

Zpracování (geo)dat v R

Jan Caha

Mendelova univerzita v Brně

jan.caha@mendelu.cz

GIS Ostrava 2018

21. 3. 2018

① Úvod

② Práce s daty v R

③ Práce s geodaty

④ Další užitečné balíky

⑤ Závěr

Něco o mně

- studium: na katedře Geoinformatiky, Univerzita Palackého v Olomouci (Bc., Mgr., Ph.D, RNDr.)
- praxe: VŠB-TUO, Mendelova Univerzita v Brně
- odborné zájmy: zpracování nejistoty v GIS, fuzzy logika a fuzzy arimetika, analýzy viditelnosti, data science (se zaměřením na prostorová data/analýzy), open source (obecně + GIS)
- jazyky:
 - dříve: Java, C#
 - aktuálně: R, Python
- moje práce:
 - Github
 - ResearchGate

Seminář

- o čem to bude: filozofie práce s daty v R, aktuální balíky a trendy v přístupech k analýzám, užitečné balíky
- co to nebude: konkrétní a specifické operace a/nebo analýzy
- co je důležité: obecně lze v R všechno řešit mnoha způsoby a u každého rozhodnutí máme mnoho možností, jak ho provést, nelze zahrnout vše a ani konstatovat, co je nejlepší

Úvodní dotazník

- kdo:
 - pravidelně používá R?
 - pracuje s RStudiem?
 - zná balíky dplyr, ggplot2 nebo tidyverse?
 - a geoinformatické sp, raster, sf a tmap?

Motivace

- jako geoinformatik věnujete 20-80% pracovního času zpracování dat
- každou analýzu budete 5x-10x předělávat, než se dopracujete ke konečnému výsledku
- důraz na automatizaci práce a opakovatelnost postupů a procesů (jak opakovatelný je klik myši???)
- většina standardních “GUI” nástrojů příliš na automatizaci a opakovatelnosti postupů nelpí a uživatele k nim systematicky nevede

Jazyk R

- implementace jazyka **S** pod open source licencí
- určený primárně ke zpracování dat a jejich vizualizaci, s výrazným zaměřením na statistiku
- interpretovaný programovací jazyk přístupný v základu především skrze příkazový řádek
- základní funkcionality není příliš široká, velká část funkcionality přichází z tzv. balíčků
- centrální sklad balíčků - **CRAN** (Comprehensive R Archive Network)
- historie: vývoj od 1993, prezentace v 1996 (Ihaka and Gentleman, 1996), hlavní verze 1.0 (2000), 2.0 (2004), 3.0 (2013), aktuálně 3.4.3
- cca od roku 2014 nesmírně dynamický vývoj různých oblastí nasazení a využití R
- aktuálně na CRAN dostupných přes 12 000 balíčků

Data Science

- datová věda
- definice: interdisciplinární obor propojující věděcké metody, procesy, algoritmy a systémy za účelem extrakce znalostí a pohledů na data, ať už strukturovaná či nikoliv (Dhar, 2013)
- většinou značnou část práce zahrnuje předzpracování, zpracování a úprava dat (data wrangling)
- další činnosti jsou obvykle vizualizace a prezentace dat, propojování dat z nerůznějších zdrojů
- důraz na opakovatelnost postupů a otevřenost
- právě data science je aktuálně hlavní hybnou silou ve vývoji R

Zdroje informací

- návod R + viněty
- knihy např. oficiální seznam na [webu](#)
- seznam volně dostupných (open source) knih týkajících se R a vytvořených přímo v R - [bookdown](#)
- R4DS, Advanced R, R Programming for Data Science, Spatial Microsimulation with R, Geocomputation with R atd.
- blogy - např. [R-bloggers](#)
- twitter - obzvláště např. [Mara Averick](#), ale i větší množství dalších (vývojáři, data scientists)
- Stack overflow a jeho varianty

