



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

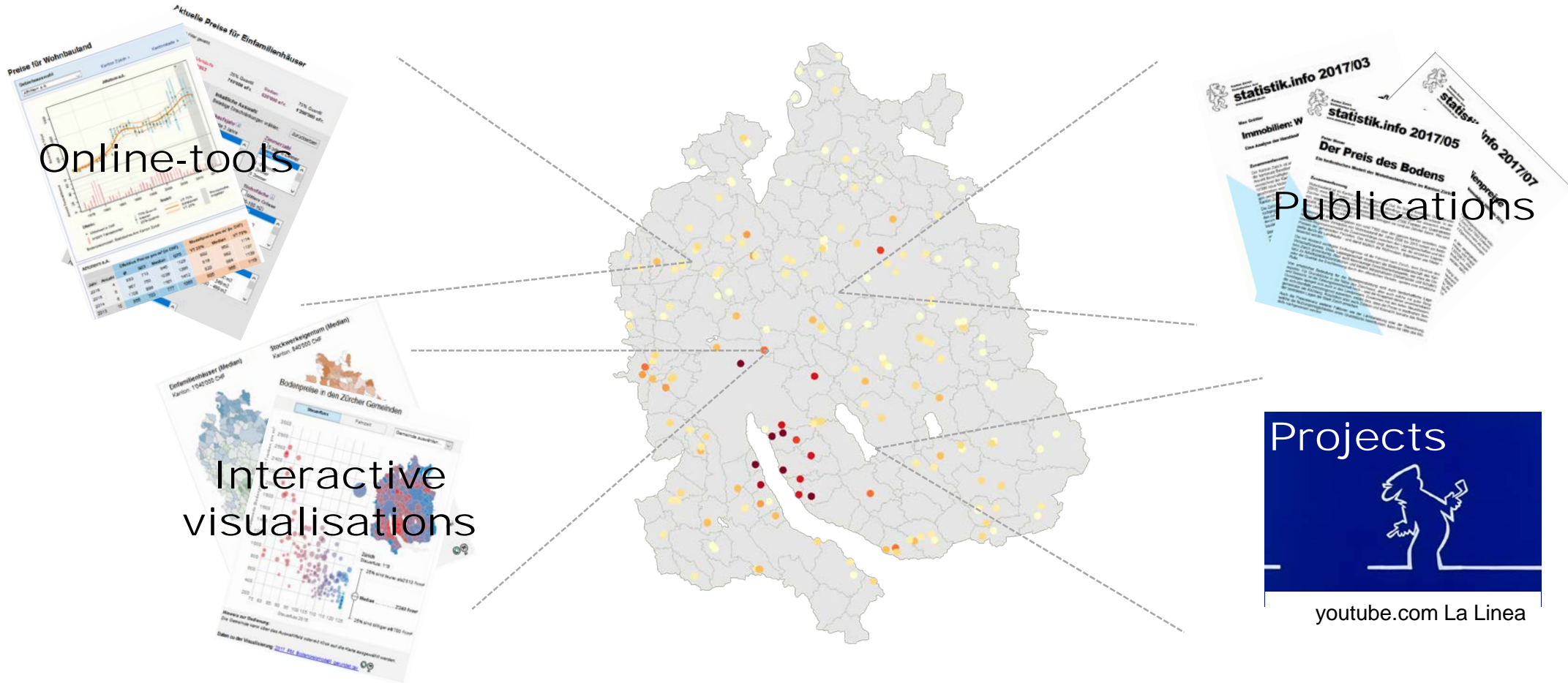
Motivation

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



Statistics ZH keeps the official statistics on real estate sales

Our journey today

Read each document



The whole file in one string



Write a
data.frame



Extract & reshape
the core information

[illegible][illegible]

