labassignment2

June 25, 2024

1 Lab Assignment 2: Qais Youssef (qmy6cv)

1.1 DS 6001: Practice and Application of Data Science

1.1.1 Instructions

Please answer the following questions as completely as possible using text, code, and the results of code as needed. Format your answers in a Jupyter notebook. To receive full credit, make sure you address every part of the problem, and make sure your document is formatted in a clean and professional way.

There are 11 data files attached to this lab assignment, with different extensions. First, download all of these data files, and save them in the same folder on your local machine. Your task in the following questions is to load each file into Python correctly, so that you can begin the process of data cleaning. If the variable names are included in the file, use those names to name the columns. If the variable names are not included, use these names in order:

[ ]:
```
column_names = ["Country", "Happiness score", "Whisker-high", "Whisker-low",
  "Dystopia (1.92) + residual", "Explained by: GDP per capita",
  "Explained by: Social support", "Explained by: Healthy life expectancy",
  "Explained by: Freedom to make life choices", "Explained by: Generosity",
  "Explained by: Perceptions of corruption" ]
```

If you loaded the data correctly, it will look like `data_clean.csv`, which is also attached to this lab.

1.2 Problem 0

Import the libraries you will need. Then write code to change the working directory to the folder in which you saved the data files, run the code displayed above to create the `column_names` list, load `data_clean.csv`, and display the output of the `.info()` method of `data_clean`. (1 point)

[ ]:
```
import os
import pandas as pd

data_folder = r"C:\Users\qaism\OneDrive - University of␣
 ↪Virginia\Documents\GitHub\MSDS\DS 6001\Lab2\lab2data"
os.chdir(data_folder)
data_clean = pd.read_csv("data_clean.csv")
print(data_clean.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 156 entries, 0 to 155
Data columns (total 11 columns):
 #   Column                                     Non-Null Count  Dtype
---  ------                                     --------------  -----
 0   Country                                    156 non-null    object
 1   Happiness score                            156 non-null    float64
 2   Whisker-high                               156 non-null    float64
 3   Whisker-low                                156 non-null    float64
 4   Dystopia (1.92) + residual                 156 non-null    float64
 5   Explained by: GDP per capita               156 non-null    float64
 6   Explained by: Social support               156 non-null    float64
 7   Explained by: Healthy life expectancy      156 non-null    float64
 8   Explained by: Freedom to make life choices 156 non-null    float64
 9   Explained by: Generosity                   156 non-null    float64
 10  Explained by: Perceptions of corruption    156 non-null    float64
dtypes: float64(10), object(1)
memory usage: 13.5+ KB
None
```

1.3 Problem 1

Load `data1.csv`. Use the tools we discussed in class to decide whether the data file loaded correctly, and include that code in your lab report. In one or two sentences, describe how you decided on the right combination of parameters needed to load the data. (1 point)

