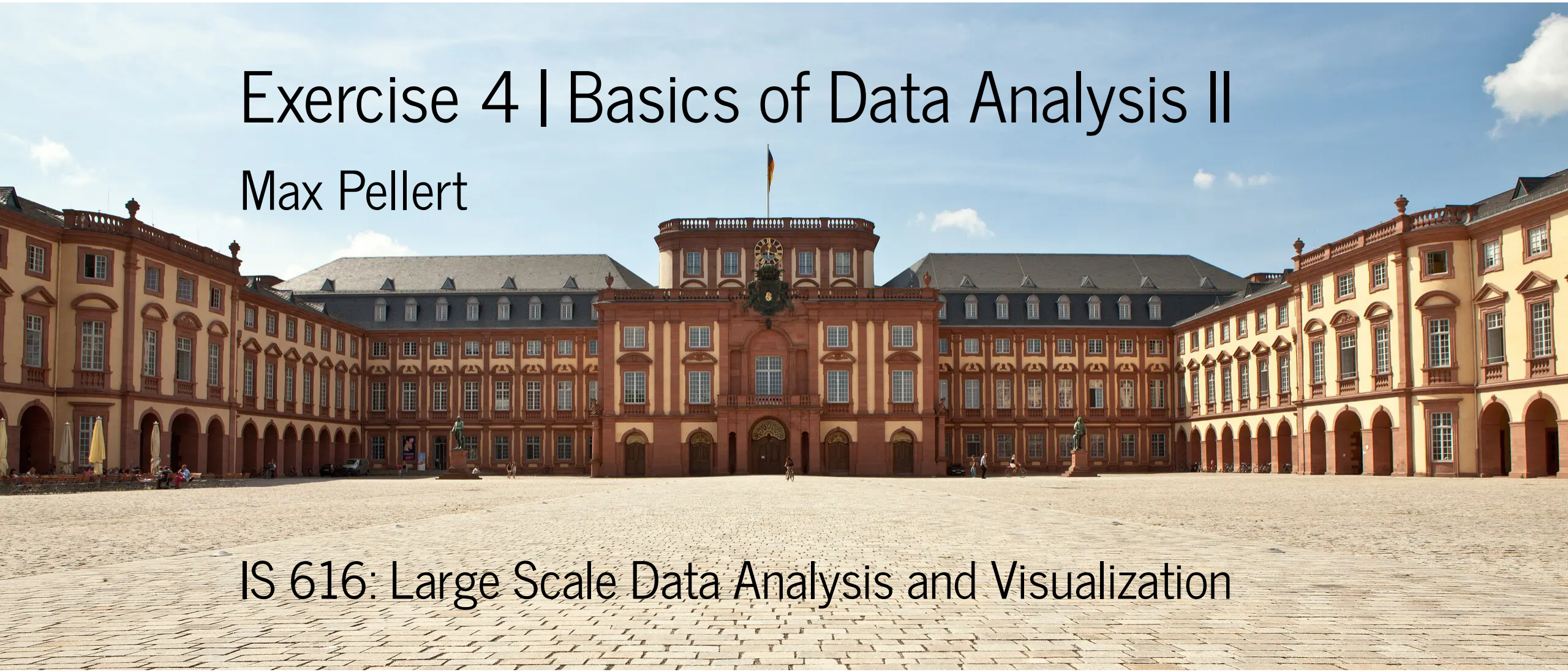
Exercise 4 | Basics of Data Analysis II

Max Pellert



IS 616: Large Scale Data Analysis and Visualization

Spreadsheets?

Excel?

Excel spreadsheet error blamed for UK's 16,000 missing coronavirus cases / The case went missing after the spreadsheet hit its filesize limit

Excel: Why using Microsoft's tool caused Covid-19 results to be lost

Does Contact Tracing Work? Quasi-Experimental Evidence from an Excel Error in England

