

# Zadania 2019

## Pravidlá:

1. Môžete použiť ľubovoľné dáta (zo svojej bakalárky či diplomovky, z Internetu, dáta z R, niektoré z odporúčaných datasetov, viď nižšie).
2. Máte za úlohu niečo podstatné o svojich dátach v R zistiť a prezentovať to. *Prezentovať* znamená, že predložíte R notebook s funkčným kódom, zrozumiteľným komentárom a výrečnou grafikou.
3. Použijete *tidyverse*, teda *ggplot2* na kreslenie obrázkov, *tidyr* / *dplyr* na manipuláciu s dátami a pod.
4. Víťané sú aplikácie bootstrapu a permutačných testov.
5. Na kurze sme toho stihli málo, využijete príležitosť naučiť sa niečo nové.
6. Pre svoj notebook a dáta si vytvoríte GitHub repozitár, a pošlete mi naň link. Ak sa pokúsite poslať mi nejaké súbory v e-maili, buď budem taký e-mail ignorovať, alebo vám pošlem toto poučenie. Nižšie nájdete podrobný postup, čo máte urobiť.

## Kde nájsť dáta

Zdrojov je viac. Vyhľadajte, čo sa vám páči, alebo použijete vlastné dáta.

## Štandardné dáta z R

Ak nenájdete vhodné dáta, môžete použiť datasety zabudované v balíčkoch R. Prakticky každý balíček má pribalené nejaké dáta, na ktorých sa štandardne demonštruje jeho funkčnosť.

## Príklad: dataset `anscombe`

Tak napríklad R base obsahuje slávny dataset `anscombe`, na ktorom sa môžete do systosti vyblázniť. Skúsíte `help("anscombe")`, potom

```
anscombe

##      x1 x2 x3 x4      y1      y2      y3      y4
## 1   10 10 10  8   8.04  9.14   7.46   6.58
## 2    8  8  8  8   6.95  8.14   6.77   5.76
## 3   13 13 13  8   7.58  8.74  12.74   7.71
## 4    9  9  9  8   8.81  8.77   7.11   8.84
## 5   11 11 11  8   8.33  9.26   7.81   8.47
## 6   14 14 14  8   9.96  8.10   8.84   7.04
## 7    6  6  6  8   7.24  6.13   6.08   5.25
## 8    4  4  4 19   4.26  3.10   5.39  12.50
## 9   12 12 12  8  10.84  9.13   8.15   5.56
## 10   7  7  7  8   4.82  7.26   6.42   7.91
## 11   5  5  5  8   5.68  4.74   5.73   6.89
```

a správne usúdite, že aby ste dáta elegantne nakreslili, potrebujete ich najprv preorganizovať. Napríklad takto:

```
library(tidyverse)

## -- Attaching packages ----- tidyverse 1.2.1 --
## v ggplot2 3.2.1      v purrr   0.3.3
## v tibble  2.1.3      v dplyr   0.8.3
## v tidyr   1.0.0      v stringr 1.4.0
## v readr   1.3.1      v forcats 0.4.0
```

```
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag() masks stats::lag()
```

```
anscombe %>%
  gather(label, value) %>% # label: názvy stĺpca (x1, x2, ... y4), value: hodnoty
  mutate(i = rep(1:11,8)) %>% # nový stĺpec s poradím hodnôt
  separate(label, into = c("xy", "set"), sep = 1) %>% # oddelíme z label 1. znak, čo je 'x' alebo 'y'
  spread(xy, value) %>% # teraz rozdelíme stĺpec value do dvoch stĺpcov x a y podľa stĺpca xy.
  select(-i) -> my_anscombe # stĺpec i zahodíme.
my_anscombe
```

