Data Science: Tools and Techniques
Lab Exercise (Week 2)
Prepared By
Dr Muhammad Atif Tahir (Spring 2018)

1. Do all examples provided in
   http://www.slideshare.net/AndrewHenshaw1/pandas-22984889

2. In this lab, we will focus on taking input from different sources. Data may be incomplete and there may be missing values. The objective is to combine data and save it as one dataframe object. Do the following steps

   a. Read data1.csv and data2.csv as df1 and df2 dataFrames objects respectively. Use Pandas I/O libraries. Use column 0 as index for df1 and column 1 as index for df2. Print df1 and df2

```
   A    B  C
0  1  2.0  3
1  4  5.0  6
4  2  NaN  5

         A  B  C
2        2  5  6
3  Hello  3  4
```

   b. Modify (a) to concatenate df1 and df2 and assign it to df3. Print df3

```
       A    B  C
0      1  2.0  3
1      4  5.0  6
4      2  NaN  5
2      2  5.0  6
3  Hello  3.0  4
```

   c. Read data3.csv as df4 dataframe object and print df4 (not shown below). There are 2 new column ‘D’ and ‘E’ in this file. Merge df4 with df3 so that new dataframe (df5) has total 5 columns (A, B, C, D, E)

```
       A    B  C    D    E
0      1  2.0  3  NaN  NaN
1      4  5.0  6  1.0  7.0
2      2  5.0  6  NaN  NaN
3  Hello  3.0  4  NaN  NaN
4      2  NaN  5  0.0  8.0
```

d. Read data.json as df6 and concatenate with df5. Use df7 as name of dataframe

	A	B	C	D	E
0	1	2.0	3.0	NaN	NaN
1	4	5.0	6.0	1.0	7.0
2	2	5.0	6.0	NaN	NaN
3	Hello	3.0	4.0	NaN	NaN
4	2	NaN	5.0	0.0	8.0
5	11	9.0	NaN	NaN	NaN
6	22	7.0	NaN	NaN	NaN
7	33	8.0	NaN	NaN	NaN

e. Replace Hello with NaN. Make it as general as possible so that all strings in dataframe df7 automatically becomes NaN

	A	B	C	D	E
0	1.0	2.0	3.0	NaN	NaN
1	4.0	5.0	6.0	1.0	7.0
2	2.0	5.0	6.0	NaN	NaN
3	NaN	3.0	4.0	NaN	NaN
4	2.0	NaN	5.0	0.0	8.0
5	11.0	9.0	NaN	NaN	NaN
6	22.0	7.0	NaN	NaN	NaN
7	33.0	8.0	NaN	NaN	NaN

f. Replace NaN with mean values of the columns. Save the final dataframe as "newdata.csv"

	A	B	C	D	E
0	1.00	2.00	3.0	0.5	7.5
1	4.00	5.00	6.0	1.0	7.0
2	2.00	5.00	6.0	0.5	7.5
3	10.71	3.00	4.0	0.5	7.5
4	2.00	5.57	5.0	0.0	8.0
5	11.00	9.00	4.8	0.5	7.5
6	22.00	7.00	4.8	0.5	7.5
7	33.00	8.00	4.8	0.5	7.5

3. The following website provides a very good exercise about real world data cleaning. Follow the steps and complete the exercise below
https://mashimo.wordpress.com/2013/09/28/read-and-clean-data-with-python-pandas/