lab2

January 23, 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:

[1]:
```
column_names = ["Country", "Happiness score", "Whisker-high", "Whisker-low",
  "Dystopia (1.92) + residual", "Explained by: GDP per capita",
  "Explained by: Social support", "Explained by: Healthy life expectancy",
  "Explained by: Freedom to make life choices", "Explained by: Generosity",
  "Explained by: Perceptions of corruption" ]
```

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

1.2 Problem 0

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

[4]:
```
import pandas as pd
import numpy as np

data_clean = pd.read_csv('lab data/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)

[13]:
```
data1 = pd.read_csv('lab data/data1.csv',header=2)
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
```

- When I first loaded the data I noticed immediately that there was an issue with the column names. Upon investigation, I realized the issue was that there were a couple rows of text above the column names, so I used the parameter `header = 2` to tell `read_csv()` where the column names were located.

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)

[29]:
```
data2 = pd.read_csv('lab data/data2.txt',header=2,comment='/')
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
```

- For this dataset, I saw that the header was on the third row, so I used `header=2`, and finally I noticed a comment in one of the rows, so I used `comment='/'` to account for that.

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)

[31]:
```
data3 = pd.read_csv('lab data/data3.txt',delimiter='\t',header=2)
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
```

- The changes needed for this dataset were using `delimiter='\t'` since the data is tab-delimited and `header=2` since the column names are in the third line.

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)

[38]:
```
data4 = pd.read_csv('lab data/data4.txt',delimiter='$',names=column_names)
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
```

- This dataset was delimited by $, so I used `delimiter='$'` here. Also this dataset was missing column names so these were applied using `names=column_names`.

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)

[47]:
```
data5 = pd.read_csv('lab data/data5.csv',skipfooter=2)
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
```

```
/tmp/ipykernel_42579/549364903.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('lab data/data5.csv',skipfooter=2)
```

- Here I used `skipfooter=2` since the metadata was located at the bottom two rows of the dataset this time.

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)

[53]:
```
data6 = pd.read_csv('lab data/data6.dat',na_values=999)
data6.head(3).T
```

[53]:

	0	1	2
Country	Finland	Norway	Denmark
Happiness score	7.632	7.594	7.555
Whisker-high	7.695	7.657	7.623
Whisker-low	7.569	7.53	7.487
Dystopia (1.92) + residual	2.595	NaN	2.37
Explained by: GDP per capita	NaN	NaN	1.351
Explained by: Social support	NaN	1.582	1.59
Explained by: Healthy life expectancy	NaN	NaN	NaN
Explained by: Freedom to make life choices	0.681	0.686	0.683

```
Explained by: Generosity                    0.192   0.286    0.284
Explained by: Perceptions of corruption     0.393    0.34    0.408
```

- I used `na_values=999` here since there were many values of 999 throughout the data. This is not likely to be real data, so `NaN` was substituted here.

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)

[58]:
```
data7 = pd.read_excel('lab data/data7.xlsx',sheet_name='Data')
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
```

- I used `sheet_name='Data'` in combination with `pd.read_excel()` since this function is the easiest way to read data from an excel document.

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)

[59]:
```
data8 = pd.read_stata('lab data/data8.dta',)
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
```

- I used `pd.read_stata()` here since this is a Stata file.

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)

[65]:
```
data9 = pd.read_spss('lab data/data9.sav')
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
```

- I used `pd.read_spss()` here since this is an SPSS file.

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)

[79]:
```
data10 = pd.read_sas('lab data/data10.xpt')
data10.head(3).T
```

[79]:

	0	1	2
COUNTRY	b'Finland'	b'Norway'	b'Denmark'
HAPPINES	7.632	7.594	7.555
WHISKERH	7.695	7.657	7.623
WHISKERL	7.569	7.53	7.487
DYSTOPIA	2.595	2.383	2.37
EXPLAINE	1.305	1.456	1.351
EXPLAIN2	1.592	1.582	1.59
EXPLAIN3	0.874	0.861	0.868
EXPLAIN4	0.681	0.686	0.683
EXPLAIN5	0.192	0.286	0.284
EXPLAIN6	0.393	0.34	0.408

- I used `pd.read_sas()` here since this is an SAS file.

1.13 Problem 11

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

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

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

[80]:
```
colwidths = [24,5,5,5,5,5,5,5,5,5,5]
```

```
data11 = pd.read_fwf('lab data/data11.txt',widths =␣
 ↪colwidths,names=column_names)

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
```

[85]:
```
data11.to_csv('lab data/newdata11.csv',index=False)

pd.read_csv('lab data/newdata11.csv').head(3)
```

[85]:

	Country	Happiness score	Whisker-high	Whisker-low	Dystopia (1.92) + residual	Explained by: GDP per capita	Explained by: Social support	Explained by: Healthy life expectancy	Explained by: Freedom to make life choices	Explained by: Generosity	Explained by: Perceptions of corruption
0	Finland	7.632	7.695	7.569	2.595	1.305	1.592	0.874	0.681	0.192	0.393
1	Norway	7.594	7.657	7.530	2.383	1.456	1.582	0.861	0.686	0.286	0.340
2	Denmark	7.555	7.623	7.487	2.370	1.351	1.590	0.868	0.683	0.284	0.408

```
   Explained by: Perceptions of corruption
0                                   0.393
1                                   0.340
2                                   0.408
```

- I used `index=False` in `to_csv()` so that the row number wouldn't be saved as well.