Video Games project: R beadandó - Pálos Péter

Beadandóm alapjául a https://www.kaggle.com/ oldalról választottam egy nagyon izgalmasnak tűnő adatbázist, mely videójátékokról származó adatokat tartalmaz.

forrás: Video Game Sales with Ratings

Az elemzéshez használt csomagok:

```
library(dplyr)
library(ggplot2)
library(tidyr)
library(lubridate)
```

Ismerkedés az adattáblával

Miután beolvastam a nyers adattáblát

```
dat <- read.csv("Video_Games_Sales.csv", header=TRUE, na.strings=c("", " ", "NA", "N/A"))
```

kicsit "körbe szaglászom" azt:

```
glimpse(dat)
```

```
## Rows: 16,719
## Columns: 16
## $ Name            <chr> "Wii Sports", "Super Mario Bros.", "Mario Kart Wii"...
## $ Platform        <chr> "Wii", "NES", "Wii", "Wii", "GB", "GB", "DS", "Wii"...
## $ Year_of_Release <int> 2006, 1985, 2008, 2009, 1996, 1989, 2006, 2006, 200...
## $ Genre           <chr> "Sports", "Platform", "Racing", "Sports", "Role-Pla...
## $ Publisher       <chr> "Nintendo", "Nintendo", "Nintendo", "Nintendo", "Ni...
## $ NA_Sales        <dbl> 41.36, 29.08, 15.68, 15.61, 11.27, 23.20, 11.28, 13...
## $ EU_Sales        <dbl> 28.96, 3.58, 12.76, 10.93, 8.89, 2.26, 9.14, 9.18, ...
## $ JP_Sales        <dbl> 3.77, 6.81, 3.79, 3.28, 10.22, 4.22, 6.50, 2.93, 4....
## $ Other_Sales     <dbl> 8.45, 0.77, 3.29, 2.95, 1.00, 0.58, 2.88, 2.84, 2.2...
## $ Global_Sales    <dbl> 82.53, 40.24, 35.52, 32.77, 31.37, 30.26, 29.80, 28...
## $ Critic_Score    <int> 76, NA, 82, 80, NA, NA, 89, 58, 87, NA, NA, 91, NA,...
## $ Critic_Count    <int> 51, NA, 73, 73, NA, NA, 65, 41, 80, NA, NA, 64, NA,...
## $ User_Score      <chr> "8", NA, "8.3", "8", NA, NA, "8.5", "6.6", "8.4", N...
## $ User_Count      <int> 322, NA, 709, 192, NA, NA, 431, 129, 594, NA, NA, 4...
## $ Developer       <chr> "Nintendo", NA, "Nintendo", "Nintendo", NA, NA, "Ni...
## $ Rating          <chr> "E", NA, "E", "E", NA, NA, "E", "E", "E", NA, NA, "...
```

Látjuk, hogy 16719 játékot tartalmaz, melyeknek 16 paraméterét ismerjük.

Ezek közül a legtöbb elnevezése beszédes, ami fejtörést okozhat az az NA_Sales (Észak Amerikai eladások) és JP_Sales (Japán eladások). Az eladási számok mindegyike milliós mértékegységgel rendelkezik.

Amit érdemes még kiemelni, hogy minden vélemény adat Metacriticről származik, valamint a Rating oszlop az ESRB tartalom alapú besorolási szabványokat jelzi.

Két orvosolandó probléma jelentkezett:

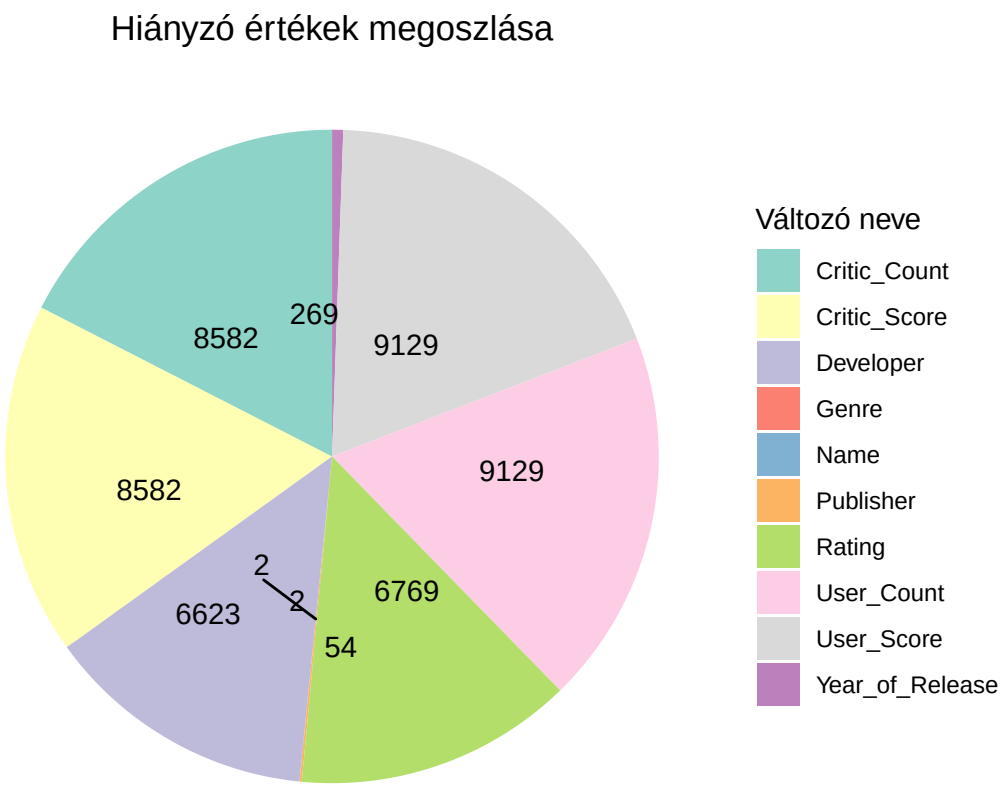
```
dat$Year_of_Release <- as.Date(paste(dat$Year_of_Release, 1, 1, sep="-"))
dat$User_Score <- as.numeric(dat$User_Score)
```

A kiadás évét átkonvertáljuk a jobb funkcionalitás érdekében dátum formátumra, valamint a user vélemények számát a helytelen szöveg típusról számmá.

Hiányzó értékek kezelése

Mivel mindenki jobban szereti a vizuális szemléltetést mint a numerikusat, ezért megér egy kevés plusz átalakítást a tábla, hogy ggplot kompatibilis legyen:

```
dat %>%
  summarise_all(list(~sum(is.na(.)))) %>%
  gather() %>%
  filter(value!=0) %>%
  ggplot(aes(x=1, y=value, fill=key)) +
  geom_col() +
  ggrepel::geom_text_repel(aes(label=value), position=position_stack(vjust=0.5))+
  coord_polar(theta="y") +
  theme_void()+
  scale_fill_brewer(palette="Set3")+
  labs(fill="Változó neve")+
  ggtitle("Hiányzó értékek megoszlása")+
  theme(plot.title=element_text(hjust=0.5))
```



A kis összecsúszás ellenére, ami

- Name: 2
- Genre: 2
- Publisher: 54

szépen felfedték magukat a hiányzó értékek. Ez, ha nem vagyunk szerencsések, akár az egész táblát lefedheti, vizsgáljuk meg az átfedésüket (valamint ha rájöttünk, hogy van egy random 2020-as játék is benne, szűkítsük le a dátumot a file létrehozásának dátumára)

```
dat <- filter(dat, year(Year_of_Release)<=2017)

dat %>%
  filter(!complete.cases(.)) %>%
  nrow()
```

```
## [1] 9624
```

Ezzel egyébként a hiányzó dátumú sorok is törlődtek, de mivel ez csak 269 elemet érintett, és szinte minden elemzésem alapja a dátum, így nem probléma.

Látjuk továbbá, hogy szerencsések vagyunk, és "csak" 9624 hiányzó sorunk van. Nem jó de, nem is tragikus. Készítünk belőle egy új, szűrt adathalmazt, és leellenőrizzük, valóban nincs-e további hiányzó érték.

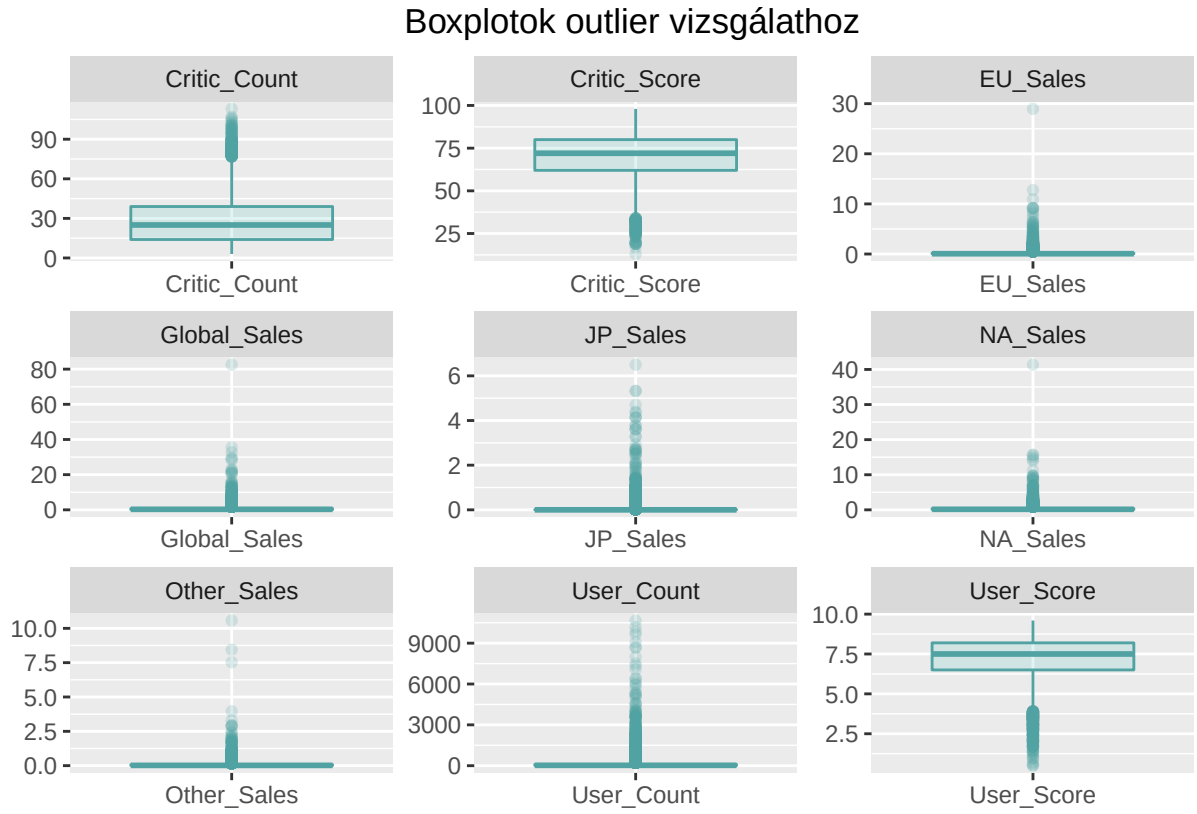
```
no_NA <- dat[complete.cases(dat), ]

sum(!complete.cases(no_NA))
```

```
## [1] 0
```