[ ]:
```
csv_path = r"C:\Users\qaism\OneDrive - University of␣
 ↪Virginia\Documents\GitHub\MSDS\DS 6001\Lab2\lab2data\data1.csv"

data1 = pd.read_csv(csv_path, skiprows=1, names=column_names)
print(data1.head())
print(data1.info())
```

```
                                    Country  Happiness score  Whisker-high  \
0  URL: http://worldhappiness.report/ed/2018              NaN           NaN
1                                    Country  Happiness score  Whisker-high
2                                    Finland            7.632         7.695
3                                     Norway            7.594         7.657
4                                    Denmark            7.555         7.623

   Whisker-low  Dystopia (1.92) + residual  Explained by: GDP per capita  \
0          NaN                         NaN                           NaN
1  Whisker-low  Dystopia (1.92) + residual  Explained by: GDP per capita
2        7.569                       2.595                         1.305
3        7.530                       2.383                         1.456
4        7.487                       2.370                         1.351

   Explained by: Social support  Explained by: Healthy life expectancy  \
```

```
0                           NaN                                    NaN
1  Explained by: Social support  Explained by: Healthy life expectancy
2                         1.592                                  0.874
3                         1.582                                  0.861
4                         1.590                                  0.868

   Explained by: Freedom to make life choices  Explained by: Generosity  \
0                                         NaN                       NaN
1  Explained by: Freedom to make life choices  Explained by: Generosity
2                                       0.681                     0.192
3                                       0.686                     0.286
4                                       0.683                     0.284

   Explained by: Perceptions of corruption
0                                      NaN
1  Explained by: Perceptions of corruption
2                                    0.393
3                                    0.340
4                                    0.408
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 158 entries, 0 to 157
Data columns (total 11 columns):
 #   Column                                     Non-Null Count  Dtype
---  ------                                     --------------  -----
 0   Country                                    158 non-null    object
 1   Happiness score                            157 non-null    object
 2   Whisker-high                               157 non-null    object
 3   Whisker-low                                157 non-null    object
 4   Dystopia (1.92) + residual                 157 non-null    object
 5   Explained by: GDP per capita               157 non-null    object
 6   Explained by: Social support               157 non-null    object
 7   Explained by: Healthy life expectancy      157 non-null    object
 8   Explained by: Freedom to make life choices  157 non-null    object
 9   Explained by: Generosity                   157 non-null    object
 10  Explained by: Perceptions of corruption    157 non-null    object
dtypes: object(11)
memory usage: 13.7+ KB
None
```

Answer: I saw that the first row contained metadata of some sort, so I skipped it using skiprows=1. The names=column_names parameter assigns the correct column names to the DataFrame.

1.4 Problem 2

Load `data2.txt`. Use the tools we discussed in class to decide whether the data file loaded correctly, and include that code in your lab report. In one or two sentences, describe how you decided on the right combination of parameters needed to load the data. (1 point)

```
[ ]: txt_path = r"C:\Users\qaism\OneDrive - University of␣
      ↪Virginia\Documents\GitHub\MSDS\DS 6001\Lab2\lab2data\data2.txt"
     data2 = pd.read_csv(txt_path, skiprows=2)
     print(data2.head())
     print(data2.info())
```

```
                                              Country  Happiness score  \
0  /The following countries comprise the "very ha…              NaN
1                                            Finland            7.632
2                                             Norway            7.594
3                                            Denmark            7.555
4                                            Iceland            7.495

   Whisker-high  Whisker-low  Dystopia (1.92) + residual  \
0           NaN          NaN                         NaN
1         7.695        7.569                       2.595
2         7.657        7.530                       2.383
3         7.623        7.487                       2.370
4         7.593        7.398                       2.426

   Explained by: GDP per capita  Explained by: Social support  \
0                           NaN                           NaN
1                         1.305                         1.592
2                         1.456                         1.582
3                         1.351                         1.590
4                         1.343                         1.644

   Explained by: Healthy life expectancy  \
0                                    NaN
1                                  0.874
2                                  0.861
3                                  0.868
4                                  0.914

   Explained by: Freedom to make life choices  Explained by: Generosity  \
0                                         NaN                       NaN
1                                       0.681                     0.192
2                                       0.686                     0.286
3                                       0.683                     0.284
4                                       0.677                     0.353

   Explained by: Perceptions of corruption
0                                      NaN
1                                    0.393
2                                    0.340
3                                    0.408
4                                    0.138
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 159 entries, 0 to 158
Data columns (total 11 columns):
 #   Column                                     Non-Null Count  Dtype
---  ------                                     --------------  -----
 0   Country                                    159 non-null    object
 1   Happiness score                            156 non-null    float64
 2   Whisker-high                               156 non-null    float64
 3   Whisker-low                                156 non-null    float64
 4   Dystopia (1.92) + residual                 156 non-null    float64
 5   Explained by: GDP per capita               156 non-null    float64
 6   Explained by: Social support               156 non-null    float64
 7   Explained by: Healthy life expectancy      156 non-null    float64
 8   Explained by: Freedom to make life choices 156 non-null    float64
 9   Explained by: Generosity                   156 non-null    float64
 10  Explained by: Perceptions of corruption    156 non-null    float64
dtypes: float64(10), object(1)
memory usage: 13.8+ KB
None
```

Answer: To load data2.txt, I skipped the first two rows containing metadata using skiprows=2 and let pandas use the column names provided in the third row of the file.

