# Interactive exercise week #7a

Liping Wu 300-958-061

In this exercise we will do the following:

- Handle various kind of data importing scenarios that is importing various kind of datasets (.csv, .txt), different kind of delimiters (comma, tab, pipe), and different methods (read_csv, read_table)

Pre-requisites:

1- Install Anoconda
2- We will be using a lot of Public datasets these datasets are available at:

https://goo.gl/zjS4C6

Under a folder named "Datasets for Predictive Modelling with Python", the datasets are organized in the order of the third text book chapters:

Python: Advanced Predictive Analytics, by Joseph Babcock and Ashish Kumar. Published by Packt Publishing Ltd ISBN: 9781788992367.(12/2017) For this exercise we need the files of chapter # 2.

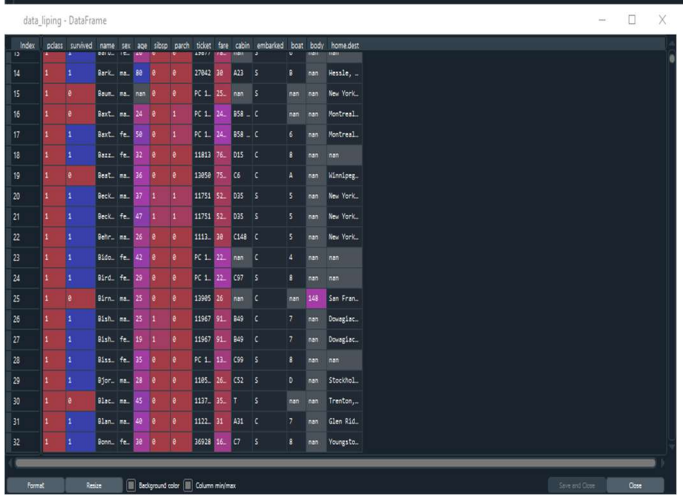
**Steps for handling various kinds of data import:**

1- Open your spider IDE
2- Follow the steps in chapter #2 to load the 'titanic3.csv' file into a dataframe name the dataframe data_firstname where first name is your first name
Following is the code, *make sure you update the path to the correct path where you placed the files*:

```
#author liping

import pandas as pd
import os
path = "D:/CentennialWu/2020Fall/COMP309Data/Assignments/Lab06DataLoading&Wrangling"
filename = 'titanic3.csv'
fullpath = os.path.join(path,filename)
## read
data_liping = pd.read_csv(fullpath)
#print (data_liping.tail(10))
print (data_liping.head(10))
```

```python
#author liping

import pandas as pd
import os
path = "D:/CentennialWu/2020Fall/COMP309Data/Assignments/Lab06DataLoading&Wrangling"
filename = 'titanic3.csv'
fullpath = os.path.join(path,filename)
## read
data_liping = pd.read_csv(fullpath)
print (data_liping.tail(10))
print (data_liping.head(10))
```

```
In [12]: import pandas as pd
    ...: import os
    ...: path = "D:/CentennialWu/2020Fall/COMP309Data/Assignments/Lab06DataLoading&Wrangling"
    ...: filename = 'titanic3.csv'
    ...: fullpath = os.path.join(path,filename)
    ...: ## read
    ...: data_liping = pd.read_csv(fullpath)
    ...: print (data_liping.tail(10))
    ...: print (data_liping.head(10))
      pclass  survived  ...   body home.dest
1300     3.0       1.0  ...    NaN       NaN
1301     3.0       0.0  ...  312.0       NaN
1302     3.0       0.0  ...    NaN       NaN
1303     3.0       0.0  ...    NaN       NaN
1304     3.0       0.0  ...  328.0       NaN
1305     3.0       0.0  ...    NaN       NaN
1306     3.0       0.0  ...  304.0       NaN
1307     3.0       0.0  ...    NaN       NaN
1308     3.0       0.0  ...    NaN       NaN
1309     NaN       NaN  ...    NaN       NaN

[10 rows x 14 columns]
   pclass  survived  ...   body                        home.dest
0     1.0       1.0  ...    NaN                     St Louis, MO
1     1.0       1.0  ...    NaN  Montreal, PQ / Chesterville, ON
2     1.0       0.0  ...    NaN  Montreal, PQ / Chesterville, ON
3     1.0       0.0  ...  135.0  Montreal, PQ / Chesterville, ON
4     1.0       0.0  ...    NaN  Montreal, PQ / Chesterville, ON
5     1.0       1.0  ...    NaN                    New York, NY
6     1.0       1.0  ...    NaN                      Hudson, NY
7     1.0       0.0  ...    NaN                     Belfast, NI
8     1.0       1.0  ...    NaN             Bayside, Queens, NY
9     1.0       0.0  ...   22.0             Montevideo, Uruguay

[10 rows x 14 columns]

In [13]:
```