Outlierek vizsgálata

Mielőtt elemzésekbe kezdenénk, a torzítások megelőzése érdekében megvizsgáljuk a kiugró értékeket, vagyis outliereket. Ezt boxplot segítségével tesszük meg.

```
no_NA %>%
  select_if(is.numeric) %>%
  gather() %>%
  ggplot(aes(factor(key), value))+
  geom_boxplot(color="#51a3a3", fill="#66CCCC", alpha=0.2)+
  facet_wrap(~key, scale="free")+
  labs(x="", y="")+
  ggtitle("Boxplotok outlier vizsgálathoz")+
  theme(plot.title=element_text(hjust=0.5))
```



Három eladási darabszám esetén láthatunk egy egész erősen kiugró értéket. Nem tudhatjuk, hogy ez valami hiba, vagy az egyik legérdekesebb tényezője az adatbázisunknak.

```
cbind(
no_NA %>%
  select(Name, NA_Sales) %>%
  arrange(desc(NA_Sales)) %>%
  top_n(2),

no_NA %>%
  select(Name, EU_Sales) %>%
  arrange(desc(EU_Sales)) %>%
  top_n(2),

no_NA %>%
  select(Name, Global_Sales) %>%
  arrange(desc(Global_Sales)) %>%
  top_n(2))
```

```
## Selecting by NA_Sales

## Selecting by EU_Sales

## Selecting by Global_Sales
```

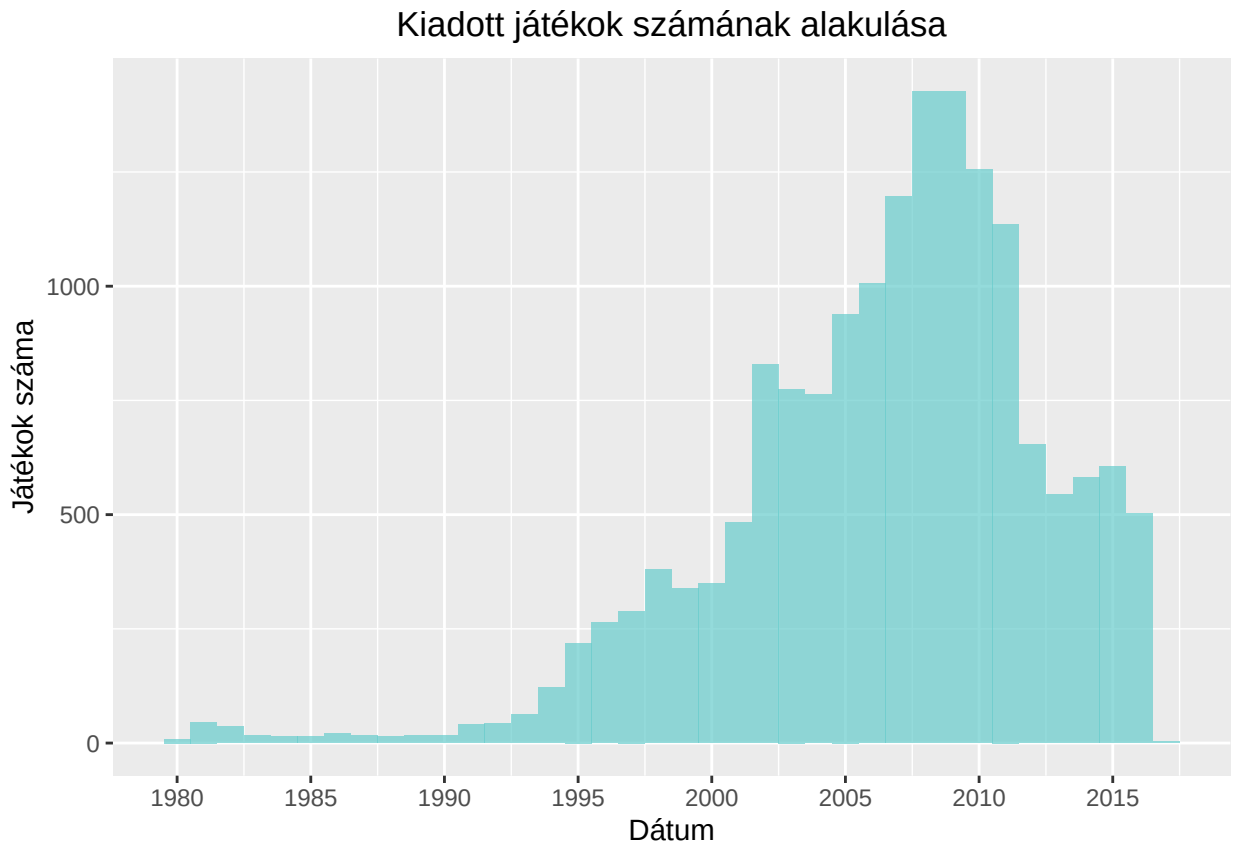
```
##             Name NA_Sales           Name EU_Sales           Name Global_Sales
## 1     Wii Sports    41.36     Wii Sports    28.96     Wii Sports        82.53
## 2 Mario Kart Wii    15.68 Mario Kart Wii    12.76 Mario Kart Wii        35.52
```

Ha mindhárom változónál megnézzük a maximális értékeket, láthatjuk, hogy minden esetben a Wii Sports-hoz tartozik a kiugró érték. És valóban, ennek a játéknak a megjelenése nagy sikerrel járt.

Adatok feltáró elemzése

Kezdjük az elején, nézzük meg a kiadott játékok számának alakulását az idő előrahaladtával.

```
ggplot(dat) +
  geom_bar(aes(year(Year_of_Release)), width=1, fill="#66CCCC", alpha=0.7)+
  labs(x="Dátum", y="Játékok száma")+
  ggtitle("Kiadott játékok számának alakulása")+
  theme(plot.title=element_text(hjust=0.5))+
  scale_x_continuous(breaks = scales::pretty_breaks(10))
```

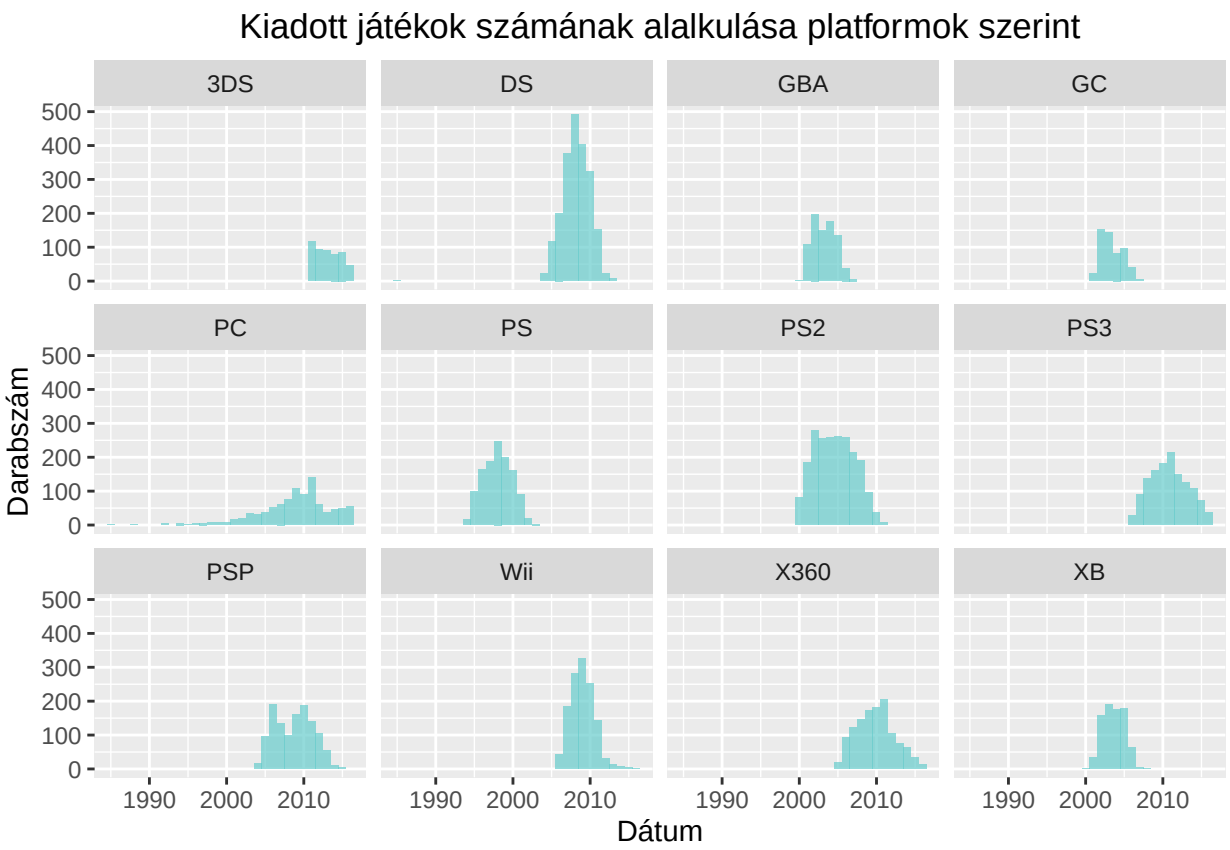


2009-ig közel exponenciáls ütemű növekedés volt tapasztalható, majd a lendület ennél is gyorsabban esett vissza a 2002-es szintre. Ezt egészen biztos, hogy érdemes tovább boncolgatni.

