To Get Started

- Paper sheet and online at: http://www.eecs.qmul.ac.uk/~william/CAS-London-2020.html
- Download sample notebooks and data
  - Create directory (N:\session3\fullname)and unzip notebooks and data
- Login to Google Colab: https://colab.research.google.com/
  - Create a new notebook
  - Use the file ‘upload’ menu to upload the ‘example’ notebook

Introduction to Data Analysis

William Marsh

Queen Mary
University of London

institute of
CODING

COMPUTING AT SCHOOL
EDUCATE · ENGAGE · ENCOURAGE
Part of BCS –The Chartered Institute for IT

How This Session Works

- Introduce concepts
- Practical work
  - Collaboration: *practice teaching!*
- Repeat
- Conclude

Probably Not Enough Time

Outline

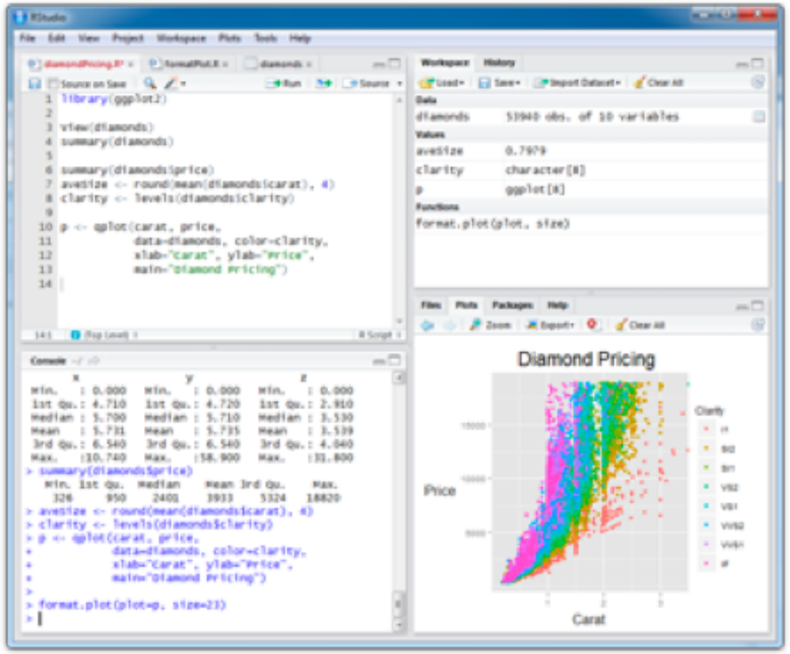
- Aims
- Introducing the Python notebook using Google Colab
- Part 1: the dataframe
- Part 2: transforming data and the pivot table
- Part 3: adding columns
- Conclusion and discussion

Session Aims

Aims

- Introduce Python Pandas
  - Popular library for data
  - Interactive notebook – Google's Colab
- Introduce key concepts
  - Dataframe
  - Filter and select
  - Pivot table
  - Visualisation

- Pandas is very complex
- Concepts common in other environments
  - Excel
  - RStudio
- Help develop pedagogy



```r
library(ggplot2)

view(diamonds)
summary(diamonds)

summary(diamonds$price)
avesize <- round(mean(diamonds$carat), 4)
clarity <- levels(diamonds$clarity)

p <- qplot(carat, price,
           data=diamonds, color=clarity,
           xlab="Carat", ylab="Price",
           main="Diamond Pricing")

```

```
       x                y                z
 Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
 1st Qu.: 4.710   1st Qu.: 4.720   1st Qu.: 2.910
 Median : 5.700   Median : 5.710   Median : 3.530
 Mean   : 5.731   Mean   : 5.735   Mean   : 3.539
 3rd Qu.: 6.540   3rd Qu.: 6.540   3rd Qu.: 4.040
 Max.   :10.740   Max.   :58.900   Max.   :31.800
> summary(diamonds$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    326     950    2401    3933    5324   18820
> avesize <- round(mean(diamonds$carat), 4)
> clarity <- levels(diamonds$clarity)
> p <- qplot(carat, price,
+            data=diamonds, color=clarity,
+            xlab="Carat", ylab="Price",
+            main="Diamond Pricing")
>
> format.plot(plot=p, size=23)
> 
```