3- Load the data in the "Customer Churn Model.txt" into a dataframe named data1_firstname where firstname is your fristname. Check the column names and types.

Following is the code, *make sure you update the path to the correct path where you placed the files.*

```python
import pandas as pd

fullpath1 = 'D:/CentennialWu/2020Fall/COMP309Data/Assignments/Lab06DataLoading&Wrangling/Customer Churn Model.txt'
```

```
data1_liping= pd.read_csv(fullpath1)

data1_liping.columns.values

print(data1_liping.columns.values)

data1_liping.dtypes

for col in data1_liping.columns:

    print(col)
```

```
import pandas as pd
fullpath1  = 'D:/CentennialWu/2020Fall/COMP309Data/Assi
data1_liping= pd.read_csv(fullpath1)
data1_liping.columns.values
print(data1_liping.columns.values)
data1_liping.dtypes
for col in data1_liping.columns:
    print(col)
```

```
In [14]: import pandas as pd
    ...: fullpath1  = 'D:/CentennialWu/2020Fall/COMP309Data/Assignments/Lab06DataLoading&Wrangling/Customer Churn Model.txt'
    ...: data1_liping= pd.read_csv(fullpath1)
    ...: data1_liping.columns.values
    ...: print(data1_liping.columns.values)
    ...: data1_liping.dtypes
    ...: for col in data1_liping.columns:
    ...:     print(col)
['State' 'Account Length' 'Area Code' 'Phone' "Int'l Plan" 'VMail Plan'
 'VMail Message' 'Day Mins' 'Day Calls' 'Day Charge' 'Eve Mins'
 'Eve Calls' 'Eve Charge' 'Night Mins' 'Night Calls' 'Night Charge'
 'Intl Mins' 'Intl Calls' 'Intl Charge' 'CustServ Calls' 'Churn?']
State
Account Length
Area Code
Phone
Int'l Plan
VMail Plan
VMail Message
Day Mins
Day Calls
Day Charge
Eve Mins
Eve Calls
Eve Charge
Night Mins
Night Calls
Night Charge
Intl Mins
Intl Calls
Intl Charge
CustServ Calls
Churn?

In [15]:
```

