The UK Price Data

Richard Davies, October 2022

www.richarddavies.io

This guide is intended for use by anyone wishing to use the UK price data. If you use the data, please reference my LSE paper, Davies (2021) and most importantly, let me know how your project goes. A list of projects that have used the data can be found at: https://richarddavies.io/research/prices

Contents

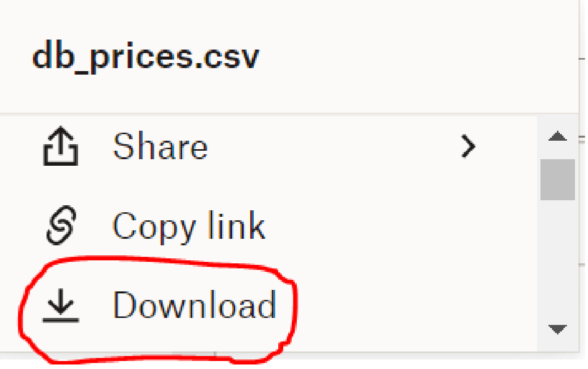
Accessing the data

Data can be found by clicking on the following link:

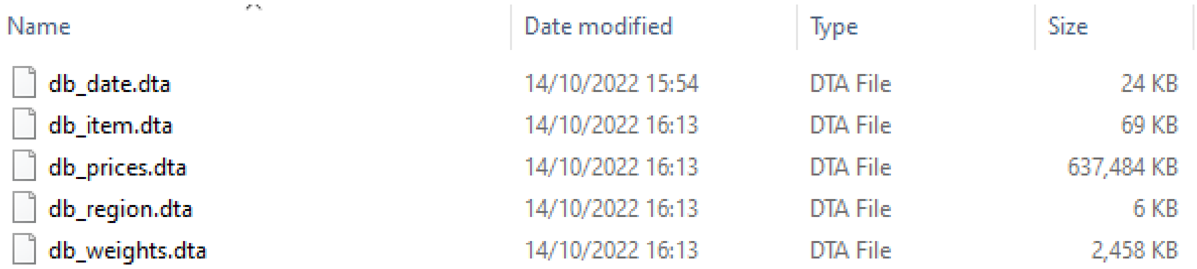
UK Price Data

From the folder you should create copies of the files, stored locally on your machine. This could be in any folder—desktop, documents—that you like.

Important: make copies of the files by downloading them. **Do not** make copies by opening them in Excel, and then saving them locally. The prices file is too large to be handled by Excel, but Excel may not warn you about this. Rather it will open the first 1 million observations, and you will lose the rest. By clicking on the three small dots next to the file names, you should get some options, one of which will be download.



The rest of this guide assumes that you have saved all of the files in a folder, on your desktop, that is called "prices". If you have done this, you should have a set of files that looks like this:



Name	Date modified	Type	Size
db_date.dta	14/10/2022 15:54	DTA File	24 KB
db_item.dta	14/10/2022 16:13	DTA File	69 KB
db_prices.dta	14/10/2022 16:13	DTA File	637,484 KB
db_region.dta	14/10/2022 16:13	DTA File	6 KB
db_weights.dta	14/10/2022 16:13	DTA File	2,458 KB

Description of databases

The data are stored as a set of relational databases. The reason for this is that there are over 40 million prices. Rather than store repetitious detail of what these prices are—"Loaf of White Bread"—for example, we use numeric codes. This is to keep the size of the files small and manageable.

The content of the databases is set out below.

Table 1 UK price data – description of databases

File	*Variables*	*Description*
db_date	date	Date in numeric format, running from 1 to 415.
	quote_date	YYYYMM format, from 198802 to 202208
	dateStata	Date stored in Stata format.[1]
	dateISO	Date stored in ISO format, yyyy-mm-dd
	obs	The number of price observations at that date
	year, month, quarter	The year, month, quarter
db_item	item_id	ONS item id. A 6-digit number.
	description	Detailed description of the item, a string variable.
	date_quote_s	Date of first price
	date_quote_e	Date of last price
	n_obs	Number of prices recorded for this item
db_prices	quote_date	*As above*
	shop_code	A unique shop identifier
	region	The region the price recorded in. Numeric. From 1 to 13.
	price	The price quote.
	item_id	*As above*
db_region	region	*As above*
	regionDesc	Description of the region
	country	England, Wales, NI or Scotland
	obs	The number of observations for that region
db_weights	quote_date	*As above*
	item_id	*As above*
	coicop_weight	Weight of this item in consumption.

