Tipologia i Cicle de Vida de les Dades. Pràctica 2

Autor: Marta Montclus, Jose Montufo
Assignatura: Tipologia i cicle de vida de les dades
Codi d'assignatura: M2.951
Aula: 1
Curs: 2020-21, semestre 2
Professor: Xavier Vivancos Garcia

Taula de continguts:

1. Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre? 3

2. Integració i selecció de les dades d'interès a analitzar. 4

3. Neteja de les dades. 5

4. Anàlisi de les dades. 18

Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar). 18

Comprovació de la normalitat i homogeneïtat de la variància. 19

Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents. 22

Anàlisi 1. Quins atributs quantitatius afecten més al preu i a la puntuació? Són els mateixos? 22

Anàlisi 2. Com afecta la saga del joc al seu preu? Com afecta l'editorial a la puntuació dels jocs? 23

Anàlisi 3. Model de regressió lineal. Predicció del preu. 25

5. Representació dels resultats a partir de taules i gràfiques. 31

6. Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema? 34

Contribucions 36

1. Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?

El dataset utilitzat és **"Catàleg de jocs de taula de la botiga online de Zacatrus.es"**, el creat a la Pràctica 1.
És un catàleg de jocs de taula amb les seves característiques, extret de la pàgina web Zacatrus.es i complementat amb dades de la base de dades BGG (Board Game Geek), la més gran que existeix a la xarxa sobre jocs de taula.

- Relació entre la puntuació dels jocs i el seu preu.
- Relació entre altres camps (nombre de jugadors, tipus de joc, durada de la partida, etc.) i el preu del joc.
- Relació entre altres camps i la puntuació del joc a la BGG.
- Fer una valoració del preu d'un joc a partir de la resta de camps.

2. Integració i selecció de les dades d'interès a analitzar.

Al procés de captura de les dades es va dur a terme una tasca d'integració de dades. D'una banda, disposem de les dades dels jocs de taula que Zacatrus publicava a la seva botiga online. A aquestes dades, mitjançant l'identificador de la BGG que es proporciona per a part dels jocs de taula, es va integrar la puntuació mitjana dels usuaris de la BGG per a cadascun d'ells. Més enllà d'aquesta integració prèvia, no és procedent integrar més fonts de dades, l'anàlisi correspon només a dades del dataset indicat.

Carreguem les dades de la següent manera:

```
zacatrus_data <- read.csv(".\\games_wih_rating.csv", sep=';',
header=TRUE)
```

De les dades del dataset, fem una selecció inicial de quines poden ser útils per a les futures anàlisis que realitzem, i quines podem descartar d'entrada.

Si analitzem les files, com que tenim dues variables objectius diferents per a les nostres anàlisis (preu i puntuació). Depenent en cada moment de l'anàlisi a realitzar, haurem de filtrar els jocs que no tinguin preu (o sigui 0), i/o els que no tinguin puntuació.

Alguns jocs es troben repetits, en la seva versió a estrenar, i en la seva versió "Kilometro 0", que són de segona mà i per tant el seu preu és menor. Aquests últims els haurem de retirar del dataset perquè introdueixen valors repetits a la puntuació, i valors incorrectes als preus.

```
zacatrus_data <- dplyr::filter(zacatrus_data, !grepl("Kilómetro
0",Nombre))
```

Respecte a les columnes, trobem les següents columnes que eliminarem del dataset:

- **Num. ID:** Aquesta columna va ser creada exclusivament per relacionar els jocs amb les seves imatges extretes de la pàgina de Zacatrus. És un identificador que no aporta cap mena d'informació sobre el joc. La podem eliminar.
- **Disponibilitat (Disponibilidad)**: El fet que un joc es trobi disponible o no a Zacatrus en el moment de recopilar les dades no influeix ni en el seu preu ni en la valoració a la BGG. Eliminarem aquest camp.
- **BGG**: És un identificador, i per tant no proporciona informació a l'anàlisi. L'eliminem.
- **Idioma**: La llengua en la qual Zacatrus ven el joc no hauria d'afectar el preu ni la valoració a la BGG. Per tant, eliminem aquest camp.

```
zacatrus_data <- select(zacatrus_data,-Num..Id, -Disponibilidad,
-BGG, -Idioma)
```

3. Neteja de les dades.

Per a la resta de columnes, detallem una a una el procés de neteja realitzat

Nom (Nombre): El nom del joc pot proporcionar informació útil, però s'han de crear nous atributs a partir d'ell. Per exemple, de cara a les anàlisis és important identificar jocs que pertanyen a un mateix joc amb les seves expansions (per exemple, Everdell), o a una franquicia en concret (per exemple, Star Wars). Tot i que no és l'opció més acurada per crear aquest camp, el construirem a partir de la primera paraula de més de tres lletres del títol.

Comprovem que després de la transformació hi ha valors nulls. Els substituïm pel valor "Desconegut"

Com que és un valor textual, no aplica trobar valors extrems.