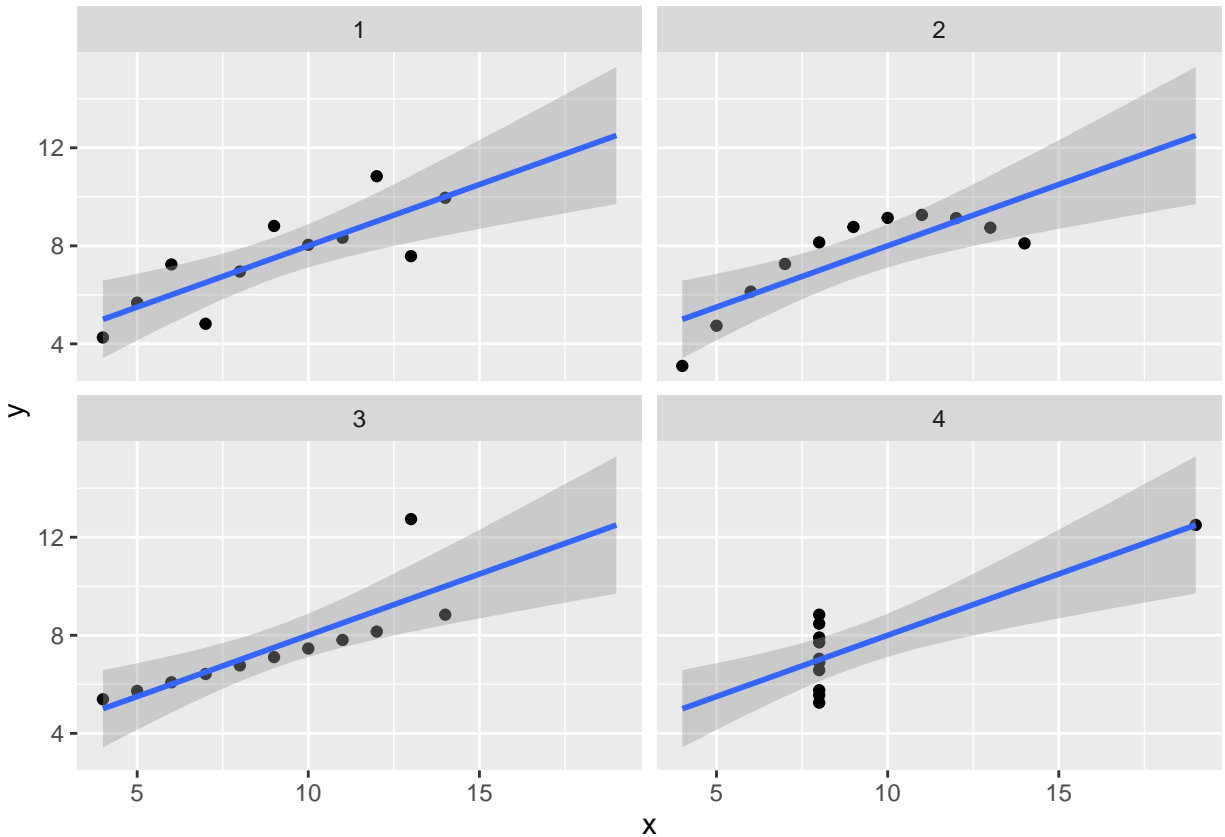
```
##   set  x    y
## 1    1 10  8.04
## 2    1  8  6.95
## 3    1 13  7.58
## 4    1  9  8.81
## 5    1 11  8.33
## 6    1 14  9.96
## 7    1  6  7.24
## 8    1  4  4.26
## 9    1 12 10.84
## 10   1  7  4.82
## 11   1  5  5.68
## 12   2 10  9.14
## 13   2  8  8.14
## 14   2 13  8.74
## 15   2  9  8.77
## 16   2 11  9.26
## 17   2 14  8.10
## 18   2  6  6.13
## 19   2  4  3.10
## 20   2 12  9.13
## 21   2  7  7.26
## 22   2  5  4.74
## 23   3 10  7.46
## 24   3  8  6.77
## 25   3 13 12.74
## 26   3  9  7.11
## 27   3 11  7.81
## 28   3 14  8.84
## 29   3  6  6.08
## 30   3  4  5.39
## 31   3 12  8.15
## 32   3  7  6.42
## 33   3  5  5.73
## 34   4  8  6.58
## 35   4  8  5.76
## 36   4  8  7.71
## 37   4  8  8.84
## 38   4  8  8.47
## 39   4  8  7.04
## 40   4  8  5.25
## 41   4 19 12.50
## 42   4  8  5.56
```

```
## 43  4  8  7.91
## 44  4  8  6.89
```