1.5 Problem 3

Load `data3.txt`. Use the tools we discussed in class to decide whether the data file loaded correctly, and include that code in your lab report. In one or two sentences, describe how you decided on the right combination of parameters needed to load the data. (1 point)

```
[ ]: txt_path = r"C:\Users\qaism\OneDrive - University of␣
      ↪Virginia\Documents\GitHub\MSDS\DS 6001\Lab2\lab2data\data3.txt"
data3 = pd.read_csv(txt_path, skiprows=2, sep='\t')
print(data3.head())
print(data3.info())
```

```
       Country  Happiness score  Whisker-high  Whisker-low  \
0      Finland            7.632         7.695        7.569
1       Norway            7.594         7.657        7.530
2      Denmark            7.555         7.623        7.487
3      Iceland            7.495         7.593        7.398
4  Switzerland            7.487         7.570        7.405

   Dystopia (1.92) + residual  Explained by: GDP per capita  \
0                       2.595                         1.305
1                       2.383                         1.456
2                       2.370                         1.351
3                       2.426                         1.343
4                       2.320                         1.420

   Explained by: Social support  Explained by: Healthy life expectancy  \
```

```
0                          1.592                                    0.874
1                          1.582                                    0.861
2                          1.590                                    0.868
3                          1.644                                    0.914
4                          1.549                                    0.927

   Explained by: Freedom to make life choices  Explained by: Generosity  \
0                                       0.681                     0.192
1                                       0.686                     0.286
2                                       0.683                     0.284
3                                       0.677                     0.353
4                                       0.660                     0.256

   Explained by: Perceptions of corruption
0                                    0.393
1                                    0.340
2                                    0.408
3                                    0.138
4                                    0.357
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 156 entries, 0 to 155
Data columns (total 11 columns):
 #   Column                                     Non-Null Count  Dtype
---  ------                                     --------------  -----
 0   Country                                    156 non-null    object
 1   Happiness score                            156 non-null    float64
 2   Whisker-high                               156 non-null    float64
 3   Whisker-low                                156 non-null    float64
 4   Dystopia (1.92) + residual                 156 non-null    float64
 5   Explained by: GDP per capita               156 non-null    float64
 6   Explained by: Social support               156 non-null    float64
 7   Explained by: Healthy life expectancy      156 non-null    float64
 8   Explained by: Freedom to make life choices  156 non-null    float64
 9   Explained by: Generosity                   156 non-null    float64
 10  Explained by: Perceptions of corruption    156 non-null    float64
dtypes: float64(10), object(1)
memory usage: 13.5+ KB
None
```

Answer: To load data3.txt, I skipped the first two rows containing metadata using skiprows=2 and specified the file as tab-separated using sep='\t'.

1.6 Problem 4

Load `data4.txt`. Use the tools we discussed in class to decide whether the data file loaded correctly, and include that code in your lab report. In one or two sentences, describe how you decided on the right combination of parameters needed to load the data. (1 point)

