labassignment2

January 28, 2025

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:

[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" ]
```

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)

[3]:

```
import os
import pandas as pd

os.chdir(r'/Users/queclay/Documents/MSDS/DS6001/M02/lab data/lab data')

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"]

data_clean = pd.read_csv('data_clean.csv')

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
```

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)

[4]:

```
data1 = pd.read_csv('data1.csv')
data1.info()
data1.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 158 entries, 0 to 157
Data columns (total 11 columns):
 #   Column
Non-Null Count  Dtype
---  ------
--------------  -----
 0   Source: The World Happiness Report (2018), The Sustainable Development
Solutions Network (SDSN)  158 non-null    object
 1   Unnamed: 1
157 non-null    object
 2   Unnamed: 2
```

```
157 non-null    object
 3   Unnamed: 3
157 non-null    object
 4   Unnamed: 4
157 non-null    object
 5   Unnamed: 5
157 non-null    object
 6   Unnamed: 6
157 non-null    object
 7   Unnamed: 7
157 non-null    object
 8   Unnamed: 8
157 non-null    object
 9   Unnamed: 9
157 non-null    object
 10  Unnamed: 10
157 non-null    object
dtypes: object(11)
memory usage: 13.7+ KB
```

[4]:

```
  Source: The World Happiness Report (2018), The Sustainable Development
Solutions Network (SDSN)  \
0           URL: http://worldhappiness.report/ed/2018
1                                             Country
2                                             Finland
3                                              Norway
4                                             Denmark

        Unnamed: 1    Unnamed: 2   Unnamed: 3                  Unnamed: 4  \
0              NaN           NaN          NaN                         NaN
1  Happiness score  Whisker-high  Whisker-low  Dystopia (1.92) + residual
2            7.632         7.695        7.569                       2.595
3            7.594         7.657        7.530                       2.383
4            7.555         7.623        7.487                       2.370

                     Unnamed: 5                    Unnamed: 6  \
0                           NaN                           NaN
1  Explained by: GDP per capita  Explained by: Social support
2                         1.305                         1.592
3                         1.456                         1.582
4                         1.351                         1.590

                              Unnamed: 7  \
0                                    NaN
1  Explained by: Healthy life expectancy
2                                  0.874
3                                  0.861
```

```
4                                  0.868

                                   Unnamed: 8                Unnamed: 9  \
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

                               Unnamed: 10
0                                      NaN
1  Explained by: Perceptions of corruption
2                                    0.393
3                                    0.340
4                                    0.408
```

[5]:

```
data1 = pd.read_csv('data1.csv', skiprows=2)
data1.info()
data1.head(3)
data1.tail(3)
```

```
<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
```

[5]:

```
                      Country  Happiness score  Whisker-high  Whisker-low  \
153               South Sudan            3.254         3.385        3.123
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  \
153                       1.691                         0.337
154                       2.487                         0.024
```

```
155                       1.752                         0.091

     Explained by: Social support  Explained by: Healthy life expectancy  \
153                         0.608                                  0.177
154                         0.000                                  0.010
155                         0.627                                  0.145

     Explained by: Freedom to make life choices  Explained by: Generosity  \
153                                       0.112                     0.224
154                                       0.305                     0.218
155                                       0.065                     0.149

     Explained by: Perceptions of corruption
153                                    0.106
154                                    0.038
155                                    0.076
```

To determine the correct parameters for loading the data, I used the `info` and `head` methods to identify and remove the two additional rows citing the data source at the top of the CSV with the `skiprows` parameter. After reloading the data and verifying with `info`, `head`, and `tail`, I confirmed that the data was correctly loaded and matched `data_clean`.

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)

[6]:

```
data2 = pd.read_csv('data2.txt')
data2.info()
data2.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 161 entries, 0 to 160
Data columns (total 11 columns):
 #   Column
Non-Null Count  Dtype
---  ------
--------------  -----
 0   Source: The World Happiness Report (2018), The Sustainable Development
Solutions Network (SDSN)  161 non-null    object
 1   Unnamed: 1
157 non-null    object
 2   Unnamed: 2
157 non-null    object
 3   Unnamed: 3
157 non-null    object
 4   Unnamed: 4
157 non-null    object
```

```
 5   Unnamed: 5
157 non-null    object
 6   Unnamed: 6
157 non-null    object
 7   Unnamed: 7
157 non-null    object
 8   Unnamed: 8
157 non-null    object
 9   Unnamed: 9
157 non-null    object
 10  Unnamed: 10
157 non-null    object
dtypes: object(11)
memory usage: 14.0+ KB
```

[6]:

```
  Source: The World Happiness Report (2018), The Sustainable Development
Solutions Network (SDSN)  \
0          URL: http://worldhappiness.report/ed/2018
1                                            Country
2  /The following countries comprise the "very ha…
3                                            Finland
4                                             Norway

        Unnamed: 1    Unnamed: 2   Unnamed: 3                  Unnamed: 4  \
0              NaN           NaN          NaN                         NaN
1  Happiness score  Whisker-high  Whisker-low  Dystopia (1.92) + residual
2              NaN           NaN          NaN                         NaN
3            7.632         7.695        7.569                       2.595
4            7.594         7.657        7.530                       2.383

                     Unnamed: 5                    Unnamed: 6  \
0                           NaN                           NaN
1  Explained by: GDP per capita  Explained by: Social support
2                           NaN                           NaN
3                         1.305                         1.592
4                         1.456                         1.582

                              Unnamed: 7  \
0                                    NaN
1  Explained by: Healthy life expectancy
2                                    NaN
3                                  0.874
4                                  0.861

                                   Unnamed: 8                Unnamed: 9  \
0                                         NaN                       NaN
1  Explained by: Freedom to make life choices  Explained by: Generosity
```

```
2                                            NaN                          NaN
3                                          0.681                        0.192
4                                          0.686                        0.286

                            Unnamed: 10
0                                   NaN
1  Explained by: Perceptions of corruption
2                                   NaN
3                                 0.393
4                                 0.340
```

[7]:

```
data2 = pd.read_csv('data2.txt', header=2, comment='/')
data2.info()
data2.head(3)
data2.tail(3)
```

```
<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
```

[7]:

```
                      Country  Happiness score  Whisker-high  Whisker-low  \
153               South Sudan            3.254         3.385        3.123
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  \
153                       1.691                         0.337
154                       2.487                         0.024
155                       1.752                         0.091

     Explained by: Social support  Explained by: Healthy life expectancy  \
153                         0.608                                  0.177
154                         0.000                                  0.010
```

```
155                          0.627                                       0.145

     Explained by: Freedom to make life choices  Explained by: Generosity  \
153                                       0.112                     0.224
154                                       0.305                     0.218
155                                       0.065                     0.149