Nr.	Name
Handelstempel Datum Beurkundung Datum H...	Handelstempel Datum Beurkundung Datum H...
Erstellung — 30.12.2008 30.12.2008	Erstellung — 30.12.2008 30.12.2008
Gemeinde bzw. Stadt/quartier Betrag Auftragsnummer	Gemeinde bzw. Stadt/quartier Betrag Auftragsnummer
Helmberg 428 A28-002094	Helmberg 428 A28-002094
Heimleitung	Heimleitung
Legenschaft	Legenschaft
Gd Br. 4312	Gd Br. 4312
Kat.Nr. 4396	Kat.Nr. 4396
297 m2	297 m2
Fundaentis	Fundaentis
Angaben Gebäudemessung	Angaben Gebäudemessung
Gebäude Stosse	Gebäude Stosse
Pächte Gebäude Nr. Baucostent Fr.	Pächte Gebäude Nr. Baucostent Fr.
Zeichenmerkmalen Ringwert 27 3408	Zeichenmerkmalen Ringwert 27 3408
Untermerkmale Ringwert 3407 179 000...	Untermerkmale Ringwert 3407 179 000...
Notenpunkte v. d. Ringwert 3416 83 000...	Notenpunkte v. d. Ringwert 3416 83 000...
Satzstellen Ringwert 3418 19 000...	Satzstellen Ringwert 3418 19 000...
Angaben Geometrie	Angaben Geometrie

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



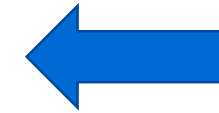
Append all files



20'000 documents
per year

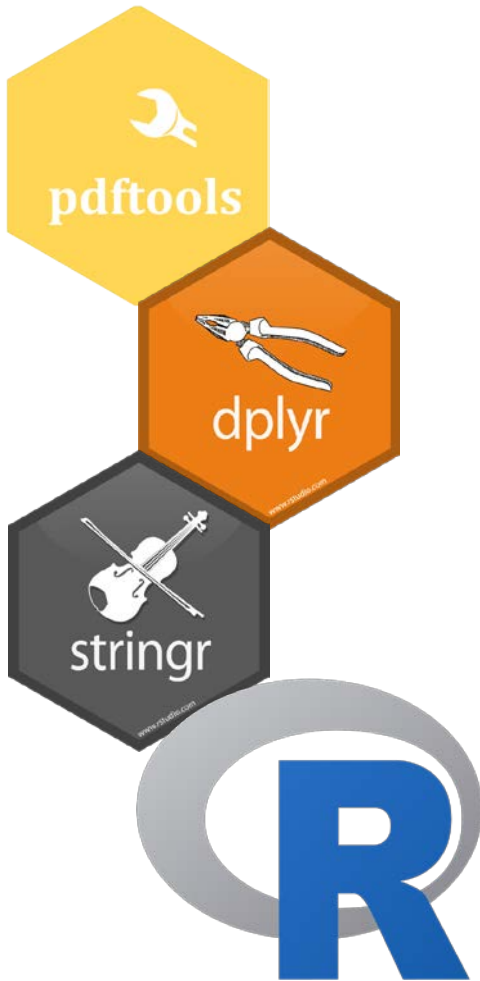
The remaining steps

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



nr	vt	vt	vt	vt
1	von Rottum (2)	unverändert	Bismarckstrasse 70	8857 Zuckow
1,2	Stationsgebäude - Baugeschossungen von Dabow 50	unverändert	Stationsstrasse 1	8857 Zuckow
1	von Kirsch 08	nicht verändert	Ordnungsstr. 55	8706 Eichen
1,2	Klosterhof	unverändert	Ordnung 9	8710 Marnsdorf
1,2	Klosterhof	unverändert	Ordnung 9	8710 Marnsdorf
1	von Witzl 10	nicht verändert	Stationsstrasse 12	8710 Marnsdorf
1	von Hermsing (2)	unverändert	Stationsstrasse 12	8636 Ritz
1	Stallhof 58	nicht verändert	Friedrichstrasse 10	8630 Ritz
1,2	Stationsgebäude (2)	unverändert	Erbsenstr. 20	8707 Ritz
1,2	Stationsgebäude (2) Station 10-130.000 186 4	unverändert	Stationsstrasse 12	8710 Marnsdorf
1	Schweisserei Fabrikgebäude	Bismarckstrasse 70	Stationsstrasse 12	8630 Ritz
1	von Egidius 40	nicht verändert	Stationsstrasse 9	8710 Marnsdorf
1	von Egidius 10	nicht verändert	Stationsstrasse 9	8710 Marnsdorf
1	von Egidius 10	unverändert	Stationsstrasse 12	8710 Marnsdorf
1,2	von Brand 18	unverändert	Ordnungsstrasse 110	8710 Marnsdorf
1	von Fuchsen (2)	unverändert	Ordnungsstrasse 110	8710 Marnsdorf
1	von Fuchsen 10	nicht verändert	Ordnungsstrasse 110	8710 Marnsdorf
1	von Fuchsen 10	nicht verändert	Ordnungsstrasse 110	8710 Marnsdorf
1	Klosterhofgebäude	10-130.000 186 4	Ordnungsstrasse 110	8710 Marnsdorf
1	Klosterhofgebäude	10-130.000 186 4	Ordnungsstrasse 110	8710 Marnsdorf
1	Stationsgebäude (2) Station	unverändert	Stationsstrasse 12	8630 Ritz
1	von Hermsing (2)	unverändert	Stationsstrasse 12	8630 Ritz
1	von Hermsing (2)	unverändert	Stationsstrasse 12	8630 Ritz