521/2020 Thiemo Fetzer and Thomas Graeber

Working Paper

Research Theme: Political Economy, Coronavirus

Contact tracing has been a central pillar of the public health response to the COVID-19 pandemic. Yet, contact tracing measures face substantive challenges in practice and well-identified evidence about their effectiveness re-mains scarce. This paper exploits quasi-random variation in COVID-19 con-tact tracing. Between September 25 and October 2, 2020, a total of 15,841COVID-19 cases in England (around 15 to 20% of all cases) were not immediately referred to the contact tracing system due to a data processing error. Case information was truncated from an Excel spreadsheet after the row limit had been reached, which was discovered on October 3. There is substantial variation in the degree to which different parts of England areas were exposed– by chance – to delayed referrals of COVID-19 cases to to the contact tracing system. We show that more affected areas subsequently experienced a drastic rise in new COVID-19 infections and deaths alongside an increase in the positivity rate and the number of test performed, as well as a decline in the performance of the contact tracing system. Conservative estimates suggest that the failure of timely contact tracing due to the data glitch is associated with more than 125,000 additional infections and over 1,500 additional COVID-19-related deaths. Our finding provide strong quasi-experimental evidence for the effectiveness of contact tracing.

Sources

https://www.theverge.com/2020/10/5/21502141/uk-missing-coronavirus-cases-excel-spreadsheet-error

https://www.bbc.com/news/technology-54423988

https://warwick.ac.uk/fac/soc/economics/research/centres/cage/publications/workingpapers/2020/does_contact_tracing_work_quasi_experimental_evidence_from_an_excel_error_in_england/

and many more…

Scientists rename human genes to stop Microsoft Excel from misreading them as dates / Sometimes it's easier to rewrite genetics than update Excel

There are tens of thousands of genes in the human genome: minuscule twists of DNA and RNA that combine to express all of the traits and characteristics that make each of us unique. Each gene is given a name and alphanumeric code, known as a symbol, which scientists use to coordinate research. But over the past year or so, some 27 human genes have been renamed, all because Microsoft Excel kept misreading their symbols as dates.

The problem isn't as unexpected as it first sounds. Excel is a behemoth in the spreadsheet world and is regularly used by scientists to track their work and even conduct clinical trials. But its default settings were designed with more mundane applications in mind, so when a user inputs a gene's alphanumeric symbol into a spreadsheet, like MARCH1 — short for "Membrane Associated Ring-CH-Type Finger 1" — Excel converts that into a date: 1-Mar.

Studies found a fifth of genetic data in papers was affected by Excel errors

This is extremely frustrating, even dangerous, corrupting data that scientists have to sort through by hand to restore. It's also surprisingly widespread and affects even peer-reviewed scientific work. One study from 2016 examined genetic data shared alongside 3,597 published papers and found that roughly one-fifth had been affected by Excel errors.

"It's really, really annoying," Dezső Módos, a systems biologist at the Quadram Institute in the UK, told *The Verge*. Módos, whose job involves analyzing freshly sequenced genetic data, says Excel errors happen all the time, simply because the software is often the first thing to hand when scientists process numerical data. "It's a widespread tool and if you are a bit computationally illiterate you will use it," he says. "During my PhD studies I did as well!"

There's no easy fix, either. Excel doesn't offer the option to turn off this auto-formatting, and the only way to avoid it is to change the data type for individual columns. Even then, a scientist might fix their data but export it as a CSV file without saving the formatting. Or, another scientist might load the data without the correct formatting, changing gene symbols back into dates. The end result is that while knowledgeable Excel users can avoid this problem, it's easy for mistakes to be introduced.

Help has arrived, though, in the form of the scientific body in charge of standardizing the names of genes, the HUGO Gene Nomenclature Committee, or HGNC. This week, the HGNC published new guidelines for gene naming, including for "symbols that affect data handling and retrieval." From now on, they say, human genes and the proteins they expressed will be named with one eye on Excel's auto-formatting. That means the symbol MARCH1 has now become MARCHF1, while SEPT1 has become SEPTIN1, and so on. A record of old symbols and names will be stored by HGNC to avoid confusion in the future.