     Explained by: Perceptions of corruption
153                                    0.106
154                                    0.038
155                                    0.076
```

I saw using `head()` that the first two rows were taken up by an unwanted header, so I used `header=2` to remove them. Then I saw there was a comment after the table identified by a '/' character, so I removed that row with `comment='/'` and confirmed with `info`, `head`, and `tail` that the data now matches `data_clean`.

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)

[8]:
```
data3 = pd.read_csv('data3.txt')
data3.info()
data3.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 158 entries, 0 to 157
Data columns (total 1 columns):
 #   Column
Non-Null Count  Dtype
---  ------
--------------  -----
 0   Source: The World Happiness Report (2018), The Sustainable Development
Solutions Network (SDSN)
158 non-null    object
dtypes: object(1)
memory usage: 1.4+ KB
```

[8]:
```
   Source: The World Happiness Report (2018), The Sustainable Development
 Solutions Network (SDSN)\t\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…
```

[9]:
```
data3 = pd.read_csv('data3.txt', sep="\t")
data3.info()
data3.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 158 entries, 0 to 157
Data columns (total 11 columns):
 #   Column
Non-Null Count  Dtype
---  ------
--------------  -----
 0   Source: The World Happiness Report (2018), The Sustainable Development
Solutions Network (SDSN)  158 non-null    object
 1   Unnamed: 1
157 non-null    object
 2   Unnamed: 2
157 non-null    object
 3   Unnamed: 3
157 non-null    object
 4   Unnamed: 4
157 non-null    object
 5   Unnamed: 5
157 non-null    object
 6   Unnamed: 6
157 non-null    object
 7   Unnamed: 7
157 non-null    object
 8   Unnamed: 8
157 non-null    object
 9   Unnamed: 9
157 non-null    object
 10  Unnamed: 10
157 non-null    object
dtypes: object(11)
memory usage: 13.7+ KB
```

[9]:
```
  Source: The World Happiness Report (2018), The Sustainable Development
Solutions Network (SDSN)  \
0          URL: http://worldhappiness.report/ed/2018
1                                            Country
2                                            Finland
3                                             Norway
4                                            Denmark

        Unnamed: 1    Unnamed: 2   Unnamed: 3                Unnamed: 4  \
0              NaN           NaN          NaN                       NaN
1  Happiness score  Whisker-high  Whisker-low  Dystopia (1.92) + residual
2            7.632         7.695        7.569                     2.595
```

```
3            7.594         7.657         7.53                       2.383
4            7.555         7.623        7.487                        2.37

                     Unnamed: 5                    Unnamed: 6  \
0                           NaN                           NaN
1  Explained by: GDP per capita  Explained by: Social support
2                         1.305                         1.592
3                         1.456                         1.582
4                         1.351                          1.59

                              Unnamed: 7  \
0                                    NaN
1  Explained by: Healthy life expectancy
2                                  0.874
3                                  0.861
4                                  0.868

                                   Unnamed: 8                Unnamed: 9  \
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

                               Unnamed: 10
0                                      NaN
1  Explained by: Perceptions of corruption
2                                    0.393
3                                     0.34
4                                    0.408
```

[10]:
```
data3 = pd.read_csv('data3.txt', sep="\t", header=2)
data3.info()
data3.head()
```

```
<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
```

```
[10]:        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
```

Using `head()`, I identified that the data was tab-delimited, so I used the `sep="\t"` parameter. I also noticed two rows of comments before the header, so I used `header=2`. After these transformations, I confirmed with `info()` and `head()` that the data matched `data_clean`.

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)

```
[11]: data4 = pd.read_csv('data4.txt')
      data4.info()
      data4.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 155 entries, 0 to 154
Data columns (total 1 columns):
 #   Column                                                                      Non-
Null Count  Dtype
---  ------
--------------  -----
 0   Finland$7.632$7.695$7.569$2.595$1.305$1.592$0.874$0.681$0.192$0.393  155
non-null    object
dtypes: object(1)
memory usage: 1.3+ KB
```

```
[11]:   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…
```

```
[12]: data4 = pd.read_csv('data4.txt', sep='$')
      data4.info()
      data4.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 155 entries, 0 to 154
Data columns (total 11 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   Finland  155 non-null    object
 1   7.632    155 non-null    float64
 2   7.695    155 non-null    float64
 3   7.569    155 non-null    float64
 4   2.595    155 non-null    float64
 5   1.305    155 non-null    float64
 6   1.592    155 non-null    float64
 7   0.874    155 non-null    float64
 8   0.681    155 non-null    float64
 9   0.192    155 non-null    float64
 10  0.393    155 non-null    float64
dtypes: float64(10), object(1)
memory usage: 13.4+ KB
```