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 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

Hausumschwing 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.090.19.2008.0093

Veräusserer

Gesamteigentum, einzelne Gesellschaft: Stutz-Höhn-Rubli-GP General Building GmbH, R.H. Kuntz AG-Pedholz Immobilien AG-Höhn + Partner AG, c/o Höhn + Partner AG, Bettstrasse 35, 8600 Dübendorf
 Gesamteigentum, Pedholz Immobilien AG, AG, mit Sitz in Dübendorf, Pedholzstrasse 26A, 8600 Dübendorf
 Gesamteigentum, GP General Building GmbH, GmbH, mit Sitz in Kloten, Eggenstrasse 132, 8502 Kloten
 Gesamteigentum, Herr Höhn, Max, geb. 04.09.1938, von Zürich, verheiratet, Am Mühlstein 1A, 8117 Pfäfers
 Gesamteigentum, Höhn + Partner AG, AG, mit Sitz in Dübendorf, Bettstrasse 35, 8600 Dübendorf
 Gesamteigentum, Herr Rubli, Rudolf, geb. 30.03.1939, von Zürich, Netstal GL, verheiratet, General Gussen-Quai 34, 8002 Zürich
 Gesamteigentum, R.H. Kuntz AG, AG, mit Sitz in Rapperswil TG, Pögenweg 1, 8558 Pörschen TG
 Gesamteigentum, Herr Dutz, Heinrich, geb. 12.09.1938, von Rhododendron, verheiratet, Rütlistrasse 21, 8308 Rhau

Erwerber

Alleineigentum, Frau Büchler-Gehrer, Claudia, geb. 08.07.1967, von St. Silvester PR, Bödingen PR, Leberdorf PR, geschieden, Brunnstrasse 29, 9155 Naterschwil

Handlungsart	Datum Beurkundung	Datum Handänderung (Eintragung)	Datum Antritt mit Eigentumsübertragung
Kauf	25.07.2008	24.08.2008	
Gemeinde bzw. Stadtquartier	Beleg	Auftragsnummer	
Naterschwil	253	A08-000028	

Ziv. GSt. 5267 bzw. 5268: Autobahnplatz Nr. 9 bzw. 6.

- **Naterschwil, Stockschneidengut, GR Bl. 5267**
68'000 Miteigentum an GR Bl. 5268
mit Sonderrecht an der Wohnung Nr. C 3 im Erdgeschoss und am Keller und Backraum im Untergeschoss, in den Aufzugsgängen sowie bemalt und mit Nr. C 3 beschriftet, gemäss Begründungsänderung vom 25.11.2006, Bel. 470, samt Änderung vom 29.09.2008, Bel. 181, DRIVE-Plan 248.

Beschreibung des gemeinschaftlichen Grundstückes
Naterschwil, Liegenschaft, GR Bl. 5267, Kat.Nr. 3048, 2098 m2, Sandrainstrasse 2 & 7

Angaben Gebäudeversicherung	Gebäude	Strasse, Pol.Nr.	Gebäude Nr.	Basiswert Fr.
Angaben Geometer	Kulturart	Strasse, Pol.Nr.	Gebäude Nr.	
m2	23		554	
2375	Naterschwil			
Angaben Geometer	Kulturart	Strasse, Pol.Nr.	Gebäude Nr.	
m2	23			
59	Naterschwil			

Beschreibung des gemeinschaftlichen Grundstückes
Naterschwil, Liegenschaft, GR Bl. 5267, Kat.Nr. 3048, 2098 m2, Sandrainstrasse

Angaben Gebäudeversicherung	Gebäude	Strasse, Pol.Nr.	Gebäude Nr.	Basiswert Fr.
Angaben Geometer	Kulturart	Strasse, Pol.Nr.	Gebäude Nr.	
m2	23			
59	Naterschwil			