Autor: El nom de l'autor d'un joc pot ser important de cara a determinar el preu dels jocs de taula, i la seva valoració. No obstant això, aquest atribut presenta tres problemes:

- Els autors d'un joc poden ser més d'un. En el nostre dataset, trobem jocs amb fins a 6 autors diferents.
- Els autors és una variable categòrica amb moltíssims valors diferents possibles.
- Hi ha autors que el nom es troba escrit de manera diferent en dos jocs que ha creat.

Tot i aquests inconvenients, processarem les dades dels autors per comprovar si realment poden ser útils per a la nostra anàlisi. Per fer-ho, transformem l'atribut "Autor" en 6 columnes, a on apareixen els diferents autors de cada joc.

```
zacatrus_data <- separate(zacatrus_data, col="Autor", into =
c("Autor1", "Autor2", "Autor3", "Autor4", "Autor5", "Autor6"), sep =
",")
```

Els valors buits o nuls a les variables AutorN, els convertim en “Desconegut”. El fet que no tingui cap autor, o que només tingui 2, pot ser informatiu.

```
zacatrus_data$Autor1[zacatrus_data$Autor1 == ''] <- NA
zacatrus_data$Autor1[is.na(zacatrus_data$Autor1)] <- "Desconegut"
zacatrus_data$Autor2[is.na(zacatrus_data$Autor2)] <- "Desconegut"
zacatrus_data$Autor3[is.na(zacatrus_data$Autor3)] <- "Desconegut"
zacatrus_data$Autor4[is.na(zacatrus_data$Autor4)] <- "Desconegut"
zacatrus_data$Autor5[is.na(zacatrus_data$Autor5)] <- "Desconegut"
zacatrus_data$Autor6[is.na(zacatrus_data$Autor6)] <- "Desconegut"
```

Finalment, en ser un camp categòric, no aplica trobar valors extrems.

Temàtica: També és una variable categòrica, tot i que en aquest cas el nombre de valors possibles és molt menor. Per aquest motiu, transformem l’atribut en 42 atributs diferents, un per cadascuna de les temàtiques possibles, a on cada nou atribut de tipus booleà indicarà si el joc pertany a la temàtica o no:

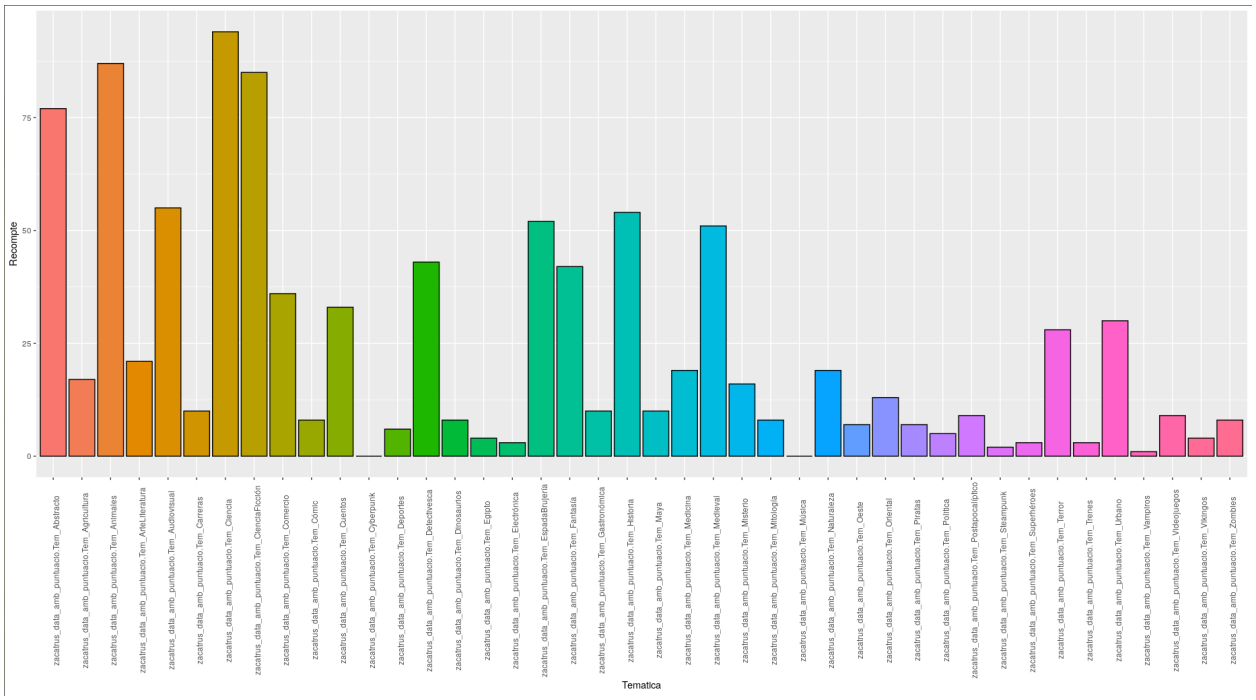
```
zacatrus_data <- transform(zacatrus_data, Tem_Abstracto =
grepl("Abstracto",Temática))
zacatrus_data <- transform(zacatrus_data, Tem_Comercio =
grepl("Comercio",Temática))
zacatrus_data <- transform(zacatrus_data, Tem_Egipto =
grepl("Egipto",Temática))
...
zacatrus_data <- transform(zacatrus_data, Tem_Terror =
grepl("Terror",Temática))
zacatrus_data <- transform(zacatrus_data, Tem_Zombies =
grepl("Zombies",Temática))

zacatrus_data <- select(zacatrus_data,-Temática)
```

Si un joc no té cap temàtica assignada, simplement tindrà tots els valors dels nous atributs a fals. Tampoc no aplica cap mena de gestió de valors extrems.