[1] See, for example: https://www.stata.com/manuals13/u24.pdf

Making your first chart

Suppose that you would like to make a chart of the price of an item—bread for example—over time. This section is a guide for how to do this. Stata code is in **blue**

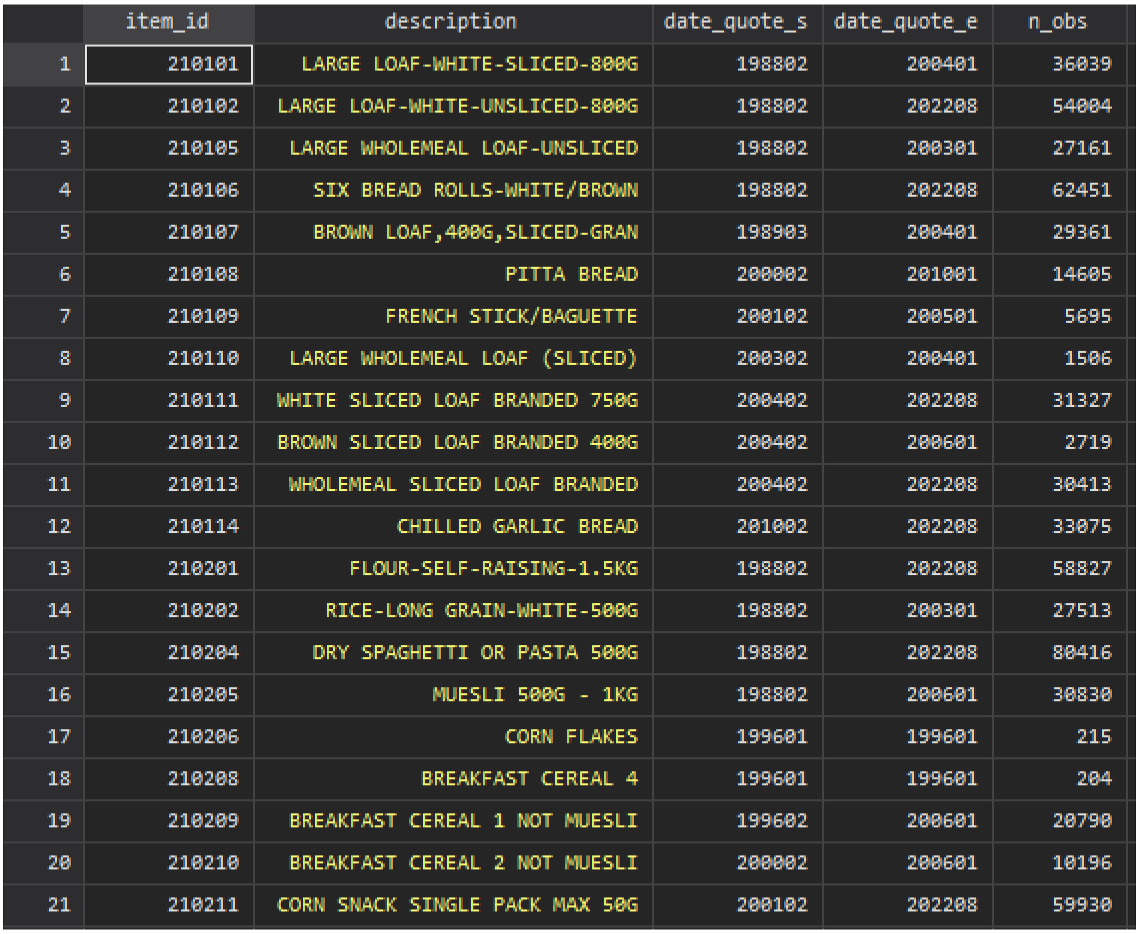
First, open the db_item.dta file.

```
use "C:\Users\{username}\desktop\prices\db_item.dta"
```

Then type “browse” to look at the data

```
browse
```

You should see something like this.