Data

diamonds   53940 obs. of 10 variables

Values

avesize   0.7979

clarity   character[8]

p   ggplot[8]

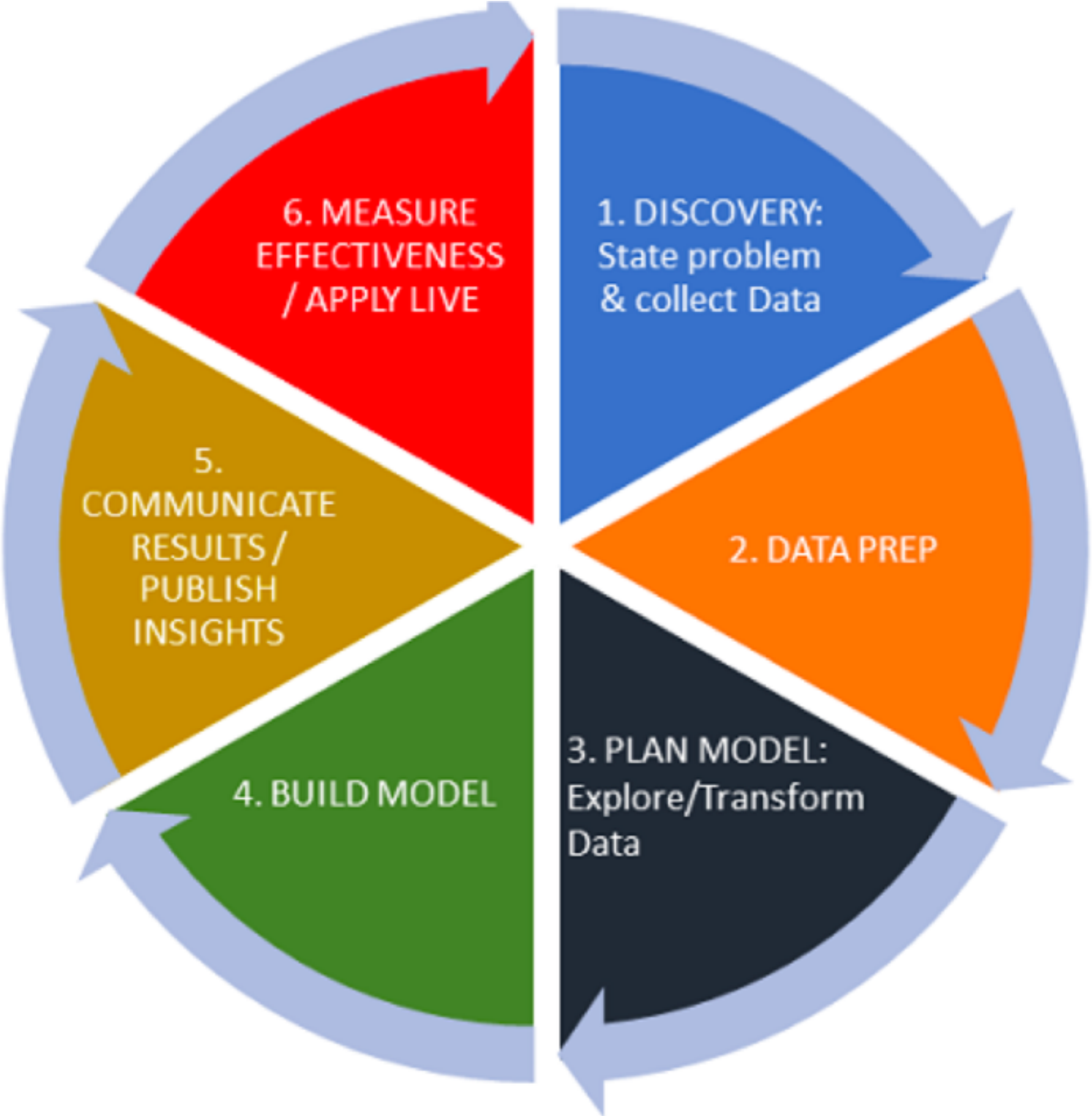
Functions

format.plot(plot, size)

The Data Life Cycle

- Using data to answer a question
    - What is the problem?
    - .. do we have the data?

