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

DS 6001: Practice and Application of Data Science

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:

In [6]:
```
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.

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)

In [7]:
```
import numpy as np
import pandas as pd
import os
import sys
sys.tracebacklimit = 0 # turn off the error tracebacks

oldpath = os.getcwd()
os.chdir("lab data")
```

In [8]:
```
clean = pd.read_csv("data_clean.csv")
clean
```

Out[8]:

	Country	Happiness score	Whisker-high	Whisker-low	Dystopia (1.92) + residual	Explained by: GDP per capita	Explained by: Social support	Ex by: exp
0	Finland	7.632	7.695	7.569	2.595	1.305	1.592	
1	Norway	7.594	7.657	7.530	2.383	1.456	1.582	
2	Denmark	7.555	7.623	7.487	2.370	1.351	1.590	
3	Iceland	7.495	7.593	7.398	2.426	1.343	1.644	
4	Switzerland	7.487	7.570	7.405	2.320	1.420	1.549	
...	...	...	...	...	...	...	...	
151	Yemen	3.355	3.448	3.262	1.106	0.442	1.073	
152	Tanzania	3.303	3.414	3.193	0.628	0.455	0.991	
153	South Sudan	3.254	3.385	3.123	1.691	0.337	0.608	
154	Central African Republic	3.083	3.227	2.939	2.487	0.024	0.000	
155	Burundi	2.905	3.074	2.735	1.752	0.091	0.627	

156 rows × 11 columns

In [9]:
```
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
```

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)

Looking at the data, it's clear that only abnormalities are the first two lines of metadata. Thus the parameter "header = 2" should correctly load it.

In [10]:
```
data1 = pd.read_csv("data1.csv")
data1
```

Out[10]:

	Source: The World Happiness Report (2018), The Sustainable Development Solutions Network (SDSN)	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5	U
0	URL: http://worldhappiness.report/ed/2018	NaN	NaN	NaN	NaN	NaN	
1	Country	Happiness score	Whisker-high	Whisker-low	Dystopia (1.92) + residual	Explained by: GDP per capita	
2	Finland	7.632	7.695	7.569	2.595	1.305	
3	Norway	7.594	7.657	7.530	2.383	1.456	
4	Denmark	7.555	7.623	7.487	2.370	1.351	
...	...	...	...	...	...	...	
153	Yemen	3.355	3.448	3.262	1.106	0.442	
154	Tanzania	3.303	3.414	3.193	0.628	0.455	
155	South Sudan	3.254	3.385	3.123	1.691	0.337	
156	Central African Republic	3.083	3.227	2.939	2.487	0.024	
157	Burundi	2.905	3.074	2.735	1.752	0.091	

158 rows × 11 columns

In [11]:
```
data1 = pd.read_csv("data1.csv", header = 2)
data1
```

Out[11]:

	Country	Happiness score	Whisker-high	Whisker-low	Dystopia (1.92) + residual	Explained by: GDP per capita	Explained by: Social support	Ex by: expe
0	Finland	7.632	7.695	7.569	2.595	1.305	1.592	
1	Norway	7.594	7.657	7.530	2.383	1.456	1.582	
2	Denmark	7.555	7.623	7.487	2.370	1.351	1.590	
3	Iceland	7.495	7.593	7.398	2.426	1.343	1.644	
4	Switzerland	7.487	7.570	7.405	2.320	1.420	1.549	
...	...	...	...	...	...	...	...	
151	Yemen	3.355	3.448	3.262	1.106	0.442	1.073	
152	Tanzania	3.303	3.414	3.193	0.628	0.455	0.991	
153	South Sudan	3.254	3.385	3.123	1.691	0.337	0.608	
154	Central African Republic	3.083	3.227	2.939	2.487	0.024	0.000	
155	Burundi	2.905	3.074	2.735	1.752	0.091	0.627	

156 rows × 11 columns

In [14]:
```
data1.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
```

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)

This dataset has the same metadata comments as data1 (thus "header = 2") as well as three extra comment rows. Looking at the dataset directly we can see that all the comment rows start with '/', so we exclude them with "comment = '/'".

In [15]:
```
data2 = pd.read_csv("data2.txt")
data2
```

Out[15]:

	Source: The World Happiness Report (2018), The Sustainable Development Solutions Network (SDSN)	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5	U
0	URL: http://worldhappiness.report/ed/2018	NaN	NaN	NaN	NaN	NaN	
1	Country	Happiness score	Whisker-high	Whisker-low	Dystopia (1.92) + residual	Explained by: GDP per capita	
2	/The following countries comprise the "very ha...	NaN	NaN	NaN	NaN	NaN	
3	Finland	7.632	7.695	7.569	2.595	1.305	
4	Norway	7.594	7.657	7.530	2.383	1.456	
...	...	...	...	...	...	...	