[ ]:
```
txt_path = r"C:\Users\qaism\OneDrive - University of␣
 ↪Virginia\Documents\GitHub\MSDS\DS 6001\Lab2\lab2data\data4.txt"
column_names = ["Country", "Happiness score", "Whisker-high", "Whisker-low",
                "Dystopia (1.92) + residual", "Explained by: GDP per capita",
                "Explained by: Social support", "Explained by: Healthy life␣
 ↪expectancy",
                "Explained by: Freedom to make life choices", "Explained by:␣
 ↪Generosity",
                "Explained by: Perceptions of corruption"]
data4 = pd.read_csv(txt_path, sep='$', names=column_names)
print(data4.head())
print(data4.info())
```

```
       Country  Happiness score  Whisker-high  Whisker-low  \
0      Finland            7.632         7.695        7.569
1       Norway            7.594         7.657        7.530
2      Denmark            7.555         7.623        7.487
3      Iceland            7.495         7.593        7.398
4  Switzerland            7.487         7.570        7.405

   Dystopia (1.92) + residual  Explained by: GDP per capita  \
0                       2.595                         1.305
1                       2.383                         1.456
2                       2.370                         1.351
3                       2.426                         1.343
4                       2.320                         1.420

   Explained by: Social support  Explained by: Healthy life expectancy  \
0                         1.592                                  0.874
1                         1.582                                  0.861
2                         1.590                                  0.868
3                         1.644                                  0.914
4                         1.549                                  0.927

   Explained by: Freedom to make life choices  Explained by: Generosity  \
0                                       0.681                    0.192
1                                       0.686                    0.286
2                                       0.683                    0.284
3                                       0.677                    0.353
4                                       0.660                    0.256

   Explained by: Perceptions of corruption
0                                    0.393
1                                    0.340
2                                    0.408
3                                    0.138
4                                    0.357
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 156 entries, 0 to 155
Data columns (total 11 columns):
 #   Column                                     Non-Null Count  Dtype
---  ------                                     --------------  -----
 0   Country                                    156 non-null    object
 1   Happiness score                            156 non-null    float64
 2   Whisker-high                               156 non-null    float64
 3   Whisker-low                                156 non-null    float64
 4   Dystopia (1.92) + residual                 156 non-null    float64
 5   Explained by: GDP per capita               156 non-null    float64
 6   Explained by: Social support               156 non-null    float64
 7   Explained by: Healthy life expectancy      156 non-null    float64
 8   Explained by: Freedom to make life choices 156 non-null    float64
 9   Explained by: Generosity                   156 non-null    float64
 10  Explained by: Perceptions of corruption    156 non-null    float64
dtypes: float64(10), object(1)
memory usage: 13.5+ KB
None
```

Answer: To load data4.txt correctly, I specified the delimiter as $ using sep='$' and assigned the appropriate column names to the DataFrame.

1.7 Problem 5

Load `data5.csv`. Use the tools we discussed in class to decide whether the data file loaded correctly, and include that code in your lab report. In one or two sentences, describe how you decided on the right combination of parameters needed to load the data. (1 point)

[ ]:
```
csv_path = r"C:\Users\qaism\OneDrive - University of␣
 ↪Virginia\Documents\GitHub\MSDS\DS 6001\Lab2\lab2data\data5.csv"
data5 = pd.read_csv(csv_path)
print(data5.head())
print(data5.info())
```

```
       Country  Happiness score  Whisker-high  Whisker-low  \
0      Finland            7.632         7.695        7.569
1       Norway            7.594         7.657        7.530
2      Denmark            7.555         7.623        7.487
3      Iceland            7.495         7.593        7.398
4  Switzerland            7.487         7.570        7.405

   Dystopia (1.92) + residual  Explained by: GDP per capita  \
0                       2.595                         1.305
1                       2.383                         1.456
2                       2.370                         1.351
3                       2.426                         1.343
4                       2.320                         1.420
```

```
   Explained by: Social support  Explained by: Healthy life expectancy  \
0                         1.592                                   0.874
1                         1.582                                   0.861
2                         1.590                                   0.868
3                         1.644                                   0.914
4                         1.549                                   0.927

   Explained by: Freedom to make life choices  Explained by: Generosity  \
0                                       0.681                    0.192
1                                       0.686                    0.286
2                                       0.683                    0.284
3                                       0.677                    0.353
4                                       0.660                    0.256

   Explained by: Perceptions of corruption
0                                    0.393
1                                    0.340
2                                    0.408
3                                    0.138
4                                    0.357
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 158 entries, 0 to 157
Data columns (total 11 columns):
 #   Column                                     Non-Null Count  Dtype
---  ------                                     --------------  -----
 0   Country                                    158 non-null    object
 1   Happiness score                            156 non-null    float64
 2   Whisker-high                               156 non-null    float64
 3   Whisker-low                                156 non-null    float64
 4   Dystopia (1.92) + residual                 156 non-null    float64
 5   Explained by: GDP per capita               156 non-null    float64
 6   Explained by: Social support               156 non-null    float64
 7   Explained by: Healthy life expectancy      156 non-null    float64
 8   Explained by: Freedom to make life choices  156 non-null    float64
 9   Explained by: Generosity                   156 non-null    float64
 10  Explained by: Perceptions of corruption    156 non-null    float64
dtypes: float64(10), object(1)
memory usage: 13.7+ KB
None
```

Answer: To load data5.csv, I used the pd.read_csv() function without additional parameters.