Práce s R

- většinou nepracujeme přímo s příkazovým řádkem R
- využití IDE (Integrated Development Environment), které poskytuje nástroje pro signifikantní zvýšení komfortu práce s R
- řada IDE zaměřených na R - **Rattle**, **RKWard**, **R Commander**
- nicméně téměř standardem a nejpopulárnějším řešením je **RStudio**
- R lze používat několika způsoby
 - tzv: interaktivní mód, kdy pracujeme "na živo" s daty
 - programování analýz
 - programování balíčků
 - tvorba nejrůznějších druhů výstupů, které obvykle kombinují text, kód v R a výstupy z R
- nejčastější je kombinace interaktivního módu s programovaním analýz (to budeme předpokládat)

RStudio

- popis prostředí
- tvorba a práce s projektem
- struktura projektu

Práce s daty - přístupy

- base R × Tidyverse
- **Tidyverse** - kolekce balíčků, zaměřená na import - export, zpracování, analýzu a vizualizaci dat
- vývoj výrazně podporuje RStudio (firma)
- hlavní postava: Hadley Wickham - koncept tidy data (Wickham, 2014)
- výhody - pipeline, slovesa pro práci s daty, gramatika grafiky
- proti základnímu R unifikace funkčnosti, atributů funkcí atd.

```
install.packages("tidyverse")
library(tidyverse)
```

Operace s daty

```
data("diamonds")
View(diamonds)
```

Operace s daty

```
diamonds_summarise <- diamonds %>%
  filter(price > 9000) %>%
  group_by(cut, color) %>%
  summarise(count = n(),
            price_min = min(price),
            price_max = max(price),
            carat = median(carat))
View(diamonds_summarise)
```

Operace s daty - joiny

```
library("nycflights13")
data(flights, airports)

View(flights)
View(airports)

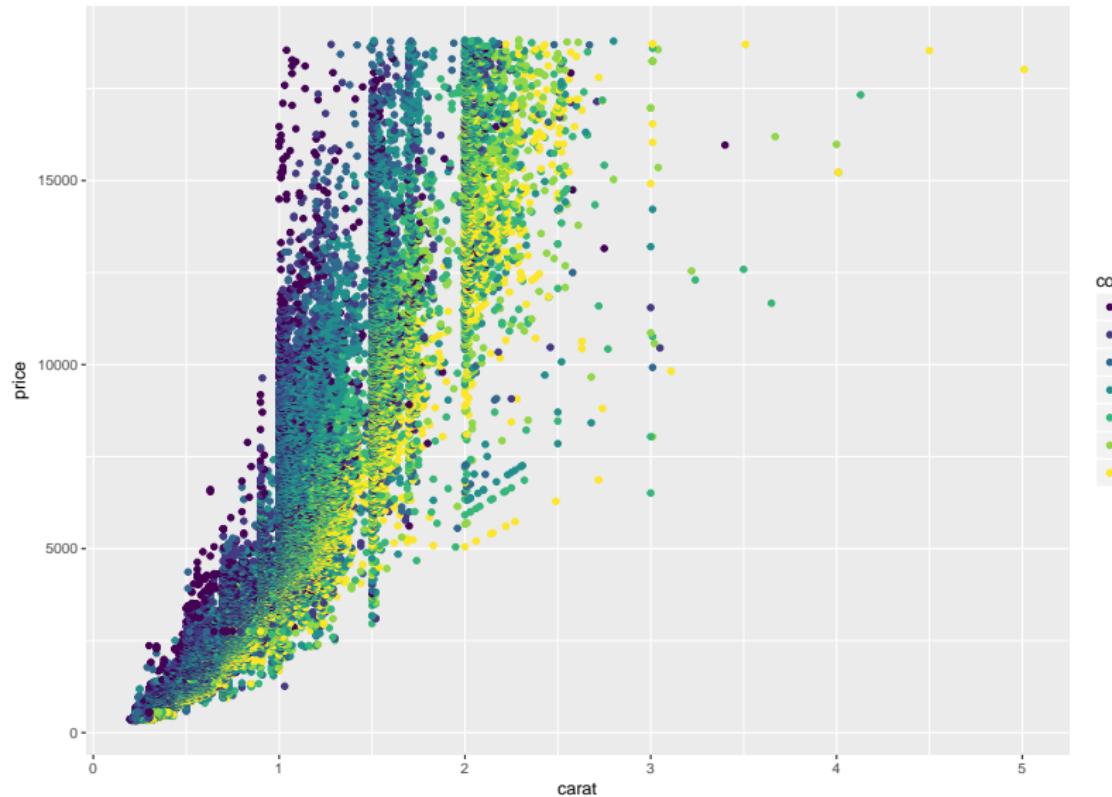
# propojení tabulek
joined_data <- flights %>%
  left_join(airports, by = c("origin" = "faa"))

View(joined_data)
```