- Data analysis not just technical
- We focus on the technical



1. DISCOVERY: State problem & collect Data
2. DATA PREP
3. PLAN MODEL: Explore/Transform Data
4. BUILD MODEL
5. COMMUNICATE RESULTS / PUBLISH INSIGHTS
6. MEASURE EFFECTIVENESS / APPLY LIVE

Introducing the Interactive Notebook

Queen Mary University of London

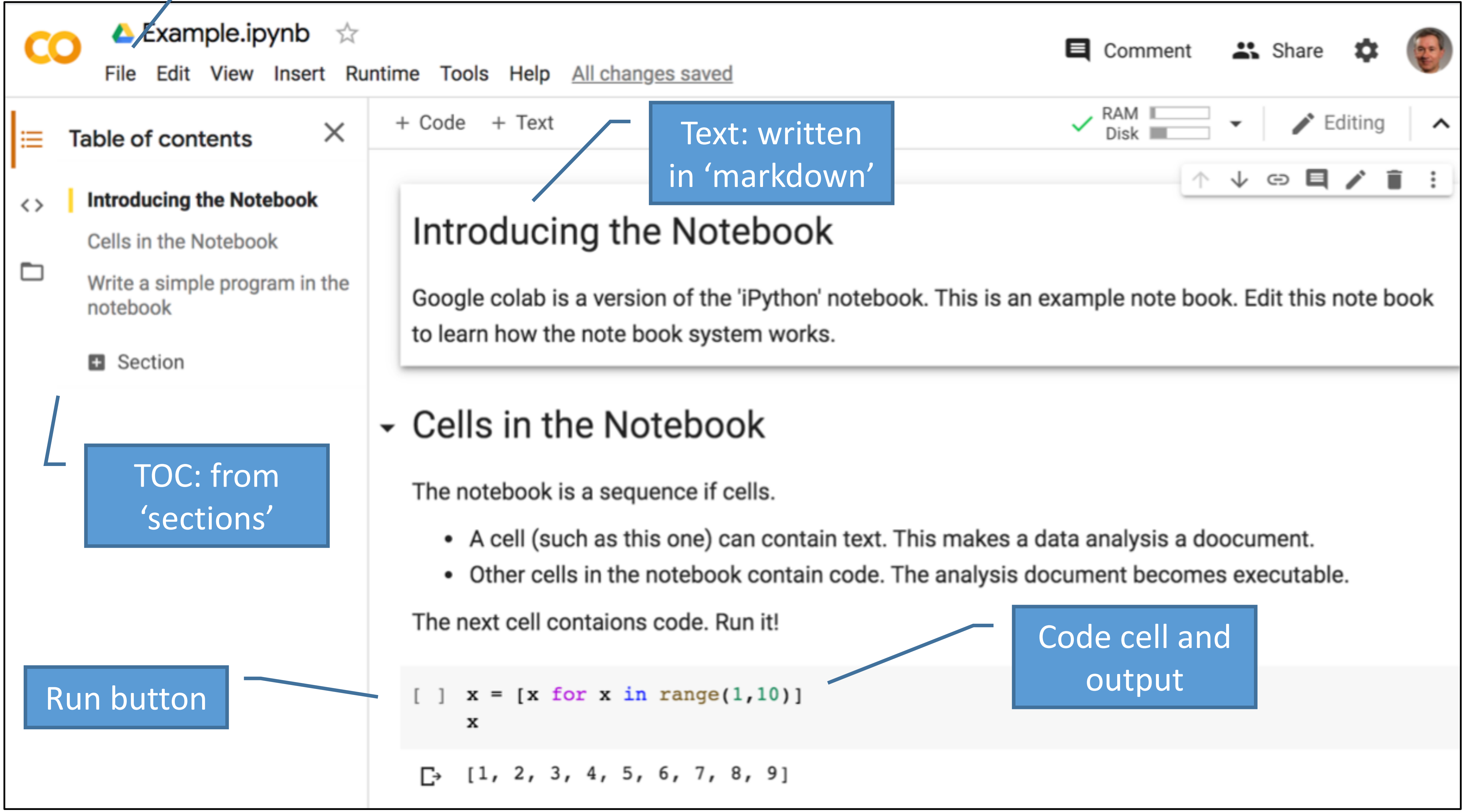
institute of CODING

COMPUTING AT SCHOOL
EDUCATE · ENGAGE · ENCOURAGE
Part of BCS –The Chartered Institute for IT

IPython Notebook

- IPython – interactive Python with graphics (2001)
- Jupyter – web based interface for IPython (2014)
  - Also supports other languages
- Google Colab
  - Hosted support for Jupyter on Google Drive
  - Better interface
- Concept: program as an executable document

New, upload, save



Introducing the Notebook

Google colab is a version of the 'iPython' notebook. This is an example note book. Edit this note book to learn how the note book system works.