1.8 Problem 6

Load `data6.dat`. Use the tools we discussed in class to decide whether the data file loaded correctly, and include that code in your lab report. In one or two sentences, describe how you decided on the right combination of parameters needed to load the data. (1 point)

```
[ ]: dat_path = r"C:\Users\qaism\OneDrive - University of␣
      ↪Virginia\Documents\GitHub\MSDS\DS 6001\Lab2\lab2data\data6.dat"
     data6 = pd.read_csv(dat_path, sep=',', names=column_names, skiprows=1)
     print(data6.head())
     print(data6.info())
```

```
       Country  Happiness score  Whisker-high  Whisker-low  \
0      Finland            7.632         7.695        7.569
1       Norway            7.594         7.657        7.530
2      Denmark            7.555         7.623        7.487
3      Iceland            7.495         7.593      999.000
4  Switzerland            7.487         7.570        7.405

   Dystopia (1.92) + residual  Explained by: GDP per capita  \
0                       2.595                       999.000
1                     999.000                       999.000
2                       2.370                         1.351
3                       2.426                         1.343
4                       2.320                         1.420

   Explained by: Social support  Explained by: Healthy life expectancy  \
0                       999.000                                999.000
1                         1.582                                999.000
2                         1.590                                999.000
3                         1.644                                  0.914
4                         1.549                                  0.927

   Explained by: Freedom to make life choices  Explained by: Generosity  \
0                                       0.681                    0.192
1                                       0.686                    0.286
2                                       0.683                    0.284
3                                       0.677                    0.353
4                                       0.660                    0.256

   Explained by: Perceptions of corruption
0                                    0.393
1                                    0.340
2                                    0.408
3                                  999.000
4                                    0.357
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 156 entries, 0 to 155
Data columns (total 11 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   Country                                 156 non-null    object
 1   Happiness score                         156 non-null    float64
 2   Whisker-high                            156 non-null    float64
```

```
 3   Whisker-low                                156 non-null    float64
 4   Dystopia (1.92) + residual                 156 non-null    float64
 5   Explained by: GDP per capita               156 non-null    float64
 6   Explained by: Social support               156 non-null    float64
 7   Explained by: Healthy life expectancy      156 non-null    float64
 8   Explained by: Freedom to make life choices 156 non-null    float64
 9   Explained by: Generosity                   156 non-null    float64
 10  Explained by: Perceptions of corruption    156 non-null    float64
dtypes: float64(10), object(1)
memory usage: 13.5+ KB
None
```

Answer: To load data6.dat correctly, I specified the delimiter as a comma using sep=',' and skipped the first row containing the column headers using skiprows=1.

1.9 Problem 7

Load `data7.xlsx`, which is an Excel file. Keep only the sheet named "Data". Use the tools we discussed in class to decide whether the data file loaded correctly, and include that code in your lab report. In one or two sentences, describe how you decided on the right combination of parameters needed to load the data. (2 points)

