

Мастер академске студије
Рачунарство и аутоматика

Рачунарство високих перформанси
у информационом инжењерингу

Анализа података кроз језик R и колекцију пакета tidyverse

(материјали за предавања)

1. Колекција пакета tidyverse
2. Анализа података
3. Извори и литература

Колекција пакета tidyverse

Колекција пакета *tidyverse*

скуп пакета који нуди широку подршку раду над подацима
учитавање, снимање, чишћење, трансформисање, визуализовање...
одабрани принципи рада над подацима уграђени у пакете
аутори

Хадли Викам и бројни сарадници

RStudio

Интернет стране

<https://www.tidyverse.org/>

<https://cran.r-project.org/web/packages/tidyverse/index.html>

актуелна верзија

tidyverse 2.0.0 (22. фебруар 2023)

две групе пакета

основни пакети

додатни пакети

Колекција пакета tidyverse

Колекција пакета *tidyverse*

ОСНОВНИ ПАКЕТИ

пакет **dplyr**

пакет **forcats**

пакет **ggplot2**

пакет **purrr**

пакет **readr**

пакет **stringr**

пакет **tibble**

пакет **tidyr**

Колекција пакета tidyverse

Колекција пакета *tidyverse*

додатни повезани пакети

пакет **broom**

пакет **cli**

пакет **crayon**

пакет **dbplyr**

пакет **haven**

пакет **hms**

пакет **httr**

пакет **jsonlite**

пакет **lubridate**

пакет **modelr**

пакет **pillar**

пакет **readxl**

Колекција пакета tidyverse

Колекција пакета *tidyverse*

додатни повезани пакети

пакет **reprex**

пакет **rlang**

пакет **rstudioapi**

пакет **rvest**

пакет **xml2**

Колекција пакета tidyverse

Модел алата за анализирање података (по Викаму и Гролемунду)

учитавање података

чишћење података

разумевање

циклични проласци кроз потфазе

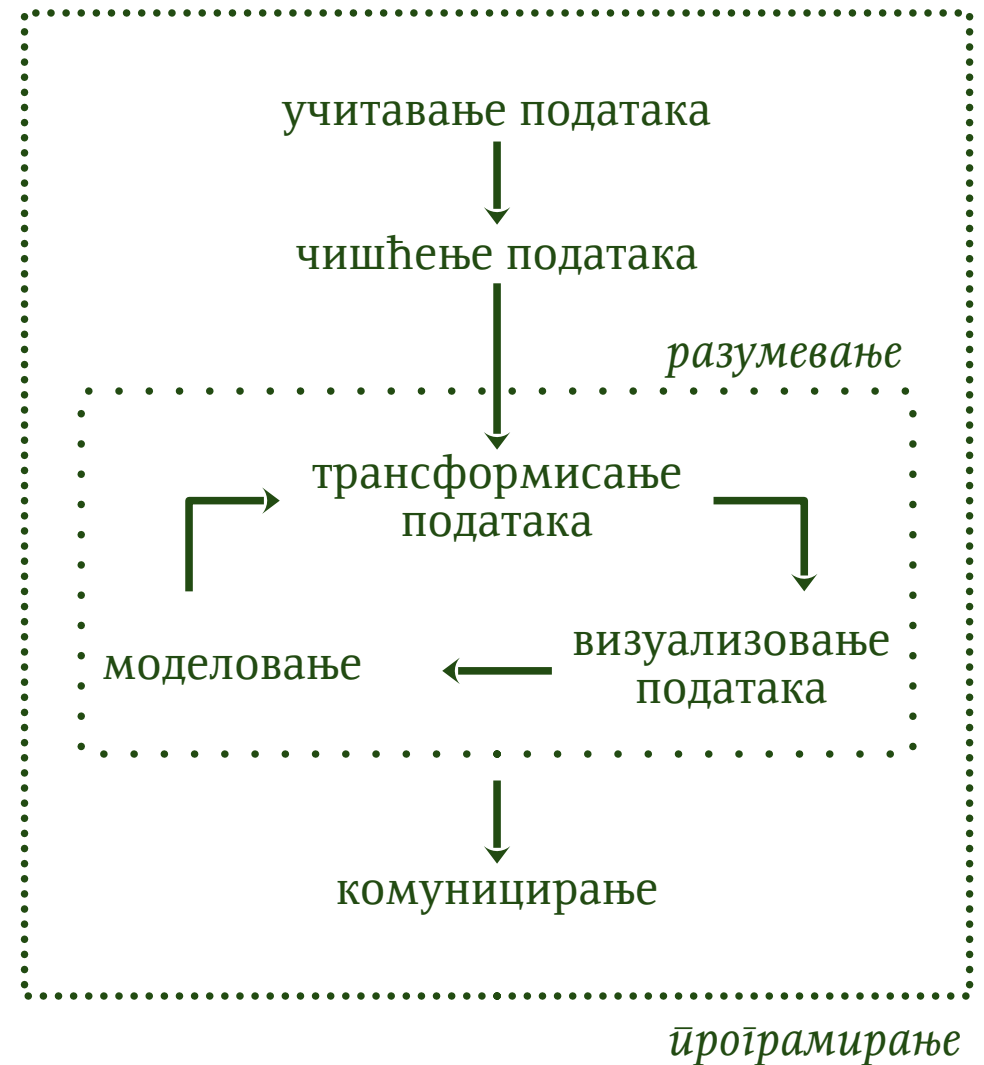
трансформисање података

визуализовање података

моделовање

комуницирање

програмирање*



Садржај

1. Колекција пакета tidyverse
- 2. Анализа података**
3. Извори и литература

Скупови података коришћени у примерима

скуп података **mostovi.svi**

подаци о мостовима Питсбурга (САД)

108 записа

16 обележја

намена, тип, распон, време изградње, дужина, број трака...

датотеке *bridges.data.version1* и *bridges.data.version2*

Pittsburgh Bridges Data Set (од 1. 8. 1990)

<http://archive.ics.uci.edu/ml/datasets/Pittsburgh+Bridges>
(преузето 24. 3. 2021)

UCI Machine Learning Repository

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository
[<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of
Information and Computer Science.

Скупови података коришћени у примерима

скуп података **mostovi**

записи из скупа података **mostovi.svi** који немају недостајућих вредности

70 записа

16 обележја

Скупови података коришћени у примерима

скупови података **mostovi.svi** и **mostovi** – припрема

```
1 # install.packages("tidyverse")
2
3 library(readr)
4 library(tidyr)
5 library(dplyr)
6 library(ggplot2)
7
8
9
10
11
12
13
14
15
16
17
18
```

УЛАЗ

Скупови података коришћени у примерима

скупови података **mostovi.svi** и **mostovi** – припрема

```
1 mostovi.v1 <- read_csv("bridges.data.version1",
2                       col_names=c("identif", "river",
3                                   "location", "erected_c",
4                                   "purpose", "length_c",
5                                   "lanes_c", "clearg",
6                                   "tord", "material",
7                                   "span", "rell",
8                                   "type"),
9                       col_types=cols(
10                      col_character(), col_factor(),
11                      col_character(), col_integer(),
12                      col_factor(), col_double(),
13                      col_double(), col_factor(),
14                      col_factor(), col_factor(),
15                      col_factor(), col_factor(),
16                      col_factor()),
17                      na="?")
18
```

УЛАЗ

Скупови података коришћени у примерима

скупови података **mostovi.svi** и **mostovi** – припрема

```
1 mostovi.v2 <- read_csv("bridges.data.version2",
2                       col_names=c("identif", "river",
3                                   "location", "erected_n",
4                                   "purpose", "length_n",
5                                   "lanes_n", "clearg",
6                                   "tord", "material",
7                                   "span", "rell",
8                                   "type"),
9                       col_types=cols(
10                      col_character(), col_factor(),
11                      col_character(), col_factor(),
12                      col_factor(), col_factor(),
13                      col_factor(), col_factor(),
14                      col_factor(), col_factor(),
15                      col_factor(), col_factor(),
16                      col_factor()),
17                      na="?")
18
```

УЛАЗ

Скупови података коришћени у примерима

скупови података **mostovi.svi** и **mostovi** – припрема

```
1 mostovi.svi <- inner_join(mostovi.v1, mostovi.v2)
2
3 mostovi <- mostovi.svi %>%
4   drop_na() %>%
5   select(identif:erected_c, erected_n, purpose,
6         length_c, length_n, lanes_c,
7         lanes_n, everything())
8
9
10
11
12
13
14
15
16
17
18
```

УЛАЗ

Пакет *tibble*

омогућава коришћење посебне структуре података за чување скупова података

таблица (енгл. *tibble*)

класа **tbl_df** из пакета **tibble**

поткласа класе **data.frame**

упрошћени скуп података

без подразумеваног увођења назива редова

без подразумеване конверзије текстуалног садржаја у факторе

флексибилност при именовању колона

прилагођени текстуални приказ

...

распрострањена употреба у пакетима колекције *tidyverse*

креирање таблице

функција **as_tibble(...)**

Пакет *tibble*

скуп података у табеларном облику

```
> typeof(mostovi)
[1] "list"
> class(mostovi)
[1] "tbl_df"      "tbl"        "data.frame"
> mostovi
# A tibble: 70 x 16
  identif river location erected_c erected_n purpose length_c length_n lanes_c lanes_n
  <chr>   <fct> <chr>      <int> <fct>    <fct>    <dbl> <fct>    <dbl> <fct>
1 E2     A     25        1819 CRAFTS   HIGHWAY   1037 MEDIUM    2 2
2 E5     A     29        1837 CRAFTS   HIGHWAY   1000 MEDIUM    2 2
3 E7     A     27        1840 CRAFTS   HIGHWAY    990 SHORT      2 2
4 E8     A     28        1844 CRAFTS   AQUEDU... 1000 MEDIUM    1 1
5 E9     M     3         1846 CRAFTS   HIGHWAY   1500 MEDIUM    2 2
6 E11    A     29        1851 CRAFTS   HIGHWAY   1000 MEDIUM    2 2
7 E14    M     6         1856 CRAFTS   HIGHWAY   1200 MEDIUM    2 2
8 E16    A     25        1859 CRAFTS   HIGHWAY   1030 MEDIUM    2 2
9 E18    A     28        1864 CRAFTS   RR        1200 MEDIUM    2 2
10 E19   A     29        1866 CRAFTS   HIGHWAY   1000 MEDIUM    2 2
# ... with 60 more rows, and 6 more variables: clearg <fct>, tord <fct>, material <fct>,
#   span <fct>, rell <fct>, type <fct>
>
```

КОНЗОЛА

Пакет *dplyr*

„граматика манипулисања подацима”

бројне функције за трансформисање скупова података

основне групе функција

- манипулисање редовима у скупу података

- манипулисање колонама у скупу података

- агрегирање података

- спајање скупова података

- ...

Пакет *dplyr*

манипулисање редовима у скупу података

функција **`slice(skup-podataka, ...)`**

издвајање редова на основу задатих позиција
позиције дате кроз додатне аргументе

функција **`filter(skup-podataka, ...)`**

издвајање редова на основу задатих услова
издвојени редови задовољавају сваки од задатих услова
услови дати кроз додатне аргументе

функција **`arrange(skup-podataka, ...)`**

уређење поретка редова на основу вредности задатих колона
колоне дате кроз додатне аргументе

...

Пакет *dplyr*

манипулисање редовима у скупу података

```
> slice(mostovi, 2, 5)
# A tibble: 2 x 16
  identif river location erected_c erected_n purpose length_c
  <chr>    <fct> <chr>         <int> <fct>      <fct>      <dbl>
1 E5      A     29          1837 CRAFTS     HIGHWAY     1000
2 E9      M      3          1846 CRAFTS     HIGHWAY     1500
# ... with 9 more variables: length_n <fct>, lanes_c <dbl>,
#   lanes_n <fct>, clearg <fct>, tord <fct>, material <fct>,
#   span <fct>, rell <fct>, type <fct>
>
```

КОНЗОЛА

Пакет *dplyr*

манипулисање редовима у скупу података

```
> filter(mostovi, erected_c==1887)
# A tibble: 1 x 16
  identif river location erected_c erected_n purpose length_c
  <chr>    <fct> <chr>         <int> <fct>      <fct>      <dbl>
1 E31      M      8             1887 EMERGING   RR          1161
# ... with 9 more variables: length_n <fct>, lanes_c <dbl>,
#   lanes_n <fct>, clearg <fct>, tord <fct>, material <fct>,
#   span <fct>, rell <fct>, type <fct>
>
```

КОНЗОЛА

Пакет *dplyr*

манипулисање редовима у скупу података

```
> slice(arrange(mostovi, desc(erected_c)),  
+         c(1, 5))  
# A tibble: 2 x 16  
  identif river location erected_c erected_n purpose length_c  
  <chr>   <fct> <chr>      <int> <fct>      <fct>      <dbl>  
1 E90     M      7        1978 MODERN     HIGHWAY     950  
2 E86     A     33        1961 MODERN     HIGHWAY     980  
# ... with 9 more variables: length_n <fct>, lanes_c <dbl>,  
#   lanes_n <fct>, clearg <fct>, tord <fct>, material <fct>,  
#   span <fct>, rell <fct>, type <fct>  
>
```

КОНЗОЛА

Пакет *dplyr*

манипулисање колонама у скупу података

функција **`select(skup-podataka, ...)`**

издвајање колона

избор колона на основу назива или својства

доступан напредан механизам за бирање колона

`contains()`, `starts_with()`, `ends_with()`, `matches()`, `last_col()`, `everything()`...

функција **`rename(skup-podataka, ...)`**

преименовање колона

задавање новог назива и старог назива као пар нови–стари

функција **`mutate(skup-podataka, ...)`**

формирање нових колона

задавање парова назив–вредности

могуће је искористити садржаје постојећих колона

...

Анализа података

Пакет *dplyr*

манипулисање колонама у скупу података

```
> select(mostovi, identif, river, contains("length"))
# A tibble: 70 x 4
  identif river length_c length_n
  <chr>    <fct>    <dbl> <fct>
1 E2      A      1037 MEDIUM
2 E5      A      1000 MEDIUM
3 E7      A       990 SHORT
4 E8      A      1000 MEDIUM
5 E9      M      1500 MEDIUM
6 E11     A      1000 MEDIUM
7 E14     M      1200 MEDIUM
8 E16     A      1030 MEDIUM
9 E18     A      1200 MEDIUM
10 E19    A      1000 MEDIUM
# ... with 60 more rows
>
```

КОНЗОЛА

Анализа података

Пакет *dplyr*

манипулисање колонама у скупу података

```
> rename(mostovi, id=identif, loc=location)
# A tibble: 70 x 16
   id      river loc erected_c erected_n purpose length_c
  <chr> <fct> <chr>    <int> <fct>    <fct>    <dbl>
1 E2     A     25    1819 CRAFTS    HIGHWAY    1037
2 E5     A     29    1837 CRAFTS    HIGHWAY    1000
3 E7     A     27    1840 CRAFTS    HIGHWAY     990
4 E8     A     28    1844 CRAFTS    AQUEDUCT    1000
5 E9     M      3    1846 CRAFTS    HIGHWAY    1500
6 E11    A     29    1851 CRAFTS    HIGHWAY    1000
7 E14    M      6    1856 CRAFTS    HIGHWAY    1200
8 E16    A     25    1859 CRAFTS    HIGHWAY    1030
9 E18    A     28    1864 CRAFTS    RR         1200
10 E19   A     29    1866 CRAFTS    HIGHWAY    1000
# ... with 60 more rows, and 9 more variables: length_n <fct>,
#   lanes_c <dbl>, lanes_n <fct>, clearg <fct>, tord <fct>,
#   material <fct>, span <fct>, rell <fct>, type <fct>
>
```

КОНЗОЛА

Пакет *dplyr*

манипулисање колонама у скупу података

```
> select(mutate(mostovi, lengthm_c=length_c * 0.3048),  
+         identif:location, length_c, length_n, last_col())  
# A tibble: 70 x 6  
  identif river location length_c length_n lengthm_c  
  <chr>   <fct> <chr>      <dbl> <fct>      <dbl>  
1 E2      A      25        1037 MEDIUM      316.  
2 E5      A      29        1000 MEDIUM      305.  
3 E7      A      27         990 SHORT        302.  
4 E8      A      28        1000 MEDIUM      305.  
5 E9      M       3        1500 MEDIUM      457.  
6 E11     A      29        1000 MEDIUM      305.  
7 E14     M       6        1200 MEDIUM      366.  
8 E16     A      25        1030 MEDIUM      314.  
9 E18     A      28        1200 MEDIUM      366.  
10 E19    A      29        1000 MEDIUM      305.  
# ... with 60 more rows  
>
```

КОНЗОЛА

Пакет *dplyr*

агрегирање података

функција **summarise(skup-podataka, ...)**

агрегирање података на основу група

могуће је применити сумарне функције

функција **group_by(skup-podataka, ...)**

увођење група у податке

критеријуми груписања дати кроз додатне аргументе

могуће је груписање на основу колона или израчунавања

...

Пакет *dplyr*

агрегирање података

```
> summarise(mostovi,  
+           lanes_c_median=median(lanes_c))  
# A tibble: 1 x 1  
  lanes_c_median  
    <dbl>  
1             2  
>
```

КОНЗОЛА

Анализа података

Пакет *dplyr*

агрегирање података

```
> select(group_by(mostovi, material),  
+         identif, material:type)  
# A tibble: 70 x 5  
# Groups:   material [3]  
  identif material span    rell  type  
  <chr>    <fct>    <fct> <fct> <fct>  
1 E2      WOOD      SHORT S     WOOD  
2 E5      WOOD      SHORT S     WOOD  
3 E7      WOOD      MEDIUM S     WOOD  
4 E8      IRON      SHORT S     SUSPEN  
5 E9      IRON      SHORT S     SUSPEN  
6 E11     WOOD      MEDIUM S     WOOD  
7 E14     WOOD      MEDIUM S     WOOD  
8 E16     IRON      MEDIUM S-F    SUSPEN  
9 E18     IRON      SHORT S     SIMPLE-T  
10 E19     WOOD      MEDIUM S     WOOD  
# ... with 60 more rows  
>
```

КОНЗОЛА

Пакет *dplyr*

агрегирање података

```
> summarise(group_by(mostovi, material),  
+           n=n(),  
+           length_c_mean=mean(length_c))  
# A tibble: 3 x 3  
  material      n length_c_mean  
  <fct>    <int>      <dbl>  
1 WOOD         9      1058.  
2 IRON         4      1182.  
3 STEEL        57      1712.  
>
```

КОНЗОЛА

Пакет *dplyr*

спајање скупова података

спајање по редовима

функција **bind_rows(...)**

просто спајање надодавањем редова

функција **union(a, b, ...)**

формирање уније

функција **intersect(a, b, ...)**

формирање пресека

функција **setdiff(a, b, ...)**

формирање разлике

...

Пакет *dplyr*

спајање скупова података

спајање по колонама

функција **bind_cols(...)**

просто спајање надодавањем колона

функција **inner_join(a, b, ...)**

унутрашње спајање

функција **left_join(a, b, ...)**

лево спајање

функција **right_join(a, b, ...)**

десно спајање

функција **full_join(a, b, ...)**

потпуно спајање

Пакет *ggplot2*

„граматика графике”

компоненте визуализације у складу са слојевитом граматиком (по Викаму)

- основни скуп података и естетских пресликавања променљивих
један или више слојева

- слој обухвата геометријски објекат, статистичку трансформацију, подешавање позиције

- опционо скуп података

- опционо скуп естетских пресликавања

- скала за свако естетско пресликавање

- координатни систем

- спецификација аспекта

креирање визуализација података по посебном систему

кроз ланчање наредби за подешавање појединачних компоненти

Пакет *ggplot2*

подржане бројне визуализације података

визуализације за једну, две или три променљиве

хистограми

графикони функције густине вероватноће

стубичасти графикони

графикони расејања

кутијаста графикони

виолински графикони

топлотне карте

...

графичке примитиве

дужи

криве

многоуглови

траке

...

Пакет *ggplot2*

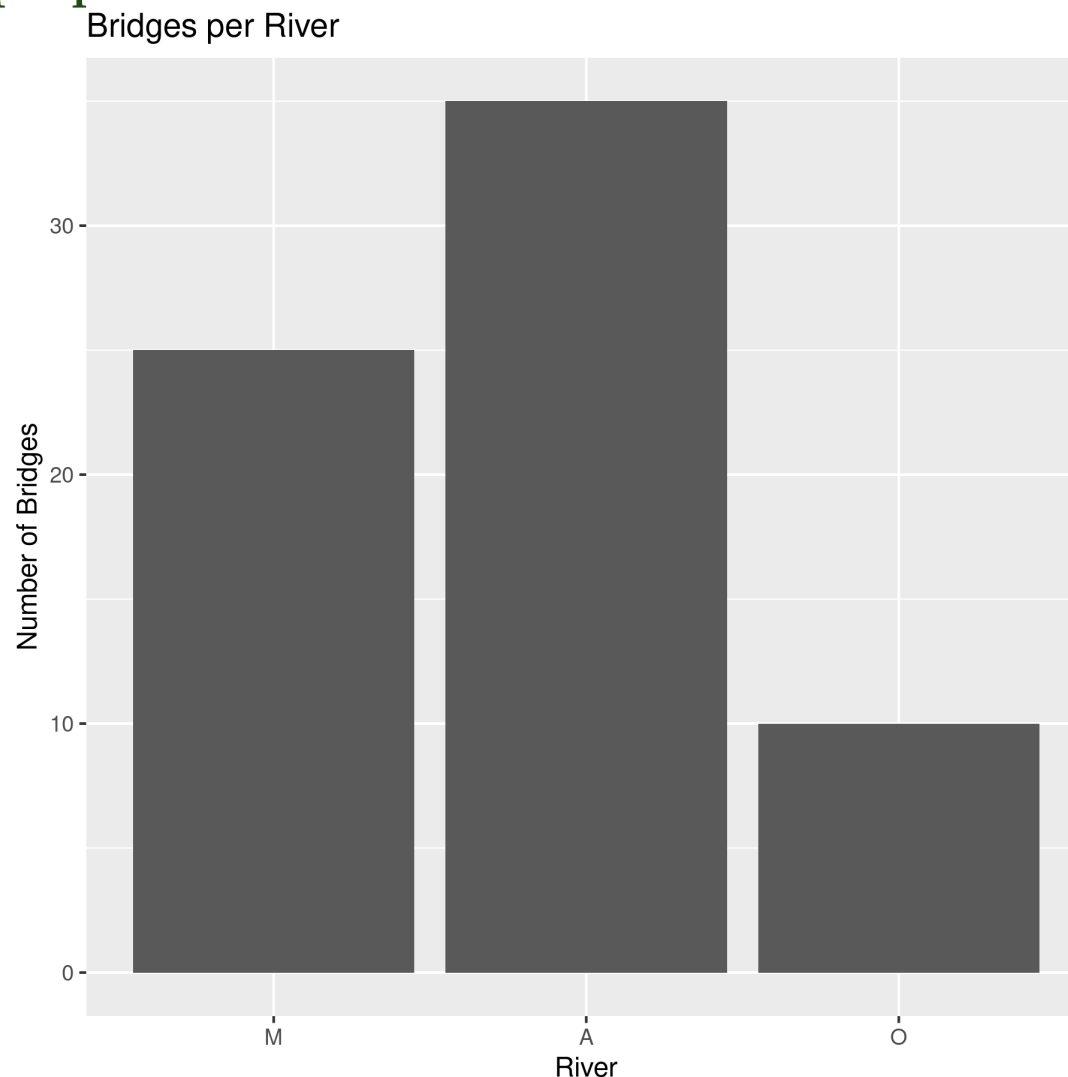
стубичасти графикон

```
1 ggplot(mostovi) +  
2   geom_bar(mapping=aes(x=river)) +  
3   labs(title="Bridges per River",  
4         x="River", y="Number of Bridges")  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18
```

УЛАЗ

Пакет *ggplot2*

стубичасти графикон



Пакет *ggplot2*

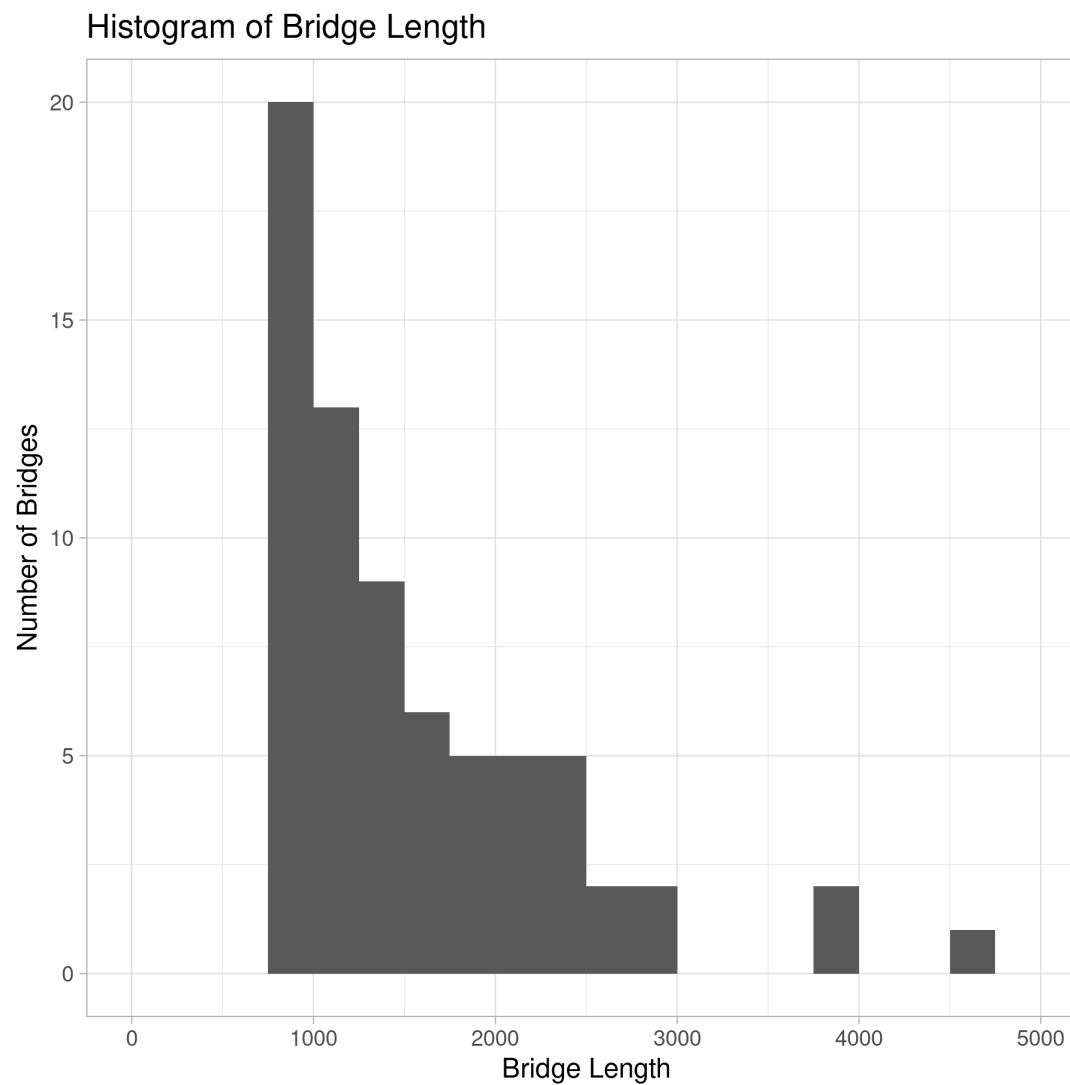
хистограм

```
1 ggplot(mostovi) +  
2   geom_histogram(mapping=aes(x=length_c),  
3                        breaks=seq(0, 5000, 250)) +  
4   labs(title="Histogram of Bridge Length",  
5         x="Bridge Length", y="Number of Bridges") +  
6   theme_light()  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18
```

УЛАЗ

Пакет *ggplot2*

хистограм



Пакет *ggplot2*

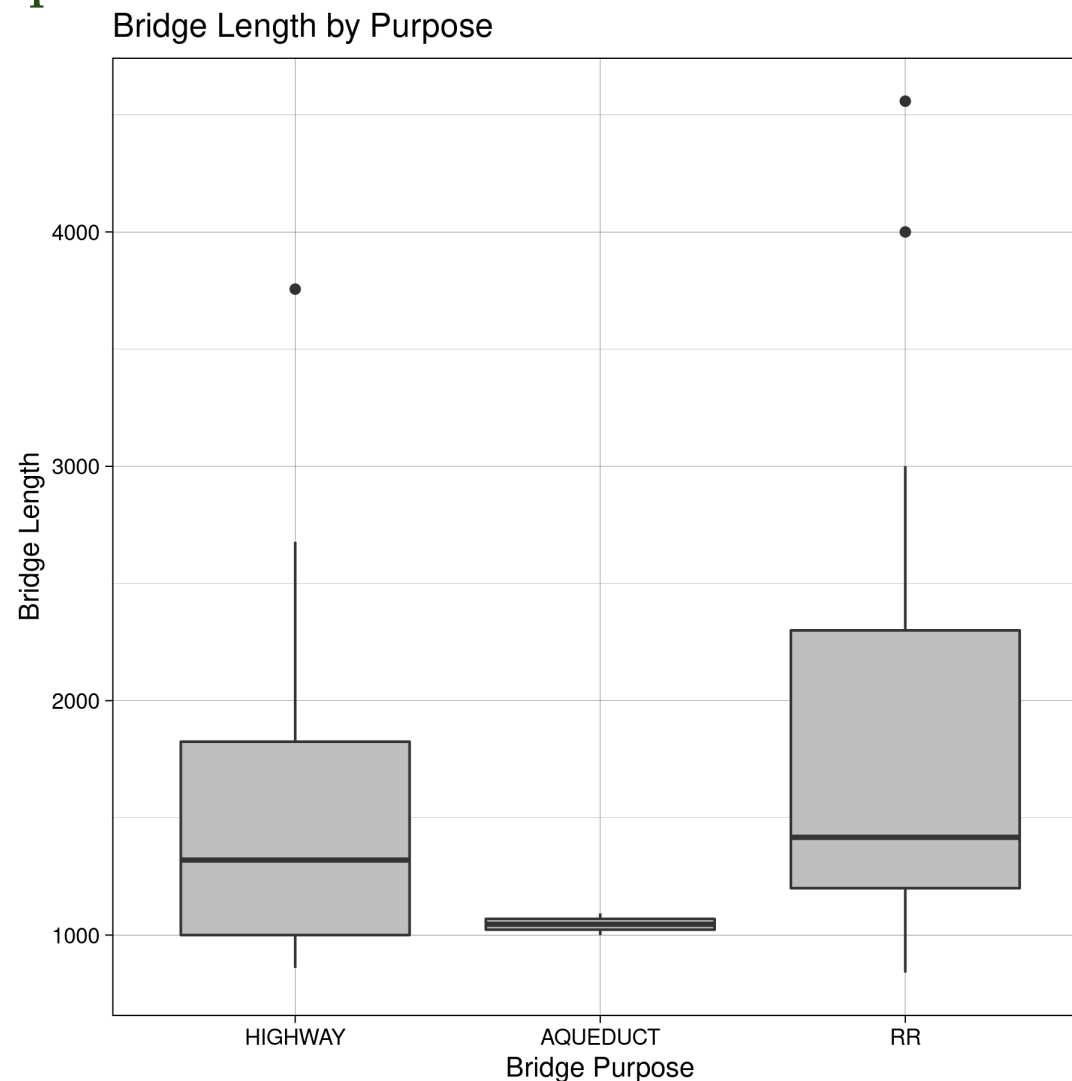
кутијасте графикон

```
1 ggplot(mostovi, mapping=aes(x=purpose, y=length_c)) +  
2   geom_boxplot(fill="gray") +  
3   labs(title="Bridge Length by Purpose",  
4         x="Bridge Purpose", y="Bridge Length") +  
5   theme_linedraw()  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18
```

УЛАЗ

Пакет *ggplot2*

кутијаста графика



Пакет *ggplot2*

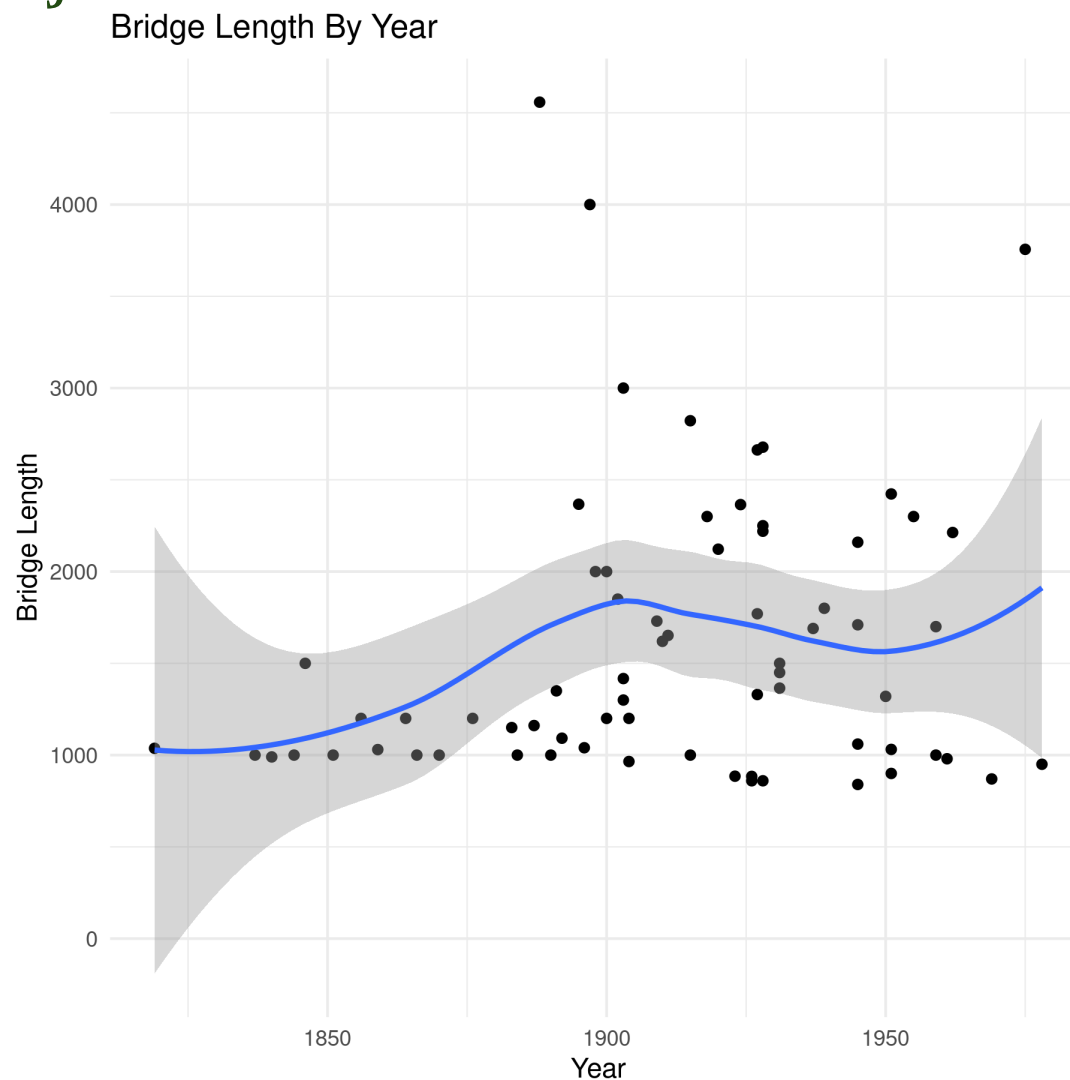
графикон расејања

```
1 ggplot(mostovi,  
2       mapping=aes(x=erected_c, y=length_c)) +  
3   geom_point() +  
4   geom_smooth(method="loess", formula=y ~ x) +  
5   labs(title="Bridge Length By Year",  
6       x="Year", y="Bridge Length") +  
7   theme_minimal()  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18
```

УЛАЗ

Пакет *ggplot2*

графикон расејања



Пакет *magrittr*

увођење помоћних оператора

могуће је другачије организовати изворни кôд
постизање веће једноставности и прегледности

помоћни оператори

оператор `%>%`

прослеђивање вредности изразу или функцији а резултат је резултат операције с десне стране

пример

уместо **`funkcija(v)`** може се користити **`v %>% funkcija`**

оператор `%T>%`

прослеђивање вредности а резултат је вредност с леве стране

оператор `%%$%`

отварање назива из објекта с леве стране према изразу с десне стране

оператор `%<>%`

прослеђивање уз доделу резултата с десне стране левој страни

Пакет *magrittr*

оператор `%>%`

```
> mostovi %>%  
+   filter(material=="WOOD") %>%  
+   summarise(length_c_mean=mean(length_c))  
# A tibble: 1 x 1  
  length_c_mean  
    <dbl>  
1       1058.  
>
```

КОНЗОЛА

Пакет *magrittr*

оператор `%>%`

```
> mostovi %>%  
+   select(identif, river, erected_c, length_c) %>%  
+   filter(erected_c > 1900) %>%  
+   group_by(river) %>%  
+   summarise(n=n(),  
+             length_c_med=median(length_c))  
# A tibble: 3 x 3  
  river      n length_c_med  
  <fct> <int>      <dbl>  
1 M         16      1374.  
2 A         20      1385  
3 0          9      1700  
>
```

КОНЗОЛА

Садржај

1. Колекција пакета tidyverse
2. Анализа података
- 3. Извори и литература**

Основни извори и литература

- ◆ R: A language and environment for statistical computing – Reference index – The R core team – Version 4.2.2 (2022-10-31). Internet:
<https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf>
- ◆ Adler J. R in a nutshell: A desktop quick reference. 2nd edition. O'Reilly; 2012.
- ◆ Tidyverse. Internet: <https://www.tidyverse.org/>
- ◆ Wickham H, Grolemund G. R for data science: Visualize, model, transform, tidy, and import data. O'Reilly; 2017. Internet:
<https://r4ds.had.co.nz/>
- ◆ Posit Cheatsheets. Internet:
<https://posit.co/resources/cheatsheets/>
- ◆ CRAN - Package tidyverse. Internet:
<https://cran.r-project.org/web/packages/tidyverse/index.html>

Основни извори и литература

- ◆ Wickham H, Averick M, Bryan J, Chang W, D'Agostino McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H. Welcome to the tidyverse. *Journal of Open Source Software*. 2019;4(43); 1686.
- ◆ Wickham H. A layered grammar of graphics. *Journal of Computational and Graphical Statistics*. 2010;19(1); 3-28.

Основни извори података

- ♦ скуп података **mostovi.svi**
 - ♦ UCI Machine Learning Repository
 - ♦ Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.
 - ♦ Подаци о мостовима Питсбурга (САД)
 - ♦ датотеке *bridges.data.version1* и *bridges.data.version2*
 - ♦ Pittsburgh Bridges Data Set (од 1. 8. 1990)
 - ♦ <http://archive.ics.uci.edu/ml/datasets/Pittsburgh+Bridges> (преузето 24. 3. 2021)

Мастер академске студије
Рачунарство и аутоматика

Рачунарство високих перформанси
у информационом инжењерингу

Анализа података кроз језик R и колекцију пакета tidyverse

(материјали за предавања)