Platformok szerinti bontás

```
dat %>%
  group_by(Platform) %>%
  filter(n()>500) %>%
  ggplot() +
    geom_bar(aes(year(Year_of_Release)), width=1, fill="#66CCCC", alpha=0.7)+
    labs(x="Dátum", y="Darabszám")+
    ggtitle("Kiadott játékok számának alalkulása platformok szerint")+
    scale_fill_brewer(palette="Set3")+
```

```
    theme(plot.title=element_text(hjust=0.5))+
    facet_wrap(.~Platform)
```

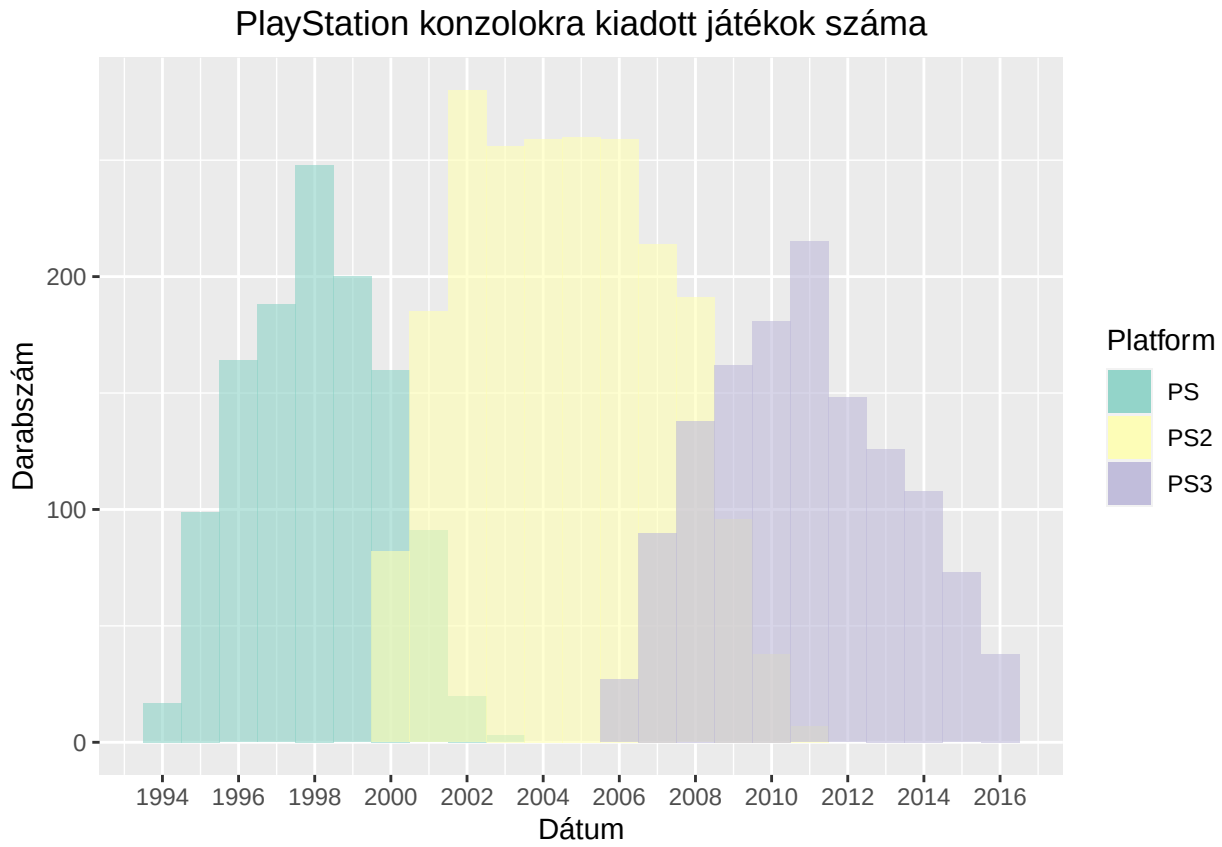


Ha megnézzük "legaktívabb" platformonként a játék kiadási számokat, láthatjuk, hogy ezért nagyrészt a Nintendo DS és a Nintendo Wii felel. A DS után pár napra a PSP is megjelent (hasonló stílus ugye), mely szintén hozzájárult a 2009-es csúcshoz.

Ami különös, hogy az állandónak tekinthető PC játékok kiadásában is visszaesés volt a 2009-es csúcsot követően, vagyis nem magyarázható pusztán az új platformok bevezetésével és lecsengésével a trend változása.

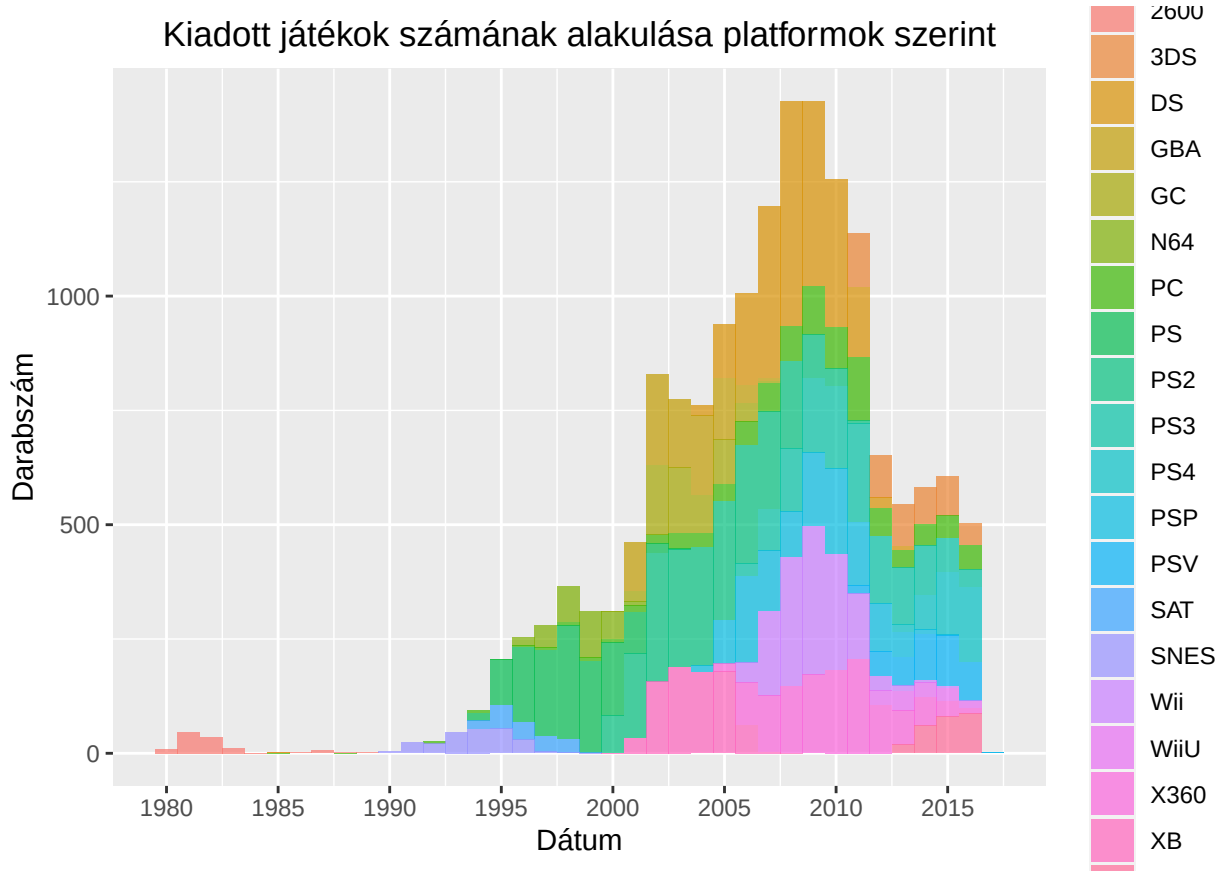
Pusztán érdekességből megnézhetjük a PlayStation konzolok élettartamát is:

```
ggplot()+
  geom_bar(data=filter(dat, Platform=="PS"), aes(year(Year_of_Release), fill=Platform), width=1, alpha=
  geom_bar(data=filter(dat, Platform=="PS2"), aes(year(Year_of_Release), fill=Platform), width=1, alpha=
  geom_bar(data=filter(dat, Platform=="PS3"), aes(year(Year_of_Release), fill=Platform), width=1, alpha=
  labs(x="Dátum", y="Darabszám")+
  ggtitle("PlayStation konzolokra kiadott játékok száma")+
  scale_fill_brewer(palette="Set3")+
  theme(plot.title=element_text(hjust=0.5))+
  scale_x_continuous(breaks = scales::pretty_breaks(10))
```



Vizsgáljuk meg a különböző platformokat egyesítve is.