[ ]:
```
import openpyxl as px

xlsx_path = r"C:\Users\qaism\OneDrive - University of␣
 ↪Virginia\Documents\GitHub\MSDS\DS 6001\Lab2\lab2data\data7.xlsx"
data7 = pd.read_excel(xlsx_path, sheet_name='Data')
print(data7.head())
print(data7.info())
```

```
       Country  Happiness score  Whisker-high  Whisker-low  \
0      Finland            7.632         7.695        7.569
1       Norway            7.594         7.657        7.530
2      Denmark            7.555         7.623        7.487
3      Iceland            7.495         7.593        7.398
4  Switzerland            7.487         7.570        7.405

   Dystopia (1.92) + residual  Explained by: GDP per capita  \
0                       2.595                         1.305
1                       2.383                         1.456
2                       2.370                         1.351
3                       2.426                         1.343
4                       2.320                         1.420

   Explained by: Social support  Explained by: Healthy life expectancy  \
0                         1.592                                  0.874
1                         1.582                                  0.861
2                         1.590                                  0.868
3                         1.644                                  0.914
```

```
4                          1.549                                    0.927

   Explained by: Freedom to make life choices  Explained by: Generosity  \
0                                       0.681                     0.192
1                                       0.686                     0.286
2                                       0.683                     0.284
3                                       0.677                     0.353
4                                       0.660                     0.256

   Explained by: Perceptions of corruption
0                                    0.393
1                                    0.340
2                                    0.408
3                                    0.138
4                                    0.357
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 156 entries, 0 to 155
Data columns (total 11 columns):
 #   Column                                    Non-Null Count  Dtype
---  ------                                    --------------  -----
 0   Country                                   156 non-null    object
 1   Happiness score                           156 non-null    float64
 2   Whisker-high                              156 non-null    float64
 3   Whisker-low                               156 non-null    float64
 4   Dystopia (1.92) + residual                156 non-null    float64
 5   Explained by: GDP per capita              156 non-null    float64
 6   Explained by: Social support              156 non-null    float64
 7   Explained by: Healthy life expectancy     156 non-null    float64
 8   Explained by: Freedom to make life choices  156 non-null    float64
 9   Explained by: Generosity                  156 non-null    float64
 10  Explained by: Perceptions of corruption   156 non-null    float64
dtypes: float64(10), object(1)
memory usage: 13.5+ KB
None
```

Answer: To load data7.xlsx correctly, I specified the sheet name as "Data" using sheet_name='Data'.

1.10 Problem 8

Load `data8.dta`, which is a Stata 13 file. Use the tools we discussed in class to decide whether the data file loaded correctly, and include that code in your lab report. In one or two sentences, describe how you decided on the right combination of parameters needed to load the data. (2 points)

[ ]:
```
dta_path = r"C:\Users\qaism\OneDrive - University of␣
 ↪Virginia\Documents\GitHub\MSDS\DS 6001\Lab2\lab2data\data8.dta"
data8 = pd.read_stata(dta_path)
```

```
print(data8.head())
print(data8.info())
```

```
       country  happinessscore  whiskerhigh  whiskerlow  dystopia192residual  \
0      Finland           7.632        7.695       7.569                2.595
1       Norway           7.594        7.657       7.530                2.383
2      Denmark           7.555        7.623       7.487                2.370
3      Iceland           7.495        7.593       7.398                2.426
4  Switzerland           7.487        7.570       7.405                2.320

   explainedbygdppercapita  explainedbysocialsupport  \
0                    1.305                     1.592
1                    1.456                     1.582
2                    1.351                     1.590
3                    1.343                     1.644
4                    1.420                     1.549

   explainedbyhealthylifeexpectancy  explainedbyfreedomtomakelifechoi  \
0                             0.874                             0.681
1                             0.861                             0.686
2                             0.868                             0.683
3                             0.914                             0.677
4                             0.927                             0.660

   explainedbygenerosity  explainedbyperceptionsofcorrupti
0                  0.192                             0.393
1                  0.286                             0.340
2                  0.284                             0.408
3                  0.353                             0.138
4                  0.256                             0.357
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 156 entries, 0 to 155
Data columns (total 11 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   country                           156 non-null    object
 1   happinessscore                    156 non-null    float32
 2   whiskerhigh                       156 non-null    float32
 3   whiskerlow                        156 non-null    float32
 4   dystopia192residual               156 non-null    float32
 5   explainedbygdppercapita           156 non-null    float32
 6   explainedbysocialsupport          156 non-null    float32
 7   explainedbyhealthylifeexpectancy  156 non-null    float32
 8   explainedbyfreedomtomakelifechoi  156 non-null    float32
 9   explainedbygenerosity             156 non-null    float32
 10  explainedbyperceptionsofcorrupti  156 non-null    float32
dtypes: float32(10), object(1)
memory usage: 7.4+ KB
```

```
None
```

Answer: To load data8.dta correctly, I used the pd.read_stata() function without additional parameters.

