Kanton Zürich
Direktion der Justiz und des Innern
Statistisches Amt

# Read and process unstructured data in R

## MeetUp Zurich R User Group

November 7, 2017

Max Grütter

# Motivation

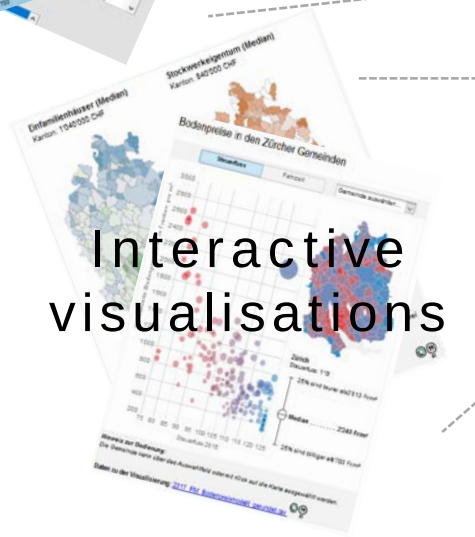
Sometimes one needs data which is only available as really raw data, e.g. as pdf files. To import them to your data environment you either can:

- type the information from the documents
- try to copy/paste them to a text- or Excel-file
- do nothing

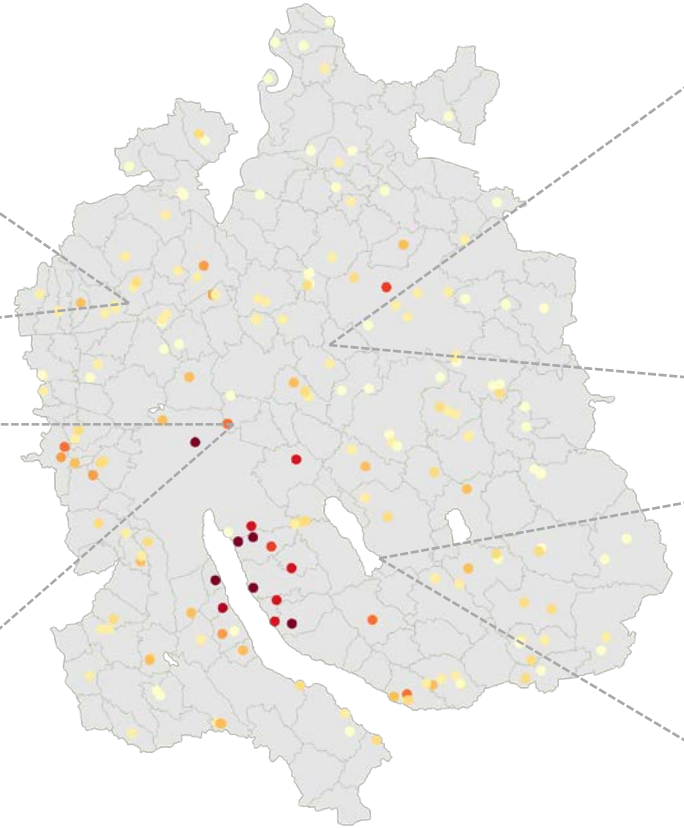
or you use one line of code with R

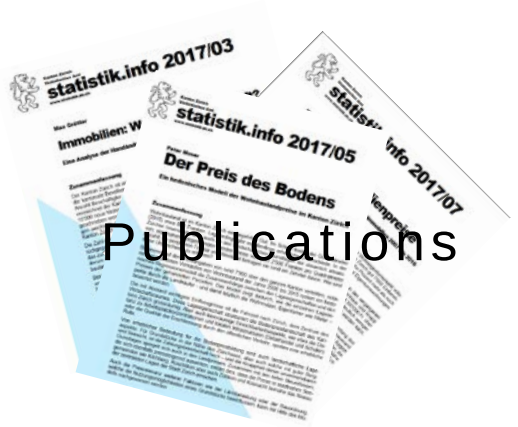
# The Data for real estate trade in Zurich


Online-tools




Interactive visualisations






Publications


youtube.com La Linea

Statistics ZH keeps the official statistics on real estate sales

# Our journey today

Read each document

20'000 documents per year

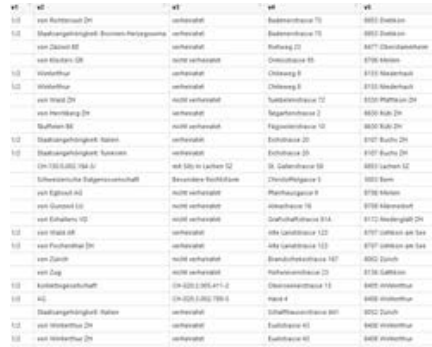
The whole file in one string

Write a data.frame

Extract & reshape the core information

| | v1 | v2 | v3 | v4 | v5 |
|---|---|---|---|---|---|
| 1 | 1/2 | von Zürich | verheiratet | Talstrasse 12 | 8125 Zollikerberg |