Vizualizace dat

```
ggplot(diamonds, aes(carat, price, color = color)) +  
  geom_point()
```

Vizualizace dat

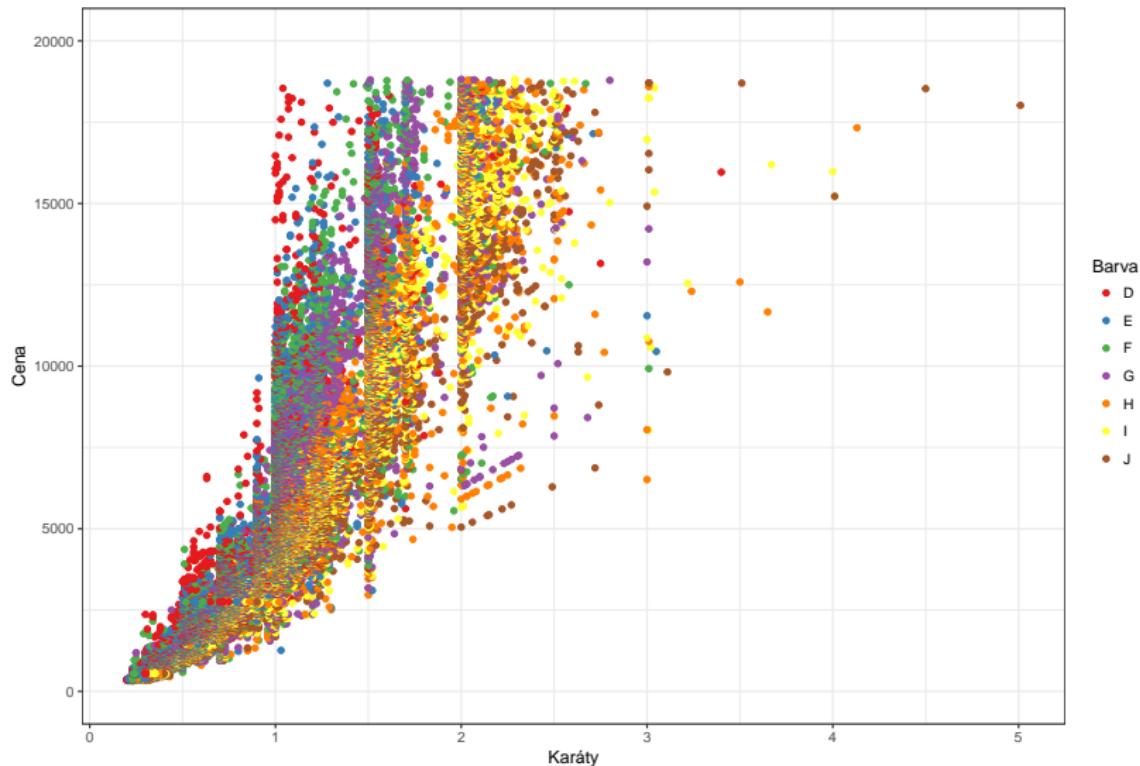


Vizualizace dat - 2

```
ggplot(diamonds, aes(carat, price, color = color)) +
  geom_point() +
  scale_color_brewer(palette = "Set1") +
  scale_y_continuous(breaks = seq(0, 20000, by = 5000),
                     limits = c(0, 20000)) +
  labs(x = "Karáty", y = "Cena", fill = "Barva",
       title = "Porovnání ceny diamantů podle karátů a barvy") +
  theme_bw()
```