156	Yemen	3.355	3.448	3.262	1.106	0.442	
157	Tanzania	3.303	3.414	3.193	0.628	0.455	
158	South Sudan	3.254	3.385	3.123	1.691	0.337	
159	Central African Republic	3.083	3.227	2.939	2.487	0.024	
160	Burundi	2.905	3.074	2.735	1.752	0.091	

161 rows × 11 columns

In [16]:
```
data2 = pd.read_csv("data2.txt", header = 2, comment = "/")
data2
```

Out[16]:

	Country	Happiness score	Whisker-high	Whisker-low	Dystopia (1.92) + residual	Explained by: GDP per capita	Explained by: Social support	Ex by: expe
0	Finland	7.632	7.695	7.569	2.595	1.305	1.592	
1	Norway	7.594	7.657	7.530	2.383	1.456	1.582	
2	Denmark	7.555	7.623	7.487	2.370	1.351	1.590	
3	Iceland	7.495	7.593	7.398	2.426	1.343	1.644	
4	Switzerland	7.487	7.570	7.405	2.320	1.420	1.549	
...	...	...	...	...	...	...	...	
151	Yemen	3.355	3.448	3.262	1.106	0.442	1.073	
152	Tanzania	3.303	3.414	3.193	0.628	0.455	0.991	
153	South Sudan	3.254	3.385	3.123	1.691	0.337	0.608	
154	Central African Republic	3.083	3.227	2.939	2.487	0.024	0.000	
155	Burundi	2.905	3.074	2.735	1.752	0.091	0.627	

156 rows × 11 columns

In [17]:
```
data2.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
```

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)

Below we can see that this dataset has the same metadata comments as data1 (thus "header = 2"), as well as using '\t' as a delimiter, so we account for that as a parameter as well.

In [18]:
```
data3 = pd.read_csv("data3.txt")
data3
```

Out[18]:

	Source: The World Happiness Report (2018), The Sustainable Development Solutions Network (SDSN)\t\t\t\t\t\t\t\t\t\t
0	URL: http://worldhappiness.report/ed/2018\t\t\...
1	Country\tHappiness score\tWhisker-high\tWhiske...
2	Finland\t7.632\t7.695\t7.569\t2.595\t1.305\t1....
3	Norway\t7.594\t7.657\t7.53\t2.383\t1.456\t1.58...
4	Denmark\t7.555\t7.623\t7.487\t2.37\t1.351\t1.5...
...	...
153	Yemen\t3.355\t3.448\t3.262\t1.106\t0.442\t1.07...
154	Tanzania\t3.303\t3.414\t3.193\t0.628\t0.455\t0...
155	South Sudan\t3.254\t3.385\t3.123\t1.691\t0.337...
156	Central African Republic\t3.083\t3.227\t2.939\...
157	Burundi\t2.905\t3.074\t2.735\t1.752\t0.091\t0....

158 rows × 1 columns

In [19]:
```
data3 = pd.read_csv("data3.txt", delimiter = "\t", header = 2)
data3
```

Out[19]:

	Country	Happiness score	Whisker-high	Whisker-low	Dystopia (1.92) + residual	Explained by: GDP per capita	Explained by: Social support	Ex by: exp
0	Finland	7.632	7.695	7.569	2.595	1.305	1.592	
1	Norway	7.594	7.657	7.530	2.383	1.456	1.582	
2	Denmark	7.555	7.623	7.487	2.370	1.351	1.590	
3	Iceland	7.495	7.593	7.398	2.426	1.343	1.644	
4	Switzerland	7.487	7.570	7.405	2.320	1.420	1.549	
...	...	...	...	...	...	...	...	
151	Yemen	3.355	3.448	3.262	1.106	0.442	1.073	
152	Tanzania	3.303	3.414	3.193	0.628	0.455	0.991	
153	South Sudan	3.254	3.385	3.123	1.691	0.337	0.608	
154	Central African Republic	3.083	3.227	2.939	2.487	0.024	0.000	
155	Burundi	2.905	3.074	2.735	1.752	0.091	0.627	

156 rows × 11 columns

In [20]:
```
data3.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
```

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)

This dataset, on initial load, has the first row in the header, meaning it is missing column names. In addition to adjusting for that, we also correct the delimiter parameter to '$', which we can tell by looking at the data file directly.

