Webscraping mit R

Airbnb im Kanton Zürich

Thomas Lo Russo, Statistisches Amt des Kantons Zürich

2017/11/21

Inhalt

1. Webscraping als ad-hoc Erhebungsmethode
2. R-packages & Ressourcen
3. Airbnb im Kanton Zürich

Delivering the evidence of tomorrow

Power from Statistics Conference, Brussels, 18-19.10.2017

Welche Rolle kommt der öffentlichen Statistik in Zukunft zu? Wie kann sie dem Informationsbedürfnis der Gesellschaft auch künftig gerecht werden?

Delivering the evidence of tomorrow

Power from Statistics Conference, Brussels, 18-19.10.2017

Welche Rolle kommt der öffentlichen Statistik in Zukunft zu? Wie kann sie dem Informationsbedürfnis der Gesellschaft auch künftig gerecht werden? Auch über im Internet erhobene Daten (**Web-data**) wurde debatiert:

> Web data aren't going to replace official stats but they complement what is already there @jonsteinberg #powerfromstatistics
>
> — Roeland Beerten (@roelandb) 19. Oktober 2017

Neue Herausforderungen

für die öffentliche Statistik

Im Zeitalter der Digitalisierung gibt es eine Vielzahl Phänomene, welche mittels konventioneller Erhebungsverfahren nur beschränkt eingefangen werden können.

→ **Online Plattform Economy**: (Airbnb, Uber etc.)

Neue Herausforderungen

für die öffentliche Statistik

Im Zeitalter der Digitalisierung gibt es eine Vielzahl Phänomene, welche mittels konventioneller Erhebungsverfahren nur beschränkt eingefangen werden können.

→ **Online Plattform Economy**: (Airbnb, Uber etc.)

Optionen

Neue Herausforderungen

für die öffentliche Statistik

Im Zeitalter der Digitalisierung gibt es eine Vielzahl Phänomene, welche mittels konventioneller Erhebungsverfahren nur beschränkt eingefangen werden können.

→ **Online Plattform Economy**: (Airbnb, Uber etc.)

Optionen

Status Quo beibehalten

Blick Artikel: "Tourismus-Statistiker ignorieren Airbnb",5.11.2017

Neue Herausforderungen

für die öffentliche Statistik

Im Zeitalter der Digitalisierung gibt es eine Vielzahl Phänomene, welche mittels konventioneller Erhebungsverfahren nur beschränkt eingefangen werden können.

→ **Online Plattform Economy**: (Airbnb, Uber etc.)

Optionen

Status Quo beibehalten

Plattformen zur Datenherausgabe bewegen

Blick Artikel: "Tourismus-Statistiker ignorieren Airbnb",5.11.2017

Neue Herausforderungen

für die öffentliche Statistik

Im Zeitalter der Digitalisierung gibt es eine Vielzahl Phänomene, welche mittels konventioneller Erhebungsverfahren nur beschränkt eingefangen werden können.

→ **Online Plattform Economy**: (Airbnb, Uber etc.)

Optionen

Status Quo beibehalten

Plattformen zur Datenherausgabe bewegen
Regulieren & Erhebung konzipieren

Blick Artikel: "Tourismus-Statistiker ignorieren Airbnb",5.11.2017

Neue Herausforderungen

für die öffentliche Statistik

Im Zeitalter der Digitalisierung gibt es eine Vielzahl Phänomene, welche mittels konventioneller Erhebungsverfahren nur beschränkt eingefangen werden können.