Vizualizace dat - 2

Porovnání ceny diamantu podle karátu a barvy



Vstup a výstup dat

- sada funkcí ve základním R, ale nejsou příliš logické a konzistentní z hlediska názvů a parametrů
- vhodnější je využití balíčků `readr`, `readxl`, případně i `writexl`
- načítací funkce pak mají podobu `read_typ-souboru`, zapisovací `write_typ-souboru`
- týká se hlavně tabelárních dat, ostatní typy dat se načítají přes specializované balíčky (např. `foreign`)

Vstup a výstup dat

```
diamonds_summarise %>%
  write_csv("data_sumarziované.csv")

library(writexl)

diamonds_summarise %>%
  write_xlsx("data_sumarziované.xlsx")

# identický zápis
write_xlsx(diamonds_summarise, "data_sumarziované.xlsx")
```

Vztah R k prostorovým datům a analýzám

- primární účel - statistická analýza geodat
- první pokusy už v roce 2000 (Bivand, 2000)
- v roce 2003 už existovalo minimálně 21 balíčků zaměřených na prostorová data (Bivand, 2003)
- formální definice prostorových objektů v R až s balíčkem `sp` (Pebesma and Bivand, 2005), který se stal de facto standardem
- později balíček `raster` (Hijmans, 2017)
- aktuálně snaha nahradit `sp` za nový balíček `sf` (Pebesma, 2017), který vhodněji odpovídá aktuálním standardům práce s daty v R
- vizualizace prostorových dat - přímo výše jmenované balíky, vývojová verze `ggplot2` instalovaná z GitHubu, balík `tmap`

Balíky zaměřené na geodata

- kategorie: reprezentace prostorových dat, operace s prostorovými daty, vizualizace prostorových dat
- základní balíky s tímto zaměřením jsou uvedené v CRAN Task Views - [Spatial](#) a [SpatioTemporal](#)
- v těchto Views dohromady cca 222 balíčků
- mnoho dalších, které v těchto Views zapsané nejsou (více moje zítřejší přednáška - případně [web](#))

Balík sf

- novinka od roku 2016
- narození od původního balíku sp vychází funkčně z balíku dplyr s nímž je kompatibilní -> možnost jednoduše pracovat s daty, např. `group_by()`, `summarise()` platí i pro geometrie -> značné zjednodušení oproti sp
- základní závislosti na knihovnách: GDAL, GEOS, PROJ.4, liblwgeom, udunits2 (knivny v C)
- závislosti na řadě R balíků
- značná část funkcionality používá pojmenování `st_*`

Kompatabilita sf s sp

- na sp zavisí relativně hodně balíků - prostorové operace, statistiky a jiné.
- přechod na sf bude trvat a některé balíky patrně nikdy přepsané nebudou
- funkce `as_Spatial()` zajistí konverzi z formátu balíku sf do sp
- inverzně funguje funkce `st_as_sf()`
- doporučení: zpracovávat data jako sf a konvertovat až při potřebě speciálních funkcí

Načtení dat

```
install.packages(c("sf", "tmap", "readxl"))

library(sf)
library(tmap)
library(readxl)

okresy <- st_read(here::here("data", "okresy.gpkg"),
                  stringsAsFactors = FALSE,
                  quiet = TRUE)
okresy_data <- read_xlsx(here::here("data",
                                      "socioekonomicka_data.xlsx"),
                           sheet = 1)
okresy_kody <- read_xlsx(here::here("data",
                                      "socioekonomicka_data.xlsx"),
                           sheet = 2)
```