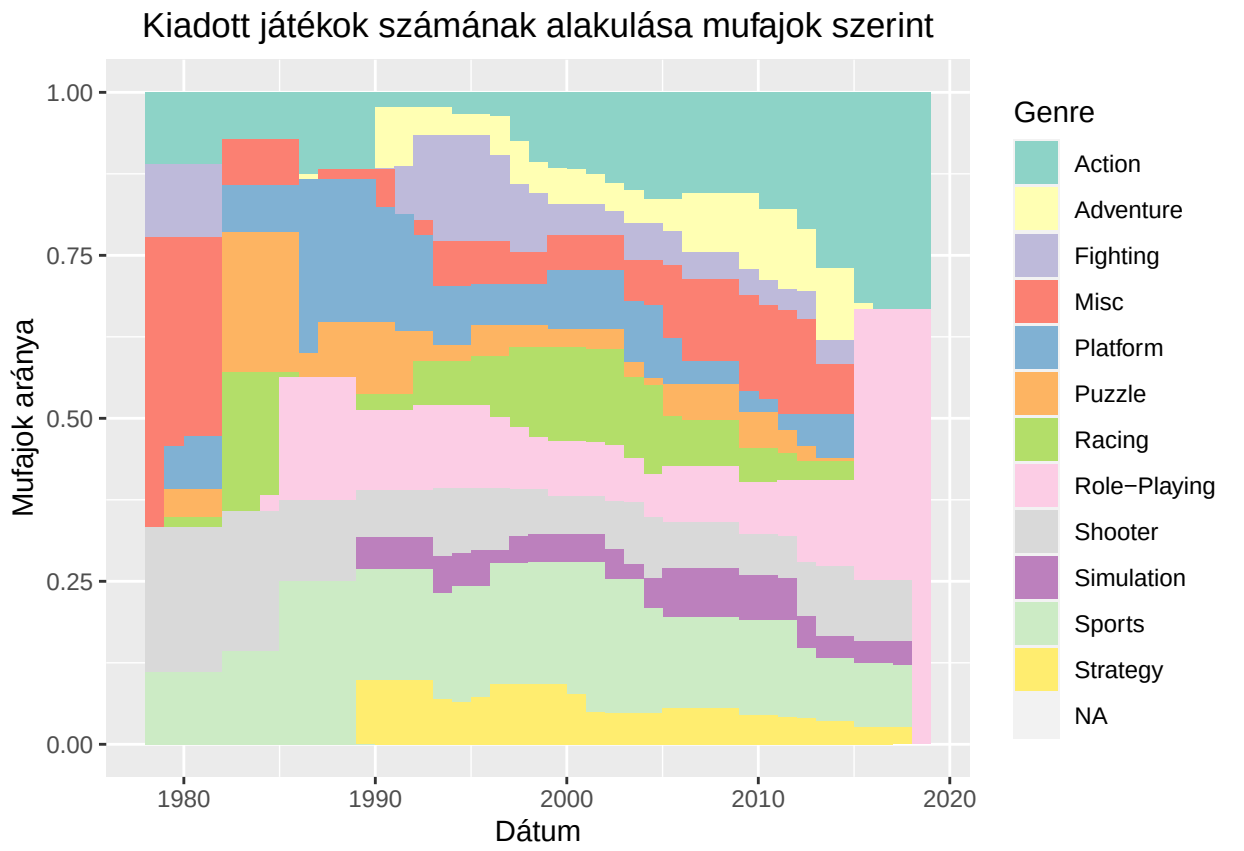
```
dat %>%
  group_by(Platform) %>%
  filter(n()>100) %>%
  ggplot(aes(fill=Platform)) +
  geom_bar(mapping = aes(year(Year_of_Release)), width=1, alpha=0.7)+
  labs(x="Dátum", y="Darabszám")+
  ggtitle("Kiadott játékok számának alakulása platformok szerint")+
  theme(plot.title=element_text(hjust=0.5))+
  scale_x_continuous(breaks = scales::pretty_breaks(10))
```



Láthatjuk, hogy a csúcs időszakában sokkal nagyobb mértékű felhozatal volt a piacon. Talán mondhatni, hogy az arcade játékok 80-as évek beli aranykora megismétlődött?

Műfajok szerinti bontás

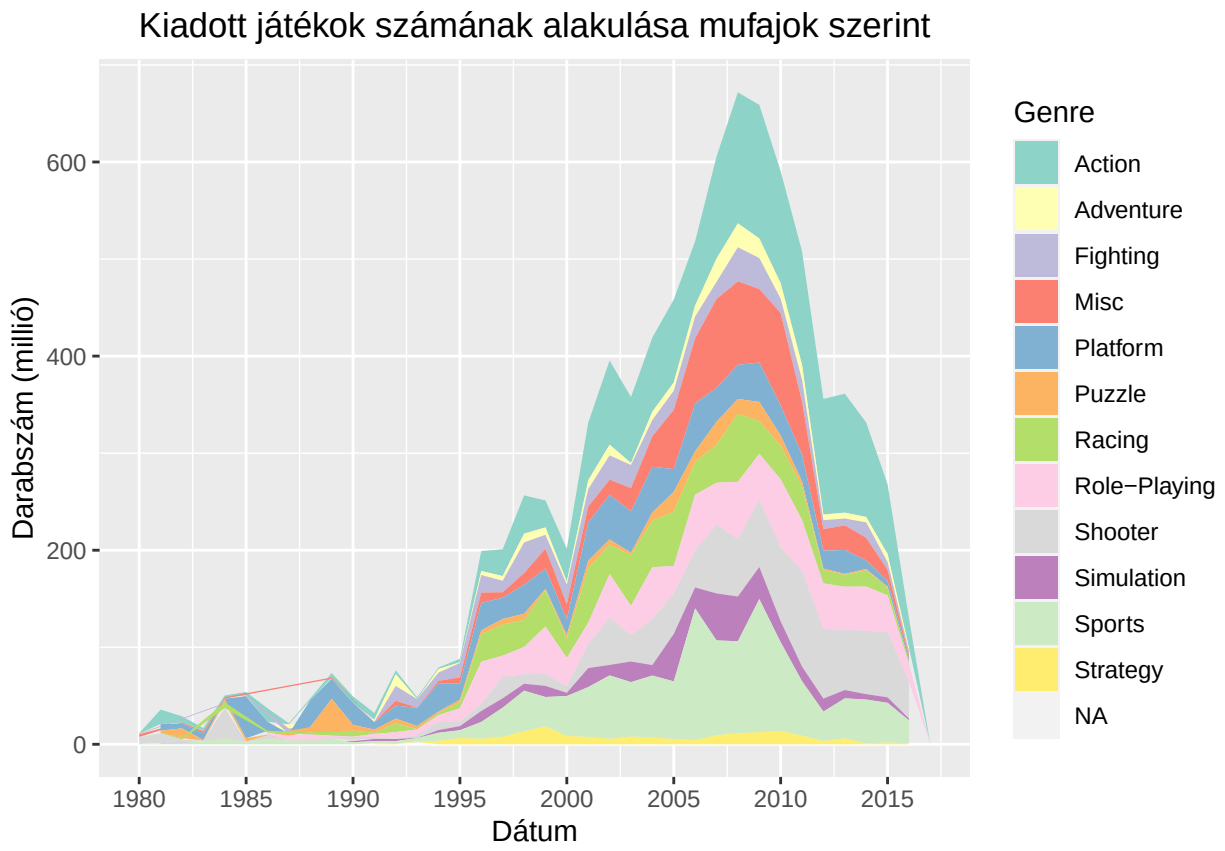
```
dat %>%
  ggplot(aes(fill=Genre)) +
  geom_bar(position="fill", mapping = aes(year(Year_of_Release)), width=4)+
  labs(x="Dátum", y="Műfajok aránya")+
  scale_fill_brewer(palette="Set3")+
  ggtitle("Kiadott játékok számának alakulása műfajok szerint")+
  theme(plot.title=element_text(hjust=0.5))
```



Ezen a 100%-ig halmozott oszlopdiagramon követni tudjuk az egyes műfajú játékok számának alakulását, megjelenését.

Amit megállapíthatunk róla, hogy az akció játékok kezdik uralni a piac nagy részét, folyamatos emelkedésben van. A Legtöbb műfaj képviselőinek aránya igazából stagnál, a stratégiai, sport, versenyző játékok vannak egyre csökkenő arányban jelen, míg mintha a szerepjátékok újra növekedésnek indultak volna.

```
dat %>%
  group_by(Year_of_Release, Genre) %>%
  summarise(Global_Sales=sum(Global_Sales)) %>%
  ggplot(aes(Year_of_Release, Global_Sales, fill=Genre))+
  geom_area()+
  xlab("Dátum")+
  ylab("Darabszám (millió)")+
  ggtitle("Kiadott játékok számának alakulása műfajok szerint")+
  theme(plot.title=element_text(hjust=0.5))+
  scale_fill_brewer(palette="Set3")+
  scale_x_date(breaks = scales::pretty_breaks(10))
```