	item_id	description	date_quote_s	date_quote_e	n_obs
1	210101	LARGE LOAF-WHITE-SLICED-800G	198802	200401	36039
2	210102	LARGE LOAF-WHITE-UNSLICED-800G	198802	202208	54004
3	210105	LARGE WHOLEMEAL LOAF-UNSLICED	198802	200301	27161
4	210106	SIX BREAD ROLLS-WHITE/BROWN	198802	202208	62451
5	210107	BROWN LOAF,400G,SLICED-GRAN	198903	200401	29361
6	210108	PITTA BREAD	200002	201001	14605
7	210109	FRENCH STICK/BAGUETTE	200102	200501	5695
8	210110	LARGE WHOLEMEAL LOAF (SLICED)	200302	200401	1506
9	210111	WHITE SLICED LOAF BRANDED 750G	200402	202208	31327
10	210112	BROWN SLICED LOAF BRANDED 400G	200402	200601	2719
11	210113	WHOLEMEAL SLICED LOAF BRANDED	200402	202208	30413
12	210114	CHILLED GARLIC BREAD	201002	202208	33075
13	210201	FLOUR-SELF-RAISING-1.5KG	198802	202208	58827
14	210202	RICE-LONG GRAIN-WHITE-500G	198802	200301	27513
15	210204	DRY SPAGHETTI OR PASTA 500G	198802	202208	80416
16	210205	MUESLI 500G - 1KG	198802	200601	30830
17	210206	CORN FLAKES	199601	199601	215
18	210208	BREAKFAST CEREAL 4	199601	199601	204
19	210209	BREAKFAST CEREAL 1 NOT MUESLI	199602	200601	20790
20	210210	BREAKFAST CEREAL 2 NOT MUESLI	200002	200601	10196
21	210211	CORN SNACK SINGLE PACK MAX 50G	200102	202208	59930

There are clearly lots of bread items that you could look at. In class you will discuss how to make an index out of many items. But for this first chart, you want to simply look at an indicative item. The first one is not ideal, since quotes end in 2004 (look at the date_quote_e variable which has value 200401, telling us that quotes end in January 2004). The second item in the list “LARGE LOAF-WHITE-UNSLICED-800G” is better since quotes are available until just last month.

So, we decide to look at this item, and take a note of its number: **210102**

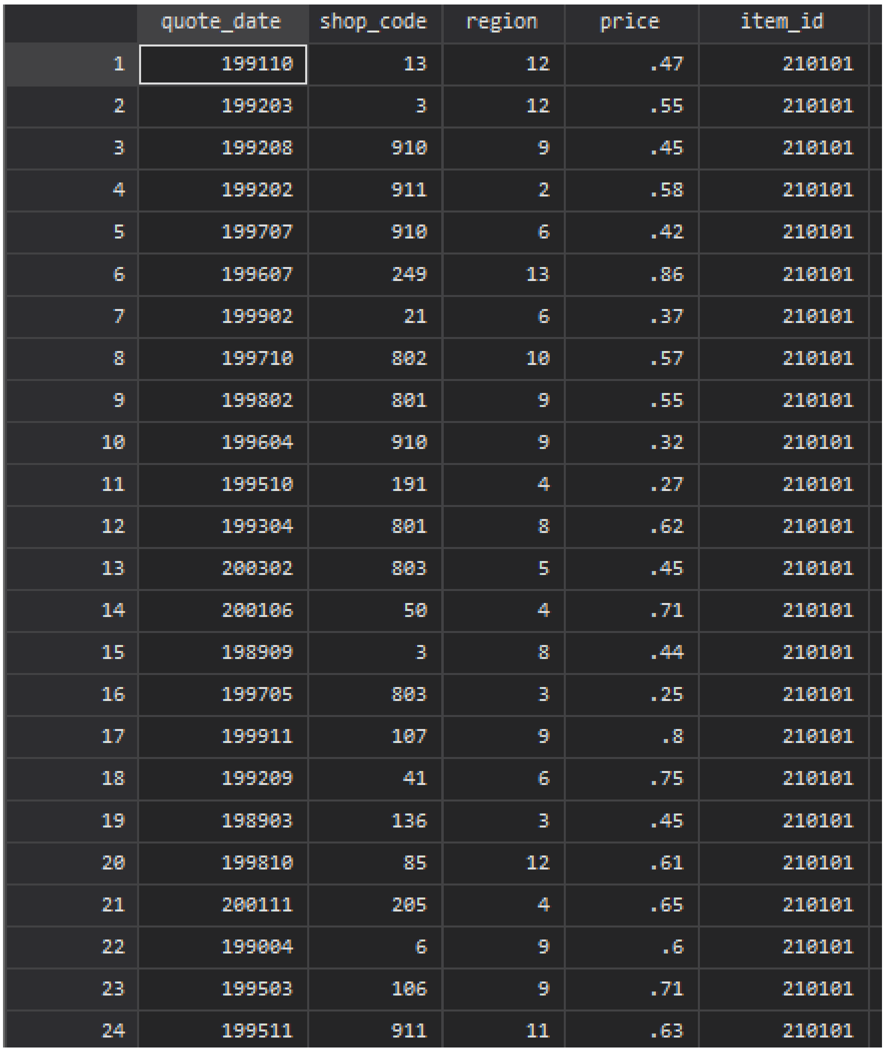
Now close your browse window, and clear the dataset.

```
clear
```