Cells in the Notebook

The notebook is a sequence if cells.

- A cell (such as this one) can contain text. This makes a data analysis a doocument.
- Other cells in the notebook contain code. The analysis document becomes executable.

The next cell contaions code. Run it!

```
[ ]  x = [x for x in range(1,10)]
     x
```

```
[1, 2, 3, 4, 5, 6, 7, 8, 9]
```

Text: written in ‘markdown’

TOC: from ‘sections’

Run button

Code cell and output

Practical Break

Please try the ‘example’ notebook

Queen Mary
University of London

institute of
CODING

COMPUTING AT SCHOOL
EDUCATE · ENGAGE · ENCOURAGE
Part of BCS –The Chartered Institute for IT

Part 1: The Dataframe

Introducing the Python Pandas Library

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
0,John,24,Arsenal
1,Mary,27,Spurs
2,Peter,31,Chelsea
```

Loaded from CSV

The Data: Country of Birth

- 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

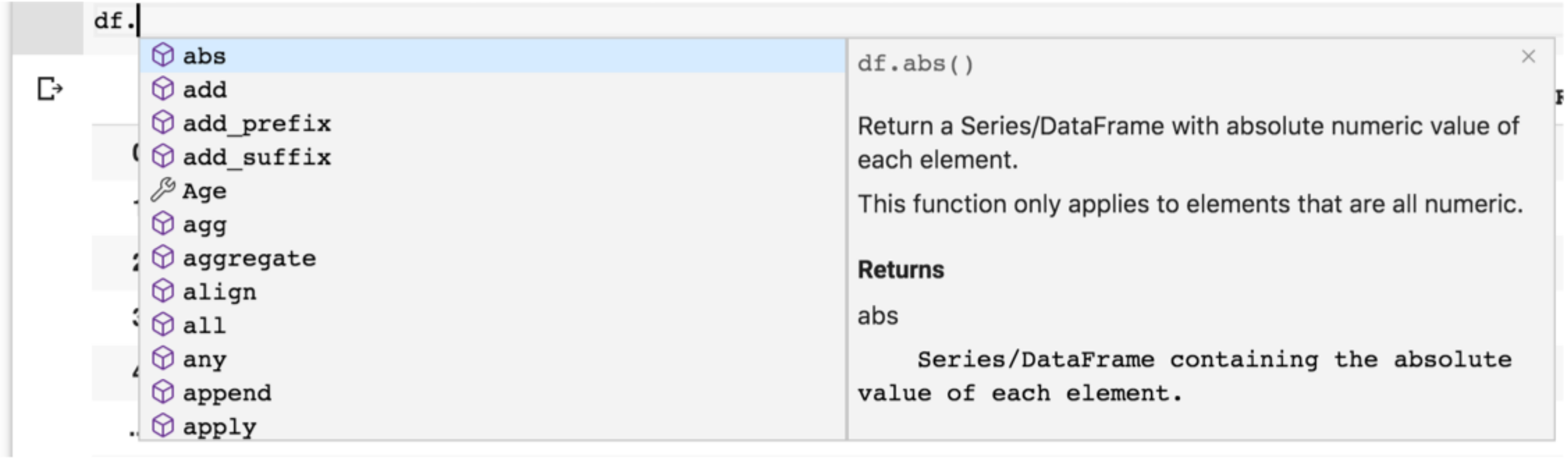
Investigating the Data

- Task 1.1: Load the data to a dataframe
- Task 1.2: 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

Getting Help

- Pop up help



```
df.
```

```
df.abs()
```

Return a Series/DataFrame with absolute numeric value of each element.

This function only applies to elements that are all numeric.

Returns

abs