Sources

https://www.theverge.com/2020/8/6/21355674/human-genes-rename-microsoft-excel-misreading-dates

Ziemann, M., Eren, Y., & El-Osta, A. (2016). Gene name errors are widespread in the scientific literature. Genome Biology, 17(1), 177. https://doi.org/10.1186/s13059-016-1044-7

https://www.nature.com/articles/d41586-021-02211-4

NEWS | 13 August 2021 | Correction 25 August 2021

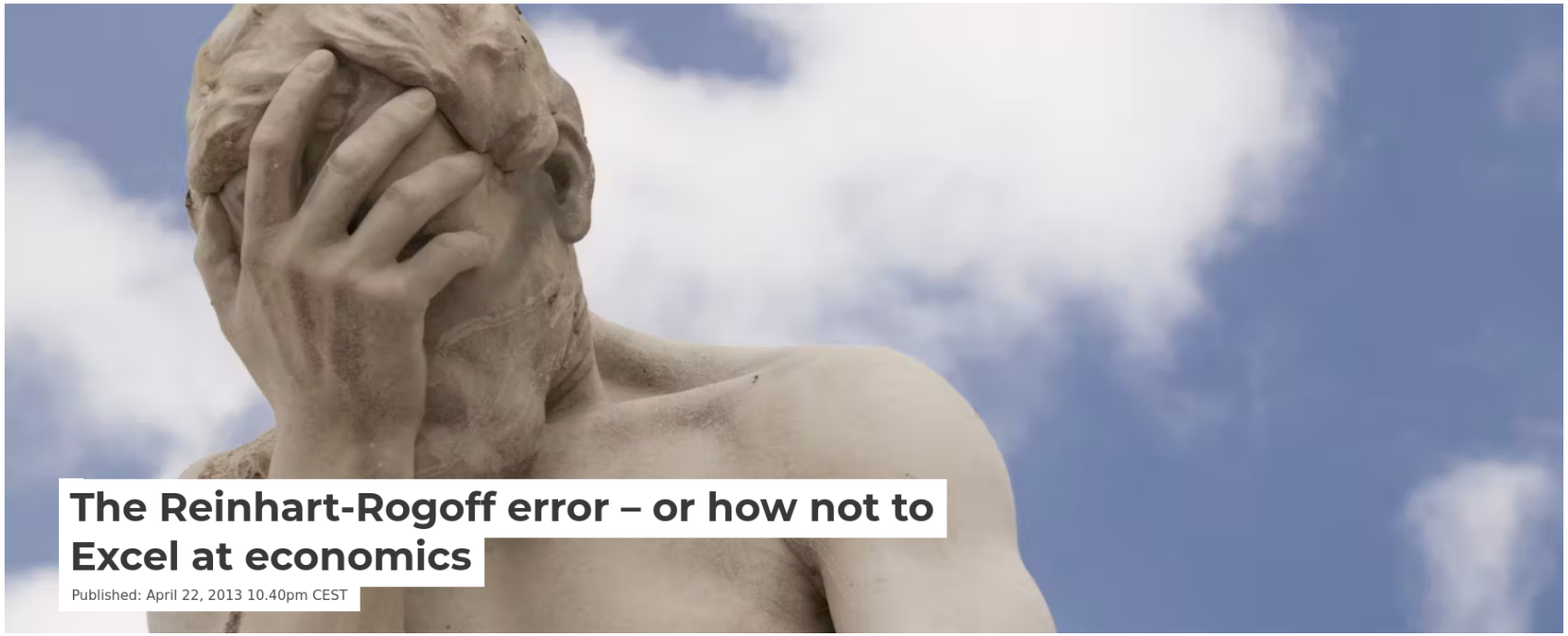
Autocorrect errors in Excel still creating genomics headache

Despite geneticists being warned about spreadsheet problems, 30% of published papers contain mangled gene names in supplementary data.

Avoid or adapt

One solution is to avoid using spreadsheets, he suggests. Although some — such as the open-source programs LibreOffice and Gnumeric — don't have the problem, spreadsheets are hard to audit. "If there's a problem, it's not readily apparent where the problem happened," because there's no record of what steps the software took, he says.

Some computational biologists use scripted computer languages, such as Python and R. These don't autocorrect gene symbols, says Ziemann, and researchers can trace the source of errors. However, they require users to learn the computer language so that they can write code to analyse data.



https://theconversation.com/the-reinhart-rogoff-error-or-how-not-to-excel-at-economics-13646

Reinhart and Rogoff's work showed average real economic growth slows (a 0.1% decline) when a country's debt rises to more than 90% of gross domestic product (GDP) – and this 90% figure was employed repeatedly in political arguments over high-profile austerity measures.

During their analysis, Herndon, Ash and Pollin obtained the actual spreadsheet that Reinhart and Rogoff used for their calculations; and after analysing this data, they identified three errors.