Next open the prices dataset.

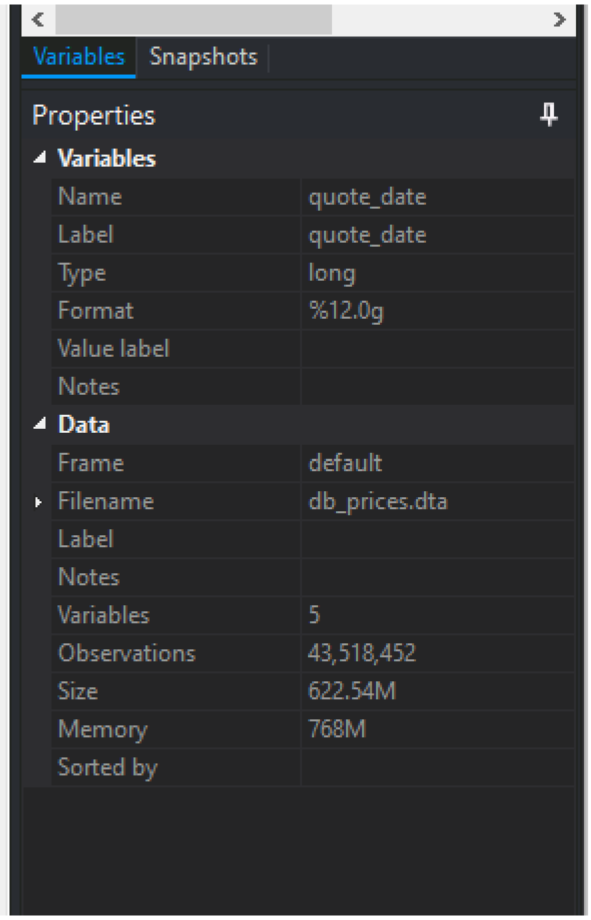
Again type browse to look at the data:

browse



	quote_date	shop_code	region	price	item_id
1	199110	13	12	.47	210101
2	199203	3	12	.55	210101
3	199208	910	9	.45	210101
4	199202	911	2	.58	210101
5	199707	910	6	.42	210101
6	199607	249	13	.86	210101
7	199902	21	6	.37	210101
8	199710	802	10	.57	210101
9	199802	801	9	.55	210101
10	199604	910	9	.32	210101
11	199510	191	4	.27	210101
12	199304	801	8	.62	210101
13	200302	803	5	.45	210101
14	200106	50	4	.71	210101
15	198909	3	8	.44	210101
16	199705	803	3	.25	210101
17	199911	107	9	.8	210101
18	199209	41	6	.75	210101
19	198903	136	3	.45	210101
20	199810	85	12	.61	210101
21	200111	205	4	.65	210101
22	199004	6	9	.6	210101
23	199503	106	9	.71	210101
24	199511	911	11	.63	210101

At this stage you should also check that you have copied the dataset properly. On the bottom right of your screen you should see a panel like the one below:



Properties

Variables	
Name	quote_date
Label	quote_date
Type	long
Format	%12.0g
Value label	
Notes	
Data	
Frame	default
Filename	db_prices.dta
Label	
Notes	
Variables	5
Observations	43,518,452
Size	622.54M
Memory	768M
Sorted by	

Note that there are over 43 million observations. The number will change (grow) as I add new data. But the key point is that if you have around 1 million observations then you have truncated the data by opening it in Excel at some point. You will need to re-download the data as a csv, and not open it in Excel.