Propojení prostorových a tabulkových dat

```
okresy_data <- okresy_data %>%
  left_join(okresy_kody)
rm(okresy_kody)

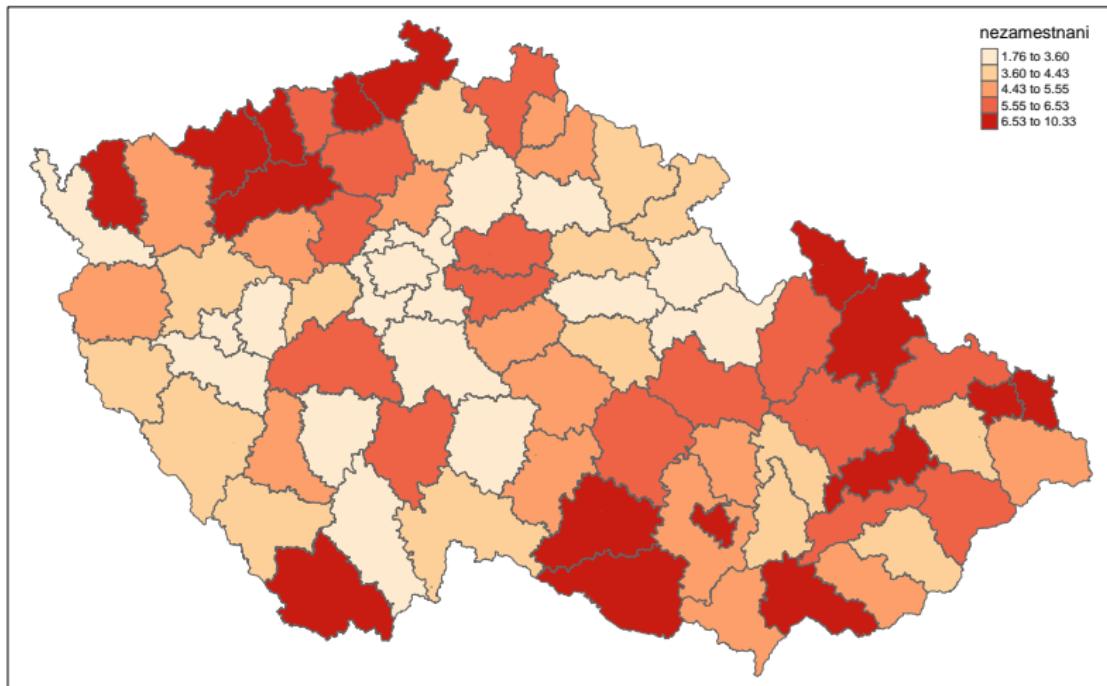
okresy <- okresy %>%
  left_join(okresy_data,
            by = c("KOD_OKRES" = "kod_okresu"))
rm(okresy_data)

okresy <- okresy %>%
  mutate(pracovni_mista_na_obyvatele =
         pracovni_mista_v_evidenci / obyvatel)
```

Vizualizace pomocí tmap

```
tm_shape(okresy) +  
  tm_polygons(col = "nezamestnani", n = 5,  
              style = "quantile", palette = "OrRd")
```

