ggplot2

Efektní vizualizace dat v prostředí jazyka R

Martin Golasowski

8. prosince 2016

Jak vizualizovat?

Požadované vlastnosti nástroje

- opakovatelnost, spolehlivost
- separace formy a obsahu
- flexibilita, jednoduchost

Jak vizualizovat?

Požadované vlastnosti nástroje

- opakovatelnost, spolehlivost
- separace formy a obsahu
- flexibilita, jednoduchost

Dostupné nástroje

- Matlab, IBM SPSS, Statgraphics

Jak vizualizovat?

Požadované vlastnosti nástroje

- opakovatelnost, spolehlivost
- separace formy a obsahu
- flexibilita, jednoduchost

Dostupné nástroje

- Matlab, IBM SPSS, Statgraphics
- MS Excel

Jak vizualizovat?

Požadované vlastnosti nástroje

- opakovatelnost, spolehlivost
- separace formy a obsahu
- flexibilita, jednoduchost

Dostupné nástroje

- Matlab, IBM SPSS, Statgraphics
- MS Excel
- GNUPlot, matplotlib, **ggplot2**

Předpoklady

- Základní znalost R (syntax, datové typy)
- Barevný monitor
- Oblíbené IDE (RStudio, RkWard) nebo textový editor

Materiály ke stažení

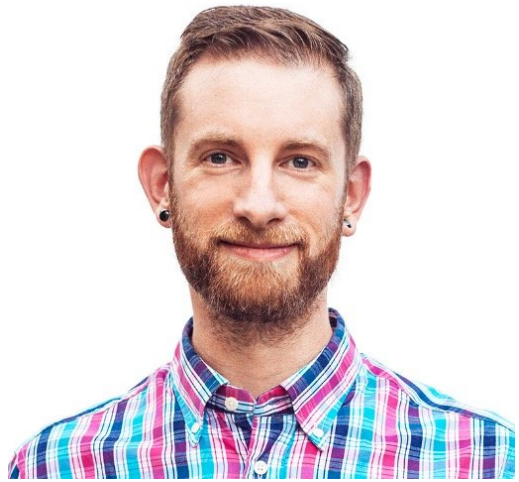
http://golasowski.com/R_workshop.zip

Vznik ggplot2

Hlavní autor: Hadley Wickham

Rice University, Stanford, Auckland, RStudio

- specifikace jazyka pro vizualizaci
- separace jednotlivých elementů grafu
- jednoduchost, znovupoužitelnost



Kniha: WICKHAM, Hadley. ggplot2: elegant graphics for data analysis. Springer Science & Business Media, 2009.

Jak získat ggplot2?

Požadavky

- R (verze alespoň 3.1)

Jak získat ggplot2?

Požadavky

- R (verze alespoň 3.1)

Instalace

- Dostupný v CRAN pro Linux, OSX, Windows, (Solaris?)
- https://cran.r-project.org/web/packages/ggplot2/index.html
- Ručně - `install.packages("ggplot2")`
- Přes vaše oblíbené IDE

Jak získat ggplot2?

Požadavky

- R (verze alespoň 3.1)

Instalace

- Dostupný v CRAN pro Linux, OSX, Windows, (Solaris?)
- https://cran.r-project.org/web/packages/ggplot2/index.html
- Ručně - `install.packages("ggplot2")`
- Přes vaše oblíbené IDE

Načtení balíčku

- `library("ggplot2")`

Toť vše.

Elementy grafu v ggplot2

- Data - std. datový formát (reshape2, stack, ...)
- Mapování proměnných - aesthetics
- Vrstvy - geoms, stats
- Styly - vlastnosti textových a grafických elementů (theme)

Vygenerování grafu

```
ggplot(data, mapping = aes(x = ...) +
         geom_point() +
         labs(x = "Sample", y = "Power")
```

Rozhraní

Zjednodušené

Základní funkce, užitečné pro jednoduché one-shot grafy

```
qplot(x, y = NULL, geom, xlab, ylab, main, ...)
```

Kompletní

Všechny funkce, komplexní parametry

```
ggplot(data = NULL, mapping = aes(), ...) +
        geom_* +
        scale_* +
        ...
```

Data

Data je nutné převést do standardního datového formátu. Ideální typ pro uložení dat je `data.frame`. Řádky obsahují jednotlivá pozorování. Jeden sloupec obsahuje hondoty všech závislých proměnných, ostatní sloupce obsahují jejich označení a další proměnné.

