ECS7024 Statistics for Artificial Intelligence and Data Science

Topic 1B: Python for Data Analysis

William Marsh

Outline

- Aim: Be Able to Use Python for Data Handling on this Module

- Pandas library and iPython (jupyter) interface
- The data frame
  - Introduction to notebook 1

Introducing Pandas and iPython

Python library for data analysis

iPython and Pandas

- iPython notebook (Jupyter)
  - Interactive Python interface running in web browser
  - Works for any Python program
  - ... good for graphics
  - *Name and brief history*

- Pandas
  - Python library for data analysis
  - Very big
  - We will learn essential features
  - ... practical work on this module

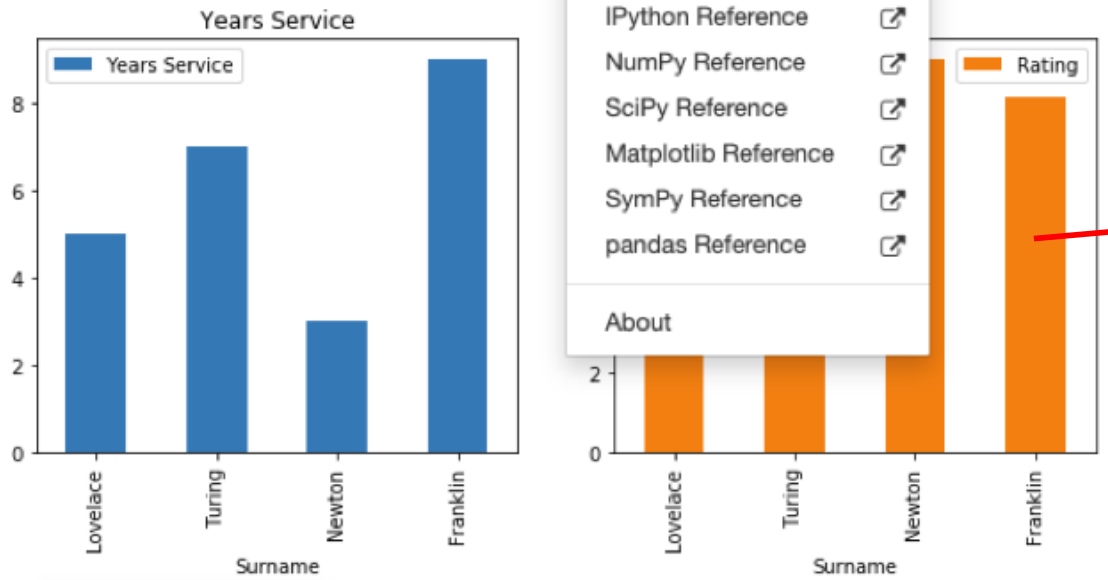
jupyter ipython-intro Last Checkpoint: 18 hours ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Run Code

In [31]:
```
import matplotlib.pyplot as plt
%matplotlib inline
```

In [32]:
```
df2.plot(kind="bar", x = 'Surname',                    ts=True, layo
                # this creates a plot b
plt.show()      # this displays the las
```



User Interface Tour
Keyboard Shortcuts
Edit Keyboard Shortcuts

Notebook Help
Markdown

Python Reference
IPython Reference
NumPy Reference
SciPy Reference
Matplotlib Reference
SymPy Reference
pandas Reference

About

Brie e of graphs available in Pandas: https://pandas.pydata.org/pandas-

Se lext

We ore detail at the start of the module. Pandas is a very large library, w
cov

```
abs
add
add_prefix
add_suffix
agg
aggregate
align
all
any
append
```

In [ ]:
```
df.
```

iPython intro available

Link to web sites

- Pandas user guide best

Graphics!

Contextual help: tab

Programming with Jupyter Notebook

- The notebook file (.ipynb) contains ‘cells’
  - Formatted text (using ‘markdown’)
  - Program fragments
- Combines document and program
  - Look at the form of the examples provided
- Program cells
  - Run cell by cell, any order
  - Run whole notebook: stops if an error occurs
  - Run all cells above / below
  - Clear output on any / all cells

