

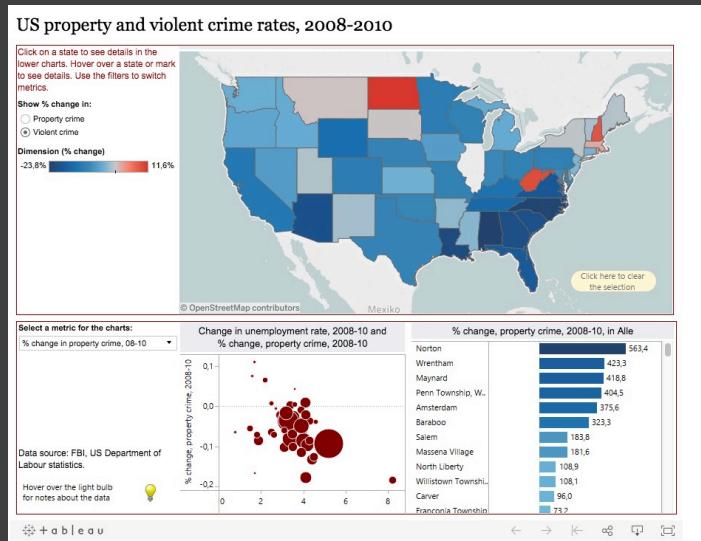
Hamburg R User Group
07.12.2017

A more structured approach to **data journalism** through the usage of **R Notebooks**

@PatrickStotz
<https://patrickstotz.github.io/>
@SPIEGEL_data

What's data journalism

fancy maps, data dashboards

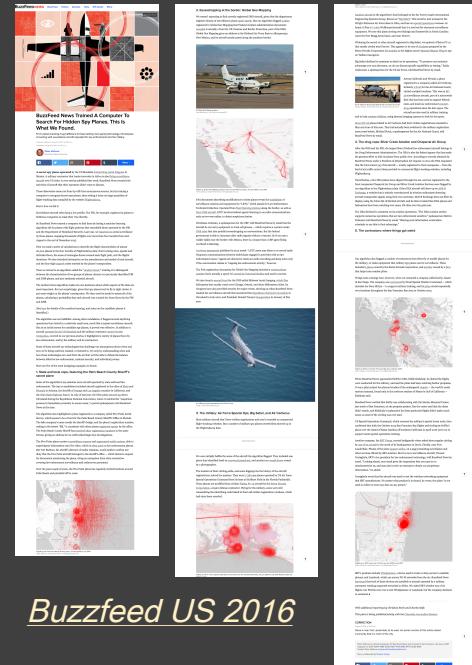


Guardian data blog 2011

No offence, they've paved the way!



FiveThirtyEight 2016



Buzzfeed US 2016

„and much more“ me

If we want it to be the latter,
we'll have to structure our workflows accordingly

Our project structure

Folder structure

- **article** [Text, drafts. Could be Microsoft Office, Office365 or Markdown format]
- **data**
 - raw [unchanged(!) raw data, extracted data (e.g.: .csv from .pdf), scraped HTML-data]
 - processed [saved output from data processing]
 - final [final data sets as used for article, charts etc.]
- **figures**
 - preliminary [exploratory visualizations, non styled figures]
 - final [final figures for publishing]
- **release-extern** [release on github]
- **ressources** [inspiration, background material. Could be links, figures, texts,...]
- **scripts** [all scripts and notebooks. Could be QGIS-files as well]
 - lib [local copy of repetedly used code snippets or self-written libraries. Local copy to ensure version compatibility]
 - reports [rendered noteboos (HTML-files)]
 - 01_scraping.py [skripts/notebooks in processing order]
 - 02_analysis.Rmd [separate data mining/wrangling from analysis]
- README.md [this file]

!!! work in progress !!!

Titel des leeren Beispiel-Projekts

Am Ende hier gerne ein Titelbild einfügen

Worum geht es?