Původní data

	time.id	O4BRUS00	O4PETR00	O4VLCO00	O4FULN00	O4VRES00	O4NOJI00	O4257000
1	2016-04-22 18:00:00	4.758770	3.150053	1.700243	0.783445	16.90137	0.940827	17.00000
2	2016-04-22 19:00:00	4.750885	3.145676	1.697545	0.781079	16.89888	0.937986	16.99782
3	2016-04-22 20:00:00	4.741676	3.141312	1.693571	0.778721	16.89228	0.935154	16.99194
4	2016-04-22 21:00:00	4.732494	3.136961	1.688462	0.776370	16.87974	0.932331	16.98057
5	2016-04-22 22:00:00	4.723339	3.132623	1.683368	0.774026	16.86306	0.929516	16.96462
6	2016-04-22 23:00:00	4.714213	3.128299	1.678289	0.771689	16.84178	0.926709	16.94418
7	2016-04-23 00:00:00	4.705115	3.123989	1.673226	0.769359	16.81725	0.923912	16.92016
8	2016-04-23 01:00:00	4.695123	3.118358	1.668178	0.767036	16.79003	0.921122	16.89332
9	2016-04-23 02:00:00	4.683352	3.110718	1.663146	0.764720	16.75993	0.918341	16.86364
10	2016-04-23 03:00:00	4.670178	3.101926	1.658128	0.762411	16.72630	0.915568	16.83060

Po převodu

	time.id	Station	value
1	2016-04-22 18:00:00	O4BRUS00	4.758770
2	2016-04-22 19:00:00	O4BRUS00	4.750885
3	2016-04-22 20:00:00	O4BRUS00	4.741676
4	2016-04-22 21:00:00	O4BRUS00	4.732494
5	2016-04-22 22:00:00	O4BRUS00	4.723339
6	2016-04-22 23:00:00	O4BRUS00	4.714213
7	2016-04-23 00:00:00	O4BRUS00	4.705115
8	2016-04-23 01:00:00	O4BRUS00	4.695123
9	2016-04-23 02:00:00	O4BRUS00	4.683352
10	2016-04-23 03:00:00	O4BRUS00	4.670178

Převod dat pomocí balíku reshape

	time.id	O4BRUS00	O4PETR00	O4VLCO00	O4FULN00	O4VRES00	O4NOJI00	O4257000
1	2016-04-22 18:00:00	4.758770	3.150053	1.700243	0.783445	16.90137	0.940827	17.00000
2	2016-04-22 19:00:00	4.750885	3.145676	1.697545	0.781079	16.89888	0.937986	16.99782
3	2016-04-22 20:00:00	4.741676	3.141312	1.693571	0.778721	16.89228	0.935154	16.99194
4	2016-04-22 21:00:00	4.732494	3.136961	1.688462	0.776370	16.87974	0.932331	16.98057
5	2016-04-22 22:00:00	4.723339	3.132623	1.683368	0.774026	16.86306	0.929516	16.96462
6	2016-04-22 23:00:00	4.714213	3.128299	1.678289	0.771689	16.84178	0.926709	16.94418
7	2016-04-23 00:00:00	4.705115	3.123989	1.673226	0.769359	16.81725	0.923912	16.92016
8	2016-04-23 01:00:00	4.695123	3.118358	1.668178	0.767036	16.79003	0.921122	16.89332
9	2016-04-23 02:00:00	4.683352	3.110718	1.663146	0.764720	16.75993	0.918341	16.86364
10	2016-04-23 03:00:00	4.670178	3.101926	1.658128	0.762411	16.72630	0.915568	16.83060

```
melt(dataset, id.vars = c("time.id"), variable_name = "Station")
```