Asi by ste potrebovali poradiť, že musíte zriadiť stĺpec `i`, inak `'spread'` nebude fungovať. Ja som potreboval :-)

Teraz už môžeme kresliť:

```
ggplot(data = my_anscombe, aes(x = x, y = y)) +
  facet_wrap(~ set) +
  geom_point() +
  geom_smooth(method = "lm", fullrange = T)
```



Nafitované priamky sú rovnaké, vrátane konfidenčných hraníc. Vaša úloha by začala tu: konfidenčné pásy určite nie sú v poriadku, pretože pochádzajú z predpokladu o rozdelení chýb hodnôt  $y$ , ktoré okrem jedného prípadu neplatia. Takže by bolo elegantné použiť bootstrap reziduálov a ukázať, že chyby koeficientov sú v skutočnosti oveľa väčšie.

Ak to ukážete, máte skúšku za sebou.

Ak sa rozumne pohráte s diagnostikou regresie a dokázate vysvetliť, čo je na jednotlivých regresiach zle a ako by sa to dalo napraviť, takisto máte skúšku za sebou.

**(Koniec príkladu)**

### Balíček `datasets`

Zbierku dát obsahuje balíček `datasets`. Pozrite si dokumentáciu a vyberte si vhodné dáta. Nájdete pomerne veľa dát s biologickou alebo fyzikálno-chemickou tematikou.

```
library(datasets)
library(help = "datasets")
```

Ak treba, balíček si doinštalujte. Ku každým dátam existuje help, dokonca s demonštráciou kódu. Kludne tento kód využijete, ale skúste ho prepísať tak, aby ste použili metódy z `tidyverse` a hlavne pridali hodnotný vlastný príspevok.

## Dáta z iných zdrojov

Dáta dnes nájdete na mnohých miestach, napríklad v novinách.

### Príklad: Kandidáti do parlamentných volieb 2020

Zoznam všetkých kandidátov nájdete tu: <https://www.minv.sk/?nr20-kandidati3>. Dáta sú ale v pdf, takže ako prvý krok potrebujete nájsť spôsob, ako z nich vyextrahovať údaje. Môžete napríklad skúsiť Google.

V repozitári kurzu v podadresári `data/kandidati_2020` nájdete niekoľko vyextrahovaných tabuliek, obsahujúcich dáta, použité v článku Daniela Kerekesa v Denníku N (<https://dennikn.sk/1684137/najviac-pravnikov-ide-za-ps-spolu-ucitelov-a-lekarov-za-kdh-najmenej-titulov-maju-u-kotlebu-volebne-infografiky/?ref=suv> - je možné, že článok je pod paywallom). Sú to menšie súbory dát extrahované z kandidátnych listín. U nich budete musieť vyriešiť problém s kódovaním, a ešte vymyslieť nejaký inovatívny pohľad či porovnanie nad rámec toho, čo je v novinách - napríklad zoskupenie strán podľa podobnosti nejakých parametrov. Upozorňujem, že ak chcete ukázať svoje štatistické svaly na testoch hypotéz, potom toto nie sú dáta pre vás.

### (Koniec príkladu)

## Odkazy

Pridávam zoznam odkazov, kde môžete nájsť zaujímavú látku na skúmanie. Poradie vyzerá čudne, pretože je úplne náhodné:

Štatistické testy ako lineárne modely: <https://lindeloev.github.io/tests-as-linear/>

Simpsonov paradox: <https://paulvanderlaken.com/2017/09/27/simpsons-paradox-two-hr-examples-with-r-code/>  
<https://towardsdatascience.com/simpsons-paradox-d2f4d8f08d42>

Kvantilové grafy (Q-Q plot): <https://seankross.com/2016/02/29/A-Q-Q-Plot-Dissection-Kit.html>

