



**Kanton Zürich
Direktion der Justiz und des Innern
Statistisches Amt**

Read and prozess unstructured data in R

MeetUp Zurich R User Group

7 November 2017

Max Grütter

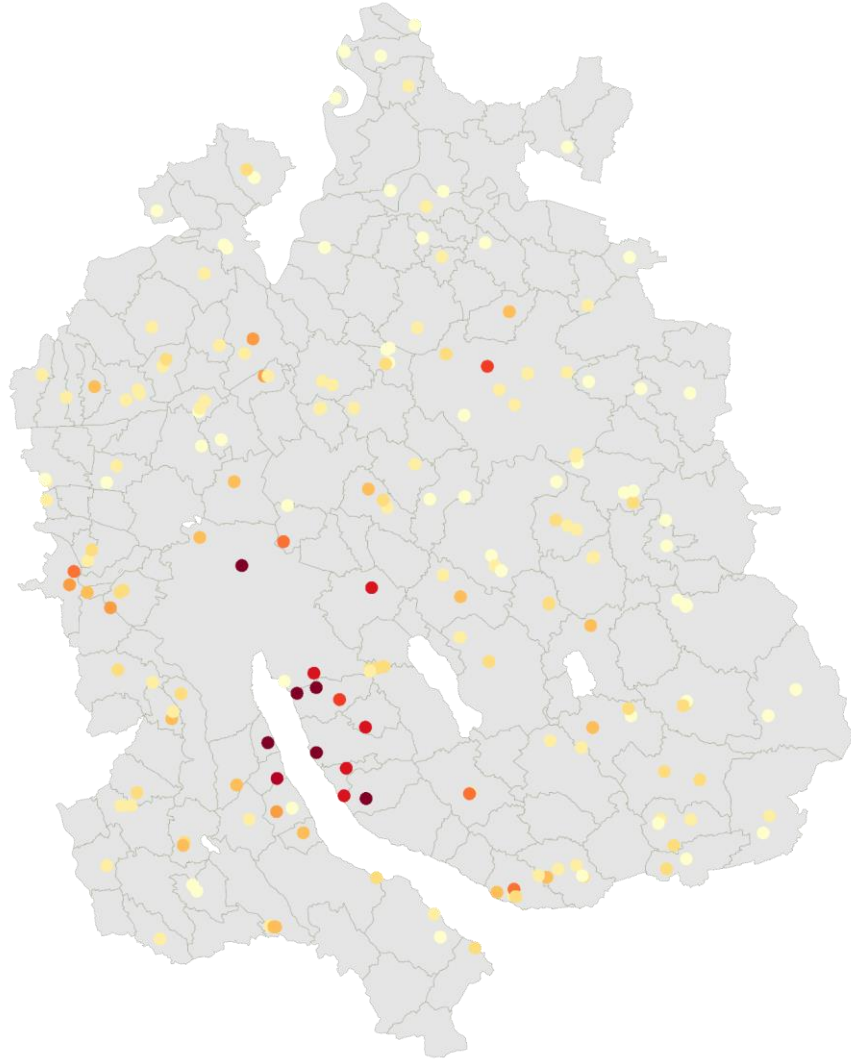
Motivation

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 market for real estate in Zurich



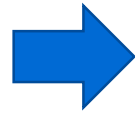
Statistics ZH keeps the **official statistics on real estate sales**

- 20'000 transactions per year (sales, inheritance, etc.)
- approx. 2'300 single-family houses sold per year
- approx. 4'500 apartments sold per year

Today I would like to know in which area someone is buying a house

Our journey today

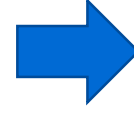
Read each document



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



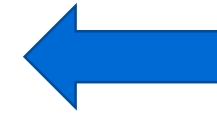
Append all files



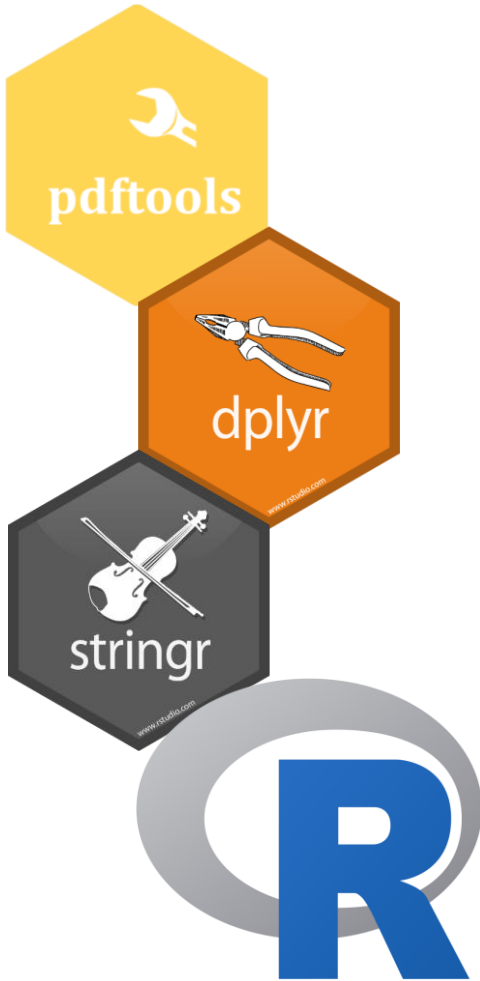
**20'000 documents
per year**

The remaining steps

- Clean the data
- Enrich the data
- analyse the data

[illegible]