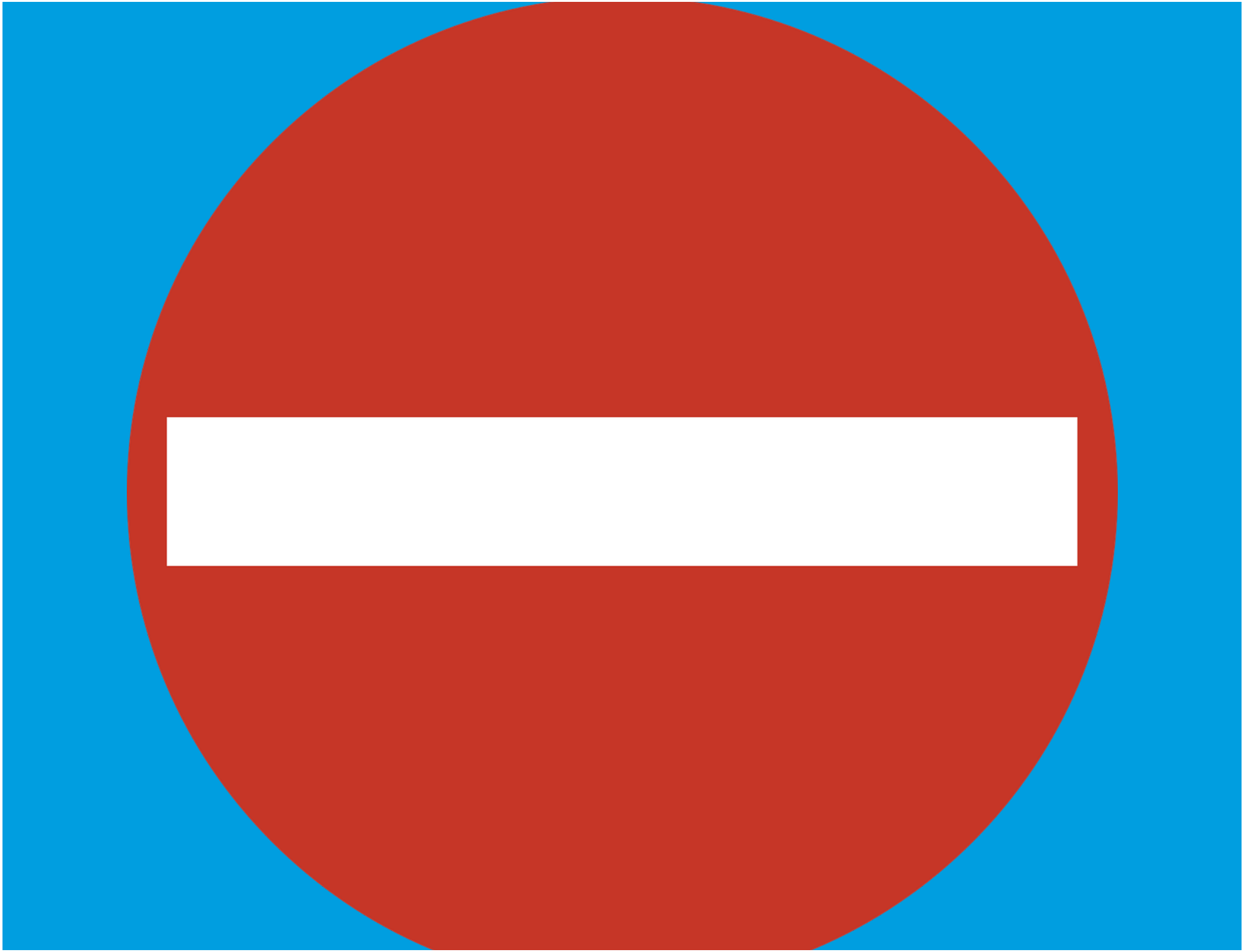
→ **Online Plattform Economy**: (Airbnb, Uber etc.)

Optionen

Status Quo beibehalten

Plattformen zur Datenherausgabe bewegen
Regulieren & Erhebung konzipieren
→ Web-Scraping

Blick Artikel: "Tourismus-Statistiker ignorieren Airbnb",5.11.2017



WebData - was gilt es zu beachten

Ist es überhaupt zulässig im Netz Daten zu sammeln?

