Class 6, Problem Set 3

Introduction to Programming and Numerical Analysis

Plan for today

UNIVERSITY OF
COPENHAGEN

1. pandas
    - The go to package when handling data
    - Some syntax
2. Work on PS3

pandas

"`pandas` is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language."

Extra help can be found here

pandas

- When working with data we will be working with pandas' DataFrame: "Two-dimensional, size-mutable, potentially heterogeneous tabular data"
    - "Classic" dataset
    - A class in python with many methods!

pandas

- What does a DataFrame look like?
    - You can read csv, xlsx, sas etc. but also create DataFrames from dictionaries

pandas

- What does a DataFrame look like?
    - You can read csv, xlsx, sas etc. but also create DataFrames from dictionaries

In [ ]:

```
import pandas as pd

# Reading data using built in method read_csv. For other formats there are read_excel() and
pd.read_csv('random_txt.txt', sep='\t')
```

In [ ]:

```
# Using a dict to return a DataFrame
data_dict = {'random_integers' : [1, 45, 54, 12498134, 666]
            , 'random_strings' : ['one', 'forty-five', 'abc', 'Pandas is cool', '9.11 was an
            }
print(data_dict,'\n')

# Convert dict to df
df = pd.DataFrame(data_dict)

print(type(df))

# Look at the first 5 rows
df.head(5) # .tail() will give you the last x rows
```

pandas

- Now we know how to load data - how do wrangle it? Lets go through som basics
    - Accesing/creating columns
    - Subsetting DataFrames

In [ ]:

```
# You can acces columns in many ways
df['random_integers']
df.random_integers
df.loc[:,'random_integers']
df.iloc[:,0]
```

In [ ]:

```
# Add new columns
df['new_column'] = df['random_integers']/2 # math is allowed
df['new_column_list'] = [int(i*2) for i in df.new_column] # lists can be added as columns
df.head()
```

In [ ]:

```
# Subset DataFrames
boolean_array = df['random_integers'] > 100
print(boolean_array)

df_new = df.loc[boolean_array, ['random_strings']]
df_new
```

In [ ]:

```
# Subset DataFrames - pandas is very flexible
# Apply condition
df_new = df[df.random_integers > 100]
df_new = df[~ (df['random_integers'] < 100)]
#df_new.head()
# Subset data
df_new = df_new[['random_strings']]
df_new = df_new.drop(columns=['random_integers', 'new_column', 'new_column_list'])
df_new = df_new.iloc[:,[1]]
df_new.head()
```

pandas

- We have now seen that all roads lead to rome
- Remember the answers to the PS is suggested answers: what matters is the right result
- However, don't over complicate things

Problem set 3

Let's go!