New York Times: What's going on in this graph? Vyberte si graf a skúste odpovedať na otázky: Čo vidím? Čo by som ešte chcel vedieť? Čo sa tu deje? <https://www.nytimes.com/2019/08/27/learning/looking-for-graphs-to-use-in-the-classroom-here-are-34.html>

Lavielle: Statistics in Action with R - učebné materiály. Naštudujte teóriu (napríklad viacnásobné porovnania) a uplatnite na dátach z cvičenia alebo ľubovoľných iných. <http://sia.webpopix.org/index.html>

Moja Pinterestová nástenka `My_R` obsahuje množstvo odkazov, kludne si nájdite inšpiráciu. [https://sk.pinterest.com/peterkvasnika/my\\_r/](https://sk.pinterest.com/peterkvasnika/my_r/)

## Ja ale fakt neviem, čo si mám vybrať!

Tuho sa zamyslite, či by ste nevedeli využiť nejaké svoje dáta. Je to najužitočnejšie a máte najväčšiu motiváciu sa niečo naučiť.

Nájdite si dáta o niečom, čo vás zaujíma. Trebárs o uhlíkových emisiách, futbalové štatistiky, prieskumy preferencií, čo len chcete, nájsť dáta na Internete je dnes ľahké, stačí sa spýtať Googlu.

Ak stále neviete, pozrite si zdroje, uvedené vyššie. Osvedčený postup je skombinovať dve témy, napríklad

- Anscombe dáta a článok o vizualizácii reziduálov z môjho Pinterestu, <https://www.r-bloggers.com/visualising-residuals/>

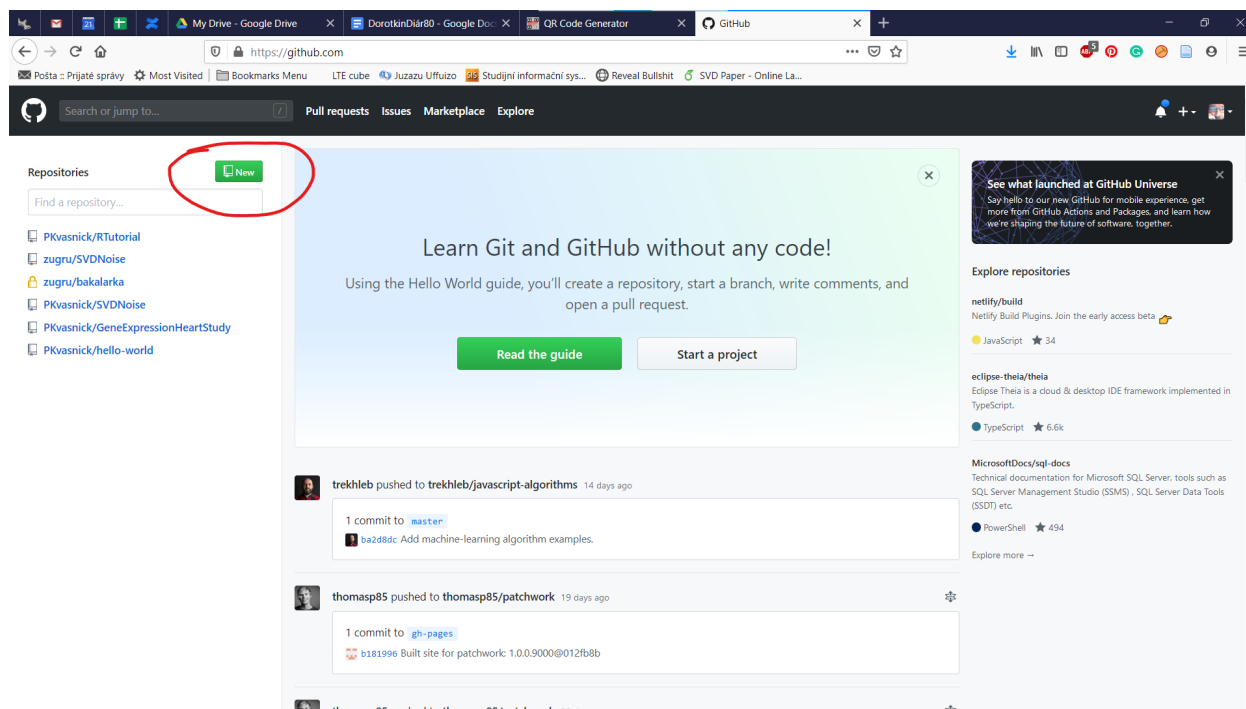


Figure 1: Kliknutím na “New” vytvoríte nový repozitár.