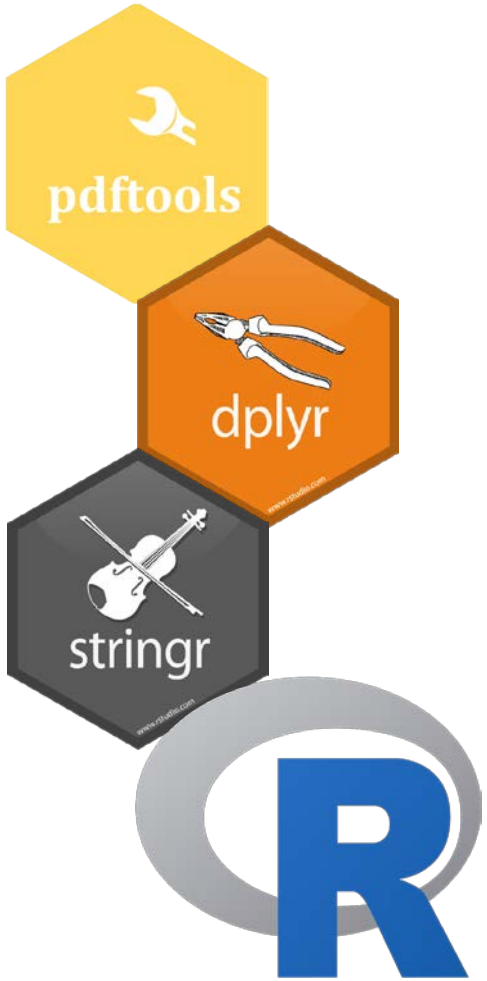
Append all files



The remaining steps

- Clean the data
- Enriche the data
- analyse the data

# The main packages




For reading the pdf-files

For data transformation:

- Manipulate variables and cases
- Combine tables

... and for piping!

For working with strings:

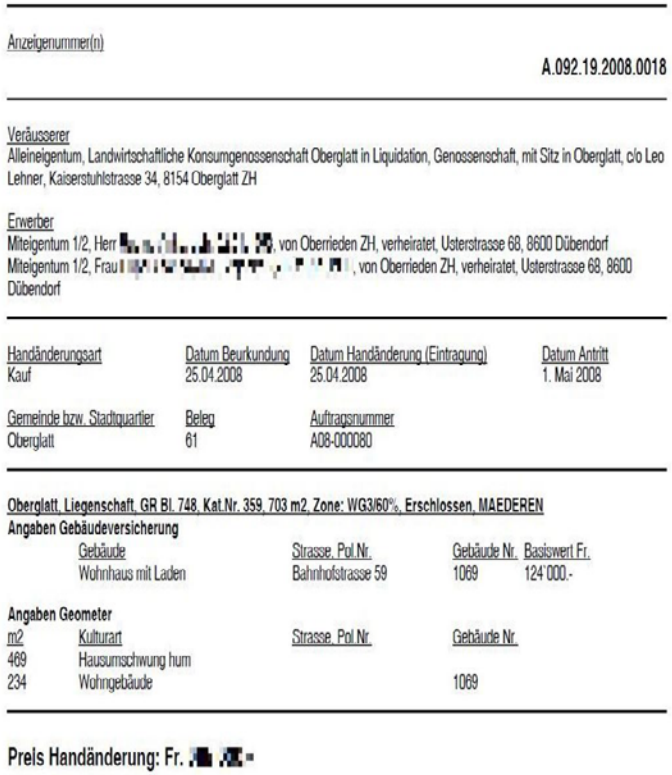
- Subsetting, splitting, joining and mutating strings

For all the basic R stuff

- list.dirs(), for loops, ifelse(), if(){} else(){}, paste(), assign(), names(), fill(), …

# Starting point: the unstructured raw data

The files are structured differently, depending on the type of transaction



Anzeigenummer(n)

A.092.19.2008.0018

Veräusserer
Alleineigentum, Landwirtschaftliche Konsumgenossenschaft Oberglatt in Liquidation, Genossenschaft, mit Sitz in Oberglatt, c/o Leo Lehner, Kaiserstuhlstrasse 34, 8154 Oberglatt ZH

Erwerber
Miteigentum 1/2, Herr [redacted], von Oberrieden ZH, verheiratet, Usterstrasse 68, 8600 Dübendorf
Miteigentum 1/2, Frau [redacted], von Oberrieden ZH, verheiratet, Usterstrasse 68, 8600 Dübendorf

| Handänderungsart | Datum Beurkundung | Datum Handänderung (Eintragung) | Datum Antritt |
|---|---|---|---|
| Kauf | 25.04.2008 | 25.04.2008 | 1. Mai 2008 |

| Gemeinde bzw. Stadtquartier | Beleg | Auftragsnummer |
|---|---|---|
| Oberglatt | 61 | A08-000080 |

