][df['Per capita personal income'] == df['Per capita personal
income'].max()].to_string(
        index=False).replace(" ", "")
    Income_max = df['Per capita personal income'][
        df['Per capita personal income'] == df['Per capita personal income'].max()].apply(
        lambda x: "${:,.2f}".format(x)).to_string(index=False)


    state_min2 = df['State'][df['Per capita personal income'] == df['Per capita personal
income'].min()].to_string(
        index=False).replace(" ", "")
    Income_min = df['Per capita personal income'][
        df['Per capita personal income'] == df['Per capita personal income'].min()].apply(
        lambda x: "${:,.2f}".format(x)).to_string(index=False)


    pd.options.display.float_format = '{:,.2f}'.format  # add format(commas every thousands
and round to 2 decimal places) to floats in df
    pd.set_option("display.max_rows", None,"display.max_columns", None)  # set option to show
all columns of data in the final output
    pd.set_option('max_colwidth', 1000)

    print("This is the data for the", thisdic[data], "region: ")
    print('')
    print(state_max1, "has the highest GDP per Capita of the region at: ", GDP_max)
    print(state_min1, "has the lowest GDP per Capita of the region at: ", GDP_min)
    print('')
    print(state_max2, "has the highest per capita personal income of the region at: ",
Income_max)
    print(state_min2, "has the highest per capita personal income of the region at: ",
Income_min)
    print('')
    print("Data for all states in the", thisdic[data], "region: ")
    print('')
    print(df)



#Create function to get plot inputs
def get_plot_input():
  while True:
    try:
      PROMPT2 = input("Specify x and y values, space separated from Pop, GDP, PI, Sub, CE,
TPI, GDPp, PIp: ").split(" ") #split the input
      x = PROMPT2[0] #First input
      y = PROMPT2[1] #Second input
      if x and y in VALUES_LIST:  #If both inputs are in Values_List, output x and y
          return x, y
      else:                       #If not, keep promting
          print("Please enter values exactly as suggested")
    except IndexError:            #If only one value is entered, ask to enter again
      print("Please type in two inputs")



#Create function for annotations
def label_point(x, y, val, ax):
    a = pd.concat({'x': x, 'y': y, 'val': val}, axis=1)
    for i, point in a.iterrows():
        ax.text(point['x'], point['y'], str(point['val']), size = 10)


#Create function for plotting scatter plot and regressino line
def plot_sct_reg(x,y,df):
    #scatter plot
    plt.scatter(df[thisdic1[x]], df[thisdic1[y]],c="darkblue",s=10)  #plot the scatter plot
    pylab.xlabel(thisdic2[x])   #label x axis
    pylab.ylabel(thisdic2[y])   #label y axis
    plt.title(thisdic2[x] + " vs. " + thisdic2[y])  #add title to plot
    label_point(df[thisdic1[x]], df[thisdic1[y]], df["State"], plt) #add annotations to each
point

    #regression
    xarr = pylab.array(df[thisdic1[x]]) #numpy array
    yarr = pylab.array(df[thisdic1[y]]) #numpy arry
    m,b = pylab.polyfit(xarr,yarr, deg = 1) #creates line, only takes numpy arrays
    #as parameters
    pylab.plot(xarr,m*xarr + b, '-',c="blue") #plotting the regression line
    plt.show()



#Call all the functions in main function
def main():
    # function1:input file name and read file into dataframe
    df = read_file()

    # function2: input region name
    df = add_GDPp_PIp(df)
    df, data = get_region_data(df)

    # function3:display data with selected region.
    get_max_min(df, data)

    # function4:ask user to choose whether or not to do the plot
    while True:
        option = input("Do you want to make a plot?")