Vizualizace pomocí tmap

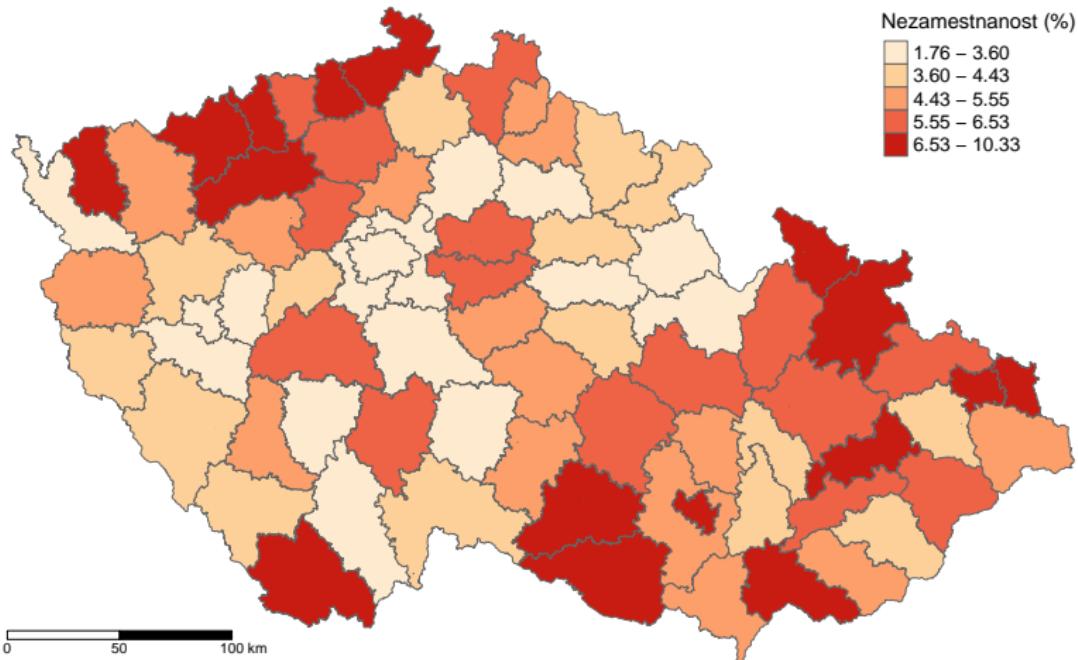


Vizualizace pomocí tmap

```
tm_shape(okresy) +
  tm_polygons(col = "nezamestnani", n = 5 ,
              style = "quantile", palette = "OrRd",
              title = "Nezaměstnanost (%)") +
  tm_scale_bar(position = c("left", "bottom"),
                breaks = c(0, 50, 100), size = 0.75) +
  tm_layout(frame = FALSE,
            legend.title.size = 1.3,
            legend.text.size = 1.0,
            legend.format = list(text.separator = "-"),
            main.title = "Nezaměstnanost v okresech ČR
k 31. 12. 2016",
            main.title.position = "center")
```

Vizualizace pomocí tmap

Nezamestnanost v okresech CR k 31. 12. 2016



Vizualizace pomocí tmap

- uložení výstupu a interaktivní vizualizace ve zdrojových kódech 02-prostorova-data.R

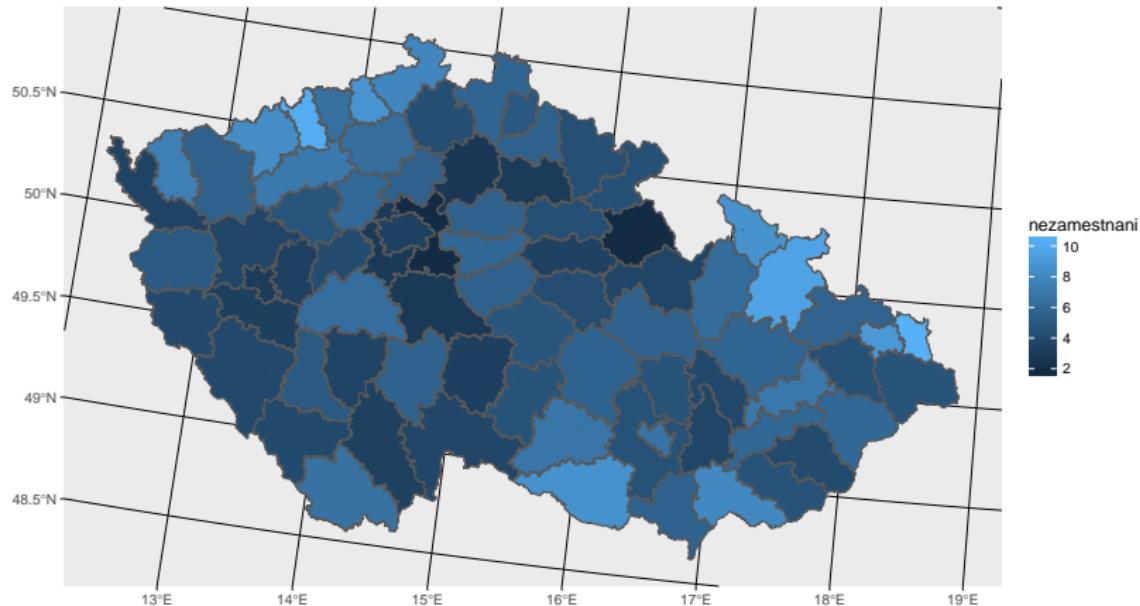
Vizualizace pomocí ggplot2

```
devtools::install_github("tidyverse/ggplot2")
```

- množství nastavení dle návodů ggplot2

```
ggplot() +
  geom_sf(data = okresy, aes(fill = nezamestnani))
```