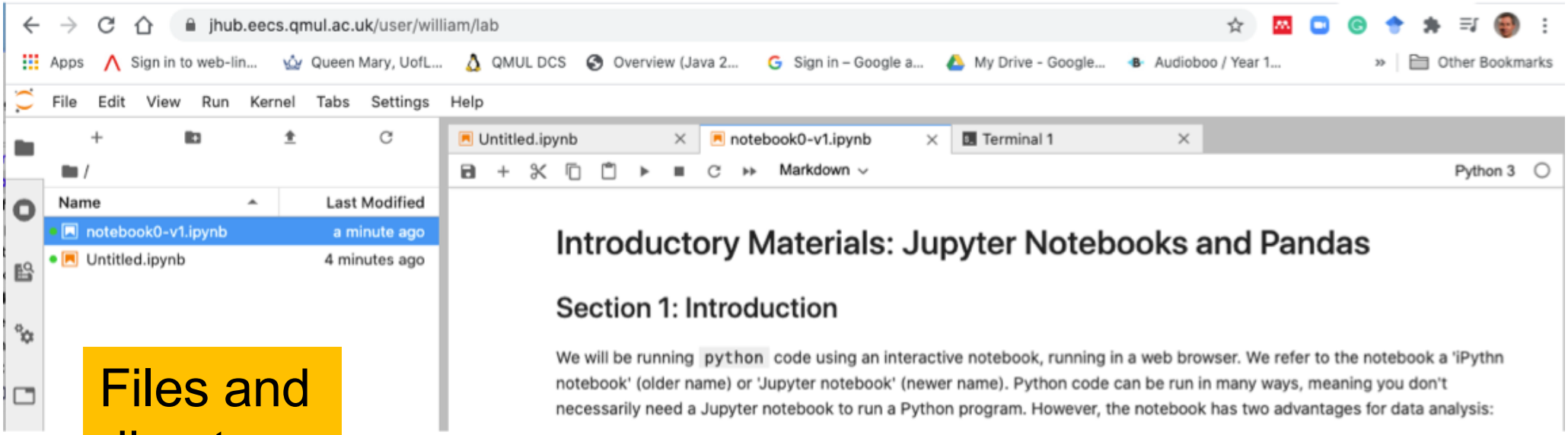
Guidance

- Develop interactively
- Use many small cells
- Alternate markdown and program
- Prefer markdown to comments
- Limited use for print
- Make a mess then tidy-up

Follow examples given

JupyterLab

- JupyterLab is an enhanced interface
  - Allows working with several notebooks at once

