InfPALS
Pandas

Pandas is a library providing data structures and data analysis tools. It allows us to load data from different sources, and then allows us to manipulate and search the data.

Enter the DataFrame

Before we get started, first we'll introduce you to the **DataFrame**, the core data structure of Pandas. It's effectively a 2-dimensional table, supporting axis names and normal table operations. For example, raw_data_1 in exercise 7 looks like this (note that Pandas automatically indexes each row for us):

```
  subject_id first_name last_name
0          1       Alex  Anderson
1          2        Amy  Ackerman
2          3      Allen       Ali
3          4      Alice      Aoni
4          5     Ayoung   Atiches
```

Below are some exercises to learn the basics of Pandas.
10 minutes to Pandas is your friend and you will find a lot of help over there to start off. After scanning through it, try to answer the questions on your own (with some help from stack overflow and the API for Pandas dataframe)– what is your motivation for studying Pandas, why we need it, what is its main functionality, how widely it can be used.

1) Download the .csv file at http://www.football-data.co.uk/mmz4281/1718/E0.csv and put it into your working directory.

2) Open a new Jupyter notebook and start setting up:

```
import pandas, numpy, matplotlib as pd, np, plt
import os
os.listdir()
```

3) Read E0.csv into a DataFrame. What if one of the parameters was header=None?

```
db = pd.read_csv("E0.csv")
```

4) Once we have some data, we want to know some information about it. Write appropriate commands to obtain:

a) number of columns

b) number of records

c) extract first 10 rows and then last 25

Also try out these commands, replacing db with whatever you assigned pd.read() to.

```
db.info()
db.shape[0]
db.shape[1]
db.head(10) // if no parameter is given, default value is 5
db.tail(25) // if no parameter is given, default value is 5
```

5) Did you see that ludicrous display last night?!

a) In E0.csv, what is the most number of goals scored by a home team (FTHG - represents Full Time Home Goals)? And for an away team (FTAG - Full Time Away Goals)?

b) Which team scored the highest number of Home Shots (HS) in one match? And the away team (AS)? Which team had the most shots across all games?

6) Pandas can do Excel-like stuff too:

a) Change type of all values in AF column to float.

b) List all the full-time results and sum the total number of distinct results.

c) When encountering larger datasets, it's helpful to reduce the space it takes up in memory by dropping unnecessary columns. Try *dropping* the 'Div', 'Date' and 'Referee' columns (note that most DataFrame functions are not inplace, hence you'll need to assign the result to itself for it to stay that way).

7) Converting new dataframes from dictionaries in Jupyter is dead simple:

```
raw_data_1 = pd.DataFrame({ 'subject_id': ['1', '2', '3', '4', '5'], 'first_name': ['Alex', 'Amy',
```

```
'Allen', 'Alice', 'Ayoung'], 'last_name': ['Anderson', 'Ackerman', 'Ali', 'Aoni', 'Atiches']})
raw_data_2 =  pd.DataFrame({ 'subject_id': ['4', '5', '6', '7', '8'], 'first_name': ['Billy', 'Brian',
'Bran', 'Bryce', 'Betty'], 'last_name': ['Bonder', 'Black', 'Balwner', 'Brice', 'Btisan']})
raw_data_3 =  pd.DataFrame({ 'subject_id': ['1', '2', '3', '4', '5', '7', '8', '9', '10', '11'], 'test_id':
[51, 15, 15, 61, 16, 14, 15, 1, 61, 16]})
```

8) You can merge dataframes in different ways (hint : refer to the API for DataFrame.merge - check the examples!).

a) Join the two dataframes along rows and assign it to allDataOne.

b) Join the first two dataframes along column and assign it to allDataTwo.

c) Merge allDataOne and third data along the subject_id.

d) Merge only the data that has the same 'subject_id' on both data1 and data2. Hint: use the how parameter e.g. exampleDF.merge(exampleDF2, how='?')

e) Merge all values in data1 and data2, with matching records from both sides where available (Hint: Notice that it's complement to d).

I hope this has made you feel more comfortable using Pandas! Please try out the next act, featuring Pandas' father, if you haven't already.