Einschätzung zur Situation in der Schweiz:

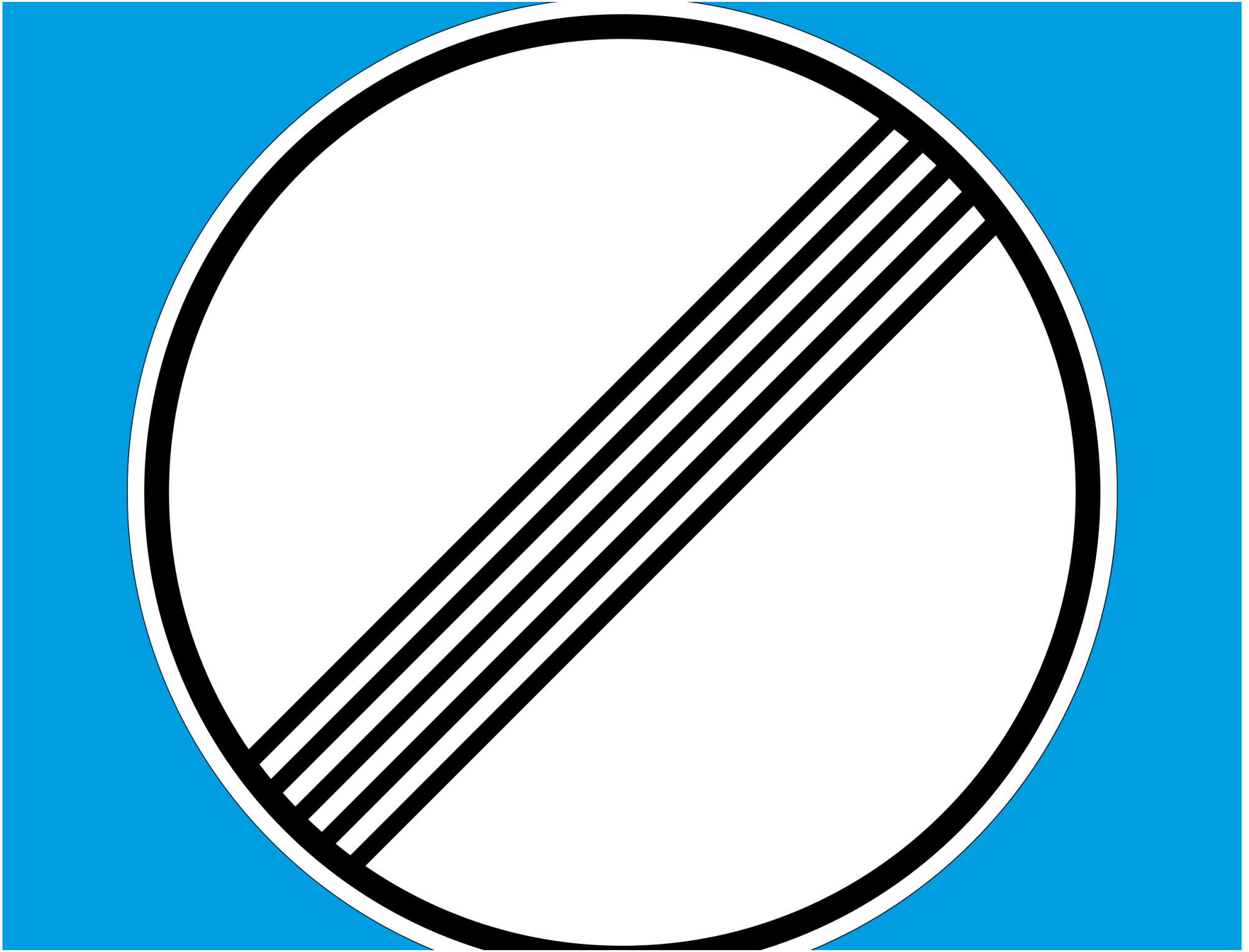
https://www.mll-news.com/verwertung-fremder-datenbanken-im-schweizer-recht/

Empfehlung: **Verhaltensregeln der Webseiten-Anbieter befolgen!**

Quasi-standard: digitale Hausordnung für Suchroboter

→ robots.txt!