The most serious was that, in their Excel spreadsheet, Reinhart and Rogoff had not selected the entire row when averaging growth figures: they omitted data from Australia, Austria, Belgium, Canada and Denmark.

In other words, they had accidentally only included 15 of the 20 countries under analysis in their key calculation.

When that error was corrected, the "0.1% decline" data became a 2.2% average increase in economic growth.

Draw your own conclusions

In the end, you are responsible for the results that you communicate at the end of a research project, in a publication, in a report written as employee of a company, …

Spreadsheets can be a useful tool for small tasks such as for example collecting expenses or annotating small data

But even for that you may find better alternatives

Be aware of the signalling power of Excel

Some example code on different ways to store and load data

Human readable

```r
library(data.table)
dt <- fread("filename.csv")

# typical parameters: sep="auto", quote="\"",
# nrows=Inf, header="auto", ...
```

```python
import pandas as pd
df = pd.read_csv("filename.csv")

# typical parameters: delimiter=None, header='infer',
# names=_NoDefault.no_default, index_col=None, usecols=None, ...
```

Binary formats

```
library("arrow")
write_parquet(df, "df.parquet")
df <- read_parquet("df.parquet")
```

https://www.r-bloggers.com/2021/09/understanding-the-parquet-file-format/

```
import numpy as np
import pandas as pd
import pyarrow as pa

df = pd.DataFrame({'one': [-1, np.nan, 2.5],
                   'two': ['foo', 'bar', 'baz'],
                   'three': [True, False, True]},
                   index=list('abc'))

table = pa.Table.from_pandas(df)
```

```
import pyarrow.parquet as pq

pq.write_table(table, 'example.parquet')

table2 = pq.read_table('example.parquet')

table2.to_pandas()
```

Replicating Anscombes Quartett

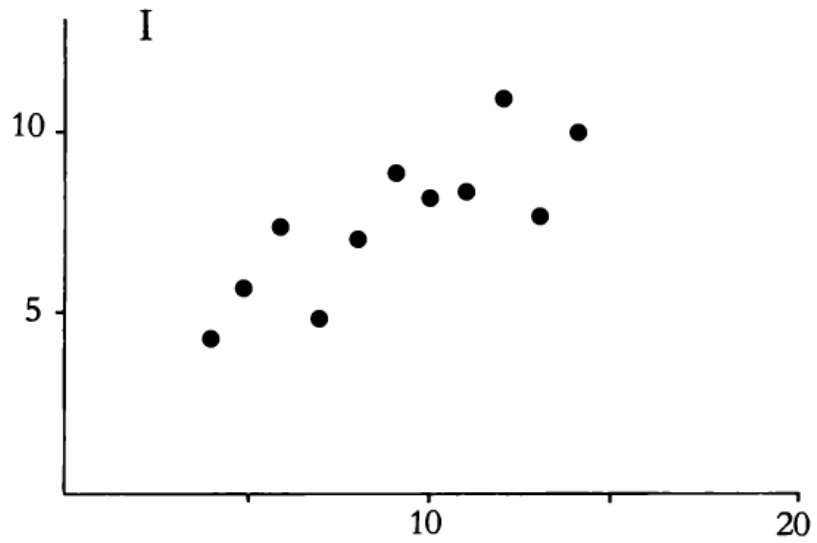
```
library(datasets)
datasets::anscombe
```

