labassignment2

June 26, 2022

1 Lab Assignment 2: How to Load CSV, ASCII, and other data into Python

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:

[1]:
```
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)

[2]:
```
import numpy as np
import pandas as pd
import os
os.chdir('/home/dfn7vs/Documents/MSDS/MSDS6001/LabAssignment_2')
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" ]
clean_data = pd.read_csv('data_clean.csv')
clean_data.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
```

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)

[3]:
```
data1 = pd.read_csv('data1.csv',skiprows=2)
data1.head(5).T
```

[3]:

	0	1	2	3	\
Country	Finland	Norway	Denmark	Iceland	
Happiness score	7.632	7.594	7.555	7.495	
Whisker-high	7.695	7.657	7.623	7.593	
Whisker-low	7.569	7.53	7.487	7.398	
Dystopia (1.92) + residual	2.595	2.383	2.37	2.426	
Explained by: GDP per capita	1.305	1.456	1.351	1.343	
Explained by: Social support	1.592	1.582	1.59	1.644	
Explained by: Healthy life expectancy	0.874	0.861	0.868	0.914	
Explained by: Freedom to make life choices	0.681	0.686	0.683	0.677	
Explained by: Generosity	0.192	0.286	0.284	0.353	
Explained by: Perceptions of corruption	0.393	0.34	0.408	0.138	

	4
Country	Switzerland

```
Happiness score                                   7.487
Whisker-high                                       7.57
Whisker-low                                       7.405
Dystopia (1.92) + residual                         2.32
Explained by: GDP per capita                       1.42
Explained by: Social support                      1.549
Explained by: Healthy life expectancy             0.927
Explained by: Freedom to make life choices         0.66
Explained by: Generosity                          0.256
Explained by: Perceptions of corruption           0.357
```

We needed to skip two rows in the csv file during the importation process becuase there were two lines of metadata before the data started.

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)

[4]:
```
data2 = pd.read_csv('data2.txt',skiprows=[1,3],header=1)
data2.head(5)
```

[4]:
```
       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
```

We knew from the last probelm that we needed to skip importing certain rows (1,3) but we needed the columns names from row 2 to be the header. Row 2 becomes row 1 after row 1 is skipped.

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)

[5]:
```
data3 = pd.read_csv("data3.txt",sep = "\t",skiprows=2)
data3.head(5)
```

[5]:
```
       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
```

We knew we needed to set the sep arugement of the read_csv function becuase the data was not comma seperated. Instead /t was used to seperate each data point.

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)

[6]:
```
data4 = pd.read_csv("data4.txt",sep = "$",names = column_names)
data4.head(5)
```

[6]:
```
       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
```

The column names needed to be added and according to the documentation the correct way to do that within the read_csv funtion is to set the names arguments.

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)

[7]:
```
data5 = pd.read_csv("data5.csv",skipfooter=2,engine="python")
data5.tail(2)
```

[7]:
```
                      Country  Happiness score  Whisker-high  Whisker-low  \
154  Central African Republic            3.083         3.227        2.939
155                   Burundi            2.905         3.074        2.735

     Dystopia (1.92) + residual  Explained by: GDP per capita  \
154                       2.487                         0.024
155                       1.752                         0.091

     Explained by: Social support  Explained by: Healthy life expectancy  \
154                         0.000                                   0.010
155                         0.627                                   0.145

     Explained by: Freedom to make life choices  Explained by: Generosity  \
154                                       0.305                     0.218
155                                       0.065                     0.149

     Explained by: Perceptions of corruption
154                                    0.038
155                                    0.076
```

The information about the data set was contained at the bottom and therefore needed to be removed.The best way to accomplish this is to set the skipfooter argument to 2.

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)

[8]:
```
data6 = pd.read_csv("data6.dat",na_values=999.000)
data6.head(5)
```

[8]:

	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	NaN	
4	Switzerland	7.487	7.570	7.405	

	Dystopia (1.92) + residual	Explained by: GDP per capita	\
0	2.595	NaN	
1	NaN	NaN	
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	NaN	NaN	
1	1.582	NaN	
2	1.590	NaN	
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	NaN
4	0.357

The values of 999.000 were not valid and needed to be changed to NAN. The best way to do this is to change the enter the value equal to the na_values argument of the read_csv function.

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)

[9]:
```
data7 = pd.read_excel("data7.xlsx",sheet_name="Data")
data7.head(5)
```

[9]:

	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

The data file is an excel sheet and therefore is best imported with the funtion read_excel. The function then allows you to specify a certain sheet to import unsuring the data is imported instead of the metadata.

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)

[10]:
```
data8 = pd.read_stata('data8.dta')
data8.columns = column_names
data8.head(5)
```

[10]:

	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
0	Finland	7.632	7.695	7.569	2.595	1.305	1.592	0.874	0.681	0.192	0.393
1	Norway	7.594	7.657	7.530	2.383	1.456	1.582	0.861	0.686	0.286	0.340
2	Denmark	7.555	7.623	7.487	2.370	1.351	1.590	0.868	0.683	0.284	0.408
3	Iceland	7.495	7.593	7.398	2.426	1.343	1.644	0.914	0.677	0.353	0.138
4	Switzerland	7.487	7.570	7.405	2.320	1.420	1.549	0.927	0.660	0.256	0.357

Read stata was used becuase of the certain data file (.dta) requires it. Since the column names were lower case they needed to be reset using the column names list.

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)

[11]:
```
data9 = pd.read_spss('data9.sav')
data9.columns = column_names
data8.head(5)
```

[11]:

	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

Read spss was used becuase of the certain data file (.sav) requires it. Since the column names were lower case they needed to be reset using the column names list.

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)

[12]:
```
data10 = pd.read_sas("data10.xpt")
data10.columns = column_names
data10.head(5)
```

[12]:

	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
0	b'Finland'	7.632	7.695	7.569	2.595	1.305	1.592	0.874	0.681	0.192	0.393
1	b'Norway'	7.594	7.657	7.530	2.383	1.456	1.582	0.861	0.686	0.286	0.340
2	b'Denmark'	7.555	7.623	7.487	2.370	1.351	1.590	0.868	0.683	0.284	0.408
3	b'Iceland'	7.495	7.593	7.398	2.426	1.343	1.644	0.914	0.677	0.353	0.138
4	b'Switzerland'	7.487	7.570	7.405	2.320	1.420	1.549	0.927	0.660	0.256	0.357

Read sas was used becuase of the certain data file (.xpt) requires it. Since the column names were not fulled displayed they needed to be reset using the column names list.

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)

[13]:
```
Width = [24,5,5,5,5,5,5,5,5,5,5]
data11 = pd.read_fwf("data11.txt", widths = Width,header = None,names =␣
↪column_names)
data11.head(5)
data11.to_csv("data11_csvformat.csv",index=False)
```