*4-* Read line by line below the code, *change my firstname to your firstname:*:

```
fullpath1  =
'D:/CentennialWu/2020Fall/COMP309Data/Assignments/Lab06DataLoading&Wrangling/Cu
stomer Churn Model.txt'

data=open(fullpath1,'r')

cols=data.readline().strip().split(',')

no_cols=len(cols)

print(no_cols)
```

```
#### Finding the number of rows
counter=0
main_dict={}
for col in cols:
    main_dict[col]=[]
    print(main_dict)
for line in data:
    values = line.strip().split(',')
    for i in range(len(cols)):
        main_dict[cols[i]].append(values[i])
    counter += 1
print ("The dataset has %d rows and %d columns" % (counter,no_cols))
import pandas as pd
df_liping=pd.DataFrame(main_dict)
print (df_liping.head(5))
```

```
    ...: import pandas as pd
    ...: df_liping=pd.DataFrame(main_dict)
    ...: print (df_liping.head(5))
21
{'State': []}
{'State': [], 'Account Length': []}
{'State': [], 'Account Length': [], 'Area Code': []}
{'State': [], 'Account Length': [], 'Area Code': [], 'Phone': []}
{'State': [], 'Account Length': [], 'Area Code': [], 'Phone': [], "Int'l Plan": []}
{'State': [], 'Account Length': [], 'Area Code': [], 'Phone': [], "Int'l Plan": [], 'VMail Plan': []}
{'State': [], 'Account Length': [], 'Area Code': [], 'Phone': [], "Int'l Plan": [], 'VMail Plan': [], 'VMail Message': []}
{'State': [], 'Account Length': [], 'Area Code': [], 'Phone': [], "Int'l Plan": [], 'VMail Plan': [], 'VMail Message': [], 'Day Mins': []}
{'State': [], 'Account Length': [], 'Area Code': [], 'Phone': [], "Int'l Plan": [], 'VMail Plan': [], 'VMail Message': [], 'Day Mins': [], 'Day Calls': []}
{'State': [], 'Account Length': [], 'Area Code': [], 'Phone': [], "Int'l Plan": [], 'VMail Plan': [], 'VMail Message': [], 'Day Mins': [], 'Day Calls': [], 'Day
Charge': []}
{'State': [], 'Account Length': [], 'Area Code': [], 'Phone': [], "Int'l Plan": [], 'VMail Plan': [], 'VMail Message': [], 'Day Mins': [], 'Day Calls': [], 'Day
Charge': [], 'Eve Mins': []}
{'State': [], 'Account Length': [], 'Area Code': [], 'Phone': [], "Int'l Plan": [], 'VMail Plan': [], 'VMail Message': [], 'Day Mins': [], 'Day Calls': [], 'Day
Charge': [], 'Eve Mins': [], 'Eve Calls': []}
{'State': [], 'Account Length': [], 'Area Code': [], 'Phone': [], "Int'l Plan": [], 'VMail Plan': [], 'VMail Message': [], 'Day Mins': [], 'Day Calls': [], 'Day
Charge': [], 'Eve Mins': [], 'Eve Calls': [], 'Eve Charge': []}
{'State': [], 'Account Length': [], 'Area Code': [], 'Phone': [], "Int'l Plan": [], 'VMail Plan': [], 'VMail Message': [], 'Day Mins': [], 'Day Calls': [], 'Day
Charge': [], 'Eve Mins': [], 'Eve Calls': [], 'Eve Charge': [], 'Night Mins': []}
{'State': [], 'Account Length': [], 'Area Code': [], 'Phone': [], "Int'l Plan": [], 'VMail Plan': [], 'VMail Message': [], 'Day Mins': [], 'Day Calls': [], 'Day
Charge': [], 'Eve Mins': [], 'Eve Calls': [], 'Eve Charge': [], 'Night Mins': [], 'Night Calls': []}
{'State': [], 'Account Length': [], 'Area Code': [], 'Phone': [], "Int'l Plan": [], 'VMail Plan': [], 'VMail Message': [], 'Day Mins': [], 'Day Calls': [], 'Day
Charge': [], 'Eve Mins': [], 'Eve Calls': [], 'Eve Charge': [], 'Night Mins': [], 'Night Calls': [], 'Night Charge': []}
{'State': [], 'Account Length': [], 'Area Code': [], 'Phone': [], "Int'l Plan": [], 'VMail Plan': [], 'VMail Message': [], 'Day Mins': [], 'Day Calls': [], 'Day
Charge': [], 'Eve Mins': [], 'Eve Calls': [], 'Eve Charge': [], 'Night Mins': [], 'Night Calls': [], 'Night Charge': [], 'Intl Mins': []}
{'State': [], 'Account Length': [], 'Area Code': [], 'Phone': [], "Int'l Plan": [], 'VMail Plan': [], 'VMail Message': [], 'Day Mins': [], 'Day Calls': [], 'Day
Charge': [], 'Eve Mins': [], 'Eve Calls': [], 'Eve Charge': [], 'Night Mins': [], 'Night Calls': [], 'Night Charge': [], 'Intl Mins': [], 'Intl Calls': []}
{'State': [], 'Account Length': [], 'Area Code': [], 'Phone': [], "Int'l Plan": [], 'VMail Plan': [], 'VMail Message': [], 'Day Mins': [], 'Day Calls': [], 'Day
Charge': [], 'Eve Mins': [], 'Eve Calls': [], 'Eve Charge': [], 'Night Mins': [], 'Night Calls': [], 'Night Charge': [], 'Intl Mins': [], 'Intl Calls': [], 'Intl
Charge': []}
{'State': [], 'Account Length': [], 'Area Code': [], 'Phone': [], "Int'l Plan": [], 'VMail Plan': [], 'VMail Message': [], 'Day Mins': [], 'Day Calls': [], 'Day
Charge': [], 'Eve Mins': [], 'Eve Calls': [], 'Eve Charge': [], 'Night Mins': [], 'Night Calls': [], 'Night Charge': [], 'Intl Mins': [], 'Intl Calls': [], 'Intl
Charge': [], 'CustServ Calls': []}
{'State': [], 'Account Length': [], 'Area Code': [], 'Phone': [], "Int'l Plan": [], 'VMail Plan': [], 'VMail Message': [], 'Day Mins': [], 'Day Calls': [], 'Day
Charge': [], 'Eve Mins': [], 'Eve Calls': [], 'Eve Charge': [], 'Night Mins': [], 'Night Calls': [], 'Night Charge': [], 'Intl Mins': [], 'Intl Calls': [], 'Intl
Charge': [], 'CustServ Calls': [], 'Churn?': []}
The dataset has 3333 rows and 21 columns
  State Account Length Area Code  ... Intl Charge CustServ Calls  Churn?
0    KS            128       415  ...    2.700000             1  False.
1    OH            107       415  ...    3.700000             1  False.
2    NJ            137       415  ...    3.290000             0  False.
3    OH             84       408  ...    1.780000             2  False.
4    OK             75       415  ...    2.730000             3  False.

[5 rows x 21 columns]

In [20]:
```