1.11 Problem 9

Load `data9.sav`, which is an SPSS file. Use the tools we discussed in class to decide whether the data file loaded correctly, and include that code in your lab report. In one or two sentences, describe how you decided on the right combination of parameters needed to load the data. (2 points)

[ ]:
```
sav_path = r"C:\Users\qaism\OneDrive - University of␣
 ↪Virginia\Documents\GitHub\MSDS\DS 6001\Lab2\lab2data\data9.sav"
data9 = pd.read_spss(sav_path)
print(data9.head())
print(data9.info())
```

```
       country  happiness  whiskerhigh  whiskerlow  dystopia  gdpPC  \
0      Finland      7.632        7.695       7.569     2.595  1.305
1       Norway      7.594        7.657       7.530     2.383  1.456
2      Denmark      7.555        7.623       7.487     2.370  1.351
3      Iceland      7.495        7.593       7.398     2.426  1.343
4  Switzerland      7.487        7.570       7.405     2.320  1.420

   socsupport  lifeexp  lifechoice  generous  corrupt
0       1.592    0.874       0.681     0.192    0.393
1       1.582    0.861       0.686     0.286    0.340
2       1.590    0.868       0.683     0.284    0.408
3       1.644    0.914       0.677     0.353    0.138
4       1.549    0.927       0.660     0.256    0.357
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 156 entries, 0 to 155
Data columns (total 11 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   country      156 non-null    object
 1   happiness    156 non-null    float64
 2   whiskerhigh  156 non-null    float64
 3   whiskerlow   156 non-null    float64
 4   dystopia     156 non-null    float64
 5   gdpPC        156 non-null    float64
 6   socsupport   156 non-null    float64
 7   lifeexp      156 non-null    float64
 8   lifechoice   156 non-null    float64
 9   generous     156 non-null    float64
 10  corrupt      156 non-null    float64
dtypes: float64(10), object(1)
memory usage: 13.5+ KB
None
```

Answer: To load data9.sav correctly, I used the pd.read_spss() function, which uses pyreadstat internally.

1.12 Problem 10

Load `data10.xpt`, which is a SAS file. Use the tools we discussed in class to decide whether the data file loaded correctly, and include that code in your lab report. In one or two sentences, describe how you decided on the right combination of parameters needed to load the data. (If some of the country names display as `b'Finland'`, don't worry aout that.) (2 points)

[ ]:
```
xpt_path = r"C:\Users\qaism\OneDrive - University of␣
 ↪Virginia\Documents\GitHub\MSDS\DS 6001\Lab2\lab2data\data10.xpt"
data10 = pd.read_sas(xpt_path, format='xport')
print(data10.head())
print(data10.info())
```

```
          COUNTRY  HAPPINES  WHISKERH  WHISKERL  DYSTOPIA  EXPLAINE  EXPLAIN2  \
0      b'Finland'     7.632     7.695     7.569     2.595     1.305     1.592
1       b'Norway'     7.594     7.657     7.530     2.383     1.456     1.582
2      b'Denmark'     7.555     7.623     7.487     2.370     1.351     1.590
3      b'Iceland'     7.495     7.593     7.398     2.426     1.343     1.644
4  b'Switzerland'     7.487     7.570     7.405     2.320     1.420     1.549

   EXPLAIN3  EXPLAIN4  EXPLAIN5  EXPLAIN6
0     0.874     0.681     0.192     0.393
1     0.861     0.686     0.286     0.340
2     0.868     0.683     0.284     0.408
3     0.914     0.677     0.353     0.138
4     0.927     0.660     0.256     0.357
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 156 entries, 0 to 155
Data columns (total 11 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   COUNTRY   156 non-null    object
 1   HAPPINES  156 non-null    float64
 2   WHISKERH  156 non-null    float64
 3   WHISKERL  156 non-null    float64
 4   DYSTOPIA  156 non-null    float64
 5   EXPLAINE  156 non-null    float64
 6   EXPLAIN2  156 non-null    float64
 7   EXPLAIN3  156 non-null    float64
 8   EXPLAIN4  156 non-null    float64
 9   EXPLAIN5  156 non-null    float64
 10  EXPLAIN6  156 non-null    float64
dtypes: float64(10), object(1)
memory usage: 13.5+ KB
None
```