Worum geht es in dem Projekt? Zusammenfassung in 2-3 Sätzen

Letzter Stand

z.B. "in Bearbeitung", "eingestellt weit...", "erschienen am..."

Zuletzt geändert durch Erika Mustermann am 22.11.2017

Autoren / Ansprechpartner

Projekt-Verantwortliche mit Kontaktinformationen

Was findet sich wo in diesem Ordner

- article [Text, Textschippe und -entwürfe. Je nach Projekt in Microsoft Office, Office365 oder Markdown-Format]
- data
 - raw [unverändert(!) Rohdaten, extrahierte Daten (z.B. .csv aus .pdf), gescrapete HTML-Daten]
 - processed [Daten aus Verarbeitungsschritten]
 - final [finale Daten, wie sie im Artikel, in Grafiken oder interaktiven Visualisierungen verwendet werden]
- figures
 - preliminary [Zwischenstände aus explorativen Arbeitsschritten, Exports vor dem Styling]
 - final [finale Abbildungen für den Artikel]
- release-extern [zur Veröffentlichung auf github]
- ressources [Inspiration und Hintergrundmaterial in Form von Links, Abbildungen, Texten]
- scripts [Alle Skripte, Notebooks oder auch QGIS-Projekte]
 - lib [eigene wiederkehrend verwendete Code-Schnipsel oder Libraries in lokalen Kopie (zwecks Versionskompatibilität)]
 - reports [gerenderte Notebooks als HTML-Dateien]
 - 01_scraping.py [Skripte/Notebooks in Reihenfolge der Ausführung]
 - 02_analysis.Rmd [Datengewinnung getrennt von Daten-Analyse]
- README.md [Diese Datei]

Datenquellen / Lizenzen

Beispieldaten_1.pdf

- Quelle: [spiegel.de](#)
- ggf. Hinweis wie: nur intern verwenden, nicht veröffentlichen

Links

hier sammeln wir alle Verknüpfungen des Projekts zum CE in tabellarischer Form

Typ	ID	Was
Artikel	1165366	SPON-Artikel
HTML-Asset	133421	Highchart Transfersummen
Tabelle	16086	Die 500 teuersten BL-Transfers
DataTable	133443	Die 500 teuersten BL-Transfers
Tabelle	16084	Die 500 teuersten Internat. Transfers
DataTable	133438	Die 500 teuersten Internat. Transfers
HTML-Asset	133444	Bindet die beiden Tabellen zusammen
HTML-Asset	133424	Highchart Durchschnittstransfer
HTML-Asset	133431	Highchart Umsatzentwicklung
Motiv	1183692	Artikel-Header Motiv

*Examples on how we use
R (Notebooks)*

Football inflation - a transfer market analysis

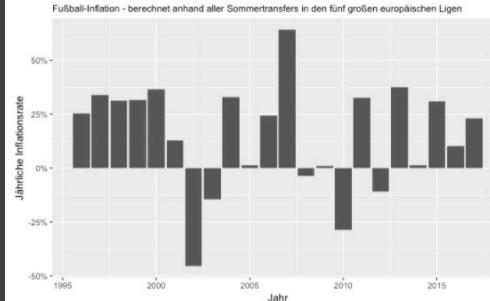
- scraping (rvest)
- data processing (tidyverse)
- exploratory dataviz (ggplot)
- communication with sports editors
- export data for charts and interactives
- updating data within minutes (rerun notebook)

Inflationsbereinigung

```
Rbind with column specification:  
cols =  
  Jahr = col_double(),  
  multi_0 = col_double()  
)
```

Bedingung hat Länge > 1 und nur das erste Element wird benutztBedingung hat Länge > 1 und nur das erste Element wird benutztBedingung hat Länge > 1 und nur das erste Element wird benutztBedingung hat Länge > 1 und nur das erste Element wird benutzt

Jährliche Inflationsrate für die 1. Bundesliga seit 1996



Die durchschnittliche jährliche Inflation in den fünf großen Ligen seit 1996 beträgt 11.3 Prozent