        # If user enter "yes", continue to plot, then break
        if option.lower() == "yes":
            x, y = get_plot_input()
            plot_sct_reg(x, y, df)
            break

        # If user enter "no", break
        if option.lower() == "no":
            break

        # If user enter string outside yes/no, keep prompting
        else:
            print("Please enter either Yes or No.")


main()
```

4 Example input and output

1 Example input and output:

```
What is the filename? State_Data.csv
Specify a region from this list --
far_west,great_lakes,mideast,new_england,plains,rocky_mountain,southeast,southwest,all: plains
This is the data for the Plains region:

North_Dakota has the highest GDP per Capita of the region at:   $46,893.07
Missouri has the lowest GDP per Capita of the region at:   $36,136.82

North_Dakota has the highest per capita personal income of the region at:   $43,235.65
Missouri has the highest per capita personal income of the region at:   $36,604.46

Data for all states in the Plains region:

           State  Region  Population (millions)  GDP (billions)  \
15          Iowa  Plains                   3.05          124.01
16        Kansas  Plains                   2.86          113.32
23     Minnesota  Plains                   5.31          240.42
25      Missouri  Plains                   6.00          216.68
27      Nebraska  Plains                   1.83           80.64
34  North_Dakota  Plains                   0.67           31.62
41  South_Dakota  Plains                   0.82           34.37

    Personal Income (billions)  Subsidies (millions)  Comp of Emp (billions)  \
15                      119.08                  1039                   72.04
16                      110.88                   777                   71.43
23                      226.32                  1327                  154.01
25                      219.48                   911                  142.84
27                       73.07                   313                   47.02
34                       29.15                   603                   18.64
41                       33.14                   617                   18.24

    Tax on Prod/Imp (billions)  GDP per Capita  Per capita personal income
15                        9.10       40,655.02                   39,038.68
16                        9.03       39,639.01                   38,787.15
23                       18.85       45,270.87                   42,615.83
25                       14.96       36,136.82                   36,604.46
27                        5.65       44,072.80                   39,935.02
34                        2.37       46,893.07                   43,235.65
41                        2.73       42,109.78                   40,597.53

Do you want to make a plot?yes
Specify x and y values, space separated from Pop, GDP, PI, Sub, CE, TPI, GDPp, PIp: Pop GDPp
```



2 Error checks for input

(1) Filename check:

```
What is the filename? State data
File not found
```

(2) Region name check:

Situation 1: input not in region lists

```
Specify a region from this list --
far_west,great_lakes,mideast,new_england,plains,rocky_mountain,southeast,southwest,all
: hjj
Sorry, your response is invalid.
```

Situation 2: Case insensitivity

```
Specify a region from this list --
far_west,great_lakes,mideast,new_england,plains,rocky_mountain,southeast,southwest,all
: souTHwest
This is the data for the Southwest region:

Texas has the highest GDP per Capita of the region at:   $44,221.50
New_Mexico has the lowest GDP per Capita of the region at:   $34,284.19

Texas has the highest per capita personal income of the region at:   $38,103.22
New_Mexico has the highest per capita personal income of the region at:   $33,169.85

Data for all states in the Southwest region:

            State     Region  Population (millions)  GDP (billions)  \
2      Arizona  Southwest                   6.41          221.02
30  New_Mexico  Southwest                   2.06           70.79
36    Oklahoma  Southwest                   3.76          132.92
43       Texas  Southwest                  25.24        1,116.27

    Personal Income (billions)  Subsidies (millions)  Comp of Emp (billions)  \
2                       217.76                   763                  135.60
30                       68.49                   302                   42.64
36                      135.06                   448                   80.06
43                      961.83                  2887                  621.10

    Tax on Prod/Imp (billions)  GDP per Capita  Per capita personal income
2                        16.94       34,476.20                   33,967.48
30                        5.55       34,284.19                   33,169.85
36                        9.23       35,355.77                   35,925.68
43                       93.06       44,221.50                   38,103.22
```

(3) Options to plot check:

Situation 1: input is not yes or no

```
Do you want to make a plot?ho
Please enter either Yes or No.
```

Situation 2: case insensitivity

```
Do you want to make a plot?yeS
Specify x and y values, space separated from Pop, GDP, PI, Sub, CE, TPI, GDPp, PIp:
```

(4) Variables for plotting check

Situation 1: only one input is entered

```
Specify x and y values, space separated from Pop, GDP, PI, Sub, CE, TPI, GDPp, PIp: Po
Please type in two inputs
```

Situation 2: inputs are not in value lists

```
Specify x and y values, space separated from Pop, GDP, PI, Sub, CE, TPI, GDPp, PIp:
GEP PIO
Please enter values exactly as suggested
```