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

- 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 I [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

Anzeigenummer(n)

A.095.19.2008.0093

Veräusserer

Gesamteigentum, einzelne Gesellschaft, StützheimVulvaGP General Building GmbH, R.H. Kurz AG/Festhol Immobilien AG/Hörn + Partner AG, c/o Höhn + Partner AG, Bettstrasse 35, 8600 Dübendorf

Gesamteigentum, Festhol Immobilien AG, AG, mit Sitz in Dübendorf, Festholstrasse 34a, 8600 Dübendorf

Gesamteigentum, GP General Building GmbH, GmbH, mit Sitz in Kloten, Eggenstrasse 132, 8302 Kloten

Gesamteigentum, Herr Höhn, Max, geb. 24.09.1936, von Zürich, verheiratet, Am Säulen 1A, 8117 Pfäfers

Gesamteigentum, Höhn + Partner AG, AG, mit Sitz in Dübendorf, Bettstrasse 35, 8600 Dübendorf

Gesamteigentum, Herr Kubi, Rudolf, geb. 30.03.1938, von Zürich, Netzel GL, verheiratet, General Gussen-Ost 34, 8002 Zürich

Gesamteigentum, R.H. Kurz AG, AG, mit Sitz in Rapperswil TG, Rigenweg 1, 8558 Pörschen TG

Gesamteigentum, Herr Dütz, Heinrich, geb. 12.08.1936, von Rheinfelden, verheiratet, Rütlistrasse 21, 8308 Rheinfelden

Erwerber

Alleineigentum, Frau Büchler-Gehrer, Claudia, geb. 08.07.1967, von DI, Gläster FR, Bödingen FR, Leberdorf FR, geschieden, Büchlerstrasse 29, 8155 Nefelbach

Handänderungsart	Datum Beurkundung	Datum Handänderung (Eintragung)	Datum Antritt
Kauf	25.07.2008	24.08.2008	
Gemeinde bzw. Stadtquartier	Beleg	Anzeigenummer	mit Eigentumsübertragung
Niederhasli	233	A08-000029	

Ziv. DBL 5267 bzw. 5268: Autobahnplatz Nr. 5 bzw. 6.

Niederhasli, Grundbesitz, GR Bl. 5267

69'000 Miteigentum an GR Bl. 5268

mit Sonderrecht an der Wohnung Nr. C 3 im Erdgeschoss und im Keller und Backraum im Untergeschoss, in den Aufzugsgängen sowie Garagen und mit Nr. C 3 beschränkt, gemäss Begründungserklärung vom 25.11.2006, Bl. 473, samt Änderung vom 25.05.2008, Bl. 181, DRIVE-Plan 248.

Beschreibung des gemeinschaftlichen Grundstücks

Niederhasli, Liegenschaft, GR Bl. 5268, Kat.Nr. 3045, 2008 mit Sonderstrasse 3 a 7

Angaben Gebäudeversicherung			
Gebäude	Strasse, Pol.Nr.	Gebäude Nr.	Basiswert Fr.
Angaben Geometer	Strasse, Pol.Nr.	Gebäude Nr.	
m2	Kulturart		
23	Hausumschwung		
2375	Auen, Wiese, Weide		954

Niederhasli, Miteigentum, GR Bl. 5267

1/37 Miteigentum an GR Bl. 5267

Niederhasli, Miteigentum, GR Bl. 5268

1/37 Miteigentum an GR Bl. 5268

Beschreibung des gemeinschaftlichen Grundstücks

Niederhasli, Liegenschaft, GR Bl. 5267, Kat.Nr. 3061, 58 mit Sonderstrasse

Angaben Gebäudeversicherung			
Gebäude	Strasse, Pol.Nr.	Gebäude Nr.	Basiswert Fr.
Angaben Geometer	Strasse, Pol.Nr.	Gebäude Nr.	
m2	Kulturart		
59	Auen, Wiese, Weide		