```
[12]:       Finland  7.632  7.695  7.569  2.595  1.305  1.592  0.874  0.681  0.192  \
     0       Norway  7.594  7.657  7.530  2.383  1.456  1.582  0.861  0.686  0.286
     1      Denmark  7.555  7.623  7.487  2.370  1.351  1.590  0.868  0.683  0.284
     2      Iceland  7.495  7.593  7.398  2.426  1.343  1.644  0.914  0.677  0.353
     3  Switzerland  7.487  7.570  7.405  2.320  1.420  1.549  0.927  0.660  0.256
     4  Netherlands  7.441  7.498  7.384  2.448  1.361  1.488  0.878  0.638  0.333

        0.393
     0  0.340
     1  0.408
     2  0.138
     3  0.357
     4  0.295
```

```
[13]: data4 = pd.read_csv('data4.txt', sep='$', names=column_names)
      data4.info()
      data4.head()
```

```
<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
```

```
[13]:       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
```

Using `head()`, I saw the values were separated by the $ character, so I added the `sep='$'` parameter. Then I noticed there were no column names, so I added the `names=column_names` parameter (as we had defined them earlier). After adding these two parameters, I confirmed that the data matched `data_clean`.

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)

[14]:

```
data5 = pd.read_csv('data5.csv')
data5.info()
data5.head()
data5.tail()
```

```
<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
```

[14]:

```
                                                Country  Happiness score  \
153                                         South Sudan            3.254
154                            Central African Republic            3.083
155                                             Burundi            2.905
156  Source: The World Happiness Report (2018), The…              NaN
157           URL: http://worldhappiness.report/ed/2018              NaN

     Whisker-high  Whisker-low  Dystopia (1.92) + residual  \
153         3.385        3.123                       1.691
154         3.227        2.939                       2.487
155         3.074        2.735                       1.752
156           NaN          NaN                         NaN
157           NaN          NaN                         NaN

     Explained by: GDP per capita  Explained by: Social support  \
153                         0.337                         0.608
154                         0.024                         0.000
155                         0.091                         0.627
156                           NaN                           NaN
157                           NaN                           NaN

     Explained by: Healthy life expectancy  \
153                                  0.177
154                                  0.010
155                                  0.145
156                                    NaN
157                                    NaN

     Explained by: Freedom to make life choices  Explained by: Generosity  \
153                                       0.112                     0.224
154                                       0.305                     0.218
155                                       0.065                     0.149
156                                         NaN                       NaN
157                                         NaN                       NaN

     Explained by: Perceptions of corruption
153                                    0.106
```

```
154                                       0.038
155                                       0.076
156                                         NaN
157                                         NaN
```

[15]:
```
data5 = pd.read_csv('data5.csv', skipfooter=2)
data5.info()
data5.head()
data5.tail()
```

```
<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

/var/folders/fn/36dz4z514cd0cmgsz1j3_f340000gn/T/ipykernel_10817/2606128970.py:1
: ParserWarning: Falling back to the 'python' engine because the 'c' engine does
not support skipfooter; you can avoid this warning by specifying
engine='python'.
  data5 = pd.read_csv('data5.csv', skipfooter=2)
```

[15]:
```
                      Country  Happiness score  Whisker-high  Whisker-low  \
151                     Yemen            3.355         3.448        3.262
152                  Tanzania            3.303         3.414        3.193
153               South Sudan            3.254         3.385        3.123
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  \
151                       1.106                         0.442
152                       0.628                         0.455
153                       1.691                         0.337
154                       2.487                         0.024
155                       1.752                         0.091
```

```
     Explained by: Social support  Explained by: Healthy life expectancy  \
151                         1.073                                  0.343
152                         0.991                                  0.381
153                         0.608                                  0.177
154                         0.000                                  0.010
155                         0.627                                  0.145

     Explained by: Freedom to make life choices  Explained by: Generosity  \
151                                       0.244                     0.083
152                                       0.481                     0.270
153                                       0.112                     0.224
154                                       0.305                     0.218
155                                       0.065                     0.149