Die teuersten Transfers in 2017er-Beträgen (alle 5 Ligen)

Ablöse = gezahlter Betrag laut Transfermarkt.de, zum damaligen Kurs in Euro umgerechnet
Ablöse_inf_Foot = Heutiger Betrag unter Berücksichtigung der Fußball-Inflation
Ablöse_inf_D = Heutiger Betrag unter Berücksichtigung des Verbraucherpreisindex

Fußball-Inflation ermittelt anhand der Gesamtsumme der Transfersumsätze im jeweiligen Jahr, dividiert durch die 2017er Summe. Da das Transferfenster 2017 noch offen ist, werden sich die Beträge noch ändern.

Name	Position	Ablöse	Ablöse_inf_Foot	Ablöse_inf_D	Saison
Neymar	Linksaußen	222,00 Mio. €	222,00 Mio. €	22200 Tsd. €	2017
Rio Ferdinand	Innenverteidiger	46,00 Mio. €	20,43 Mio. €	55761 Tsd. €	2002
Ronaldo	Mittelstürmer	45,00 Mio. €	198,03 Mio. €	54549 Tsd. €	2002
Cristiano Ronaldo	Linksaußen	94,00 Mio. €	182,82 Mio. €	102079 Tsd. €	2009
David Beckham	Rechtes Mittelfeld	37,50 Mio. €	176,10 Mio. €	44950 Tsd. €	2003
Zinédine Zidane	Offensives Mittelfeld	73,50 Mio. €	165,96 Mio. €	90319 Tsd. €	2001
Alan Shearer	Mittelstürmer	21,00 Mio. €	163,81 Mio. €	27640 Tsd. €	1996
Ronaldo	Mittelstürmer	28,00 Mio. €	160,26 Mio. €	36144 Tsd. €	1997
Hernán Crespo	Mittelstürmer	36,00 Mio. €	158,42 Mio. €	43539 Tsd. €	2002
Gareth Bale	Rechtsaußen	101,00 Mio. €	158,09 Mio. €	102624 Tsd. €	2013