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
A.002.01.2008.0149\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, Teststrasse 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: nicht\r\nbekanntgegeben\r\nAngaben Geometer\r\nnm2      Gebäudeart / Kulturart
Strasse, Pol.Nr.      Gebäude Nr.\r\nGebäude(AVGBS)\r\nn458      Gebäude Industrie
Alte Dorfstrasse 23      00201689\r\nBodenbedeckungen\r\nn458      Gebäude\r\nn384      befestigte Fläche\r\nn344
Gartenanlage\r\nPreis Handänderung: Fr. xxxxxxxx.-\r\nBemerkungen:\r\nBemerkungen Empfänger:\r\nDatum/Benutzer: 30.11.2008 / Sven Hodel\r\nnGr
undbuchamt Affoltern\r\nn"]
```

`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 this for each
pdf-file
for (x in 1:i)

<code>stringr::str_replace_all(..., "\r\n", " ")</code>	eliminates linebreaks
<code>stringr::str_split(..., keyword)</code>	splits a string at a keyword
<code>stringr::str_count(..., keyword)</code>	counts the apperance of a keyword
<code>dplyr::bind_rows()</code>	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.

- erase bothersome parts (e. g. whitespaces)
- correct typing errors
- removed German umlauts

```
stringr::str_trim()      erases the whitespaces  
stringr::str_sub()       separates disturbing parts  
sub("str[.]|strasse|strase|strsse|straasse|strass$", "strasse", str_name))  
                           replaces a pattern with a single expression
```



Enriche the data

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

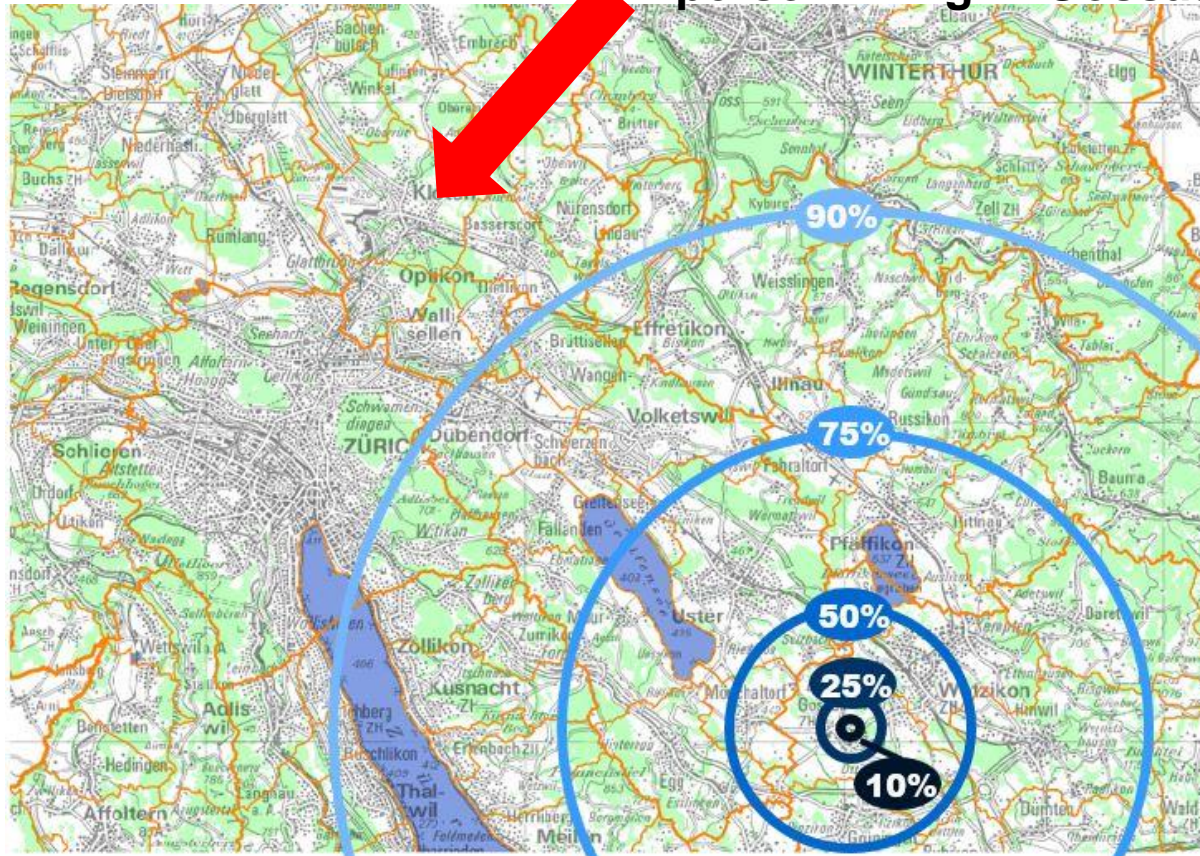
- Add and select variables
- Merge with other data.frames

<code>dplyr::mutate()</code>	create and mutate variable
<code>dplyr::select()</code>	select variables
<code>dplyr::join()</code>	Merge with other data
<code>dplyr::filter()</code>	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

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

Questions???

Max Grütter

max.gruetter@statistik.ji.zh.ch

@MaxGruetter

statistik.zh.ch

@statistik_zh