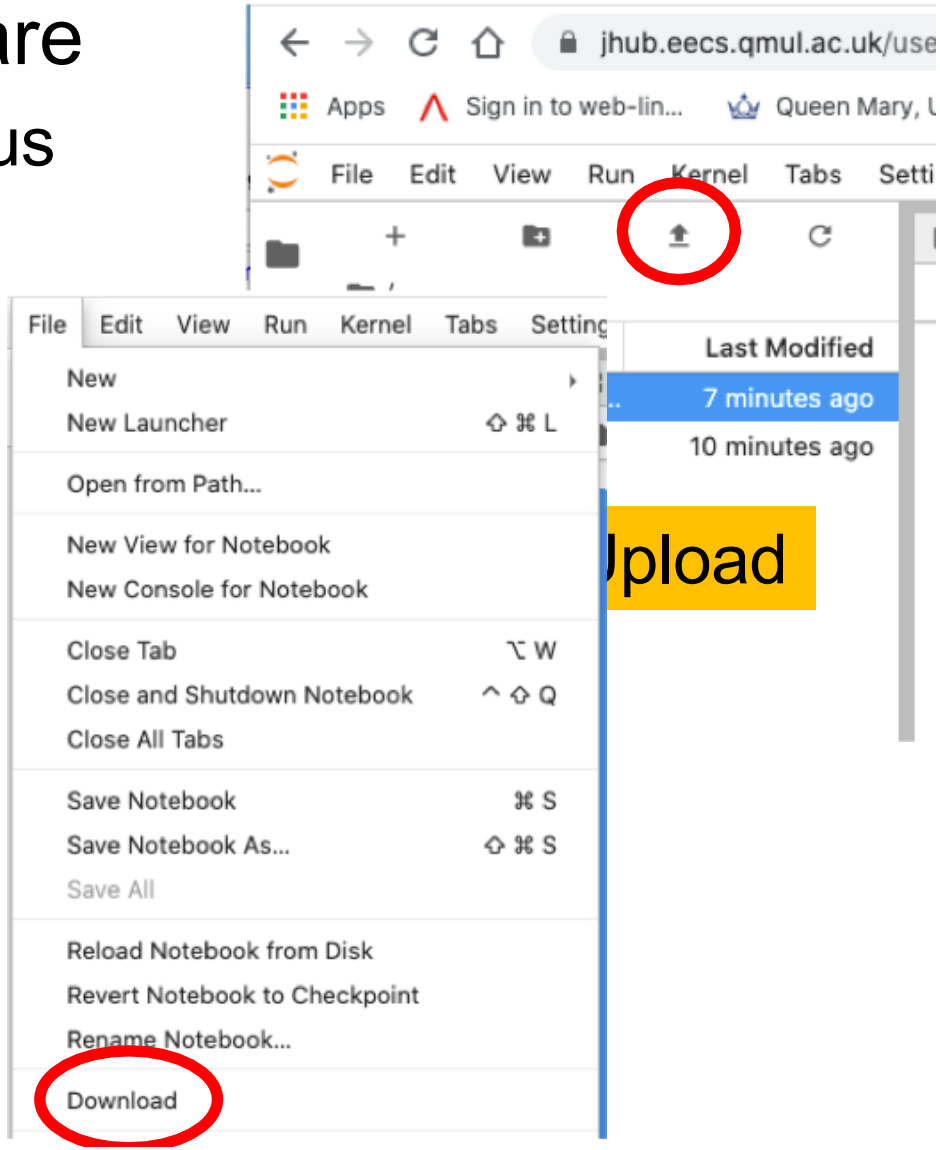


Files and directory

Notebook

Using jhub.eecs.qmul.ac.uk

- Web hosted system as an alternative to installing your own software
  - Download files from QMPlus to your PC
  - Upload to jhub ... modify
  - Download to your PC
  - ... submit on QMPlus



Menti Quiz 1

Pandas: data frame

Most important data type for data analysis

The Data Frame

- Header row
  - Shows the columns
- Rows
  - Shows individuals
- Tidy data
  - All columns have headings
  - All columns same ‘type’ (e.g. numbers)
  - No blanks

	Name	Age	Team
0	John	24	Arsenal
1	Mary	27	Spurs
2	Peter	31	Chelsea

```
Name,Age,Team
John,24,Arsenal
Mary,27,Spurs
Peter,31,Chelsea
```

Loaded from CSV

The Data: Country of Birth

- London Boroughs
- Taken from 2011 census
  - 67,252 row
  - Example of ‘narrow’ or ‘tall’ data

Area	Age	Sex	Usual Residents	Birth Country	Birth Region
Tower Hamlets	Age 0 to 4	Females	3	Ghana	Africa
Tower Hamlets	Age 5 to 9	Females	2	Ghana	Africa
Tower Hamlets	Age 10 to 15	Females	4	Ghana	Africa

London Boroughs

- Details not important
  - QMUL in Tower Hamlets
  - Smaller near centre
  - River is a
  - boundary

Investigating the Data

- Load the data to a dataframe
- Look at values the unique values in each column
  - Answer some questions about the data

Column	Description
Area	Includes London Boroughs
Age	The ages in a number of bands
Sex	Males and Females
Usual Residents	An integer
BirthCountry	A country
BirthRegion	A region e.g. Africa or Europe

Selecting and Transforming Data

- Selecting some data
  - Data for one borough / age
- Multiple conditions
- *No loops*

```
df.loc[(df.Area == 'Tower Hamlets')]
```

```
df.loc[(df.Area == 'Tower Hamlets’) &
       (df.Age == 'Age 0 to 4')]
```

See practical sheet for code examples

Learning Pandas: Our Approach

- Very large library
  - Many ways to achieve some end
- Learning
  - Task 1: "*What am I trying to do?*" Lectures
  - Task 2: "*How do I achieve it in Pandas?*" Practical work
- Generally avoid code on lecture slides
- Small subset of Pandas features
  - Develop an understanding
  - Sufficient for this module
  - Expand in future (see documentation)
  - 'Internet examples' problem

The Pivot Table

- Transform data shape
- Also in spreadsheets

Person	Genre	Rating
Andy	Classic	Like
Andy	Jazz	Hate
Andy	Folk	Hate
Bill	Classic	Hate
Bill	Jazz	Like
Bill	Folk	Like
Charlie	Classic	Like
Charlie	Jazz	Like
Charlie	Folk	Hate

Person	Genre		
	Classic	Jazz	Folk
Andy	Like	Hate	Hate
Bill	Hate	Like	Like
Charlie	Like	Like	Hate

Genre	Person		
	Andy	Bill	Charlie
Classic	Like	Hate	Like
Jazz	Hate	Like	Like
Folk	Hate	Like	Hate

The Pivot Table

- Transform data with aggregation

Person	Place	Purpose	Visits
Andy	Berlin	Hols	1
Andy	Berlin	Work	2
Andy	Paris	Hols	2
Andy	NY	Work	3
Andy	Madrid	Hols	1
Bill	Berlin	Work	4
Bill	Paris	Work	3
Charlie	Paris	Hols	1
Charlie	Rome	Hols	1
Charlie	Zurich	Hols	1

Aggregate over the places

Purpose	Visits		
	Andy	Bill	Charlie
Hols	4	0	3
Work	5	7	0

Aggregate over the person

Purpose	Visits					
	Berlin	Madrid	NY	Paris	Rome	Zurich
Hols	1	0	3	3	1	1
Work	6	1	0	3	0	0