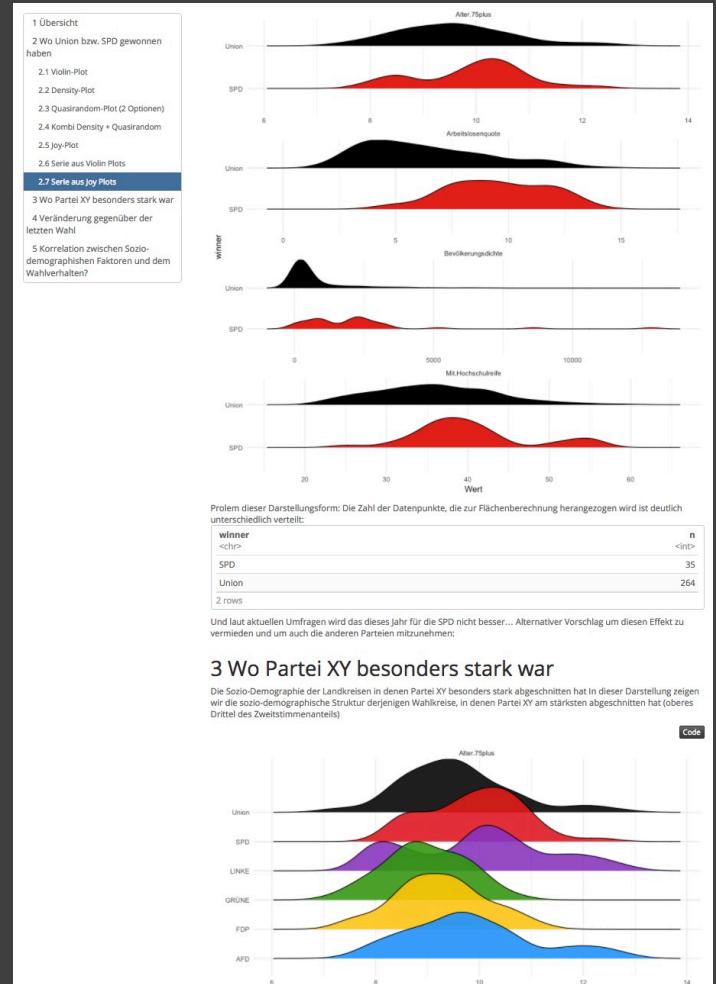
1-10 of 31,691 rows | 1-6 of 8 columns

Previous 1 2 3 4 5 6 ... 100 Next

Notebook (not public)

The socio-demography of party strongholds

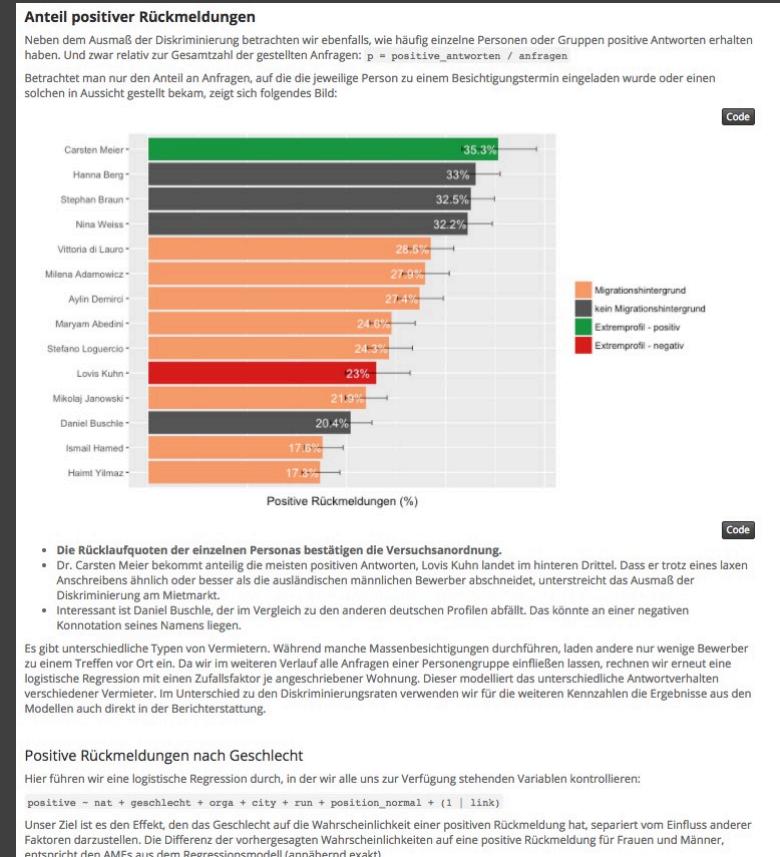
- data processing (tidyverse)
- exploratory data viz (ggplot + ggbeeswarm + ggridges)
- communication with politics editors
- exporting final figures (ggplot)
- entirely prepared before election night. Figures were ready within minutes



[link to notebook](#)

Measuring rental housing discrimination in Germany

- data processing (tidyverse)
- statistical tests and modeling
- communication with editors from different departments
- export data for charts and interactives
- publishing source code, sample data and methodology → reproducibility & transparency



[link to notebook](#)

Why R Notebooks?

BTW: Python + Jupyter Notebooks is also great

- Code forces us to be accurate
- Scripting makes us fast and enables us to prepare ahead
- Rendered notebooks are great for discussing preliminary results with (non tech-savvy) colleagues
- Data journalism workflows are iterative. Analysis and reporting are highly entangled. Notebooks let us keep everything in one place

Questions?

Want to work with us? Looking for an internship? Get in touch!

Patrick.Stotz@spiegel.de
@PatrickStotz