



CAS London Conference: February 2020

Practical Sheet: Introduction to Data Analysis

1 Getting Started

1.1 Login to Google Colab

Go to <https://colab.research.google.com/> The system is free but you need to create an account, using your google logon.

1.2 Download the Data and Analysis Notebook

Visit <http://www.eecs.qmul.ac.uk/~william/CAS-London-2020.html>

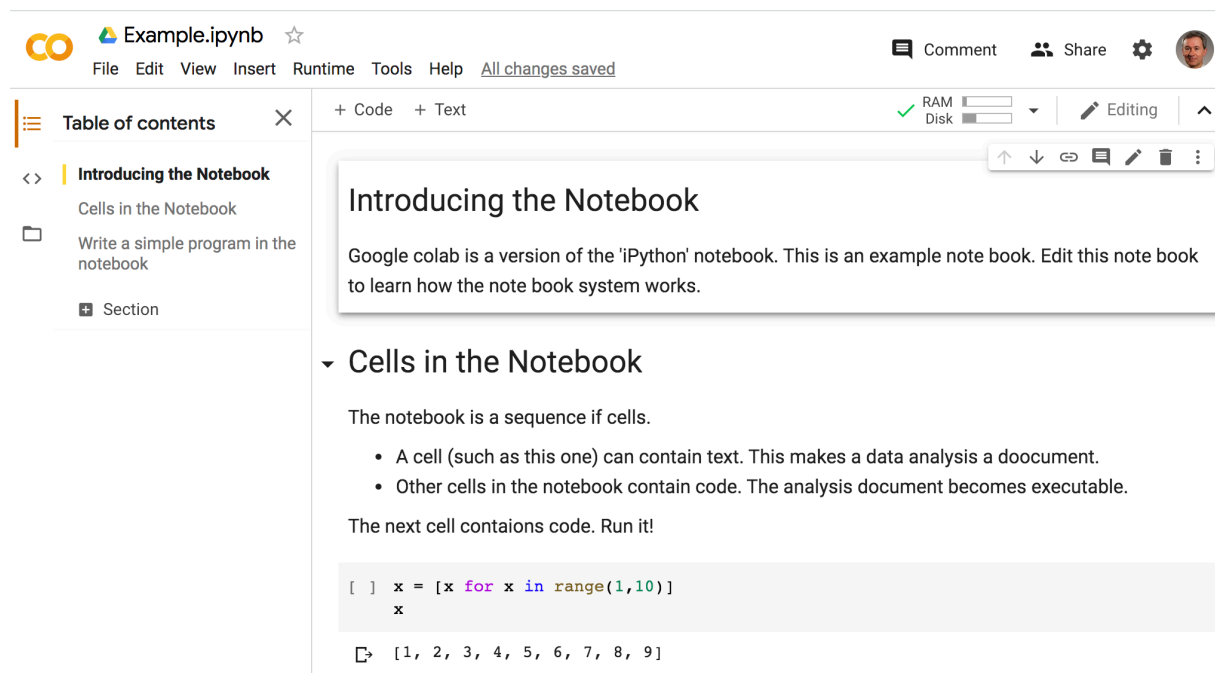
1. Download the notebooks: 'example' and 'CAS-London-data-analysis'
2. Download and unzip the data file.
3. Other resources are available to view if you wish.

2 Using Google Colab Notebook

Google colab is a system that hosts a version of the 'iPython notebook' that is widely used for data analysis. The notebook runs inside a browser and is able to execute Python code and display plots.

2.1 Open the 'Example Notebook'

Create a new notebook. Then use the file 'upload' menu to upload the 'example' notebook. The example should look like this:



The screenshot shows a Google Colab notebook titled 'Example.ipynb'. The interface includes a top menu bar with 'File', 'Edit', 'View', 'Insert', 'Runtime', 'Tools', and 'Help'. A 'Table of contents' sidebar on the left lists sections: 'Introducing the Notebook', 'Cells in the Notebook', 'Write a simple program in the notebook', and 'Section'. The main content area displays the text 'Introducing the Notebook' followed by '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.' Below this is a section titled 'Cells in the Notebook' with the text 'The notebook is a sequence if cells.' and a bulleted list: 'A cell (such as this one) can contain text. This makes a data analysis a document.' and 'Other cells in the notebook contain code. The analysis document becomes executable.' The next cell contains the code 'x = [x for x in range(1,10)]' and the output '[1, 2, 3, 4, 5, 6, 7, 8, 9]'.

2.2 Documentation

There is an introductory video (*probably no time to watch today*):

<https://colab.research.google.com/notebooks/welcome.ipynb>

Colab is based on the Jupyter project (<https://jupyter.org/>) This system can run on your laptop (i.e. without a web connection). The notebooks for both Colab and Jupyter are stored in the same format, although the interface is slightly different. Think of Colab as 'Jupyter hosted on Google Drive'.

Text cells are written in markdown. For more details, see

https://colab.research.google.com/notebooks/markdown_guide.ipynb

3 Introducing Data Analysis

Upload the CAS_london_data_analysis.ipynb notebook into Colab. The instructions are in the notebook. You should:

- Follow the steps in order, running the code. In some cases, you may need to 'uncomment' the provided code.
- Insert additional cells comments and variations of the code. For example, trying the same steps with different data.

3.1 About the Data

The data we are using is about the country of Birth of people in different age groups in London Boroughs.,

- The data is from the 2011 census. The specific data set is <https://data.london.gov.uk/dataset/country-of-birth---population-pyramid-tool> and it is taken from <https://data.london.gov.uk/dataset/census-2011-small-population-tables>
- The data has been transformed into a form that is easier to work within. The provided table is an example of 'tall' (or 'narrow' data – see https://en.wikipedia.org/wiki/Wide_and_narrow_data).
- The data is a part of the London Datastore, a repository of lots of large datasets. See <https://data.london.gov.uk/>

The analysis notebook describes the variables in the data.

3.2 About the Pandas Library

The analysis uses the Pandas library for data: see <https://pandas.pydata.org/> The library is very capable but also very complex! The documentation includes:

- A user guide: https://pandas.pydata.org/docs/user_guide/index.html#user-guide and
- API reference <https://pandas.pydata.org/docs/reference/index.html>