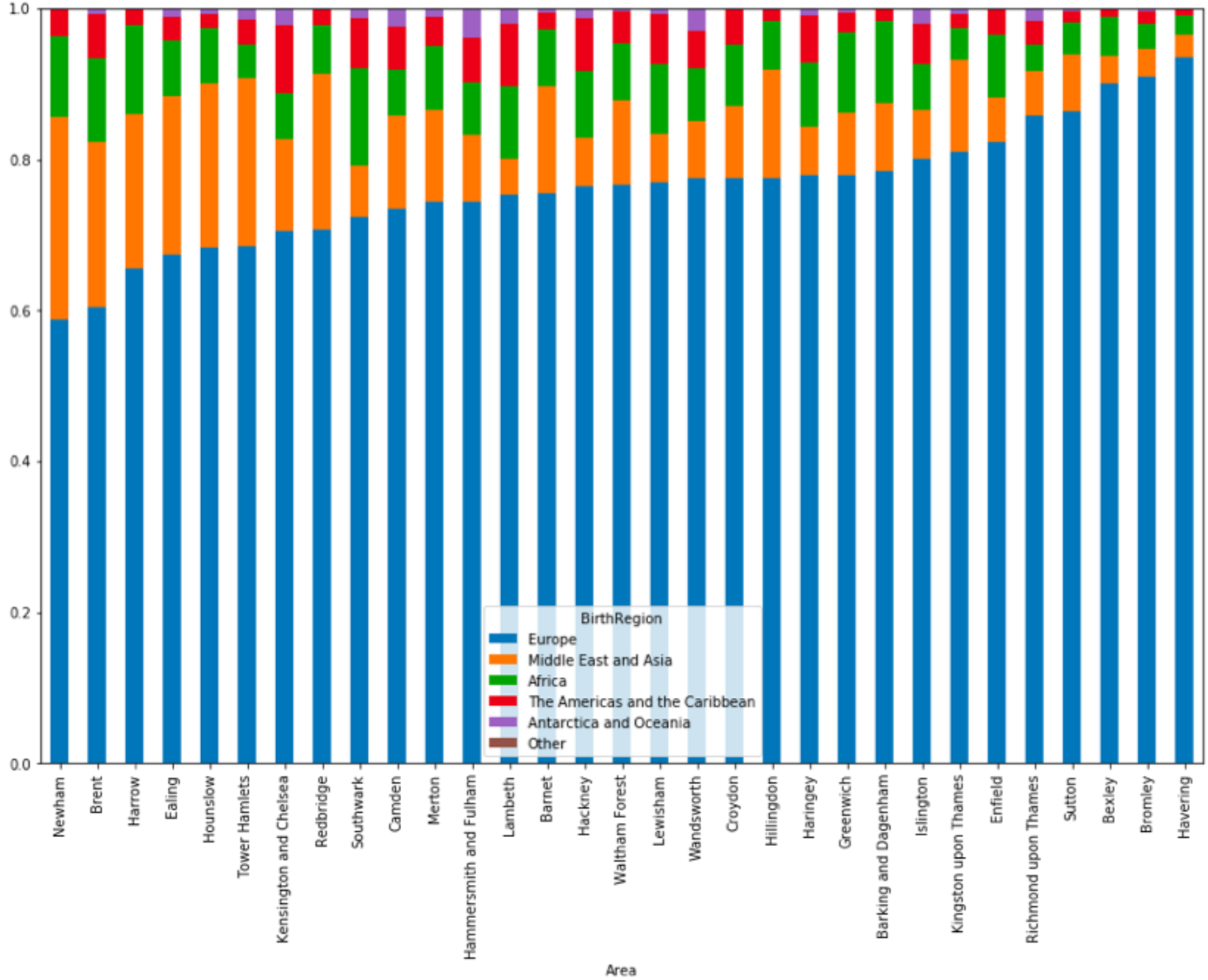
The Problem of Comparing Boroughs

- What does this question mean?

  Which Borough has more people from ...?

- Boroughs vary in population
- Need to transform the data into ‘proportion’
  - Use a Pivot table to find total Borough populations
  - Add a new column with ‘proportion’ rather than count

Proportion of Borough Born in Each Region



Summary

- Python notebook
  - Originally ‘iPython’, then Jupyter notebook and now JupyterLab
  - Interactive development
  - Document not just program
- Pandas
  - Complex library for data analysis in Python
  - ‘Dataframe’ for holding data
  - Notebook 0 introduces essential features
  - Notebook 1 looks at our first dataset

Menti Quiz 2