Digital humanities

Cvičení: transformace a analýza dat v R

Jindřich Marek

Konvence v tomto souboru

- skript v R
  - zapisovaný do souboru

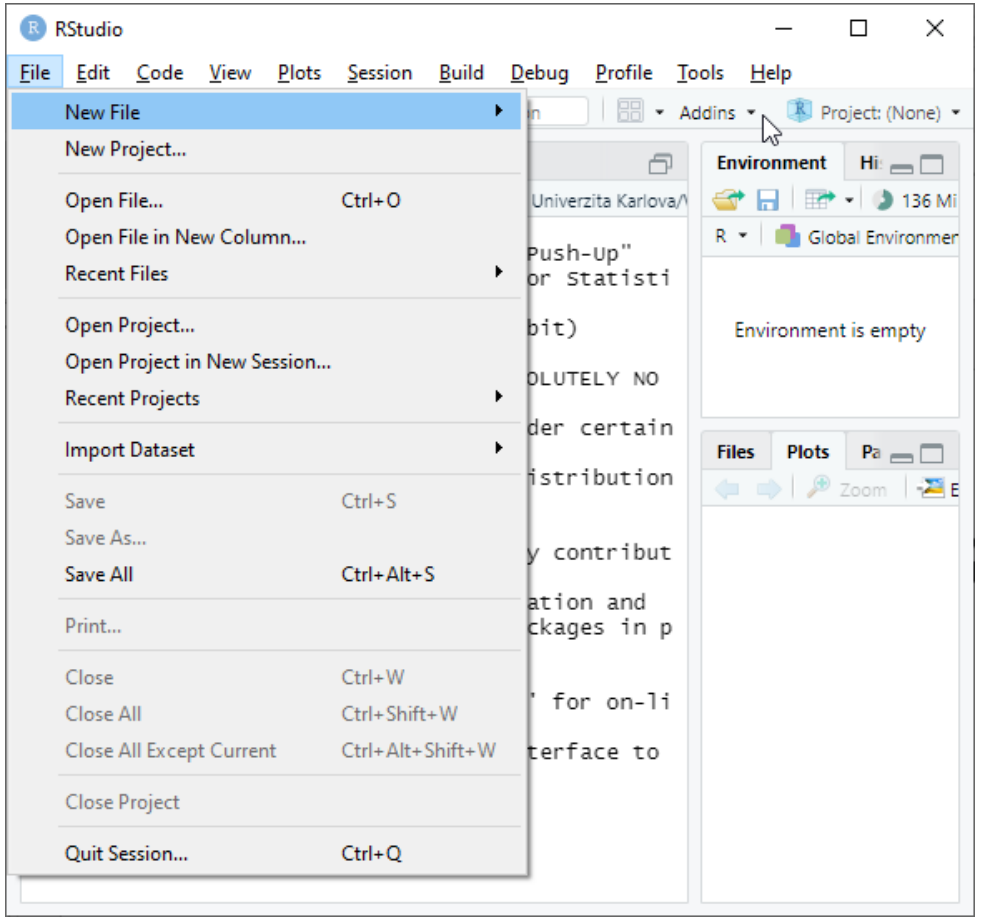
```
ggsave("rk_ggplot.png")
```

- příkazy v R
  - zapisované na konzoli

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

Založení nového souboru

- RStudio
  - File > New File > R Script
  - File > Save As…
  - název souboru: analyza.R
    - /cloud/project
  - online: nahrajte soubor
  - offline: do stejné složky umístěte soubor ve formátu MS Excel, který jsme minule upravovali v OpenRefine



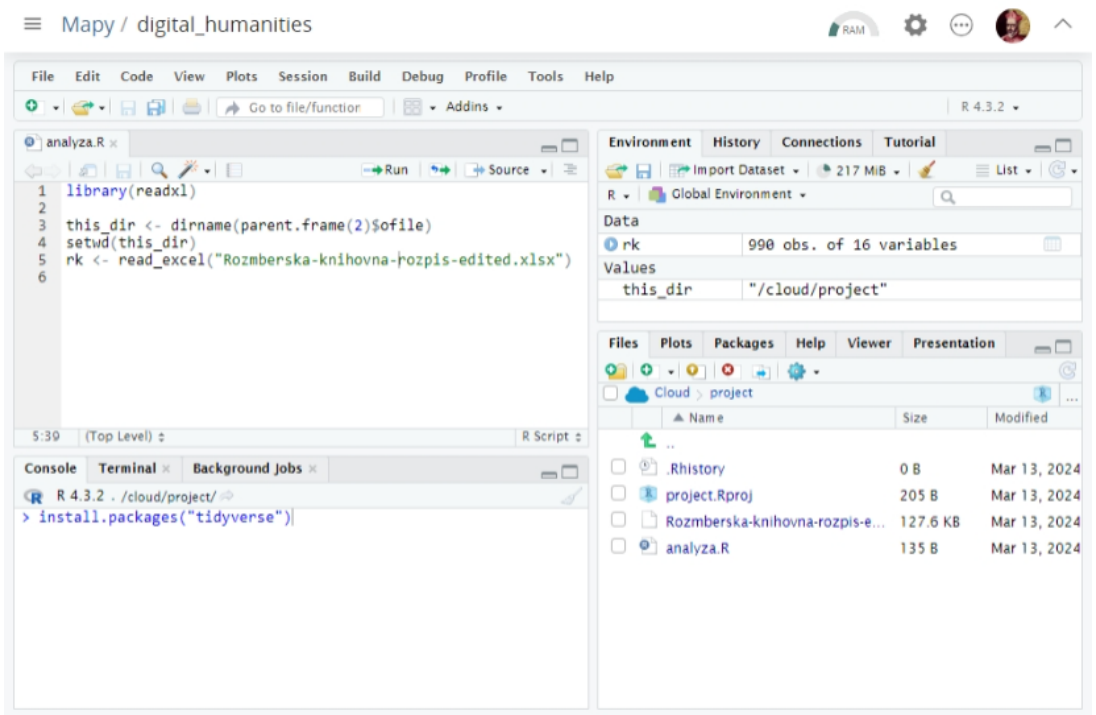
Instalace balíků

- balík, který umí načíst soubor ve formátu MS Excel

> install.packages("readxl")

- tidyverse

> install.packages("tidyverse")

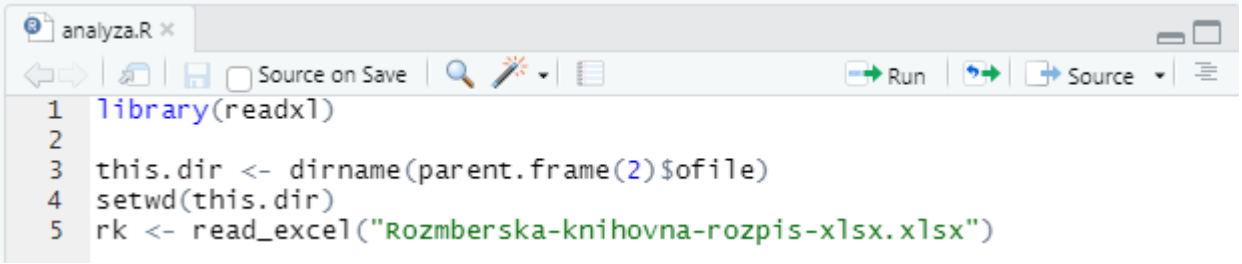


Načtení dat vyčištěných v OpenRefine

```
library(readxl)
this_dir <- dirname(parent.frame(2)$ofile)
setwd(this_dir)
rk <- read_excel("Rozmberska-knihovna-rozpis-edited.xlsx")
```

pozor na název souboru: musí se shodovat (+ MS Windows ve výchozím nastavení skrývají přípony souborů)

- kliknout na Source



File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/functior | Addins | R 4.3.2

analyza.R | rk

Filter

	Č.	Katalog	Obor/číslo	Fol.	Autor	Název	Datace	D1	D2
1	1.0	Březan	Teologie	10r	NA	Bibl' staročeské psané na pergamenė (jenž někdy náležela p...	1400-1450	1400	
2	2.0	Březan	Teologie	10v	NA	Biblia Hebraica	NA	NA	
3	3.0	Březan	Teologie	10v	NA	Bibliorum pars una, complectens Pentateuchum, Josue, Judi...	1350-1450	1350	
4	4.0	Březan	Teologie	10v	NA	Bibliorum volumen integrum, crassum, scriptum in papyro 1...	1473	-1	
5	5.0	Březan	Teologie	10v	NA	Bibliorum Veteris testamenti pars prior 1461. In fine genealo...	1461	-1	
6	6.0	Březan	Teologie	10v	NA	Codex unus confuse scriptus, habet quosdam libros Veteris t...	15. století	1401	
7	7.0	Březan	Teologie	10v	NA	Novum testamentum scriptum quater I. II. III. IV.	15. století	1401	
8	8.0	Březan	Teologie	13v	Augustin, svatý, 354-430	Ejusdem [Augustini] libri manuscripti ante annos CC - Com...	1350-1450	1350	
9	9.0	Březan	Teologie	13v	Augustin, svatý, 354-430	Homeliae in D. Johannem Evangelistam scriptae 1381	1381	-1	
10	10.0	Březan	Teologie	13v	Bartolomeus de Urbino	Milleloquiorum sancti Augustini Pars I. II. III.	1387	-1	
11	11.0	Březan	Teologie	13v	Augustin, svatý, 354-430	De civitate Dei liber	1381	-1	
12	12.0	Březan	Teologie	13v	Augustin, svatý, 354-430	Codex continens epistolam ad comitem et libros de Trinitate	14. století	1301	
13	13.0	Březan	Teologie	13v	NA	D. Augustini vita, in principio desiderantur quaedam, una cu...	NA	NA	
14	14.0	Březan	Teologie	14r	Ambrož, svatý, 339-397	Commentarius in psalmum Beati immaculati etc. 118	NA	NA	
15	15.0	Březan	Teologie	14v	Bernard z Clairvaux, svatý, asi 1090-1153	Homeliae in Cantica canticorum, duobus voluminibus compl...	NA	NA	
16	16.0	Březan	Teologie	14v	Bernard z Clairvaux, svatý, asi 1090-1153	Epistolae et quidam tractatus	NA	NA	
17	17.0	Březan	Teologie	14v	Bernard z Clairvaux, svatý, asi 1090-1153	Meditatio super Salve regina, Flores excerpti ex operibus Be...	NA	NA	
18	18.0	Březan	Teologie	15r	Chrysostomus, Ioannes	Expositiones quaedam super Matthaeum	15. století	1401	
19	19.0	Březan	Teologie	15v	Cassiodorus, Flavius Magnus Aurelius, asi 490-asi 583	In septem psalmos poenitentiales, vide De Trinitate Theol.	NA	NA	
20	20.0	Březan	Teologie	16r	Řehoř, I., papež, asi 540-604	Super Cantica canticorum. Tractatus dictus Lignum vitae. Qu...	14. století	1301	
21	21.0	Březan	Teologie	16r	Řehoř, I., papež, asi 540-604	Enarrationem super Ezechielem et aliquot vitae sanctorum, ...	1200-1300	1200	
22	22.0	Březan	Teologie	16r	Řehoř, I., papež, asi 540-604	Moralia seu Commentarii super Job	12./13. století	1180	
23	23.0	Březan	Teologie	16r	Řehoř, I., papež, asi 540-604	Moralium pars prima desideratur / II. a capite 12 usque ad 23	NA	NA	
24	24.0	Březan	Teologie	16r	Řehoř, I., papež, asi 540-604	[Moralium pars] III. Reliqua usque ad finem	1372	-1	
25	25.0	Březan	Teologie	16r	Řehoř, I., papež, asi 540-604	Moralia	1383	-1	
26	26.0	Březan	Teologie	16r	Řehoř, I., papež, asi 540-604	Pastorale Gregorii, vide inter Tractatus	NA	NA	
27	27.0	Březan	Teologie	16v	Hieronymus, Sophronius Eusebius, 345-416	Super Psalmos. Huic quaedam alia adjuncta ut epistola beati...	1387	-1	

Showing 1 to 27 of 990 entries, 16 total columns

Console | Terminal | Background Jobs

R 4.3.2 · /cloud/project/

```
> View(rk)
>
```

Environment | History | Connections | Tutorial

Import Dataset | 232 MiB | List

R | Global Environment

Data

rk	990 obs. of 16 variables

Values

this_dir	"/cloud/project"

Files | Plots | Packages | Help | Viewer | Presentation

New Folder | New Blank File | Upload | Delete | Rename | More

Cloud › project

Name	Size	Modified
..		
.Rhistory	0 B	Mar 13, 2024, 2:38 PM
project.Rproj	205 B	Mar 13, 2024, 2:38 PM
Rozmberska-knihovna-rozpis-edited.xlsx	127.6 KB	Mar 13, 2024, 2:41 PM
analyza.R	135 B	Mar 13, 2024, 2:46 PM

Dataset

- rukopisy Rožmberské knihovny dle katalogu 1608

Otázky k datasetu

- obecné
  - jaký je rozsah dat?
  - jak jsou data strukturována?
  - jaké obsahují datové typy?
  - odpovídají datové typy sloupců jejich obsahu?
  - jsou data konzistentní?
  - jsou v datech zjevné chyby?
- konkrétní
  - na základě analýzy
  - kvantitativní distribuce hodnot, odchylky
  - vzájemná závislost jednotlivých údajů
  - …
  - vizualizace celku

Statistika

```
> summary(rk)
      Č.              Katalog          Obor/číslo           Fol.              Autor              Název          
 Length:990         Length:990         Length:990         Length:990         Length:990         Length:990        
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character  


    Datace                D1                D2          Místo vzniku          Jazyk           Ps.l., vazba      
 Length:990         Min.   :  -1.0   Min.   :  -1.0   Length:990         Length:990         Length:990        
 Class :character   1st Qu.:  -1.0   1st Qu.:  -1.0   Class :character   Class :character   Class :character  
 Mode  :character   Median :1380.0   Median :1420.0   Mode  :character   Mode  :character   Mode  :character  
                    Mean   : 928.9   Mean   : 970.7                                                          
                    3rd Qu.:1401.0   3rd Qu.:1500.0                                                          
                    Max.   :1610.0   Max.   :1612.0                                                          
                    NA's   :673      NA's   :674                                                             
     Obor            Poznámka          Identifikace       Provenience       
 Length:990         Length:990         Length:990         Length:990        
 Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character  
```

Statistika

```
> str(rk)
tibble [990 × 16] (S3: tbl_df/tbl/data.frame)
 $ Č.           : chr [1:990] "1.0" "2.0" "3.0" "4.0" ...
 $ Katalog      : chr [1:990] "Březan" "Březan" "Březan" "Březan" ...
 $ Obor/číslo   : chr [1:990] "Teologie" "Teologie" "Teologie" "Teologie" ...
 $ Fol.         : chr [1:990] "10r" "10v" "10v" "10v" ...
 $ Autor        : chr [1:990] NA NA NA NA ...
 $ Název        : chr [1:990] "Biblí staročeské psané na pergameně (jenž někdy náležela panům z Kunstátu) dílové dva I. II." "Biblia
Hebraica" "Bibliorum pars una, complectens Pentateuchum, Josue, Judicum, Ruth et quatuor lib. Regum ante CC annos scripta" "Biblior
um volumen integrum, crassum, scriptum in papyro 1473" ...
 $ Datace       : chr [1:990] "1400-1450" NA "1350-1450" "1473" ...
 $ D1           : num [1:990] 1400 NA 1350 -1 -1 ...
 $ D2           : num [1:990] 1500 NA 1450 -1 -1 1500 1500 1450 -1 -1 ...
 $ Místo vzniku: chr [1:990] "Morava?" NA NA NA ...
 $ Jazyk        : chr [1:990] "ces" "heb" "lat" "lat" ...
 $ Ps.l., vazba: chr [1:990] "perg; ilum, neúpl.; obsahuje i Žďárskou kroniku. Vazba prvního i druhého svazku shodná (převazba rožm
berské pro"| __truncated__ "perg" "perg" "pap" ...
 $ Obor         : chr [1:990] "bible" "bible" "bible" "bible" ...
 $ Poznámka     : chr [1:990] "Rukopisný záznam Toto Sau Knihy Panie Bozkowy dany Sau do kostela Swateho Matiege W miestie Bechynij
(1. sv., f"| __truncated__ NA NA NA ...
 $ Identifikace: chr [1:990] "Brno, MZA, G 10, č. 123/1-2; https://krigsbyte.lib.cas.cz/Record/3298000000516153" "možná BAV: 437 ne
bo 439 (v katalogu KB 1650, v jiném zdroji heb. Pentatuch nebyl)" NA NA ...
 $ Provenience : chr [1:990] "rodový; Petr Vok - Bechyně; Bočkovská bible vznikla v první polovině 15. stoletím, jejím objednavatel
em byl pra"| __truncated__ NA NA "Petr Vok - Bechyně 1573 \"Sacra Biblia Latina scripta in fol\" - může jí i o jiný záznam" ...
```

Oprava chyb

- ve sloupcích D1 a D2 je na některých místech místo limitů datace údaj „-1“
- je třeba je přepsat číslem ze sloupce Datace
- dále je třeba změnit formát sloupců D1 a D2 (zpět) na číselný

```
> rk$D1 <- with( rk, ifelse( D1 == "-1", Datace, D1 ) )
> rk$D2 <- with( rk, ifelse( D2 == "-1", Datace, D2 ) )
```

```
> rk$D1 <- as.numeric(rk$D1)
> rk$D2 <- as.numeric(rk$D2)
```

Zastoupení oborů v rukopisech RK (Obor)

```
> str(rk$Obor)

> rk$Obor <- as.factor(rk$Obor)
> str(rk$Obor)
> levels(rk$Obor)
> table(rk$Obor)
> as.data.frame(table(rk$Obor))

> library(dplyr)
> rk %>% count(Obor, sort = TRUE)
```

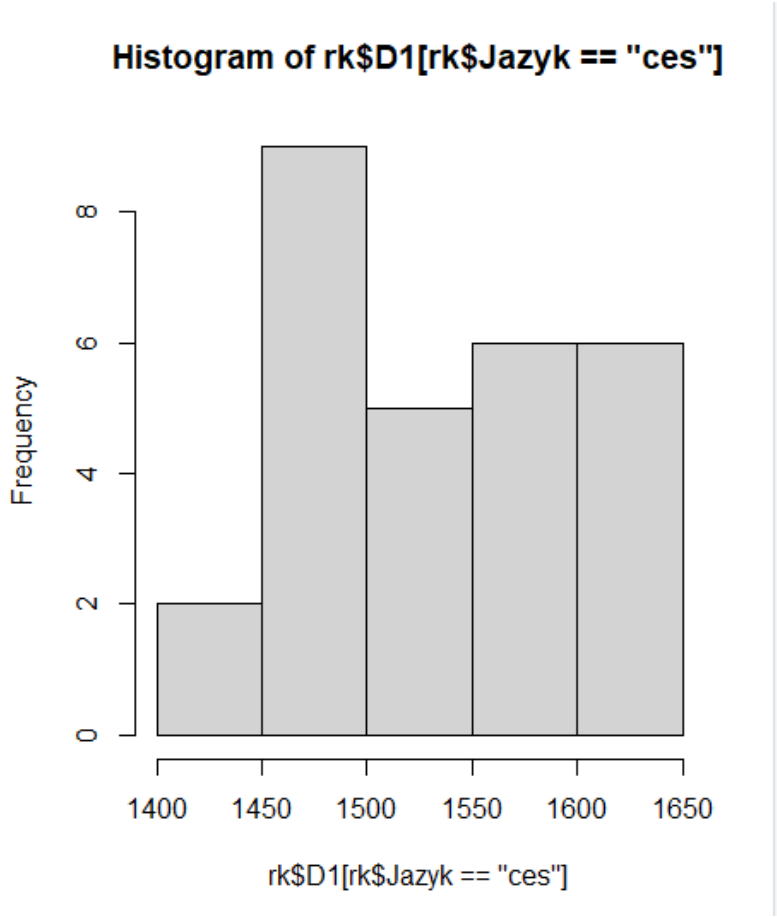
```
> rk %>% count(Obor, sort = TRUE)
# A tibble: 288 × 2
   Obor                          n
   <fct>                     <int>
 1 kazatelství                  90
 2 teologie praktická           68
 3 lékařství                    64
 4 exegeze biblická             46
 5 hagiografie                  31
 6 teologie                     30
 7 bible                        22
 8 alchymie                     21
 9 exegeze biblická, patristika 21
10 teologie dogmatická          21
# i 278 more rows
# i Use `print(n = ...)` to see more rows
```

Rozsah datace (D1, D2)

```
> mean(rk$D1, na.rm=TRUE)
> median(rk$D1, na.rm=TRUE)

> mean(rk$D2, na.rm=TRUE)
> median(rk$D2, na.rm=TRUE)

> sort(rk$D1)
> quantile(rk$D1, na.rm = TRUE)
```

Jednoduchá vizualizace datace (D1, Jazyk)

```
> hist(rk$D1[rk$Jazyk=="ces"])
> hist(rk$D1[rk$Jazyk=="deu"])
> hist(rk$D1[rk$Jazyk=="lat"])
```



Frekvence výskytu (Jazyk)

> prop.table(table(rk$Jazyk))

```
> prop.table(table(rk$Jazyk))

        ces ces, deu, lat      ces, lat          deu      deu, lat deu, lat, ces          eng      gre, lat          heb
0.097264438   0.001013171   0.004052685   0.085106383   0.003039514   0.001013171   0.001013171   0.001013171   0.001013171
        ita      ita, lat          lat     lat, ces lat, deu, ces      lat, ita          pol
0.001013171   0.001013171   0.792299899   0.007092199   0.002026342   0.001013171   0.001013171
>
```

Jedna hodnota ve sloupci (Jazyk, Obor)

```
> library(tidyr)
> rk2 <- separate_rows(rk,Jazyk,sep=",\
+")
> rk3 <- separate_rows(rk2,Obor,sep=",\
+")
> View(rk3)

> rk3$Jazyk <- as.factor(rk3$Jazyk)
> str(rk3$Jazyk)
> levels(rk3$Jazyk)
> rk3$Obor <- as.factor(rk3$Obor)
> levels (rk3$Obor)
```

- stejný proces se dá provést v OpenRefine při čištění hodnot
- některé řádky budou nově duplicitní, v každém řádku bude uveden jeden údaj pro Jazyk/Obor

Počet výskytů hodnoty: nový sloupec (Obor)

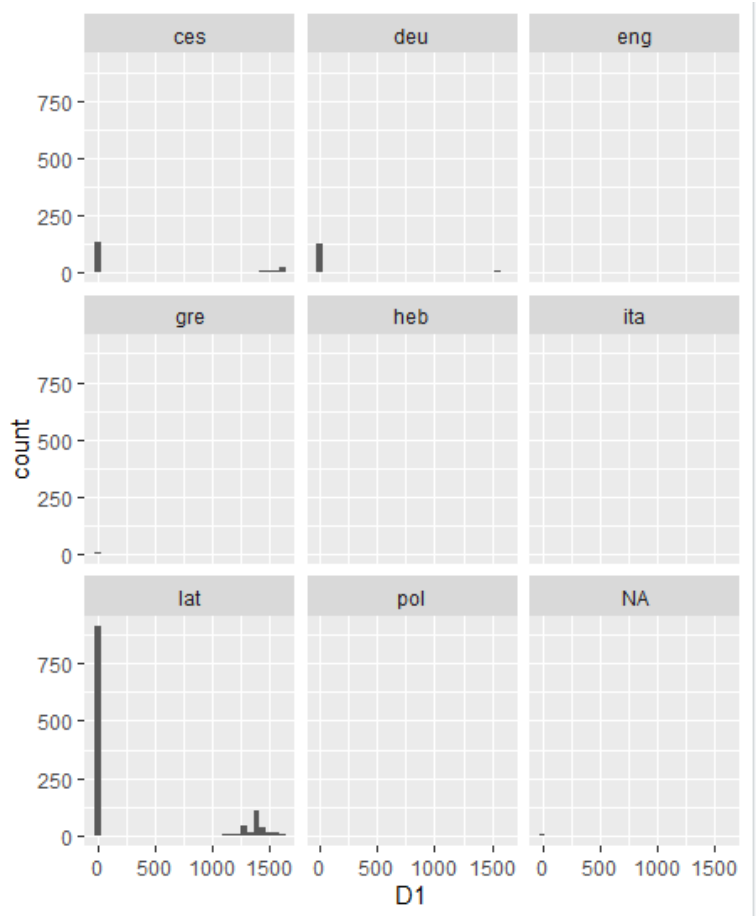
```
> rk4 <- rk3 %>% add_count(Obor, sort = TRUE)
```

Median v závislosti na dalších hodnotách (D1 ~ Jazyk)

```
> median(rk3$D1[rk3$Jazyk=="ces"], na.rm=TRUE)
> median(rk3$D1[rk3$Jazyk=="deu"], na.rm=TRUE)
> median(rk3$D1[rk3$Jazyk=="lat"], na.rm=TRUE)

> median(rk3$D1[rk3$Obor=="alchymie"], na.rm=TRUE)
```

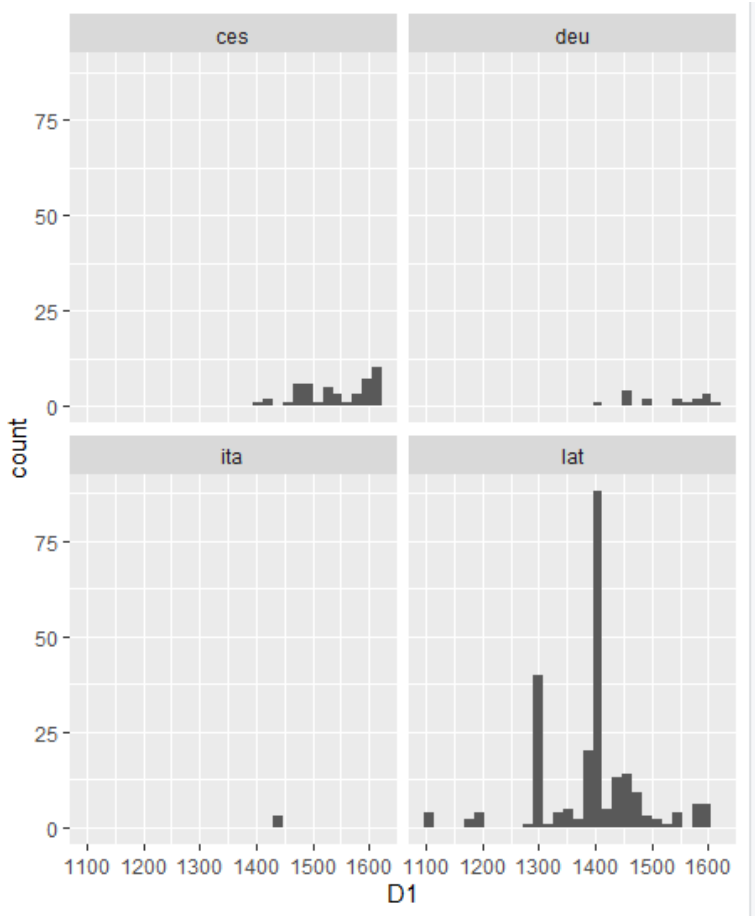
Fasetový graf (D1 ~ Jazyk, vč. nedat.)

```
library(ggplot2)

rk3$D1[is.na(rk3$D1)] = 0
rk3$D2[is.na(rk3$D2)] = 0

ggplot(data=subset(rk3,
!is.na(rk3$D1)), aes(x = D1)) +
  geom_histogram() +
  facet_wrap(~ Jazyk)
```



Fasetový graf (D1 ~ Jazyk, bez nedat.)

```
library(ggplot2)

rk3$D1[rk3$D1 == 0] <- NA
rk3$D2[rk3$D2 == 0] <- NA

ggplot(data=subset(rk3,
!is.na(rk3$D1)), aes(x = D1)) +
  geom_histogram() +
  facet_wrap(~ Jazyk)
```



Graf (D1 ~ Jazyk, n, Obor; více než 5 knih)

```
ggplot(rk4[which(rk4$n>5),],
aes(x=D1, y=Obor, color=Jazyk,
alpha=0.05)) +
  geom_count()
```