- článok o zmiešaných modeloch <https://www.r-bloggers.com/getting-started-with-mixed-effect-models-in-r/> a niektoré dáta, ktoré si pýtajú dvojfaktorovú ANOVu
- tutorial o kreslení animovaných grafov **gganimate**, <https://gganimate.com/articles/gganimate.html> s niektorými komplexnejšími dátami, napríklad datasetom z balíka **gapminder** alebo z datasetu o letoch z/do NYC, ktorým sme sa zaoberali na prednáške.

Ak si *naozaj* neviete rady, napíšte mi a niečo nájdeme či vymyslíme.

## Git

Prosím pozrite si časť *Github Desktop Workflow* v dokumente *README.md* v koreňovom adresári repozitára tohoto kurzu (<https://github.com/PKvasnick/RTutorial>)

Jediné, čo musíte urobiť nové, je založiť si vlastný repozitár a primapovať ho k adresáru na vašom počítači. Postup:

1. Chodte na <https://github.com> a prihláste sa alebo zaregistrujte.
2. V ľavom ráme, zobrazujúcom všetky vaše repozitáre, kliknite na “New”. Vyplňte formulár a kliknite na zelené tlačidlo, aby sa vytvoril nový repozitár.

Ospravedlňujem sa za rozutekané obrázky ku kroku 2, ale nechcelo sa mi to už naprávať.

## Nefunguje to!

Ak narazíte na problém a nedarí sa vám nájsť pomoc v helpe R alebo na Internete ([stackoverflow.com](https://stackoverflow.com)), neštrácajte čas a ozvite sa, [peter.kvasnicka@mff.cuni.cz](mailto:peter.kvasnicka@mff.cuni.cz). Potrebujem ale vidieť váš kód, inak vám s veľkou pravdepodobnosťou nebudem vedieť poradiť.

rkx Menu    LTE cube    Juzazu Uffuizo    Studijní informační sys...    Reveal Bullshit    SVD Paper - Online La...


## Create a new repository

A repository contains all project files, including the revision history. Already have a project repository elsewhere? [Import a repository.](#)

---

Owner

Repository name \*

 PKvasnick ▾

 / 

skuska-R čokolívek rozumné ✓

Great repository names are short and memorable. Need inspiration? How about **super-fiesta**?

Description (optional)

Kód a data ku skúške z R, 2019 alebo čokolívek iné alebo nič

---

☒ **Public** ✓  
Anyone can see this repository. You choose who can commit.

☐ **Private**  
You choose who can see and commit to this repository.

---

Skip this step if you're importing an existing repository.

☒ **Initialize this repository with a README** Dobrá vec, môžete stručne popísať svoj projekt a prečítať si markdown.  
This will let you immediately clone the repository to your computer.

Add .gitignore: **R** ▾

 | 

Add a license: **None** ▾ ✓ i


---

Create repository

Stlačením vytvoríte adresár repozitára, default je pod Documents/GitHub, a tam si môžete skopírovať data a začať vkladať kód.

---

9 GitHub, Inc.    Terms    Privacy    Security    Status    Help



Contact GitHub    Pricing    API    Training    Blog    About

Figure 2: Zadájte vlastnosti nového repozitára.

## **Hodnotenie a zápis do indexov**

Z tohoto predmetu dostanú všetci “A”. Ak nebudem s niektorou prezentáciou spokojný, vrátim ju.

Zápis do indexov sa uskutoční niekedy koncom januára, v pondelok alebo v piatok. Prosím skúste sa zhodnúť na jedinom termíne. V prípade nutnosti stačí, ak po spolužiakovi pošlete index.