In [21]:
```
data4 = pd.read_csv("data4.txt")
data4
```

Out[21]:

	Finland$7.632$7.695$7.569$2.595$1.305$1.592$0.874$0.681$0.192$0.393
0	Norway$7.594$7.657$7.530$2.383$1.456$1.582$0.8...
1	Denmark$7.555$7.623$7.487$2.370$1.351$1.590$0....
2	Iceland$7.495$7.593$7.398$2.426$1.343$1.644$0....
3	Switzerland$7.487$7.570$7.405$2.320$1.420$1.54...
4	Netherlands$7.441$7.498$7.384$2.448$1.361$1.48...
...	...
150	Yemen$3.355$3.448$3.262$1.106$0.442$1.073$0.34...
151	Tanzania$3.303$3.414$3.193$0.628$0.455$0.991$0...
152	South Sudan$3.254$3.385$3.123$1.691$0.337$0.60...
153	Central African Republic$3.083$3.227$2.939$2.4...
154	Burundi$2.905$3.074$2.735$1.752$0.091$0.627$0....

155 rows × 1 columns

In [22]:
```
data4 = pd.read_csv("data4.txt", delimiter = "$", header = None, names = col
data4
```

Out[22]:

	Country	Happiness score	Whisker-high	Whisker-low	Dystopia (1.92) + residual	Explained by: GDP per capita	Explained by: Social support	Ex by: exp
0	Finland	7.632	7.695	7.569	2.595	1.305	1.592	
1	Norway	7.594	7.657	7.530	2.383	1.456	1.582	
2	Denmark	7.555	7.623	7.487	2.370	1.351	1.590	
3	Iceland	7.495	7.593	7.398	2.426	1.343	1.644	
4	Switzerland	7.487	7.570	7.405	2.320	1.420	1.549	
...	...	...	...	...	...	...	...	
151	Yemen	3.355	3.448	3.262	1.106	0.442	1.073	
152	Tanzania	3.303	3.414	3.193	0.628	0.455	0.991	
153	South Sudan	3.254	3.385	3.123	1.691	0.337	0.608	
154	Central African Republic	3.083	3.227	2.939	2.487	0.024	0.000	
155	Burundi	2.905	3.074	2.735	1.752	0.091	0.627	

156 rows × 11 columns

In [23]:
```
data4.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
```

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)

Much like data1 had two lines of metadata at its top, data5 has two lines at its bottom. We can remove these with the "skipfooter" parameter.

In [24]:
```
data5 = pd.read_csv("data5.csv")
data5
```

Out[24]:

	Country	Happiness score	Whisker-high	Whisker-low	Dystopia (1.92) + residual	Explained by: GDP per capita	Expla by: S sup
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	
...	...	...	...	...	...	...	
153	South Sudan	3.254	3.385	3.123	1.691	0.337	
154	Central African Republic	3.083	3.227	2.939	2.487	0.024	
155	Burundi	2.905	3.074	2.735	1.752	0.091	
156	Source: The World Happiness Report (2018), The...	NaN	NaN	NaN	NaN	NaN	
157	URL: http://worldhappiness.report/ed/2018	NaN	NaN	NaN	NaN	NaN	

158 rows × 11 columns

In [25]:
```
data5 = pd.read_csv("data5.csv", skipfooter = 2, engine='python')
data5
```

Out[25]:

	Country	Happiness score	Whisker-high	Whisker-low	Dystopia (1.92) + residual	Explained by: GDP per capita	Explained by: Social support	Ex by: expe
0	Finland	7.632	7.695	7.569	2.595	1.305	1.592	
1	Norway	7.594	7.657	7.530	2.383	1.456	1.582	
2	Denmark	7.555	7.623	7.487	2.370	1.351	1.590	
3	Iceland	7.495	7.593	7.398	2.426	1.343	1.644	
4	Switzerland	7.487	7.570	7.405	2.320	1.420	1.549	
...	...	...	...	...	...	...	...	
151	Yemen	3.355	3.448	3.262	1.106	0.442	1.073	
152	Tanzania	3.303	3.414	3.193	0.628	0.455	0.991	
153	South Sudan	3.254	3.385	3.123	1.691	0.337	0.608	
154	Central African Republic	3.083	3.227	2.939	2.487	0.024	0.000	
155	Burundi	2.905	3.074	2.735	1.752	0.091	0.627	

156 rows × 11 columns

In [26]:
```
data5.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
```

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)

We can see this dataset has several "999" values, which we know are meant to signify missing data values. While we must accept these missing data values, we can adjust the "na_values" parameter to better reflect them.