5- Read data directly into a data frame from a URL, below the code, *change my firstname to your firstname:*

```
import pandas as pd
url = 'http://winterolympicsmedals.com/medals.csv'
medal_data_liping=pd.read_csv(url)
print (medal_data_liping.head(5))
```

```
In [21]: import pandas as pd
    ...: url = 'http://winterolympicsmedals.com/medals.csv'
    ...: medal_data_liping=pd.read_csv(url)
    ...: print (medal_data_liping.head(5))
   Year      City       Sport  ...       Event Event gender   Medal
0  1924  Chamonix     Skating  ...  individual            M  Silver
1  1924  Chamonix     Skating  ...  individual            W    Gold
2  1924  Chamonix     Skating  ...       pairs            X    Gold
3  1924  Chamonix   Bobsleigh  ...    four-man            M  Bronze
4  1924  Chamonix  Ice Hockey  ...  ice hockey            M    Gold

[5 rows x 8 columns]

In [22]:
```

```
In [21]: import pandas as pd
    ...: url = 'http://winterolympicsmedals.com/medals.csv'
    ...: medal_data_liping=pd.read_csv(url)
    ...: print (medal_data_liping.head(5))
   Year      City       Sport  ...       Event Event gender   Medal
0  1924  Chamonix     Skating  ...  individual            M  Silver
1  1924  Chamonix     Skating  ...  individual            W    Gold
2  1924  Chamonix     Skating  ...       pairs            X    Gold
3  1924  Chamonix   Bobsleigh  ...    four-man            M  Bronze
4  1924  Chamonix  Ice Hockey  ...  ice hockey            M    Gold

[5 rows x 8 columns]

In [22]: import pandas as pd
    ...: url = 'http://winterolympicsmedals.com/medals.csv'
    ...: medal_data_liping=pd.read_csv(url)
    ...: print (medal_data_liping)
      Year      City       Sport  ...            Event Event gender   Medal
0     1924  Chamonix     Skating  ...       individual            M  Silver
1     1924  Chamonix     Skating  ...       individual            W    Gold
2     1924  Chamonix     Skating  ...            pairs            X    Gold
3     1924  Chamonix   Bobsleigh  ...         four-man            M  Bronze
4     1924  Chamonix  Ice Hockey  ...       ice hockey            M    Gold
...    ...       ...         ...  ...              ...          ...     ...
2306  2006     Turin      Skiing  ...        Half-pipe            M  Silver
2307  2006     Turin      Skiing  ...        Half-pipe            W    Gold
2308  2006     Turin      Skiing  ...        Half-pipe            W  Silver
2309  2006     Turin      Skiing  ...  Snowboard Cross            M    Gold
2310  2006     Turin      Skiing  ...  Snowboard Cross            W  Silver

[2311 rows x 8 columns]

In [23]:
```