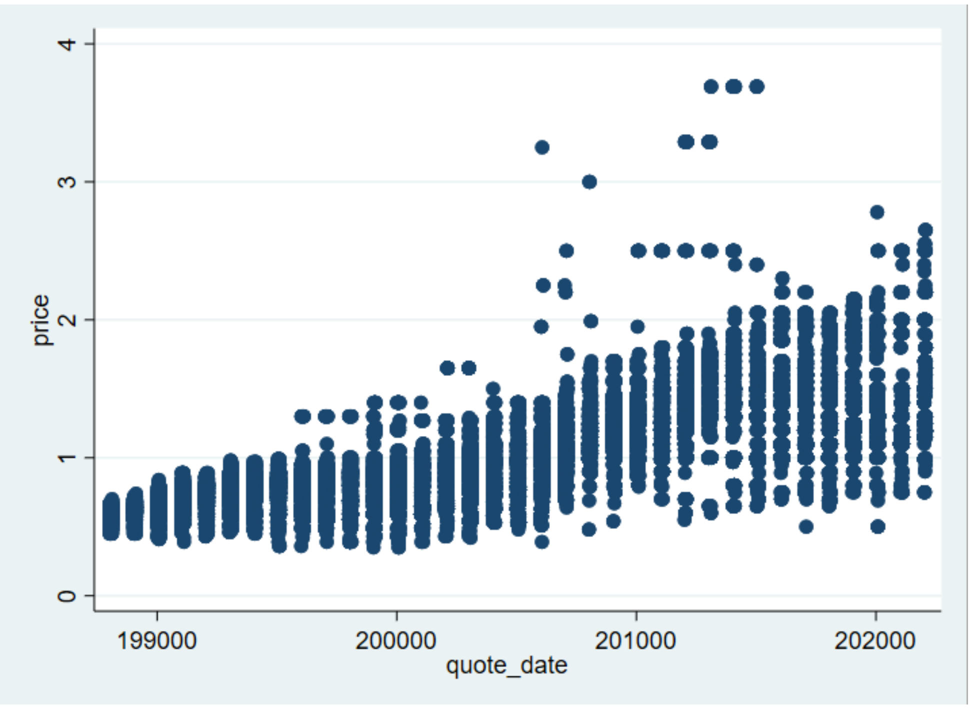
Once we are happy we have all the data, we are going to simplify (and shrink) our data by just looking at the dataset we are interested in.

```
keep if item_id==210102
```

You should now have a much smaller dataset, in this case around 54k observations.

In order to take a look at the data via a plot, we are going to draw a scatter chart, giving each observation a dot on the page.

```
scatter price quote_date
```



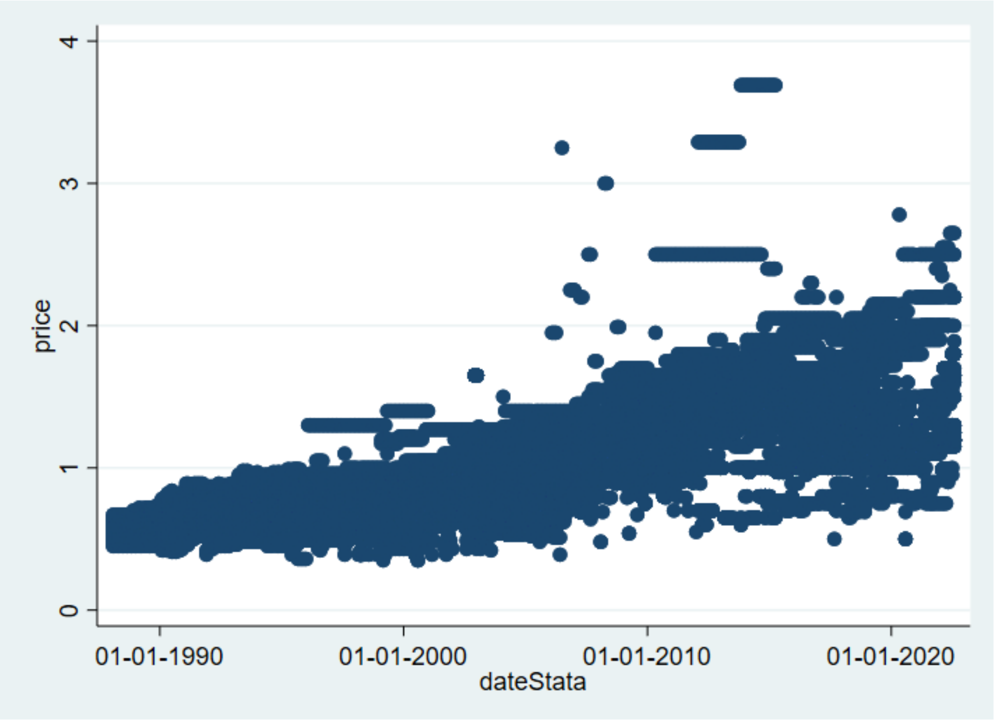
This chart is interesting, as it gives us an idea of the distribution of prices over time. But the x (time) axis is not plotted properly, since the dates used are in the form YYYYMM. In order to plot properly we need a better form of date variable. These are given to you in the file **db_date.dta.** The next step is therefore to merge the two datasets:

