PANDAS

Pandas is library providing data structures and data analysis tools. It allows us to load data from different sources, and then uses python to do something with that.

Here it is assumed that you have your Jupyter with necessary libraries downloaded. I recommend getting familiar with keyboard shortcuts in Jupyter by clicking Help on the top of the page.

Here below are some exercises to learn some pandas as there is no theory for that. Instead of googling everything, this page is your friend and you will find a lot of help over there. Scroll “10 minutes to Pandas” beforehand. After doing it, try to answer on your own – what is your motivation for studying Pandas, why we need, what is the main functionality, how widely it can be used.

1) Import two datasets - files csv and tsv - using command pd.read_csv.

2) csv file is http://www.football-data.co.uk/mmz4281/1718/E0.csv
   and tsv file is http://media.lokad.com/www/misc/salescast-tsv-sample.zip
   (Lokad_Items.tsv)
   What if one of argument was header=None?

3) Open new Jupyter notebook in the same folder, rename it and
   import pandas, numpy, matplotlib as pd, np, plt. Furthermore, type this lines:
   import os
   os.listdir()

4) Once we have some data, we want to know some information about it. Write appropriate command to obtain:
   a) number of columns,
   b) number of records,
   c) extract first 10 and then last 25,
   Check out also:
   d) db.info()
   e) db.shape[0]
   f) db.shape[1]
   where db is database file.

5) a) In first file, what is the most popular number of goals scored by home team(FTHG)? And for away team(FTAG)? Can we assume that combining these two will give us the most frequent result? If not, how to do that?
   b) List two teams that had the biggest number of shots in one match(HS/AS)? Which team had the most shots overall? Give a record with the biggest change of goal differences (and indicate its value). For instance if after first half (HTHG,HTAG) it was 2-0, and after full time(FTHG,FTAG) 3-6, change of goal differences is 5, but if after full time was 6-3, the change of goal differences is 1.

6) a) Change type of all values in AF column to float.
b) List all different full-time results and sum the total number of goals only in distinct results.

7) In second file:
a) Give mean of the values from column StockOnHand.
b) Extract all record with the least service level.

8) This might be important, since we usually don’t need to work with the full data.

Filter with service level bigger than 0.95, sort it by StockOnHand and slice it: we need to have LabelName, TagLabelCategory, TagSubcategory, ServiceLevel, LoadTime and StockOnHand of first 20 records, but not first 5.

9) Creating new data in Jupiter:

```
raw_data_1 = { 'subject_id': ['1', '2', '3', '4', '5'], 'first_name': ['Alex', 'Amy', 'Allen', 'Alice',
'Ayoung'],         'last_name': ['Anderson', 'Ackerman', 'Ali', 'Aoni', 'Atiches']}
raw_data_2 = { 'subject_id': ['4', '5', '6', '7', '8'], 'first_name': ['Billy', 'Brian', 'Bran', 'Bryce',
'Betty'],         'last_name': ['Bonder', 'Black', 'Balwner', 'Brice', 'Btisan']}
raw_data_3 = { 'subject_id': ['1', '2', '3', '4', '5', '7', '8', '9', '10', '11'], 'test_id': [51, 15, 15, 61, 16,
14, 15, 1, 61, 16]}
```

10) And merge it, in different ways.
a) Join the two 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 parameter 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).

11) Save your workspace, are you sure you know how to access the notebook later?