Beispiel: https://www.airbnb.ch/robots.txt?locale=de




60

Webscraping in R

Scraping-Strategien

- Human Copy-Paste
- DOM Parsing
- Text pattern matching (unstructured data)
- **API Interface**

Ressourcen für R

R bietet unzählige Ressourcen für Webscraping, welche unterschiedliche Funktionalitäten bieten.

- Retrieving and parsing html rvest
- Crawling, retrieving and parsing RCrawler

Weitere: scrapeR, tm.plugin.webmining

Parsing only: jsonlite, RJSONIO, tidyjson, XML, XML2

Beispiel Airbnb

Problem: Airbnb-API ist kein 'Datenservice' welcher alle Resultate für eine Abfrage liefert. Die API liefert die selben Resultate wie die Suchmaske auf der Webseite. Pro Suchabfrage: max. 300 Angebote.

Lösung: Automatisierte Abfragen über Preiskategorien, Postleitzahlen, Kapazität etc.

```
baseurl <- "https://api.airbnb.com/v2/search_results?client_id=3092nxybyb0otqw18e8nh5

library(jsonlite)
#Zur Veranschaulichung (Skript unvollständig):
#URL muss für automatisierte Abfragen dynamisch erweitert werden

for (plz in lastplz:nrow(ortzh)){ #Loop über PLZ/Ort Liste
for (p in c(seq(20,100,20),seq(200,500,100),seq(500,1000,250))){ #Loop über Preiskate
for (i in seq(0,350,50)){#Loop über Seitenindex der Suchresultate

  zhlist <- fromJSON(paste0(baseurl,"&sort=1&_offset=",i ,"&&location=",ortzh$ORTNAME

  ...
}}}
```

```
Retrieving listings 0 max price:20 PLZ:  8607, NUMBER of listings retrived: 365
Retrieving listings 50 max price:20 PLZ:  8607, NUMBER of listings retrived: 370
Retrieving listings 0 max price:40 PLZ:  8607, NUMBER of listings retrived: 372
Retrieving listings 50 max price:40 PLZ:  8607, NUMBER of listings retrived: 382
Retrieving listings 0 max price:60 PLZ:  8607, NUMBER of listings retrived: 382
Retrieving listings 50 max price:60 PLZ:  8607, NUMBER of listings retrived: 382
Retrieving listings 0 max price:80 PLZ:  8607, NUMBER of listings retrived: 385
Retrieving listings 50 max price:80 PLZ:  8607, NUMBER of listings retrived: 385
Retrieving listings 0 max price:100 PLZ:  8607, NUMBER of listings retrived: 387
Retrieving listings 50 max price:100 PLZ:  8607, NUMBER of listings retrived: 387
Retrieving listings 0 max price:200 PLZ:  8607, NUMBER of listings retrived: 388
Retrieving listings 50 max price:200 PLZ:  8607, NUMBER of listings retrived: 389
Retrieving listings 0 max price:300 PLZ:  8607, NUMBER of listings retrived: 390
Retrieving listings 50 max price:300 PLZ:  8607, NUMBER of listings retrived: 391
Retrieving listings 0 max price:400 PLZ:  8607, NUMBER of listings retrived: 393
....
```

Airbnb im Kanton Zürich

Mitteilung: Airbnb blüht vor allem in den Städten, Dez. 2016.

Vielen Dank
für Ihre Aufmerksamkeit!

https://statistikzh.github.io/SST17

Thomas Lo Russo
thomas.lorusso@statistik.ji.zh.ch
043 259 75 13

Slides created via the R package **xaringan**.