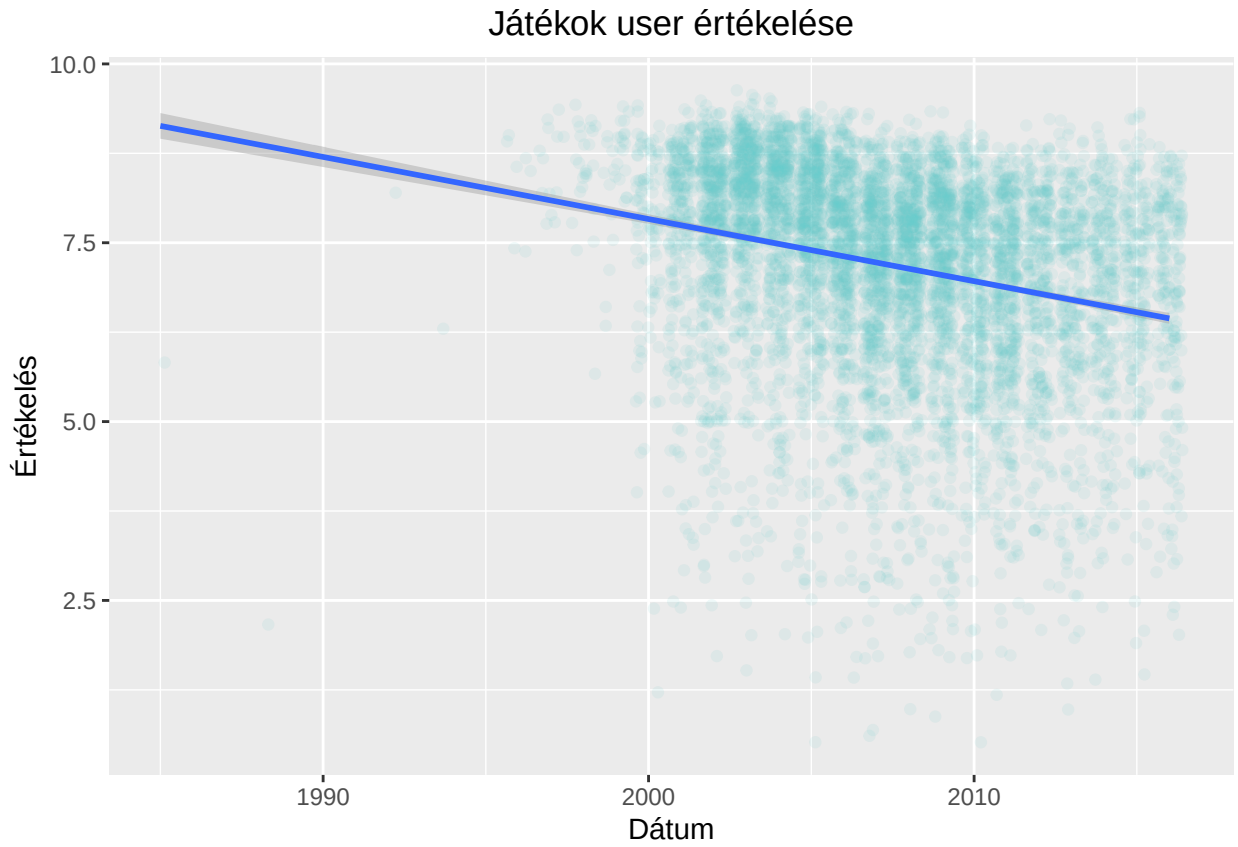
Hogy ne csak az arányokat lássuk, hanem a mértéküket is megismerjük, hasznos lehet a fenti terület diagram.

Vélemények elemzése

Nézzük meg, hogyan alakultak a Metacritic felhasználói értékelései az évek alatt, valamint illesszünk rá egy lineáris regressziós trendet.

```
no_NA %>%
  ggplot(aes(Year_of_Release, User_Score))+
  geom_jitter(alpha=0.1, color="#66CCCC")+
  geom_smooth(method="lm")+
  labs(x="Dátum", y="Értékelés")+
  ggtitle("Játékok user értékelése")+
  theme(plot.title=element_text(hjust=0.5))
```

```
## `geom_smooth()` using formula 'y ~ x'
```

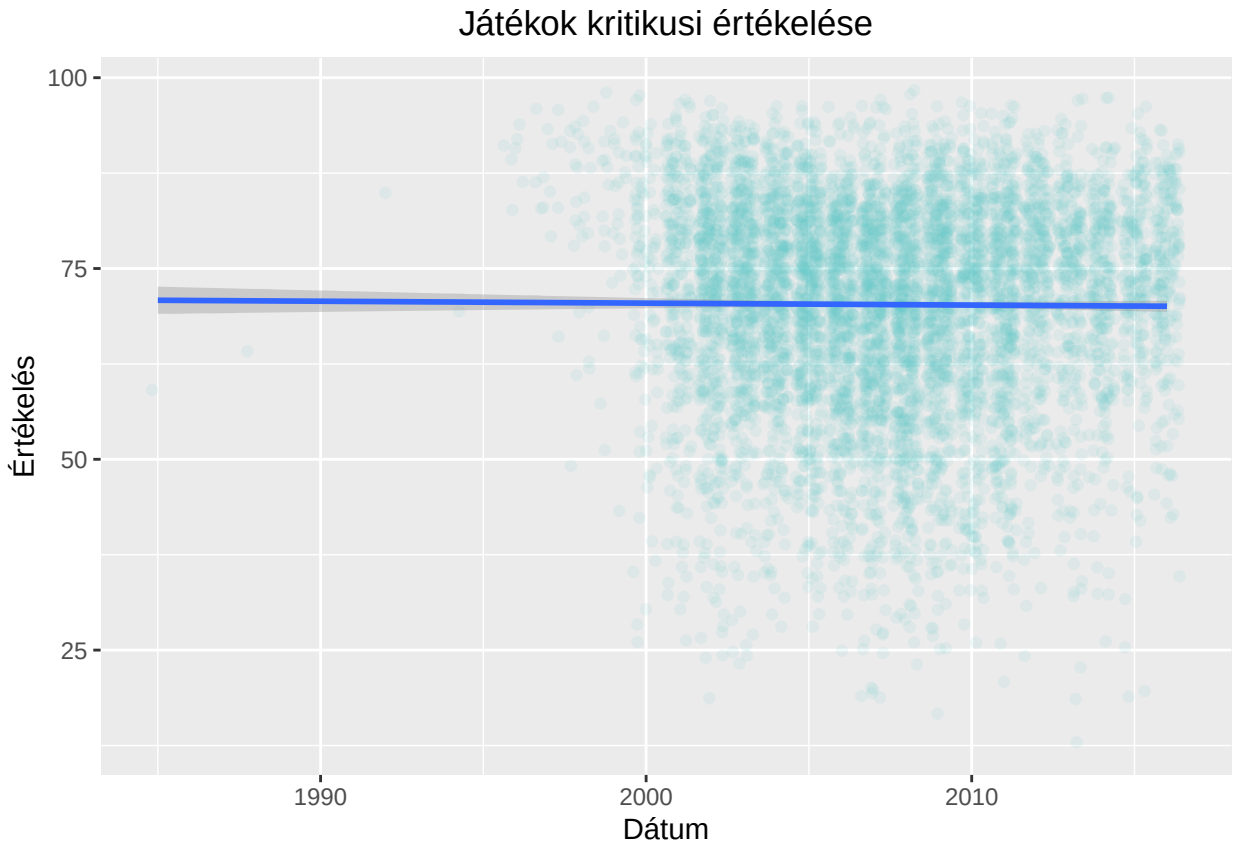


A pontdiagram és a trend alapján is úgy néz ki, csökkenés mutatkozik a játékosok elégedettségét illetően. Ez persze a játékok számának növekedésével, hígulásával is összefüggésben lehet.

Nézzük meg ugyanezt a diagramot a kritikusok véleménye alapján.

```
no_NA %>%
  ggplot(aes(Year_of_Release, Critic_Score))+
  geom_jitter(alpha=0.1, color="#66CCCC")+
  geom_smooth(method="lm")+
  labs(x="Dátum", y="Értékelés")+
  ggtitle("Játékok kritikusi értékelése")+
  theme(plot.title=element_text(hjust=0.5))
```

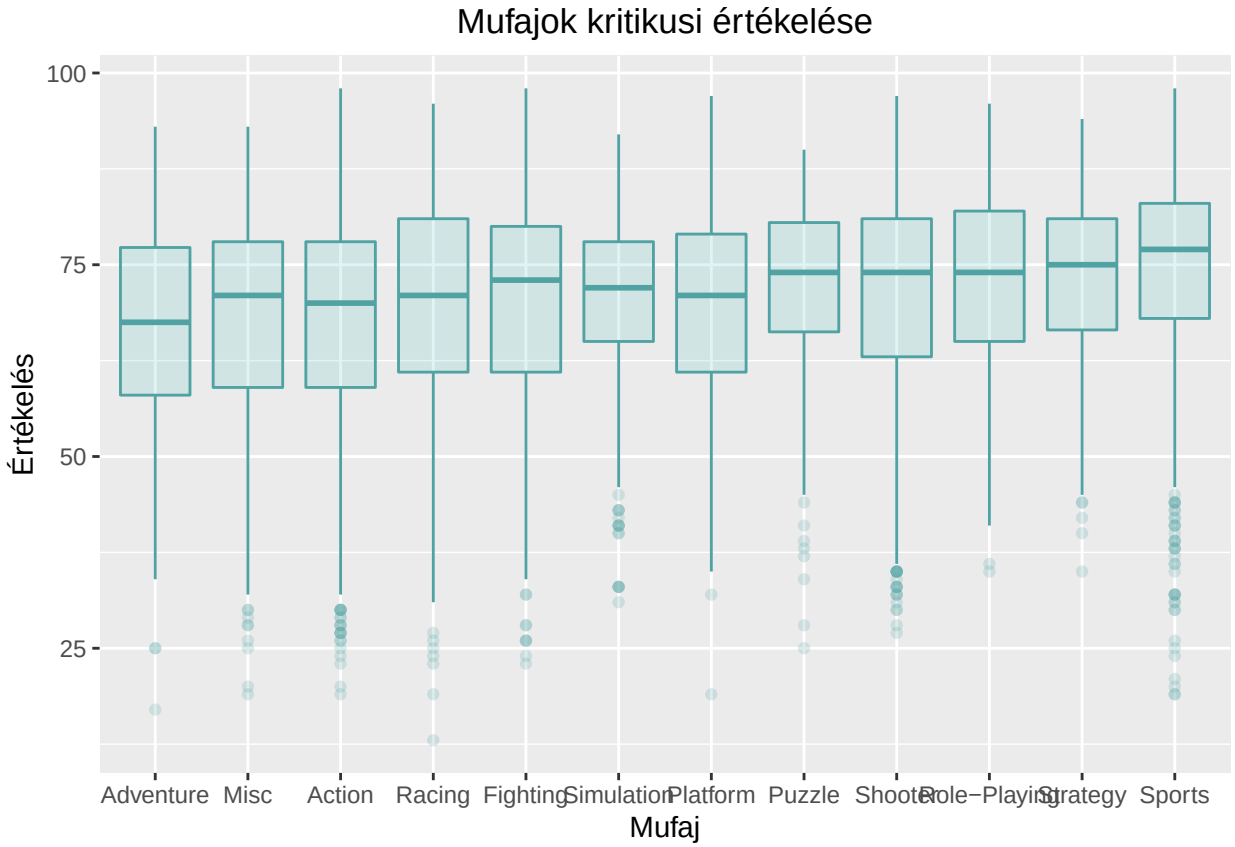
```
## `geom_smooth()` using formula 'y ~ x'
```



Érdekes módon a kritikusi vélemények átlagos értékelése az évek során jelentősen nem változott.