Preis Handänderung: Fr. [REDACTED]

Page 1

Page 2

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
28.11.2008.0149\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, Bergstrasse 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. 570\r\nBekanntgegeben\r\nAngaben Geometer\r\nm2 Gebäudeart / Kulturart
Strasse, Pol.Nr. Gebäude Nr.\r\nGebäude(AVGBS)\r\nn458 Gebäude Industrie
Alte Dorfstrasse 23 00201689\r\nDeckungen\r\nn458 Gebäude\r\nn384 befestigte Fläche\r\nn344
Gartenanlage\r\nPreis Handänderung: Fr. x\r\nx\r\nx.-\r\nBemerkungen:\r\nBemerkungen Empfänger:\r\nDatum/Benutzer: 30.11.2008 \r\n\r\nGr
undbuchamt Affoltern\r\n\r\n"
```

`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)`

<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.

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

```
sub("str[.]|strasse|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.

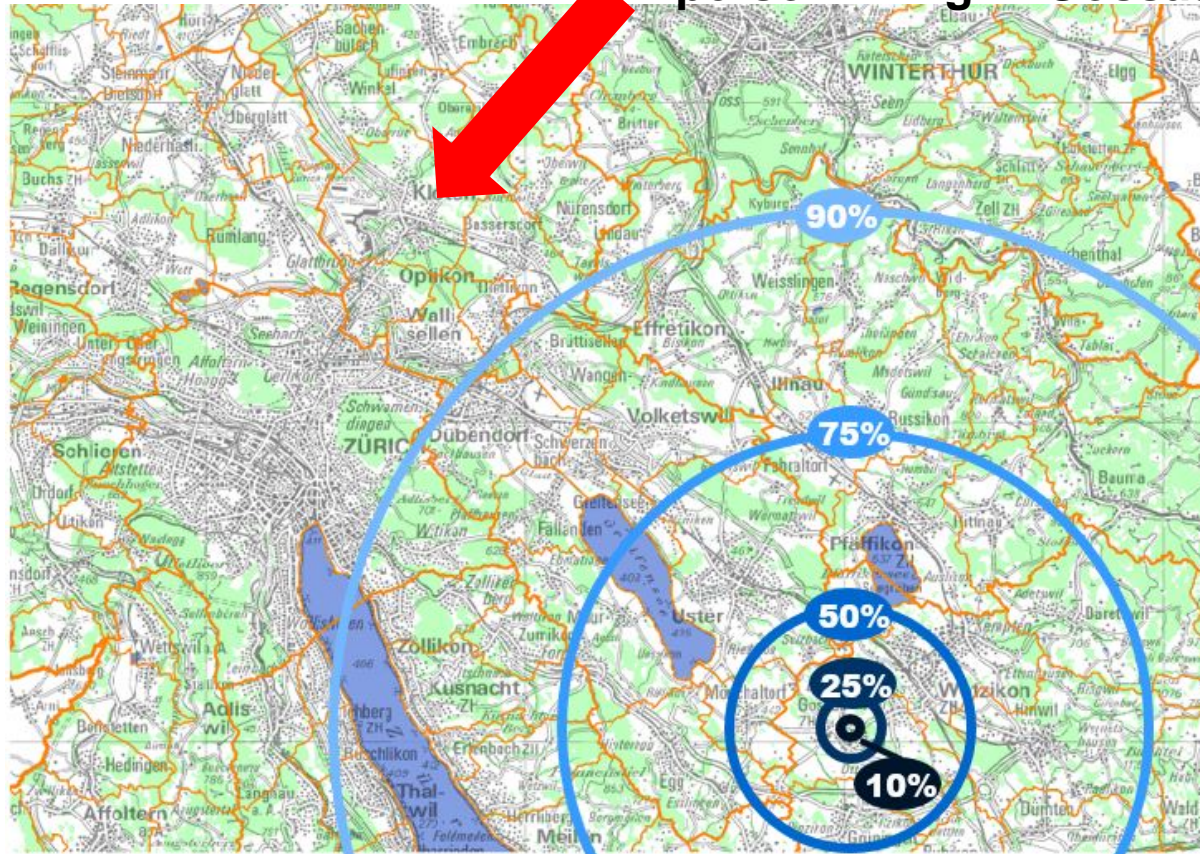
- Calculate new variables
- Select and rename variables of interest
- Merge with other data.frames
- ...

<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
- 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/unstructured_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