Vizualizace pomocí ggplot2



Příklad komplexnější vizualizace

- využití tmap
- ve zdrojových kódech 02-prostorova-data.R

Knitr

- balík, který umožňuje tvorbu dokumentů, které kombinují text, zpracování a analýzu dat
- <https://yihui.name/knitr/>
- výsledkem je potom dokument, který obsahuje komentář, kód a jeho výsledky
- soubory s příponou .Rmd - rozšíření značkovacího jazyka Markdown
- v základu jsou výstupem html dokumenty
- pokud se nainstaluje pandoc a libovolná distribuce TeXu, lze kompilovat dokumenty i do mnoha jiných formátů - např. Word, PDF, tex
- ukázka ve zdrojových kódech 03-knitr.Rmd
- hlavička ve formátu yaml

Balíky odvozené z Knitr

- např. bookdown - umožňuje tvorbu knih na základě .Rmd souborů, některé z knih zmíněných na začátku prezentace jsou takto vytvořeny
- rmarkdown vytváří různé formáty výstupů na základě konfiguračních .yaml souborů a .Rmd souborů
- např. web <https://jancaha.github.io/GIS-Ostrava-2018/> vzniká na základě 8 .Rmd a jednoho .yaml souboru
- blogdown umožňuje v R vytvářet statické blogy (založeno na Hugo)
- pkgdown z help balíčku vygeneruje web - využívají standardně balíky z tidyverse

Shiny

- tvorba interaktivních online dokumentů
- <https://shiny.rstudio.com/>
- R slouží jako backend, který zpracovává požadavky a vrací definované výsledky
- je třeba Shiny Server
- ukázka ve složce 04-shiny
- celá řada rozšiřujících balíčků - jména obvykle shiny*
- ukázky aplikací v Shiny - Dean Attali - [web](#)

Plotly

- tvorba interaktivních grafů pomocí plotly.js
- lze kombinovat s grafy nastylovanými jako ggplot2
- ukázka 05-plotly.R

Rozšíření ggplot2

- galerie extenzí
- zajímavý např. balíček geofacet
<https://hafen.github.io/geofacet/>

Scrapování dat z webu

- balíky `xml2` a `rvest`
- poměrně slušný nástroj na získávání dat z webových stránek

Hlavní plus R

- nulové vstupní náklady
- množství dostupné literatury a zdrojů + široká a otevřená komunita
- i pro geoinformatiku zajímavá funkctionalita
- množství funkctionality v balíčcích
- propojování funkctionality dohromady - online dokumenty komplilované přes knitr, které obsahují shiny aplikace

Kontakt

- v případě zájmu mě neváhejte kontaktovat
- pracovní mail: jan.caha@mendelu.cz
- soukromý mail: jan.caha@outlook.com

Literatura I

- Bivand, R. S. (2000) Using the R statistical data analysis language on GRASS 5.0 GIS database files. *Computers and Geosciences*, 26(9-10), pp. 1043–1052.
doi:10.1016/S0098-3004(00)00057-1.
- Bivand, R. S. (2003) Approaches to Classes for Spatial Data in R. In: Hornik, K., Leisch, F., and Zeileis, A. (eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, March 20-22, Vienna, Austria. Vienna, Austria.
- Dhar, V. (2013) Data science and prediction. *Communications of the ACM*, 56(12), p. 64.
- Hijmans, R. J. (2017) *raster: Geographic Data Analysis and Modeling*. R package version 2.6-7.
URL: <https://CRAN.R-project.org/package=raster>
- Ihaka, R. and Gentleman, R. (1996) R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, 5(3), pp. 299–314.
doi:10.1080/10618600.1996.10474713.
- Pebesma, E. (2017) *sf: Simple Features for R*. R package version 0.5-5.
URL: <https://CRAN.R-project.org/package=sf>
- Pebesma, E. and Bivand, R. (2005) Classes and methods for spatial data in R. *R-NEWS*, 5(2), pp. 9–13.
- Wickham, H. (2014) Tidy data. *Journal of Statistical Software*, 46(10), pp. 1–23.

Dotazy?
Děkuji za pozornost.