     Explained by: Perceptions of corruption
151                                    0.064
152                                    0.097
153                                    0.106
154                                    0.038
155                                    0.076
```

From `info()`, I saw from the `RangeIndex` that there were two additional rows, which I found at the very end using `tail()`. I used `skipfooter=2` to remove the last two rows that cite the data and then confirmed the data matched `data_clean`.

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)

[16]:
```
data6 = pd.read_csv('data6.dat')
data6.info()
data6.tail().T
```

```
<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
```

[16]:

```
                                              151       152          153  \
Country                                       999  Tanzania  South Sudan
Happiness score                             3.355     999.0        3.254
Whisker-high                                3.448     999.0        999.0
Whisker-low                                 3.262     3.193        3.123
Dystopia (1.92) + residual                  1.106     0.628        1.691
Explained by: GDP per capita                0.442     999.0        0.337
Explained by: Social support                1.073     0.991        999.0
Explained by: Healthy life expectancy       0.343     0.381        0.177
Explained by: Freedom to make life choices  0.244     0.481        0.112
Explained by: Generosity                    999.0      0.27        0.224
Explained by: Perceptions of corruption     0.064     0.097        0.106

                                                                 154      155
Country                                     Central African Republic  Burundi
Happiness score                                                3.083    2.905
Whisker-high                                                   3.227    3.074
Whisker-low                                                    2.939    999.0
Dystopia (1.92) + residual                                     2.487    1.752
Explained by: GDP per capita                                   0.024    0.091
Explained by: Social support                                     0.0    999.0
Explained by: Healthy life expectancy                           0.01    0.145
Explained by: Freedom to make life choices                     0.305    0.065
Explained by: Generosity                                       999.0    0.149
Explained by: Perceptions of corruption                        0.038    0.076
```

[17]:

```
data6 = pd.read_csv('data6.dat', na_values=999)
data6.info()
data6.tail(7).T
```

```
<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
```

[17]:

```
                                              149     150    151       152  \
Country                                     Syria  Rwanda    NaN  Tanzania
Happiness score                             3.462     NaN  3.355       NaN
Whisker-high                                3.664     3.5  3.448       NaN
Whisker-low                                  3.26   3.317  3.262     3.193
Dystopia (1.92) + residual                  1.244     0.5  1.106     0.628
Explained by: GDP per capita                0.689   0.332  0.442       NaN
Explained by: Social support                0.382   0.896  1.073     0.991
Explained by: Healthy life expectancy       0.539     NaN  0.343     0.381
Explained by: Freedom to make life choices  0.088   0.636  0.244     0.481
Explained by: Generosity                    0.376     0.2    NaN      0.27
Explained by: Perceptions of corruption     0.144   0.444  0.064     0.097

                                                    153  \
Country                                     South Sudan
Happiness score                                   3.254
Whisker-high                                        NaN
Whisker-low                                       3.123
Dystopia (1.92) + residual                        1.691
Explained by: GDP per capita                      0.337
Explained by: Social support                         NaN
Explained by: Healthy life expectancy             0.177
Explained by: Freedom to make life choices        0.112
Explained by: Generosity                          0.224
Explained by: Perceptions of corruption           0.106

                                                                154      155
Country                                     Central African Republic  Burundi
Happiness score                                                3.083    2.905
Whisker-high                                                   3.227    3.074
Whisker-low                                                    2.939      NaN
Dystopia (1.92) + residual                                     2.487    1.752
Explained by: GDP per capita                                   0.024    0.091
Explained by: Social support                                     0.0      NaN
Explained by: Healthy life expectancy                           0.01    0.145
Explained by: Freedom to make life choices                     0.305    0.065
Explained by: Generosity                                         NaN    0.149
Explained by: Perceptions of corruption                        0.038    0.076
```

When I loaded this file, I saw that it had the correct number of rows and columns, and the column names were correct. However, `tail().T` revealed a Country listed as `999` and many values reported as this number. I inferred it was a placeholder for NA, so I used the `na_values=999` argument.

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)

[19]:
```
data7 = pd.read_excel("data7.xlsx", sheet_name="Data")
data7.info()
data7.head()
data7.tail()
```