```
    Series/DataFrame containing the absolute
value of each element.
```

- Pandas documentation - https://pandas.pydata.org/docs/
  - API – many optional arguments
  - User guides

Practical Break

Part 1: ‘Data analysis’ notebook

Part 2: Selecting, Transforming and visualising Data

- Data transformation: Tall to Wide
- The Pivot Table

Queen Mary
University of London

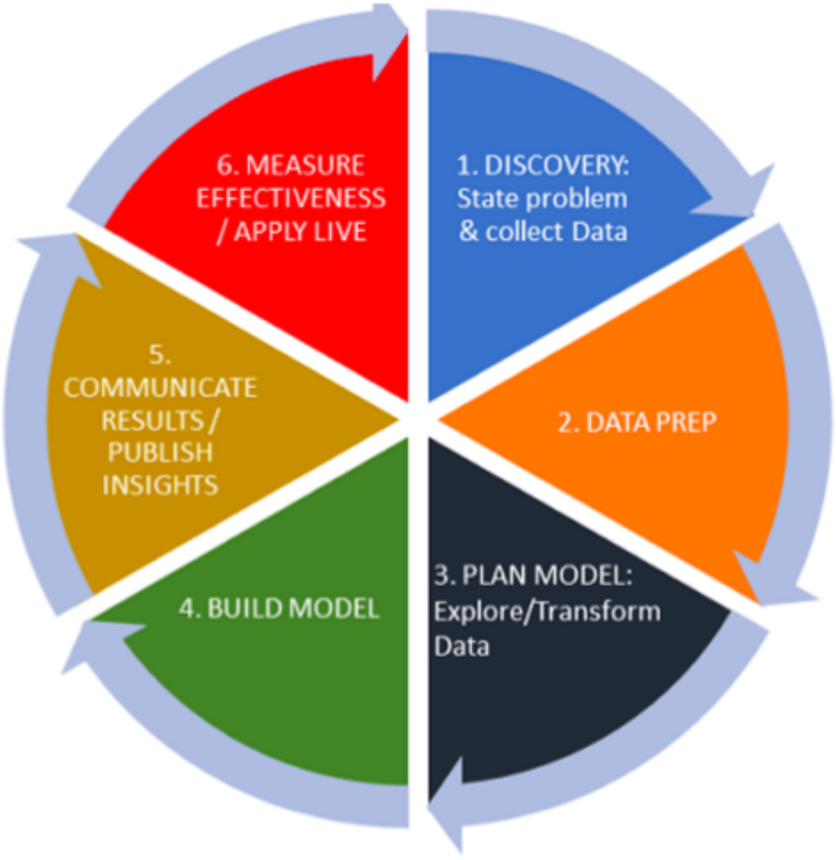
institute of CODING

COMPUTING AT SCHOOL
EDUCATE · ENGAGE · ENCOURAGE
Part of BCS –The Chartered Institute for IT

What Questions Can this Dataset Answer?

- How many people from the Americas live in Sutton?
- Which Borough has the 'most' young people?
- What age are people born outside the UK?
- …

- *What additional information is missing?*


1. DISCOVERY: State problem & collect Data
2. DATA PREP
3. PLAN MODEL: Explore/Transform Data
4. BUILD MODEL
5. COMMUNICATE RESULTS / PUBLISH INSIGHTS
6. MEASURE EFFECTIVENESS / APPLY LIVE

Selecting and Transforming Data

- Selecting some data
  - Data for one borough
  - ... or one age group
  - ... or one country or region
- Use conditions
- *No loops*

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

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