Ezt kicsit jobban is megvizsgálom:

```
no_NA %>%
  ggplot(aes(reorder(Genre, Critic_Score, function(x) + mean(x)), Critic_Score))+
  geom_boxplot(color="#51a3a3", fill="#66CCCC", alpha=0.2)+
  labs(x="Műfaj", y="Értékelés")+
  ggtitle("Műfajok kritikusi értékelése")+
  theme(plot.title=element_text(hjust=0.5))
```

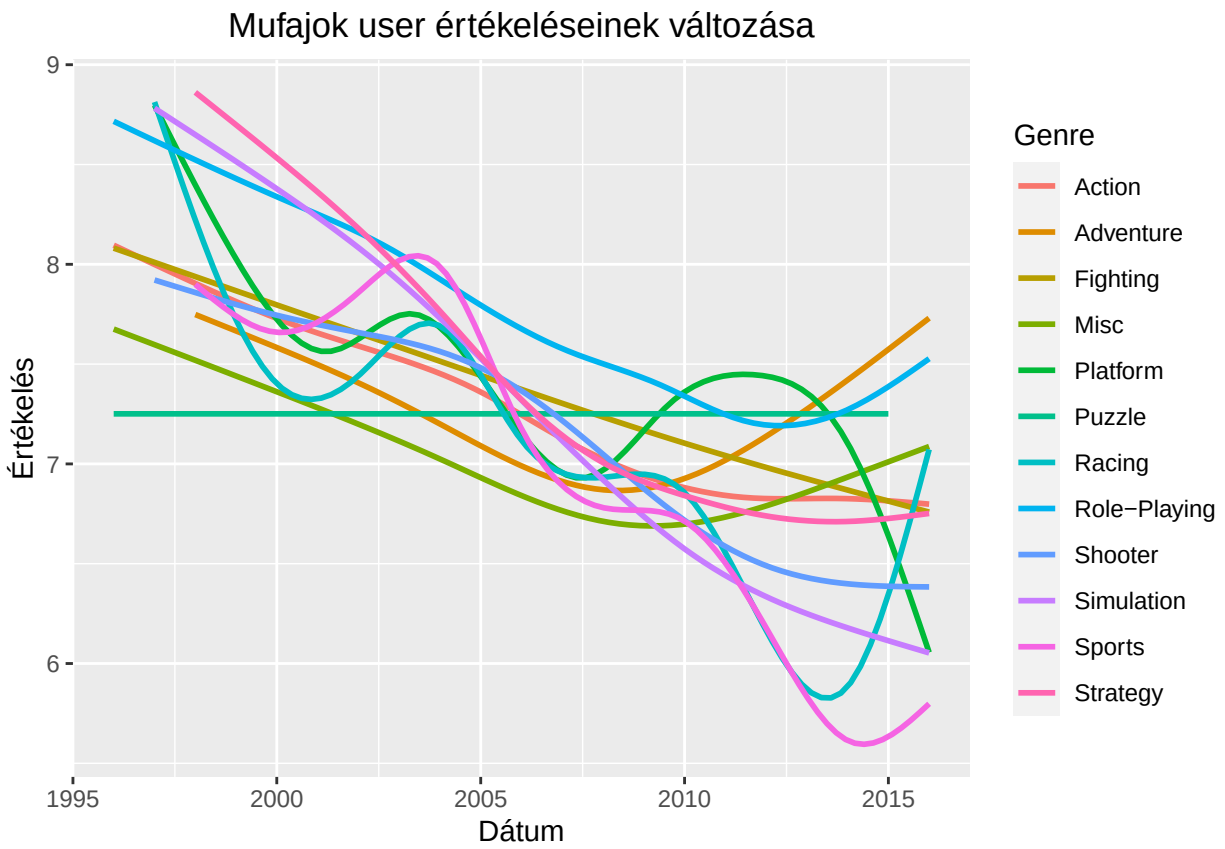


Láthatjuk, hogy minden kategóriában bőven akadnak lefelé kilógó értékek, és a nagy számban kiadott akció játékok átlagos értékelése bizony a legalacsonyabb értékek egyike, míg a sport és stratégia játékok rendelkeznek a legjobb átlag értékeléssel.

Megnézhetjük, hogyan vélekedtek ezzel szemben a játékosok évről évre.

```
no_NA %>%
  filter(year(Year_of_Release)>1995) %>%
  ggplot(aes(Year_of_Release, User_Score, color=Genre))+
  geom_smooth(se=FALSE)+
  labs(x="Dátum", y="Értékelés")+
  ggtitle("Műfajok user értékeléseinek változása")+
  theme(plot.title=element_text(hjust=0.5))
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

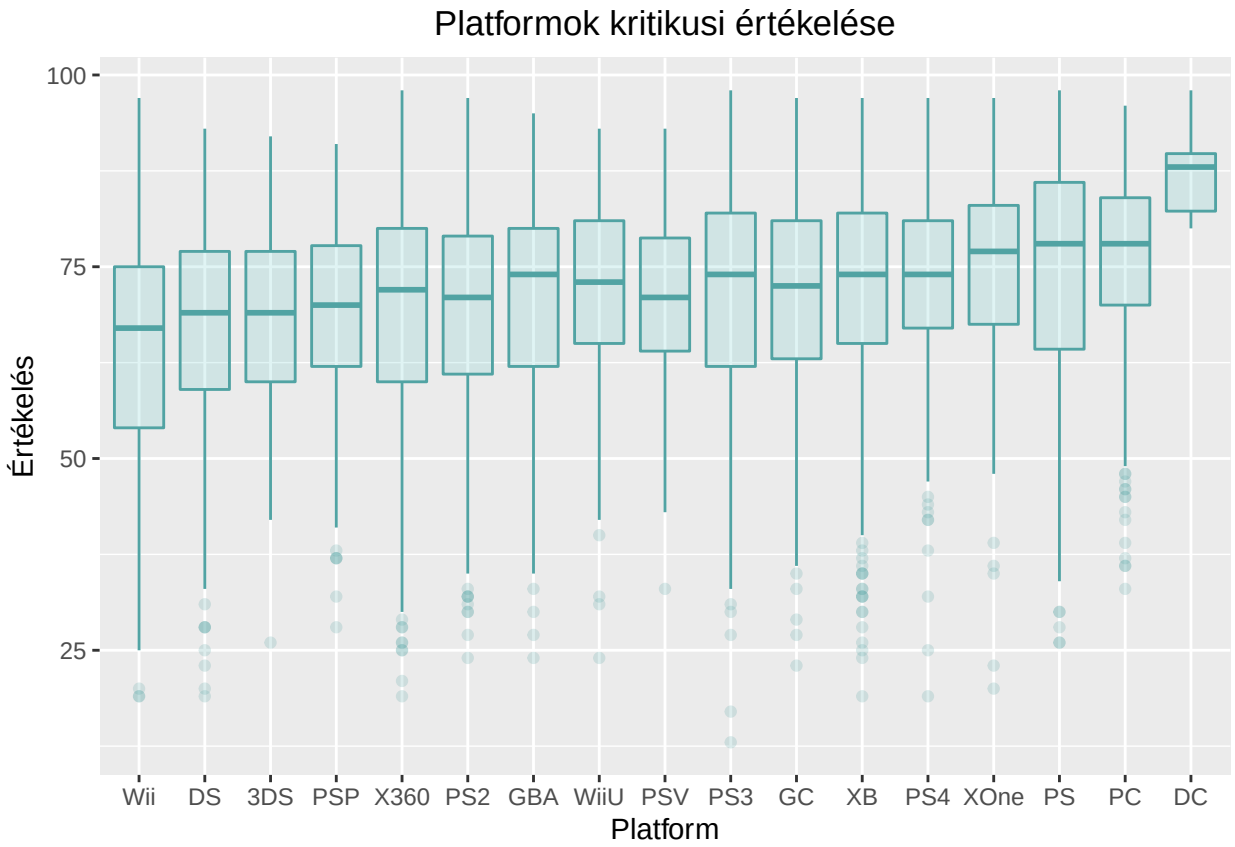


A legtöbb műfaj értékelése romlott idővel, a kalandjátékok, szerepjátékok értékelése javult az utóbbi években.

A stratégiai, platform és logikai játékok értékelése elég változó, míg a sport és szimulációs játékoké romlott a legerősebb módon.

Most hogy már tudjuk hogyan vélekednek a kritikusok és játékosok a műfajokról, nézzük meg a véleményüket a platformokat illetően.