Answer: To load data10.xpt I used the pd.read_sas() function with format='xport' to read the SAS file and then displayed the first few rows and DataFrame information

1.13 Problem 11

Please load the `data11.txt` file, which is a fixed width file. The columns are defined as follows:

Variable	Width	Start	End
Country	24	1	24
Happiness score	5	25	29
Whisker-high	5	30	34
Whisker-low	5	35	39
Dystopia (1.92) + residual	5	40	44
Explained by: GDP per capita	5	45	49
Explained by: Social support	5	50	54
Explained by: Healthy life expectancy	5	55	59
Explained by: Freedom to make life choices	5	60	64
Explained by: Generosity	5	65	69
Explained by: Perceptions of corruption	5	70	74

Then save the this loaded data frame as a CSV file on your local machine. Be sure to use a unique filename so as not to overwrite any existing files. (5 points)

[ ]:
```
colspecs = [(0, 24), (24, 29), (29, 34), (34, 39), (39, 44), (44, 49), (49, 54),
            (54, 59), (59, 64), (64, 69), (69, 74)]

data11 = pd.read_fwf(r"C:\Users\qaism\OneDrive - University of␣
 ↪Virginia\Documents\GitHub\MSDS\DS 6001\Lab2\lab2data\data11.txt",
                     colspecs=colspecs, names=column_names)
print(data11.head())
print(data11.info())
data11unique = r"C:\Users\qaism\OneDrive - University of␣
 ↪Virginia\Documents\GitHub\MSDS\DS 6001\Lab2\lab2data\data11_loaded.csv"
data11.to_csv(data11unique, index=False)
```

```
       Country  Happiness score  Whisker-high  Whisker-low  \
0      Finland            7.632         7.695        7.569
1       Norway            7.594         7.657        7.530
2      Denmark            7.555         7.623        7.487
3      Iceland            7.495         7.593        7.398
4  Switzerland            7.487         7.570        7.405

   Dystopia (1.92) + residual  Explained by: GDP per capita  \
0                       2.595                         1.305
1                       2.383                         1.456
2                       2.370                         1.351
3                       2.426                         1.343
```

```
4                      2.320                          1.420

   Explained by: Social support  Explained by: Healthy life expectancy  \
0                         1.592                                  0.874
1                         1.582                                  0.861
2                         1.590                                  0.868
3                         1.644                                  0.914
4                         1.549                                  0.927

   Explained by: Freedom to make life choices  Explained by: Generosity  \
0                                       0.681                    0.192
1                                       0.686                    0.286
2                                       0.683                    0.284
3                                       0.677                    0.353
4                                       0.660                    0.256

   Explained by: Perceptions of corruption
0                                    0.393
1                                    0.340
2                                    0.408
3                                    0.138
4                                    0.357
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 156 entries, 0 to 155
Data columns (total 11 columns):
 #   Column                                    Non-Null Count  Dtype
---  ------                                    --------------  -----
 0   Country                                   156 non-null    object
 1   Happiness score                           156 non-null    float64
 2   Whisker-high                              156 non-null    float64
 3   Whisker-low                               156 non-null    float64
 4   Dystopia (1.92) + residual                156 non-null    float64
 5   Explained by: GDP per capita              156 non-null    float64
 6   Explained by: Social support              156 non-null    float64
 7   Explained by: Healthy life expectancy     156 non-null    float64
 8   Explained by: Freedom to make life choices  156 non-null  float64
 9   Explained by: Generosity                  156 non-null    float64
 10  Explained by: Perceptions of corruption   156 non-null    float64
dtypes: float64(10), object(1)
memory usage: 13.5+ KB
None
```

To load data11.txt, I specified the fixed-width format for the columns using colspecs and assigned the appropriate column names. Colspecs follows python's indexing rules