The Pivot Table

- Transform data
- Origin 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

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

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

Aggregate over the person

Practical Break

Part 2: ‘Data analysis’ notebook

Part 3: Adding Columns

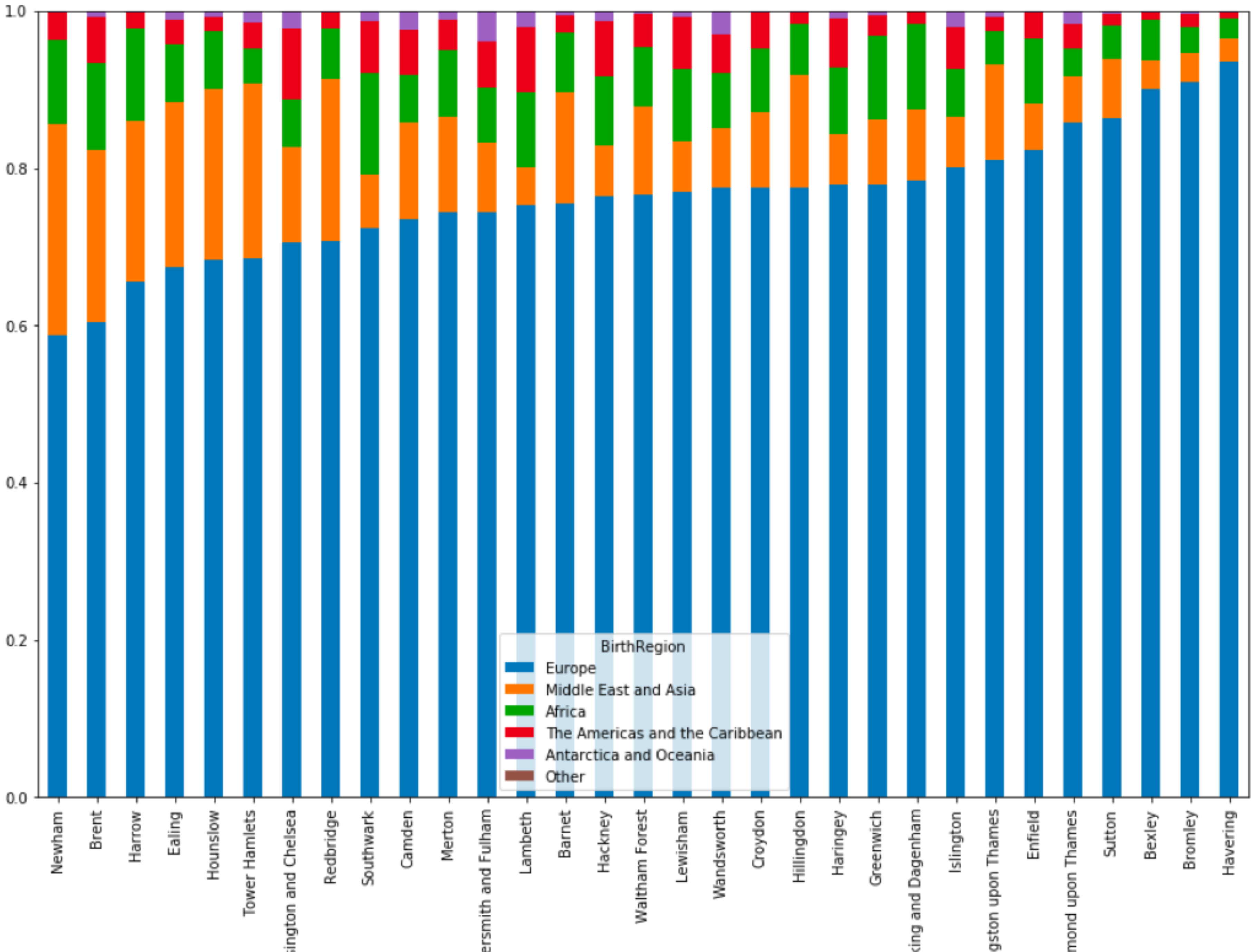
The Problem of Comparing Boroughs

- Boroughs vary in population
- What does this question mean?

Where in London do most people born in Spain live?

- Need to transform the data into ‘proportion’

Proportion of Borough Born in Each Region



Practical Break

Part 3: ‘Data analysis’ notebook

Queen Mary
University of London

institute of
CODING

COMPUTING AT SCHOOL
EDUCATE · ENGAGE · ENCOURAGE
Part of BCS –The Chartered Institute for IT

Summary and Discussion

Summary

- The environment
  - Web-based Interactive Python
  - Does not have to be hosted
- The Pandas library
  - Comprehensive but complex
- Key concepts
  - Dataframe and selecting data
  - Transforming: pivot table
  - Visualizing: plot

Discussion Questions

- What is emphasis of curriculum?
- Should we use a large data file?
- Is the complexity of Pandas manageable?
- Balance of technical versus ‘interpretation’ of data