```
no_NA %>%
  ggplot(aes(reorder(Platform, Critic_Score, function(x) + mean(x)), Critic_Score))+
  geom_boxplot(color="#51a3a3", fill="#66CCCC", alpha=0.2)+
  labs(x="Platform", y="Értékelés")+
  ggtitle("Platformok kritikusi értékelése")+
  theme(plot.title=element_text(hjust=0.5))
```



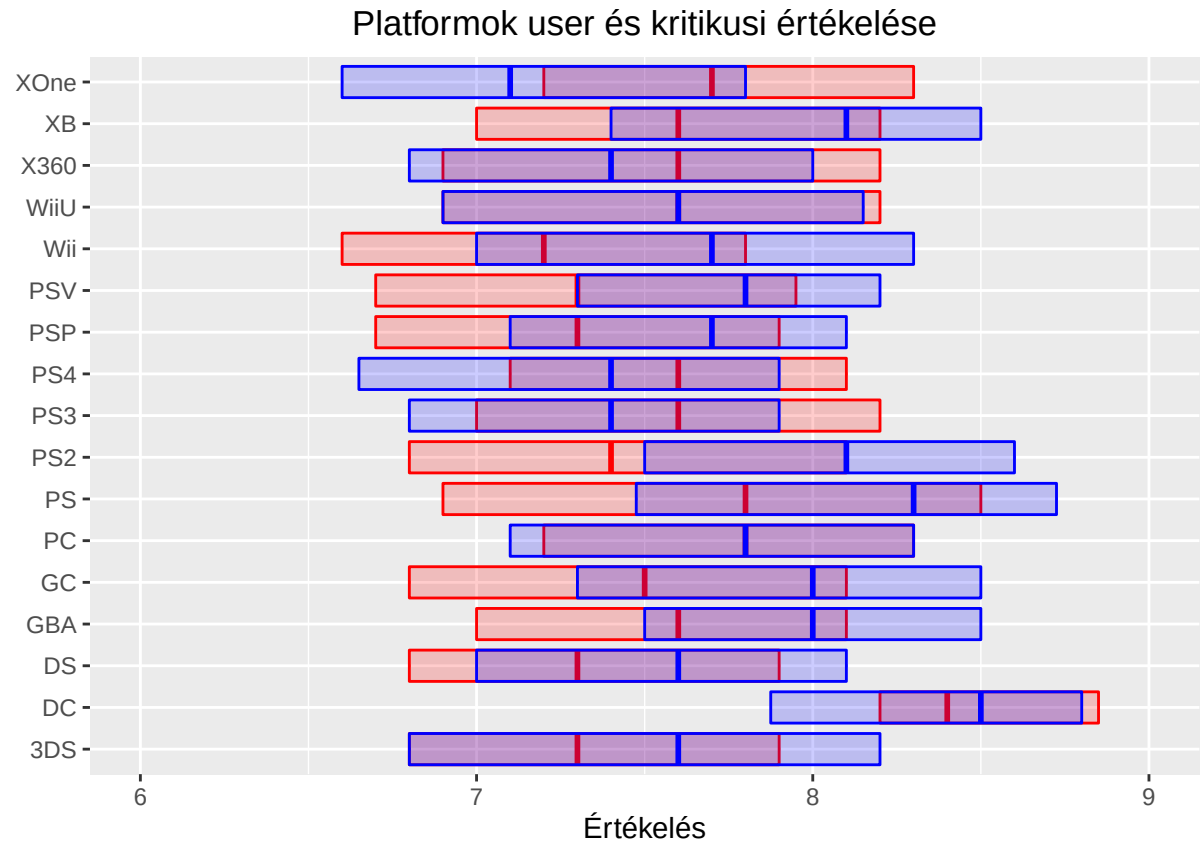
A Wii játékok hiába voltak nagy számban eladva, a kritikusok mégis a legrosszabb játékokkal rendelkező platformnak tartják. Magasan a Dreamcast platform játékai kapták a legjobb értékelést, ez a Sega 1998-as konzolja egyébként, és így nézett ki (érdekessége, hogy a kontrolleren is volt egy kis képernyő):

A jó értékeléshez persze hozzájárul, hogy jóval kevesebb játékról van szó mint a többi platform esetén, így

például a második legjobb játékokkal rendelkező PC talán még nagyobb eredménynek számít.

Érdekes lehet, hogy mennyire találkozik a kritikusok és felhasználók véleménye az egyes platformok esetén.

```
no_NA %>%
  ggplot()+
  geom_boxplot(aes(Platform, Critic_Score/10, ymin=..lower.., ymax=..upper..), fill="red", color="red",
  geom_boxplot(aes(Platform, User_Score, ymin=..lower.., ymax=..upper..), fill="blue", color="blue", al
  labs(x="", y="Értékelés")+
  ggtitle("Platformok user és kritikusi értékelése")+
  theme(plot.title=element_text(hjust=0.5))+
  coord_flip()+
  ylim(6,9)
```



Pirossal a kritikusok, kékkel a felhasználók értékeléseinek interkvartilis terjedelmeit, valamint az átlagukat látjuk. Meglepő módon egész kevésszer értenek egyet.

Az Xbox One esetében a felhasználók sokkal kevésbé elégedettek a játékokkal, míg az Xbox, Wii, PS Vita, PS2, PS és GameCube platformok esetén a kritikusoknak nyerték el kevésbé a tetszésüket.

TOP játékok

Nincs más hátra, mint az igazi gigászok vizsgálata, minden év legnagyobb példányszámban eladott játékainak (ami része az adathalmaznak). Tiszteletem jeléül álljon itt a teljes lista:

```
dat %>%
  group_by(Name) %>%
  group_by(Year_of_Release) %>%
  top_n(1, Global_Sales) %>%
```