merge m:1 quote_date using "C:\Users\hi19329\ Desktop\prices\db_date.dta"

Merging datasets is a key part of the challenge in modern economics and data science, and will be discussed in class.

With the date information in place, we can draw a new chart:

scatter price quote_date



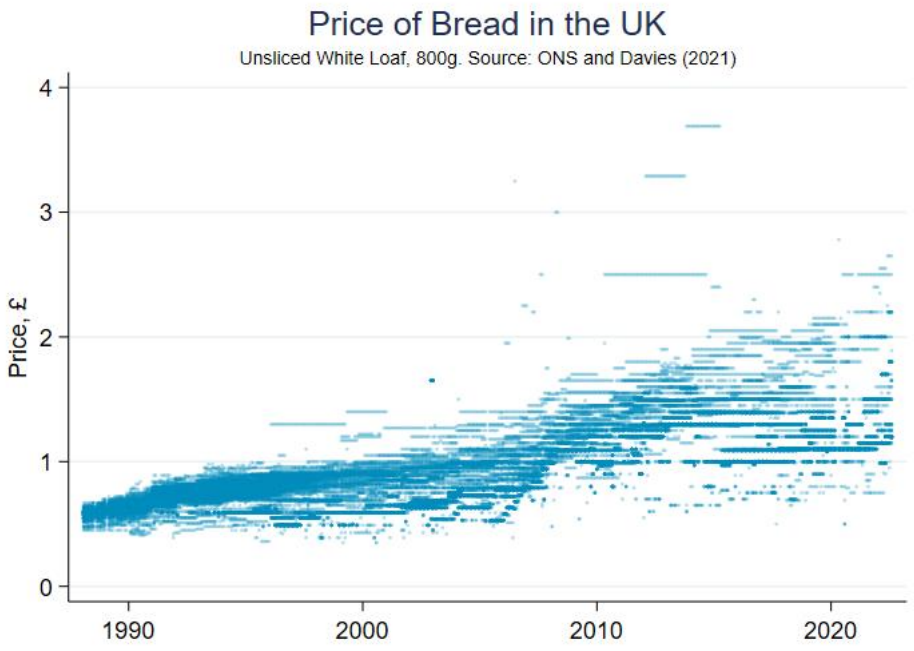
This chart is better, since there are no gaps, the dates are clearly being plotted properly.

Formatting your first chart

The remaining issues are formatting. Problems/improvements with the previous chart include:

- Labelling of x and y axis
- Size and opacity of markers. When you have lots of marker, make them opaque, as this gives the sense of the weight of the data (dots get darker when plotted on top of one another).
- Labelling of dates.
- Angle of the y-axis numbers.
- Stata-Blue background.

After improving the formatting, we end up with a chart like this:



The formatting steps that are used to produce this chart are included in a do file that you can run through step by step. This is here.

REFERENCES

Prices data

Davies, R (2021a), “Prices and inflation in the UK - A new dataset”, Centre for Economic Performance, Occasional Paper 55, February.
https://cep.lse.ac.uk/_NEW/publications/abstract.asp?index=7726

Davies, R (2021b), “Prices and inflation in a Pandemic – A Micro Data Approach”, Centre for Economic Performance, Covid-19 Analysis Series. February.
https://cep.lse.ac.uk/pubs/download/cepcovid-19-017.pdf

Playfair Prize Website, ‘Heroes’ section:

- https://www.playfairprize.com/william-playfair
- https://www.playfairprize.com/william-petty
- https://www.playfairprize.com/florence-nightingale
- https://www.playfairprize.com/edward-tufte