Oberglatt, Liegenschaft, GR Bl. 748, Kat.Nr. 359, 703 m2, Zone: WG3/60%, Erschlossen, MAEDEREN

Angaben Gebäudeversicherung

| Gebäude | Strasse, Pol.Nr. | Gebäude Nr. | Basiswert Fr. |
|---|---|---|---|
| Wohnhaus mit Laden | Bahnhofstrasse 59 | 1069 | 124'000.- |

Angaben Geometer

| m2 | Kulturart | Strasse, Pol.Nr. | Gebäude Nr. |
|---|---|---|---|
| 469 | Hausumschwung hum | | |
| 234 | Wohngebäude | | 1069 |

Preis Handänderung: Fr. [redacted]

Information about the sellers and buyers

Information about type and time of trade

Information about the traded objects

The Price

A.090.19.2008.0093

Page 1

Page 2

Preis Handänderung: Fr. [redacted]

# Reading a pdf-document: the core statement

```
library(pdftools)
dir <- "F:/meetup/"
txt <- pdf_text(paste0(dir,"filename.pdf"))
```

```
[1] "                                        Handänderungsanzeige\r\nAdressat:
          Statistisches Amt des Kantons Zürich\r\n
Postfach\r\n                                                                  8090 Zürich\r\nAnzeigenummer(n)\r\n
                                                                              A08-004136.2008.0149\r\nVeräusserer\r\nAlleineigent
um, ABC AG, Aktiengesellschaft (AG), mit Sitz in Affoltern am Albis,  Alte\r\nDorfstrasse 23, 8910 Affoltern am Albis\r\nErwerber\r\nMiteige
ntum 1/3, Herr, M, von Zürich, ledig, Bergstrasse 17, 5630 Muri AG\r\nMiteigentum 1/3, Herr, M, von Serbien, ledig, Bergstrasse 17, 5630
Muri AG\r\nMiteigentum 1/3, Herr, M, Staatsangehörigkeit: Serbien, ledig, Testgasse 11, 8006 Zürich\r\nHandänderungsart
Datum Beurkundung       Datum Handänderung (Eintragung)                       Datum Antritt\r\nKauf                                  28.11.2008
       28.11.2008                                    Eigentumsübertragung\r\nGemeinde bzw. Stadtquartier        Beleg
Auftragsnummer\r\nAffoltern am Albis                 570                      A08-004136\r\nAffoltern am Albis, Liegenschaft, GR Bl. 6527,
Kat.Nr. 5498, 1186 m2, Zone: G, , Hägeler, Wohnungs-Nr.                       \r\nbekanntgegeben\r\nAngaben Geometer\r\nm2            Gebäudeart / Kulturart
                  Strasse, Pol.Nr.                                            Gebäude Nr.\r\nGebäude(AVGBS)\r\n458           Gebäude Industrie
Alte Dorfstrasse 23           00201689\r\nBedeckungen\r\n458           Gebäude\r\n384           befestigte Fläche\r\n344
Gartenanlage\r\nPreis Handänderung: Fr. xxxxx.-\r\nBemerkungen:\r\nBemerkungen Empfänger:\r\nDatum/Benutzer: 30.11.2008            \r\nGr
undbuchamt Affoltern\r\n"
```

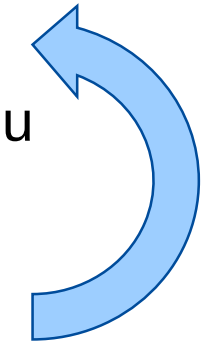
hypothetical example

`pdftools::pdf_text()` extracts the whole file with multiple pages into one string

# Trim the string

To get the information you want, you have to split the string into handy pieces.

– harmonize the string (e.g. eliminate linebreaks)
– look for keywords and isolate the parts that interest you
– format the parts as variables in a one row data.frame
– Add this row to the main data.frame



Repeat the import and all other steps for each pdf-file
`for (x in 1:i)`

| | |
|---|---|
| `stringr::str_replace_all(…,"\r\n"," ")` | eliminates linebreaks |
| `stringr::str_split(…,keyword)` | splits a string at a keyword |
| `stringr::str_count(…, keyword)` | counts the apperance of a keyword |
| `dplyr::bind_rows()` | appends a row to an existing data.frame |

# Clean the data

At this point the data is mostly in the right order, but there are still some cleanup work to be done.