```
  arrange(Year_of_Release) %>%
  mutate(year = year(Year_of_Release)) %>%
  ungroup() %>%
  select(year, Name, Global_Sales) %>%
  knitr::kable("html", col.names=c("Év", "Játék", "Nemzetközi eladás (millió darab)"), align="c")
```

Év

Játék

Nemzetközi eladás (millió darab)

1980

Asteroids

4.31

1981

Pitfall!

4.50

1982

Pac-Man

7.81

1983

Baseball

3.20

1984

Duck Hunt

28.31

1985

Super Mario Bros.

40.24

1986

The Legend of Zelda

6.51

1987

Zelda II: The Adventure of Link

4.38

1988

Super Mario Bros. 3

17.28

1989

Tetris

30.26

1990

Super Mario World

20.61

1991

The Legend of Zelda: A Link to the Past

4.61

1992

Super Mario Land 2: 6 Golden Coins

11.18

1993

Super Mario All-Stars

10.55

1994

Donkey Kong Country

9.30

1995

Donkey Kong Country 2: Diddy's Kong Quest

5.15

1996

Pokemon Red/Pokemon Blue

31.37

1997

Gran Turismo

10.95

1998

PokĂ©mon Yellow: Special Pikachu Edition

14.64

1999

Pokemon Gold/Pokemon Silver

23.10

2000

PokĂ©mon Crystal Version

6.39

2001

Gran Turismo 3: A-Spec

14.98

2002

Grand Theft Auto: Vice City

16.15

2003

Need for Speed Underground

7.20

2004

Grand Theft Auto: San Andreas

20.81

2005

Nintendogs

24.67

2006

Wii Sports

82.53

2007

Wii Fit

22.70

2008

Mario Kart Wii

35.52

2009

Wii Sports Resort

32.77

2010

Kinect Adventures!

21.81

2011

Call of Duty: Modern Warfare 3

14.73

2012

Call of Duty: Black Ops II

13.79

2013

Grand Theft Auto V

21.04

2014

Grand Theft Auto V

12.61

2015

Call of Duty: Black Ops 3

14.63

2016

FIFA 17

7.59

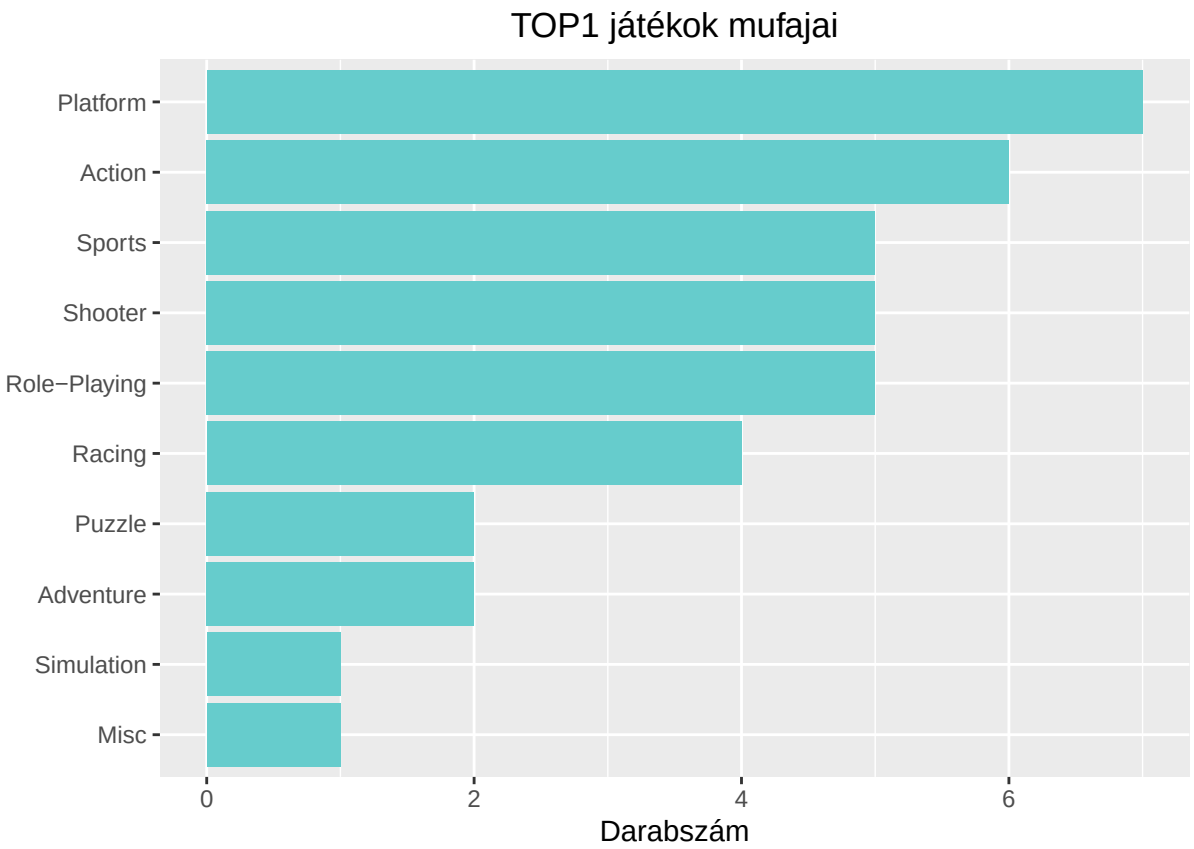
2017

Phantasy Star Online 2 Episode 4: Deluxe Package

0.04

Ha már nem bogarásszuk a neveket, vizualizáljuk az adataikat. Mondjuk kíváncsiak lehetünk rá, milyen műfajból került ki a legtöbb gigász.

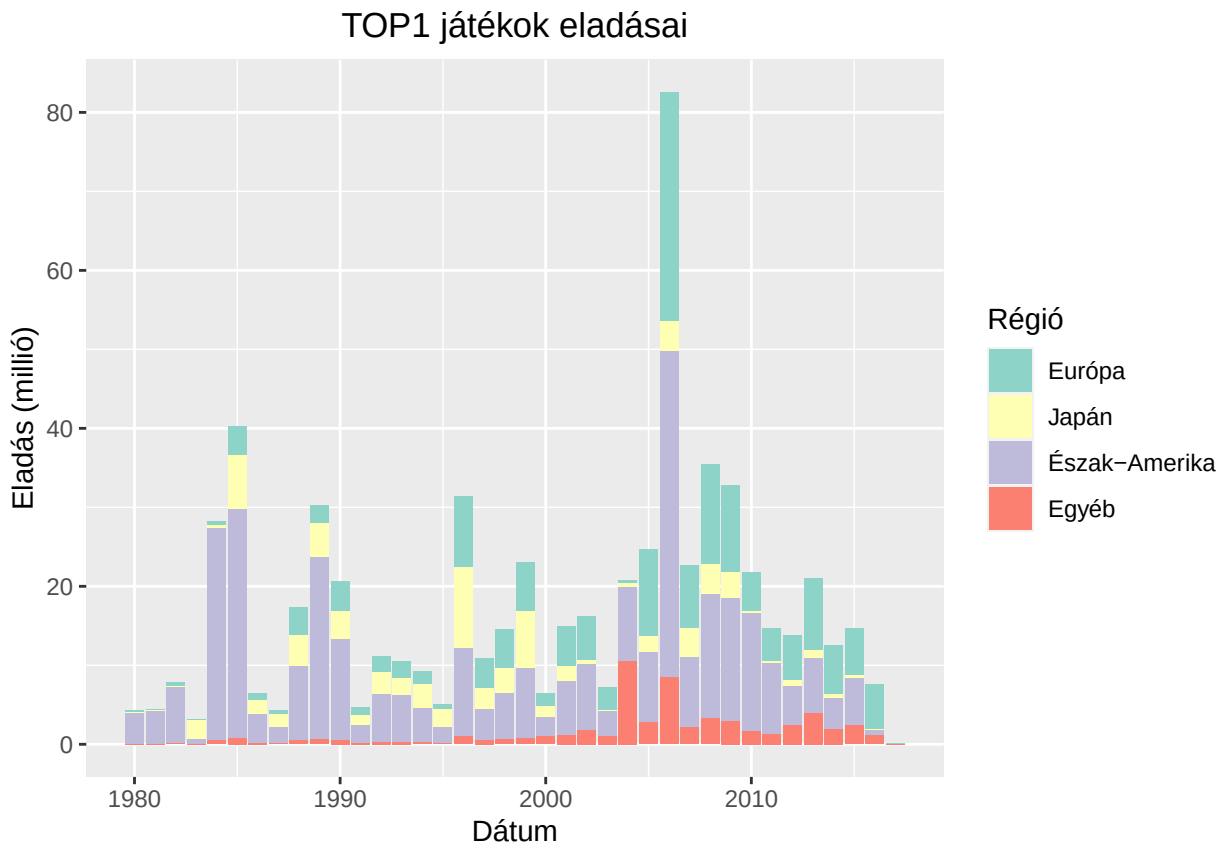
```
dat %>%
  group_by(Name) %>%
  group_by(Year_of_Release) %>%
  top_n(1, Global_Sales) %>%
  ggplot(aes(reorder(Genre,Genre,function(x)+length(x))))+
  geom_bar(fill="#66CCCC")+
  coord_flip()+
  labs(x="", y="Darabszám")+
  ggtitle("TOP1 játékok műfajai")+
  theme(plot.title=element_text(hjust=0.5))
```



Platform játékok alatt azokat az elsősorban retró játékokat értjük, ahol 2D-ben levő környezetben kell emelvényekről, vagyis platformokról ugrálni és előre (hátra) futni.

Nagy meglepetés nincs a további sorrendben, nem mond ellent a korábban elemzett kedveltségnek.

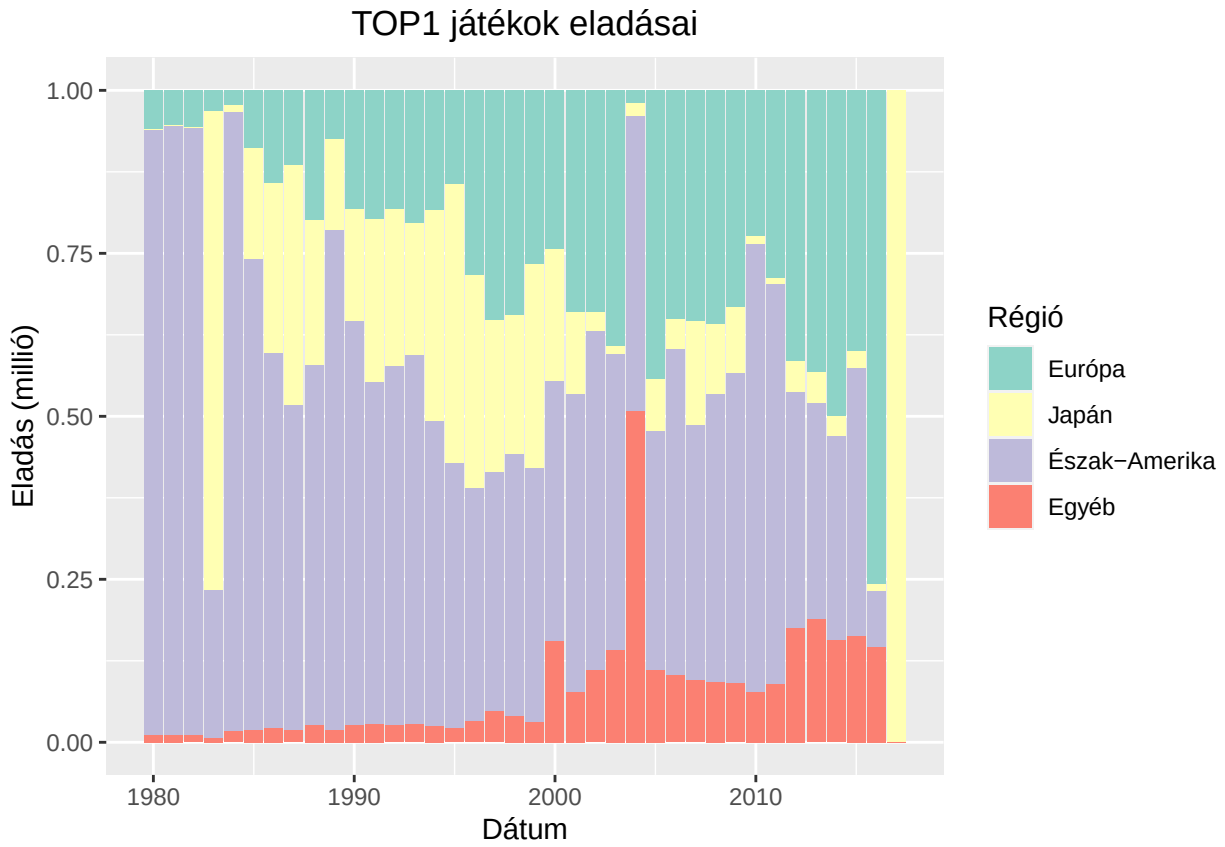
```
dat %>%
  group_by(Name) %>%
  group_by(Year_of_Release) %>%
  top_n(1, Global_Sales) %>%
  select(Year_of_Release, NA_Sales, EU_Sales, JP_Sales, Other_Sales) %>%
  gather(type, count, -Year_of_Release) %>%
  ggplot(aes(as.Date(Year_of_Release, "%Y"), count, fill=factor(type)))+
  geom_bar(stat="identity")+
  labs(x="Dátum", y="Eladás (millió)", fill="Régió")+
  ggtitle("TOP1 játékok eladásai")+
  theme(plot.title=element_text(hjust=0.5))+
  scale_fill_brewer(palette="Set3", labels=c("Európa", "Japán", "Észak-Amerika", "Egyéb"))
```



Az első diagramon látható, hogy a kétezres évek előtt az átlagos TOP1 játék eladási száma viszonylag alacsonynak mondható, míg vannak kimondottan magas számokkal rendelkező társaik is.

2000 után ez a fajta forma megtört, és egy sokkal kiszámíthatóbban mozgó ábra rajzolódik ki.

```
dat %>%
  group_by(Name) %>%
  group_by(Year_of_Release) %>%
  top_n(1, Global_Sales) %>%
  select(Year_of_Release, NA_Sales, EU_Sales, JP_Sales, Other_Sales) %>%
  gather(type, count, -Year_of_Release) %>%
  ggplot(aes(as.Date(Year_of_Release, "%Y"), count, fill=factor(type)))+
  geom_bar(stat="identity", position="fill")+
  labs(x="Dátum", y="Eladás (millió)", fill="Régió")+
  ggtitle("TOP1 játékok eladásai")+
  theme(plot.title=element_text(hjust=0.5))+
  scale_fill_brewer(palette="Set3", labels=c("Európa", "Japán", "Észak-Amerika", "Egyéb"))
```



A második, 100%-ig halmozott ábrán pedig jól megfigyelhető, ahogy Európa felvásárlási ereje egyre jelentősebbé vált, míg Japáné szinte megszűnt. Észak-Amerika viszonylagos stabilitást mutat, az egyéb régiók pedig az ezredforduló után váltak jelentőssé.

Úgy gondolom, elejétől a végéig sikerült kivesézni az adattáblát, és nagyon sok érdekes következtetést tudtunk hozni.

Köszönöm a figyelmet, remélem aki olvasta érdekesnek találta!