	time.id	Station	value
1	2016-04-22 18:00:00	O4BRUS00	4.758770
2	2016-04-22 19:00:00	O4BRUS00	4.750885
3	2016-04-22 20:00:00	O4BRUS00	4.741676
4	2016-04-22 21:00:00	O4BRUS00	4.732494
5	2016-04-22 22:00:00	O4BRUS00	4.723339
6	2016-04-22 23:00:00	O4BRUS00	4.714213
7	2016-04-23 00:00:00	O4BRUS00	4.705115
8	2016-04-23 01:00:00	O4BRUS00	4.695123
9	2016-04-23 02:00:00	O4BRUS00	4.683352
10	2016-04-23 03:00:00	O4BRUS00	4.670178

Elementy grafu - aesthetics

Přiřazení k parametru mapping pomocí parametru funkce `aes`.
Každý `geom` bere v potaz jiné elementy.

- x,y
- color - barva prvků (bodů, čar)
- shape - tvar bodů (puntíky, kolečka, ...)
- linetype - čára (čárkovná, přerušovaná, ...)
- size - velikost elementu
- alpha - průhlednost elementu
- fill - barva (typ?) výplně
- ...a mnoho dalších

```
aes(x = cut, y = depth, fill = clarity)
```

Geometrické objekty

Tvoří jednotlivé vrstvy grafu, berou v potaz mapování na elementy a mohou je předefinovat.

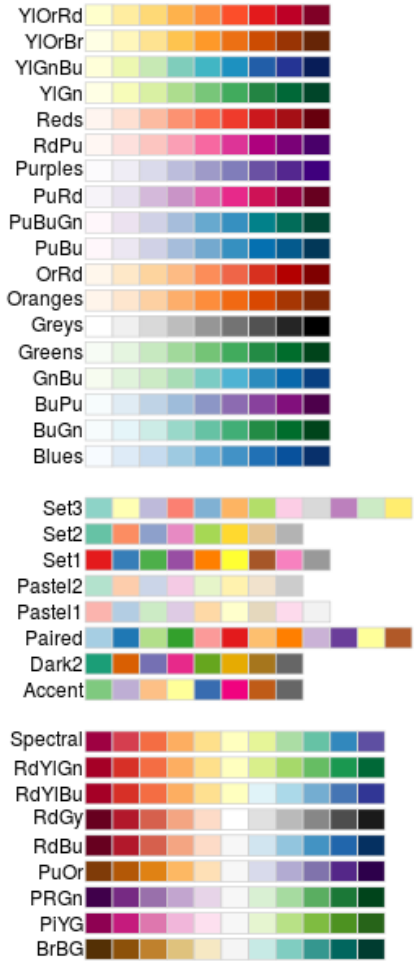
- geom_point - scatterplot, bodový graf
- geom_line - liniový graf
- geom_tile - rastry, dlaždice
- geom_bar - sloupcový graf
- geom_density - odhad hustoty pravděpodobnosti
- ...

```
geom_point(mapping = aes(x = cut, y = depth, fill =
clarity), size = 2.0)
```

Vzhled grafu

Další funkce pro ruční specifikaci vlastností grafu, které se nevztahují k datům.

- scale_* - vlastnosti os, rozsah, zobrazené hodnoty, atd.
- coord_* - systém souřadnic, otáčení, měřítko, změna
- labs - popisky
- theme - přímý přístup k elementům šablony
- annotation_* - dodatečné textové informace



Užitečné zdroje informací

R - obecně

- Advanced R by H. Wickham - http://adv-r.had.co.nz

ggplot2

- Dokumentace ggplot2 - http://docs.ggplot2.org/current/index.html
- R Graphics Cookbook by W. Chang - http://www.cookbook-r.com/Graphs/

https://www.r-bloggers.com/, Stack Overflow, http://stats.stackexchange.com/, ...