In [27]:
```
data6 = pd.read_csv("data6.dat")
data6
```

Out[27]:

	Country	Happiness score	Whisker-high	Whisker-low	Dystopia (1.92) + residual	Explained by: GDP per capita	Explained by: Social support	Ex by: expe
0	Finland	7.632	7.695	7.569	2.595	999.000	999.000	
1	Norway	7.594	7.657	7.530	999.000	999.000	1.582	
2	Denmark	7.555	7.623	7.487	2.370	1.351	1.590	
3	Iceland	7.495	7.593	999.000	2.426	1.343	1.644	
4	Switzerland	7.487	7.570	7.405	2.320	1.420	1.549	
...	...	...	...	...	...	...	...	
151	999	3.355	3.448	3.262	1.106	0.442	1.073	
152	Tanzania	999.000	999.000	3.193	0.628	999.000	0.991	
153	South Sudan	3.254	999.000	3.123	1.691	0.337	999.000	
154	Central African Republic	3.083	3.227	2.939	2.487	0.024	0.000	
155	Burundi	2.905	3.074	999.000	1.752	0.091	999.000	

156 rows × 11 columns

In [28]:
```
data6 = pd.read_csv("data6.dat", na_values = 999)
data6
```

Out[28]:

	Country	Happiness score	Whisker-high	Whisker-low	Dystopia (1.92) + residual	Explained by: GDP per capita	Explained by: Social support	Ex by: expe
0	Finland	7.632	7.695	7.569	2.595	NaN	NaN	
1	Norway	7.594	7.657	7.530	NaN	NaN	1.582	
2	Denmark	7.555	7.623	7.487	2.370	1.351	1.590	
3	Iceland	7.495	7.593	NaN	2.426	1.343	1.644	
4	Switzerland	7.487	7.570	7.405	2.320	1.420	1.549	
...	...	...	...	...	...	...	...	
151	NaN	3.355	3.448	3.262	1.106	0.442	1.073	
152	Tanzania	NaN	NaN	3.193	0.628	NaN	0.991	
153	South Sudan	3.254	NaN	3.123	1.691	0.337	NaN	
154	Central African Republic	3.083	3.227	2.939	2.487	0.024	0.000	
155	Burundi	2.905	3.074	NaN	1.752	0.091	NaN	

156 rows × 11 columns

In [29]:
```
data6.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 156 entries, 0 to 155
Data columns (total 11 columns):
 #   Column                                  Non-Null Count  Dtype  
---  ------                                  --------------  -----  
 0   Country                                 145 non-null    object 
 1   Happiness score                         142 non-null    float64
 2   Whisker-high                            135 non-null    float64
 3   Whisker-low                             136 non-null    float64
 4   Dystopia (1.92) + residual              145 non-null    float64
 5   Explained by: GDP per capita            137 non-null    float64
 6   Explained by: Social support            134 non-null    float64
 7   Explained by: Healthy life expectancy   142 non-null    float64
 8   Explained by: Freedom to make life choices  140 non-null    float64
 9   Explained by: Generosity                145 non-null    float64
 10  Explained by: Perceptions of corruption     143 non-null    float64
dtypes: float64(10), object(1)
memory usage: 13.5+ KB
```

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)

This data file seems to have no abnormalities, other than being a .xlsx file.

In [30]:
```
data7 = pd.read_excel("data7.xlsx", sheet_name="Data")
data7
```

Out[30]:

	Country	Happiness score	Whisker-high	Whisker-low	Dystopia (1.92) + residual	Explained by: GDP per capita	Explained by: Social support	Ex by: exp
0	Finland	7.632	7.695	7.569	2.595	1.305	1.592	
1	Norway	7.594	7.657	7.530	2.383	1.456	1.582	
2	Denmark	7.555	7.623	7.487	2.370	1.351	1.590	
3	Iceland	7.495	7.593	7.398	2.426	1.343	1.644	
4	Switzerland	7.487	7.570	7.405	2.320	1.420	1.549	
...	...	...	...	...	...	...	...	
151	Yemen	3.355	3.448	3.262	1.106	0.442	1.073	
152	Tanzania	3.303	3.414	3.193	0.628	0.455	0.991	
153	South Sudan	3.254	3.385	3.123	1.691	0.337	0.608	
154	Central African Republic	3.083	3.227	2.939	2.487	0.024	0.000	
155	Burundi	2.905	3.074	2.735	1.752	0.091	0.627	

156 rows × 11 columns

In [31]:
```
data7.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
```

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)

This data file seems to have no abnormalities, other than being a .dta file.