– correct typing errors

– removed German umlauts

– erase bothersome parts (e. g. whitespaces)

```
sub("str[.]|strassse|strase|strsse|straasse|strass$", "strasse", str_name))
```

| | |
|---|---|
| | replaces a pattern with a single expression |
| `stringr::str_replace_all(…,"ä","ae")` | replace german umlauts |
| `stringr::str_trim()` | erases the whitespaces |
| `stringr::str_sub()` | separates disturbing parts |

# Enriche the data

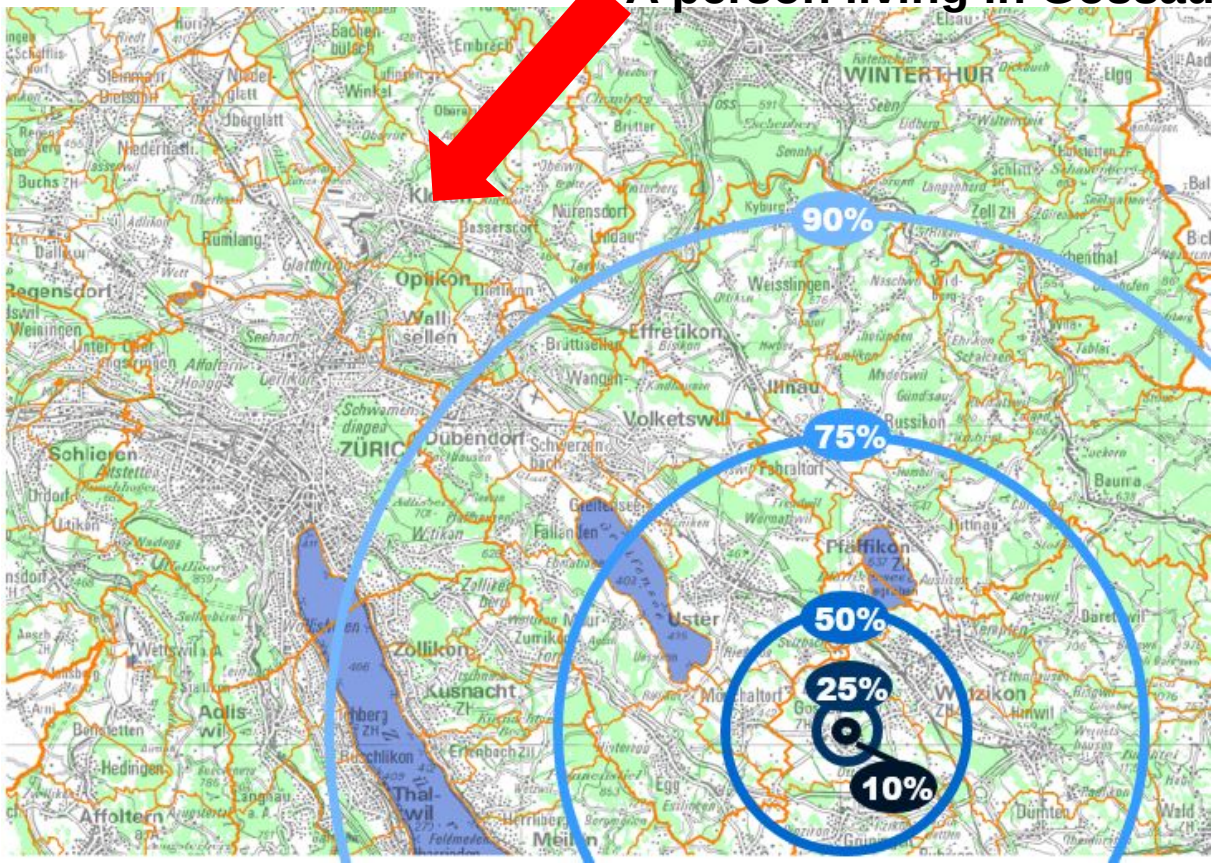
After the record has been cleaned up, it can be enriched with additional information.

– Calculate new variables
– Select and rename variables of interest
– Merge with other data.frames
– …

| | |
|---|---|
| `dplyr::mutate()` | create and mutate variable |
| `dplyr::select()` | select variables |
| `dplyr::join()` | Merge with other data |
| `dplyr::filter()` | select cases |

# Analyse the data

**A person living in Gossau will hardly buy a house in Kloten.**




**Distance between residence and new home:**

- 10 percent bought a house in the immediate neighborhood
- A quarter bought a house in an area of less than 950 meters
- Half of all buyers lived less than 3.6 km from the new home
- 90% bought a house at a distance of 15 km or less.

**Conclusion: The market for real estate seems to be very local**

# This presentation:

github.com/statistikZH/RMeetup/blob/master/unstructered_data.pdf

# A recent blog with similar tasks:

www.r-bloggers.com/taming-exam-results-in-pdf-with-pdftools/

Max Grütter
max.gruetter@statistik.ji.zh.ch
@MaxGruetter

statistik.zh.ch
@statistik_zh