```
<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
```

[19]:
```
                      Country  Happiness score  Whisker-high  Whisker-low  \
151                     Yemen            3.355         3.448        3.262   
152                  Tanzania            3.303         3.414        3.193   
153               South Sudan            3.254         3.385        3.123   
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  \
151                       1.106                         0.442   
152                       0.628                         0.455   
153                       1.691                         0.337   
154                       2.487                         0.024   
155                       1.752                         0.091   

     Explained by: Social support  Explained by: Healthy life expectancy  \
151                         1.073                                  0.343
152                         0.991                                  0.381
153                         0.608                                  0.177
154                         0.000                                  0.010
155                         0.627                                  0.145

     Explained by: Freedom to make life choices  Explained by: Generosity  \
151                                       0.244                     0.083
152                                       0.481                     0.270
153                                       0.112                     0.224
154                                       0.305                     0.218
155                                       0.065                     0.149

     Explained by: Perceptions of corruption
151                                    0.064
152                                    0.097
153                                    0.106
154                                    0.038
155                                    0.076
```

After specifying the path to the Excel file and the sheet, I confirmed that the data has the expected number of rows and columns, and the column names match `data_clean`.

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)

[20]:

```python
data8 = pd.read_stata("data8.dta")
data8.columns = column_names
data8.info()
data8.head()
```

```
<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    float32
 2   Whisker-high                                156 non-null    float32
 3   Whisker-low                                 156 non-null    float32
 4   Dystopia (1.92) + residual                  156 non-null    float32
 5   Explained by: GDP per capita                156 non-null    float32
 6   Explained by: Social support                156 non-null    float32
 7   Explained by: Healthy life expectancy       156 non-null    float32
 8   Explained by: Freedom to make life choices  156 non-null    float32
 9   Explained by: Generosity                    156 non-null    float32
 10  Explained by: Perceptions of corruption     156 non-null    float32
dtypes: float32(10), object(1)
memory usage: 7.4+ KB
```

```
[20]:        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
```

Using `pd.read_stata`, I saw the expected number of rows and columns. However, the column names were not formatted correctly, so I used `data8.columns = column_names` with our previously set column names.

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)

[22]:
```
data9 = pd.read_spss("data9.sav")
data9.columns = column_names
data9.info()
data9.head()
```

```
<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
```

[22]:
```
       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
```

After using `pd.read_spss`, I saw the expected number of rows and columns. However, the column names were not formatted correctly, so I used `data9.columns = column_names` with our previously set column names.

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)

[23]:

```
data10 = pd.read_sas("data10.xpt")
data10.columns = column_names
data10.info()
data10.head()
```

```
<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
```

[23]:

```
          Country  Happiness score  Whisker-high  Whisker-low  \
0      b'Finland'            7.632         7.695        7.569
```

```
1       b'Norway'            7.594         7.657        7.530
2      b'Denmark'            7.555         7.623        7.487
3      b'Iceland'            7.495         7.593        7.398
4  b'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
```

After using `pd.read_sas`, I saw the expected number of rows and columns. However, the column names were not formatted correctly, so I used `data10.columns = column_names` with our previously set column names.

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

Variable	Width	Start	End
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)

[24]:

```
data11 = pd.read_fwf("data11.txt", widths=[24,5,5,5,5,5,5,5,5,5,5], header=None)
data11.columns = column_names
data11.info()
data11.head()
data11.to_csv("data11.csv", sep=",")
```

```
<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
```

I used `pd.read_fwf` and input a list of the `widths` as taken from the reference table and I saw the expected number of rows and columns. However, the column names were not formatted correctly, so I used `data11.columns = column_names` with our previously set column names.