In [32]:
```
data8 = pd.read_stata("data8.dta")
data8
```

Out[32]:

	country	happinessscore	whiskerhigh	whiskerlow	dystopia192residual	expla
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	
...	...	...	...	...	...	
151	Yemen	3.355	3.448	3.262	1.106	
152	Tanzania	3.303	3.414	3.193	0.628	
153	South Sudan	3.254	3.385	3.123	1.691	
154	Central African Republic	3.083	3.227	2.939	2.487	
155	Burundi	2.905	3.074	2.735	1.752	

156 rows × 11 columns

In [33]:
```
data8.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   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
```

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)

This data file seems to have no abnormalities, other than being a .sav file.

In [34]:
```
data9 = pd.read_spss("data9.sav")
data9
```

Out[34]:

	country	happiness	whiskerhigh	whiskerlow	dystopia	gdpPC	socsupport	l
0	Finland	7.632	7.695	7.569	2.595	1.305	1.592	
1	Norway	7.594	7.657	7.530	2.383	1.456	1.582	
2	Denmark	7.555	7.623	7.487	2.370	1.351	1.590	
3	Iceland	7.495	7.593	7.398	2.426	1.343	1.644	
4	Switzerland	7.487	7.570	7.405	2.320	1.420	1.549	
...	...	...	...	...	...	...	...	...
151	Yemen	3.355	3.448	3.262	1.106	0.442	1.073	
152	Tanzania	3.303	3.414	3.193	0.628	0.455	0.991	
153	South Sudan	3.254	3.385	3.123	1.691	0.337	0.608	
154	Central African Republic	3.083	3.227	2.939	2.487	0.024	0.000	
155	Burundi	2.905	3.074	2.735	1.752	0.091	0.627	

156 rows × 11 columns

In [35]:
```
data9.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    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
```

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)

This data file seems to have no abnormalities, other than being a .xpt file (and the b' ' around the row names, of course, but we're ignoring that.

In [36]:
```
data10 = pd.read_sas("data10.xpt")
data10
```

Out[36]:

	COUNTRY	HAPPINES	WHISKERH	WHISKERL	DYSTOPIA	EXPLAINE	EXP
0	b'Finland'	7.632	7.695	7.569	2.595	1.305	1.59200
1	b'Norway'	7.594	7.657	7.530	2.383	1.456	1.58200
2	b'Denmark'	7.555	7.623	7.487	2.370	1.351	1.59000
3	b'Iceland'	7.495	7.593	7.398	2.426	1.343	1.64400
4	b'Switzerland'	7.487	7.570	7.405	2.320	1.420	1.54900
...	...	...	...	...	...	...	
151	b'Yemen'	3.355	3.448	3.262	1.106	0.442	1.07300
152	b'Tanzania'	3.303	3.414	3.193	0.628	0.455	9.9100(
153	b'South Sudan'	3.254	3.385	3.123	1.691	0.337	6.0800(
154	b'Central African Republic'	3.083	3.227	2.939	2.487	0.024	5.3976(
155	b'Burundi'	2.905	3.074	2.735	1.752	0.091	6.2700(

156 rows × 11 columns

In [37]:
```
data10.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   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
```

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)

In [38]:
```
data11_widths = [24, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5]
data11 = pd.read_fwf("data11.txt", widths = data11_widths, header = None, na
data11
```

Out[38]:

	Country	Happiness score	Whisker-high	Whisker-low	Dystopia (1.92) + residual	Explained by: GDP per capita	Explained by: Social support	Ex by: exp
0	Finland	7.632	7.695	7.569	2.595	1.305	1.592	
1	Norway	7.594	7.657	7.530	2.383	1.456	1.582	
2	Denmark	7.555	7.623	7.487	2.370	1.351	1.590	
3	Iceland	7.495	7.593	7.398	2.426	1.343	1.644	
4	Switzerland	7.487	7.570	7.405	2.320	1.420	1.549	
...	...	...	...	...	...	...	...	
151	Yemen	3.355	3.448	3.262	1.106	0.442	1.073	
152	Tanzania	3.303	3.414	3.193	0.628	0.455	0.991	
153	South Sudan	3.254	3.385	3.123	1.691	0.337	0.608	
154	Central African Republic	3.083	3.227	2.939	2.487	0.024	0.000	
155	Burundi	2.905	3.074	2.735	1.752	0.091	0.627	

156 rows × 11 columns

In [39]:
```
data11.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
```

In [40]:
```
data11.to_csv("data11_clean.csv", sep = ",")
```

In [41]:
```
os.chdir(oldpath)
```

In [ ]: