Data visualization in R

Zoltan Kekecs, Marton Kovacs

February 18, 2020

Table of Contents

[Adatvizualizáció 1](#_Toc33628419)

[Absztrakt 1](#_Toc33628420)

[Ismétlés 1](#_Toc33628421)

[Adatok ábrázolása 2](#_Toc33628422)

[Geomok eloszlás vizsgálatára 8](#_Toc33628423)

[Hisztogramm 8](#_Toc33628424)

[Sűrűségi ábra (density plot) 8](#_Toc33628425)

[Pöttydiagramm (dotplot) 8](#_Toc33628426)

[Oszlopdiagram (geom\_bar) 9](#_Toc33628427)

[Doboz ábra (boxplot) 9](#_Toc33628428)

[Hegedű ábra (violin plot) 10](#_Toc33628429)

[Adatok előkészítése ábra készítésre 11](#_Toc33628430)

[Szöveg ábrára rakása 12](#_Toc33628431)

[szóródás ábrázolása 14](#_Toc33628432)

# Adatvizualizáció

## Absztrakt

Ezen a gyakorlaton megtanuljuk hogyan készíthetünk szemléletes ábrákat az adatainkban található összefüggések megjelenítésére. A gyakorlat bemutatja az eloszlásfüggvények, hisztogram, pont-, oszlop-, vonal- és dobozdiagramok elkészítésének módját a ggplot2 package segítségével.

## Ismétlés

Az alábbi gyakorláshoz használd a Tidyverse pakcage-et és az alapvető funkciókat amit a dplyr-ből tanultunk!

*\_\_\_\_\_\_\_\_\_\_\_\_Gyakorlás\_\_\_\_\_\_\_\_\_\_\_\_\_*

* Telepítsd és töltsd be a “gapminder” csomagot!
* Töltsd be a gapminder adattáblát!
* Szűrd meg az adatokat úgy hogy csak a 2007-ből (year) származó adatokkal dolgozzunk.
* Számold ki az átalagos várható életkort (lifeExp) kontinensenként (continent) ezeken a 2007-es adatokon.
* Nézd meg, hogy hány mérés tartozik az egyes országokhoz. (Segítség: Minden mérés egy sor.)
* Hogy könnyebben átlássuk a populációt, hozz létre egy “pop\_thousand” nevű változót, amiben a meglévő populáció értékek el vannak osztva ezerrel. Az adattáblát amiben már ez az új változó is benne van mentsd el egy új objektumba amit “gapminder\_with\_pop\_thousand”-nak nevezz el.

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

## Adatok ábrázolása

A gyakorlat során egy movies nevű adatbázissal fogunk dolgozni ami filmekről szóló adatokat tartalmaz. Az adatok az IMDB és a Rottent Tomato film-review oldalakról származnak. Ezt az adatbázist betölthetjük az alábbi kóddal. A kód lefuttatása után a környezetben (environment) megjelenik a movies adattábla.

load(url("https://stat.duke.edu/~mc301/data/movies.Rdata"))

Nézzük meg az adattábla alapvető tulajdonságait a megszokott módon.

movies  
  
View(movies)  
  
str(movies)

Ennek az adatbázisnak az adatait fogjuk ábrázolni. Az ábrázoláshoz a **ggplot2** nevű csomagot használjuk majd. Töltsd be ezt a csomagot! A tidyverse csomag tartalmazza a ggplot2-t, így az alábbi kódban ezen keresztül töltöm be a ggplot2-t. Így a %>%operatárot is használni tudjuk és egyéb dplyer funkciókat.

library(tidyverse)

## -- Attaching packages ------------------------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.2.1 v purrr 0.3.3  
## v tibble 2.1.3 v dplyr 0.8.4  
## v tidyr 1.0.2 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()

A ggplot2 csomag rengeteg funciót tartalmaz. Ezek átlátásához segítséget nyújthat a ggplot cheatsheet. (Több csomaghoz is van ilyen, érdemes rájuk keresni!) <https://rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf>

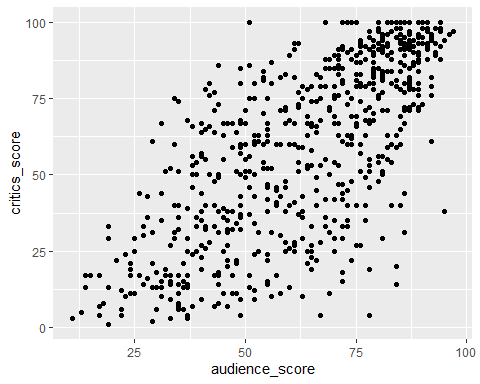
Először is vizualizáljuk mennyire értenek egyet a nézők a kritikusokkal!

Nézd hogyan lehet a pipe %>% operátort használni ahhoz, hogy meg a ggplot funkciót a movies adatbázisra alkalmazzuk.

A ggplotba a sorok végén “+” jelet használunk ahhoz, hogy új elemet adjunk hozzá az ábránkhoz. Az “aesthetics” aes() funkcióval határozzuk meg, hogy az adattáblából melyik változókat akarjuk ábrázolni és melyik tengelyeken, vagy egyéb vizualizációs elemben. A geom\_… funkciókkal határozzuk meg, milyen vizualizációs elemek szerepeljenek az ábrán.

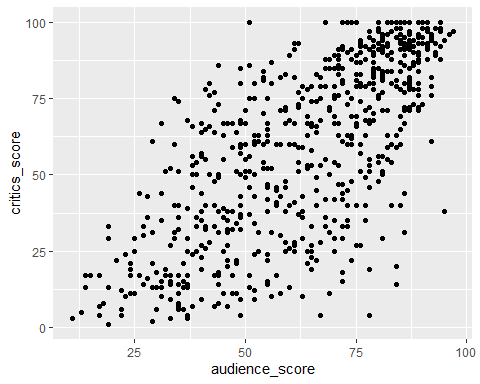
Az hogy mennyire jár együtt a nézők és a kritikusok véleménye jól látszik egy pontdiagramon, ezért most a **geom\_point()** geomot használjuk. ez minden egyes megfigyelést egy pontként ábrázol egy kétdimenziós koordinátarendszerben.

movies %>%   
 ggplot() +  
 aes(x = audience\_score,  
 y = critics\_score) +  
 geom\_point()



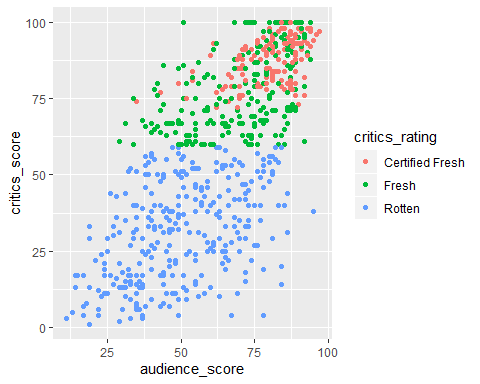
Az ábránkat is, mint minden mást R-ben elmenthetünk egy objektumba, és amikor újra lefuttatjuk ezt az objektumot, akkor az ábra újra megjelenik.

plot1 <- movies %>%   
 ggplot() +  
 aes(x = audience\_score,  
 y = critics\_score) +  
 geom\_point()  
  
plot1



Az alábbi éldán láthatjuk, hogy hogyan tudunk egy új változót bevonni a megjelenítésben. Ebben az esetben egy másik kritikus értékelést jelenítünk meg színekkel. Mivel a geom\_point-ot használjuk, ez a pontok színét fogja befolyásolni, de ha más geomot használnák, azokban is hatna ez a színezésre, hiszen az aes() általános aesthetics részben specifikáltuk.

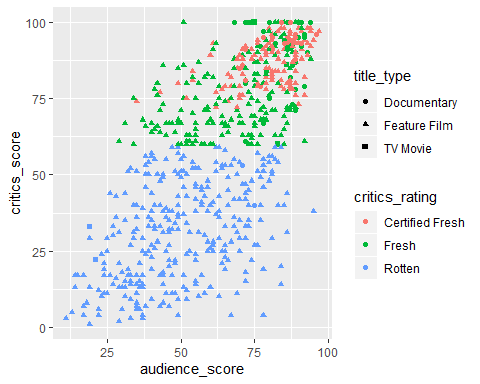
plot2 <- movies %>%   
 ggplot() +  
 aes(x = audience\_score,  
 y = critics\_score,  
 color = critics\_rating) +  
 geom\_point()  
  
plot2



Nem kell mindig kiírnunk a teljes ábra kódot amikor valam,i új elemet szeretnénk hozzáadni az ábrázoláshoz. Ha az ábrát korábban elmentettük objektumként, akkor az objektumhoz + jellel hozzáadhatjuk az új elemeket.

Erről az ábráról például látszik hogy a legjobb (100 pont) kritikus értékelást kapott filmek közül számos film dokumentumfilm.

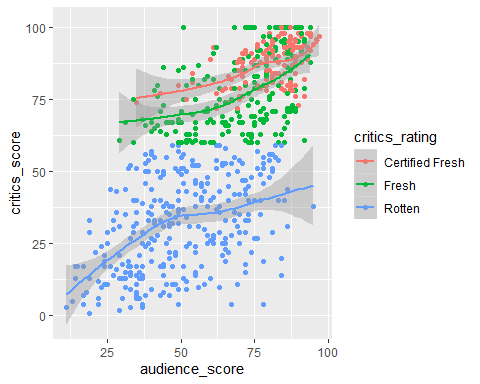
plot2 + aes(shape = title\_type)



Hozzáadhatunk új geomokat is az ábrához hasonló módon. Itt például a **geom\_smooth** geomot adtuk hozzá a korábbi ábrához, ami egy vonalat illeszt az adatpontokra, és ezzel igyekszik vizualizálni az adatokban lévő trendeket.

movies %>%   
 ggplot() +  
 aes(x = audience\_score,  
 y = critics\_score,  
 color = critics\_rating) +  
 geom\_point() +  
 geom\_smooth()

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



Mivel az egész ggplot-ra vonatkozó aes() funkció tartalmazza a color = critics\_rating részt, ezért ez minden geomra hat, így látható hogy a geom\_smooth vonalai is a critics\_rating csoportinként lettek kirajzolva, mindegyik a megfelelő színnel. Azonban megtehetjük, hogy az egyes változókat csak bizonyos geomokon jelenítjük meg. Ezt úgy tudjuk elérni ha a geom funkcióján belül specifikálunk egy új aes() függvényt. Az alábbi kódban a szín szerinti csoportosítás csak a pontokban jelenik meg, a simított vonalban nem

movies %>%   
 ggplot() +  
 aes(x = audience\_score,  
 y = critics\_score) +  
 geom\_point(aes(color = critics\_rating)) +  
 geom\_smooth()

Ha szeretnéd egyes geomok tulajdonságait konstans értékre állítani ahelyett hogy az adatok alapján változnánka (pl. szeretnéd egy geom színát átszínezni anélkül hogy ez megfigyelésenként vagy adatcsoportonként változna), az adott paramétert a geom függvényén belül kell megadni. Ha használsz aes() függvényt is, akkor fontos hogy ez a paraméter az aes() függvényen kívül legyen specifikálva.

Alább a pontok formáját és kitöltési színét, valamint a simított vonal színét állítjuk be konstans értékekre.

A pontok formájának (shape) meghatározásához számokat szoktunk használni. Az hogy melyik szám mit jelent itt találod: <http://www.sthda.com/english/wiki/ggplot2-point-shapes>

A színeket be lehet írni angolul. Egy részletesebb guide erről: <http://sape.inf.usi.ch/quick-reference/ggplot2/colour>

movies %>%   
 ggplot() +  
 aes(x = audience\_score,  
 y = critics\_score) +  
 geom\_point(aes(color = critics\_rating), shape = 21, fill = "white") +  
 geom\_smooth(color = "tomato2")

*\_\_\_\_\_\_\_\_\_\_\_\_Gyakorlás\_\_\_\_\_\_\_\_\_\_\_\_\_*

Továbbra is használjuk a movies adatbázist.

* Ábrázold az összefüggést az IMDB értékelések (imdb\_rating) és a között hogy egy adott filmre hány értékelés jött (imdb\_num\_votes).
* Nézd meg, hogy a műfaj (genre) releváns-e ebben az előbbi összefüggésnél (a műfaj változó is jelenjen meg valahogy az ábrán).

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

## Geomok eloszlás vizsgálatára

Számos fajta geom van. Alább látható néhány gyakran használt geom amit az adatok eloszlásának vizualizációjára szoktunk használni.

### Hisztogramm

movies %>%   
 ggplot() +  
 aes(x = runtime) +  
 geom\_histogram()

### Sűrűségi ábra (density plot)

Az alpha-val azt adjuk meg, hogy mennyire átlátszó az ábra, 0-1 közötti értéket vehet fel.

movies %>%   
 ggplot() +  
 aes(x = runtime, group = title\_type, fill = title\_type) +  
 geom\_density(alpha = .5)

### Pöttydiagramm (dotplot)

A sűrűségfüggvény és a hisztogramm egy változata ahol jól látszik a megfigyelések száma is.

movies %>%   
 ggplot() +  
 aes(x = audience\_score, fill = genre) +  
 geom\_dotplot()

### Oszlopdiagram (geom\_bar)

Kategorikus változók eloszlását vizsgálhatjuk az oszlopdiagrammal.

Az alábbi ábra megmutatja hogy hogyan oszlanak meg a filmek az adatbázisban a Motion Picture Association of America (MPAA) film rating system szerint.

* G – General Audiences
* PG – Parental Guidance Suggested
* PG-13 – Parents Strongly Cautioned
* R – Restricted (Under 17 requires accompanying parent or adult guardian)
* NC-17 – Adults Only

movies %>%   
 ggplot() +  
 aes(x = mpaa\_rating) +  
 geom\_bar()

*\_\_\_\_\_\_\_\_\_\_\_\_Gyakorlás\_\_\_\_\_\_\_\_\_\_\_\_\_*

Továbbra is használjuk a movies adatbázist.

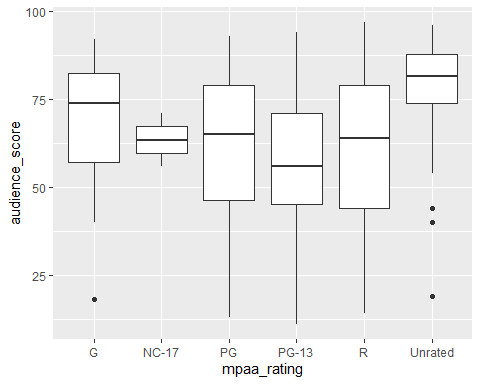
* Hozz létre egy ábrát és mentsd el “my\_first\_plot” néven.
* Vizsgáld meg kritikusok által adott értékelás (critics\_score) eloszlását
* Tetszőleges geomot használhatsz. A ggplot2 cheatsheet segíthet kitalálni, melyik a legjobb geom erre a célra. A critcs\_score egy folytonos (continuous) változó: <https://rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf>)
* Az ábrán az is legyen látható, hogy az eloszlás hogyan különbözik azoknál a filmeknél amiket jelöltek a legjobb film oszkárdíjra (best\_pic\_nom). Ehhez használj tetszőleges aes()-t, pl.: color, fill, linetype, size. A ggplot2 cheatsheet segít hogy az általad választott geomnál melyik a releváns aes()

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

### Doboz ábra (boxplot)

Az alábbi doboz ábra (boxplot) a nézői értékelést mutatja a Motion Picture Association of America (MPAA) film rating system kategóriái szerint. Ez az ábra típus a mediánt mutatja középen, és az adatok szóródását körülötte, a kvartilisek szerint felosztva.

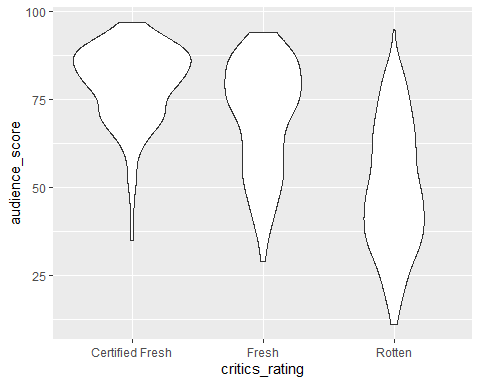
movies %>%   
 ggplot() +  
 aes(y = audience\_score, x = mpaa\_rating) +  
 geom\_boxplot()



### Hegedű ábra (violin plot)

A célja ugyanaz, mint a doboz ábrának, de jobban szemlélteti az adatok eloszlását. Gyakorlatilag a doboz ábra és a density plot keveréke.

movies %>%   
 ggplot() +  
 aes(y = audience\_score, x = critics\_rating) +  
 geom\_violin()

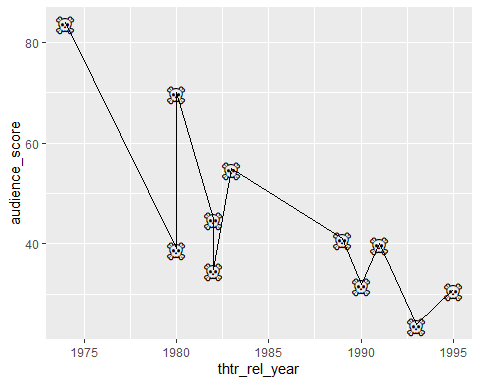


## Adatok előkészítése ábra készítésre

Ahogy azt korábban is láttuk, a ggplot egy pipe végére is berakható, így előkészítheted az adatokat, amit ábrázolni akarsz.

Például az alábbi kóddal kiválasztjuk azokat a horrorfilmeket, amelyek 1972 és 2002 között jelentek meg, és csak azokat az adatokat ábrázoljuk.

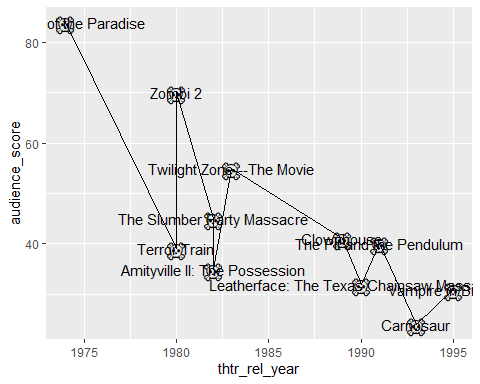
movies %>%   
 filter(genre == "Horror", thtr\_rel\_year > 1972, thtr\_rel\_year < 2002) %>%   
 ggplot() +   
 aes(x = thtr\_rel\_year,  
 y = audience\_score) +  
 geom\_line() +  
 geom\_point(shape="\u2620", fill = "white", size = 5)



## Szöveg ábrára rakása

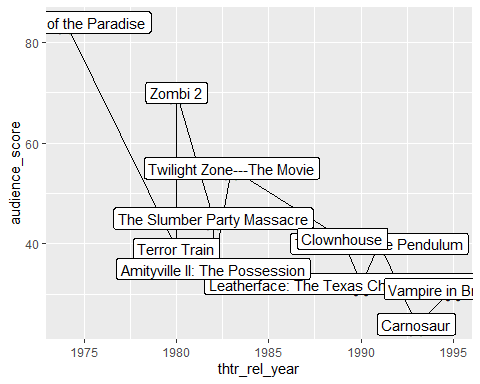
használhatjuk a geom\_text-et

movies %>%   
 filter(genre == "Horror", thtr\_rel\_year > 1972, thtr\_rel\_year < 2002) %>%   
 ggplot() +   
 aes(x = thtr\_rel\_year,  
 y = audience\_score,  
 label = title) +  
 geom\_line() +  
 geom\_point(shape="\u2620", fill = "white", size = 5) +   
 geom\_text()



vagy a geom\_label-t.

movies %>%   
 filter(genre == "Horror", thtr\_rel\_year > 1972, thtr\_rel\_year < 2002) %>%   
 ggplot() +   
 aes(x = thtr\_rel\_year,  
 y = audience\_score,  
 label = title) +  
 geom\_line() +  
 geom\_point(shape="\u2620", fill = "white", size = 5) +   
 geom\_label()



*\_\_\_\_\_\_\_\_\_\_\_\_Gyakorlás\_\_\_\_\_\_\_\_\_\_\_\_\_*

Továbbra is használjuk a movies adatbázist.

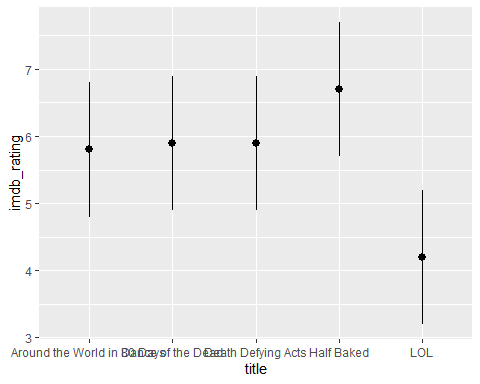
* Ábrázold a nézői értékelés (audience\_score) és a kritikusi értékelés (critics\_score) kapcsolatát egy pontdiagrammal úgy, hogy csak az 1995-ben megjelent filmek szerepeljenek az ábrán.
* Tedd rá a filmek címét az ábrára feliratként, hogy minden ponton szerepeljen a film címe.

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

## szóródás ábrázolása

Adatok szóródásának ábrázolásához használhatjuk még a geom\_pointrange

tail(movies, 5) %>%   
ggplot() +   
 aes(x = title, y = imdb\_rating, ymin = imdb\_rating - 1, ymax = imdb\_rating + 1) +  
 geom\_pointrange()



vagy a geom\_errorbar geomokat.

tail(movies, 5) %>%   
ggplot() +   
 aes(x = title, y = imdb\_rating) +  
 geom\_errorbar(aes(ymin = imdb\_rating - 1, ymax = imdb\_rating + 1))