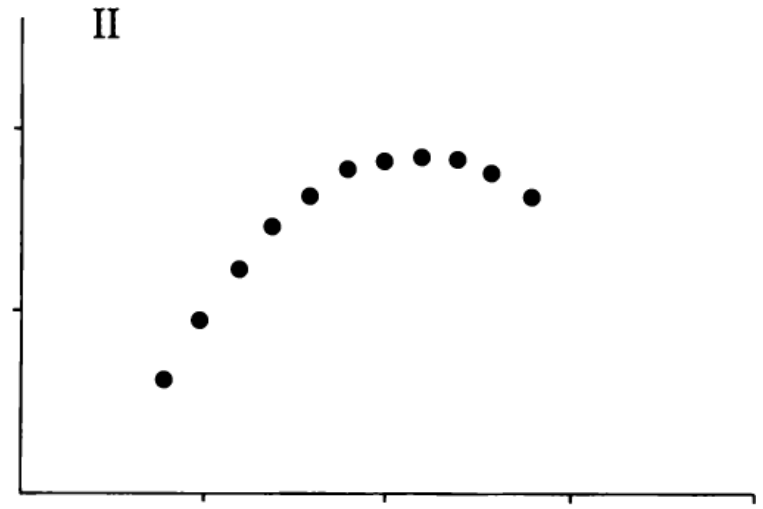
```
##    x1 x2 x3 x4    y1   y2    y3    y4
## 1  10 10 10  8  8.04 9.14  7.46  6.58
## 2   8  8  8  8  6.95 8.14  6.77  5.76
## 3  13 13 13  8  7.58 8.74 12.74  7.71
## 4   9  9  9  8  8.81 8.77  7.11  8.84
## 5  11 11 11  8  8.33 9.26  7.81  8.47
## 6  14 14 14  8  9.96 8.10  8.84  7.04
## 7   6  6  6  8  7.24 6.13  6.08  5.25
## 8   4  4  4 19  4.26 3.10  5.39 12.50
## 9  12 12 12  8 10.84 9.13  8.15  5.56
## 10  7  7  7  8  4.82 7.26  6.42  7.91
## 11  5  5  5  8  5.68 4.74  5.73  6.89
```

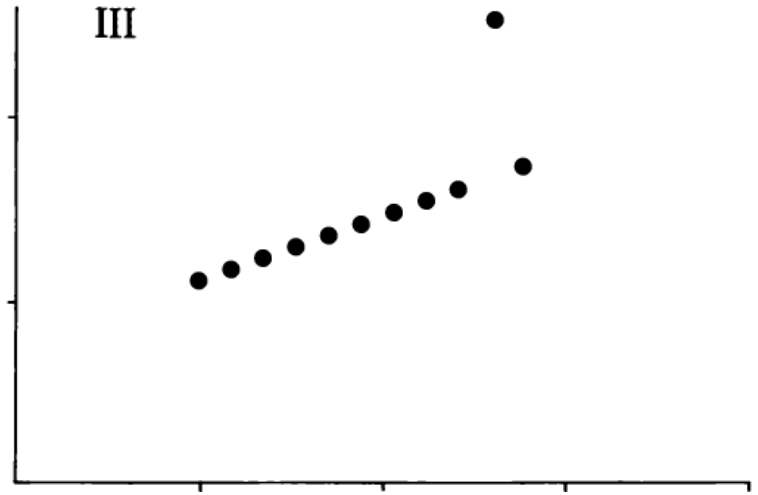
Replicating Anscombes Quartett

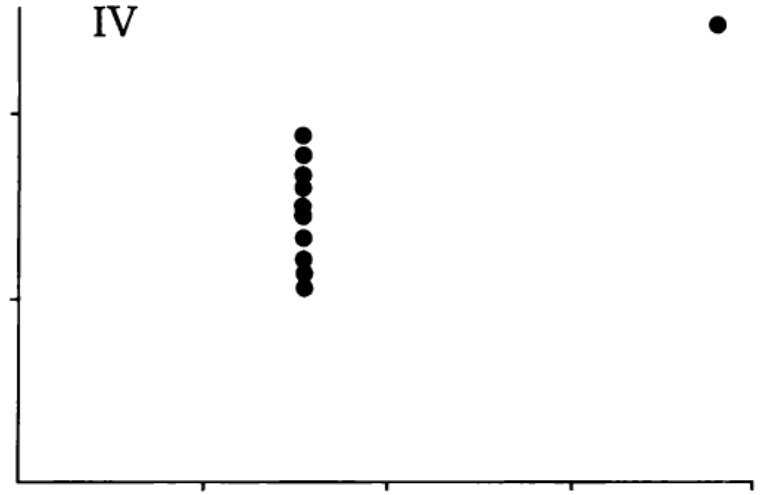
```
x = [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
y1 = [8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68]
y2 = [9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74]
y3 = [7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73]
x4 = [8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8]
y4 = [6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89]

datasets = {
    'I': (x, y1),
    'II': (x, y2),
    'III': (x, y3),
    'IV': (x4, y4)
}
```


I
10
5
10
20


II


III


IV