A continuació visualitzem amb un gràfic de barres el nombre de jocs amb cada temàtica.

```
tematica<-zacatrus_data_amb_puntuacio[,
c("Tem_Abstracto","Tem_Comercio","Tem_Egipto","Tem_Medicina","Tem_Ori
ental","Tem_Trenes","Tem_Agricutura","Tem_Cómic","Tem_Electrónica","T
```

```
em_Medieval","Tem_Piratas","Tem_Audiovisual","Tem_Animales","Tem_Cuen
tos","Tem_EspadaBrujería","Tem_Misterio","Tem_Politica","Tem_Urbano",
"Tem_ArteLiteratura","Tem_Cyberpunk","Tem_Fantasía","Tem_Mitología","
Tem_Postapocalíptico","Tem_Vampiros","Tem_Carreras","Tem_Deportes","T
em_Gastronómica","Tem_Música","Tem_Steampunk","Tem_Videojuegos","Tem_
Ciencia","Tem_Detectivesca","Tem_Historia","Tem_Naturaleza","Tem_Supe
rhéroes","Tem_Vikingos","Tem_CienciaFicción","Tem_Dinosaurios","Tem_M
aya","Tem_Oeste","Tem_Terror","Tem_Zombies")]

sumdata=data.frame(value=apply(tematica,2,sum))
sumdata$key=rownames(sumdata)
ggplot(data=sumdata, aes(x=key, y=value,
fill=key))+geom_bar(colour="black",
stat="identity")+theme(legend.position="none",axis.text.x =
element_text(angle = 90))+ylab("Recompte")+xlab("Tematica")
```



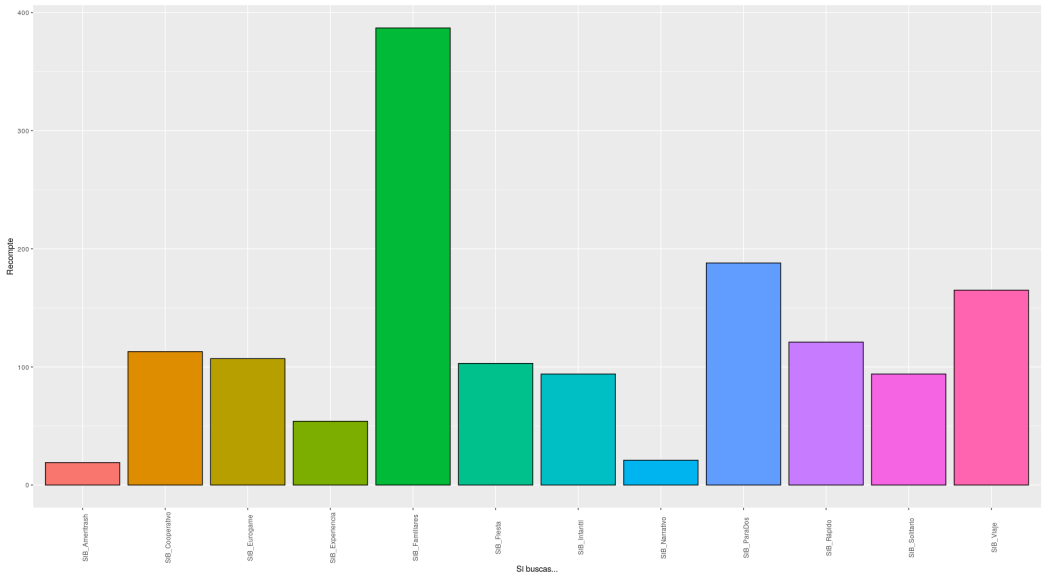
Si Buscas... : Aquest camp conté una llista amb les tipologies en les quals es pot encabir del joc. Les característiques d'aquest atribut són exactament les mateixes que les de la temàtica, i per tant apliquem la mateixa transformació:

```
zacatrus_data <- transform(zacatrus_data, SiB_Ameritrash =
grepl("Ameritrash",Si.Buscas...))
zacatrus_data <- transform(zacatrus_data, SiB_Cooperativo =
grepl("Cooperativo",Si.Buscas...))
...
```

```
zacatrus_data <- transform(zacatrus_data, SiB_Solitario =
grepl("Solitario",Si.Buscas...))
zacatrus_data <- transform(zacatrus_data, SiB_Viaje =
grepl("Viaje",Si.Buscas...))

zacatrus_data <- select(zacatrus_data,-Si.Buscas...)
```

A continuació visualitzem amb un gràfic de barres el nombre de jocs de cada grup "Si buscas...".

```
sibuscas<-zacatrus_data_amb_puntuacio[,
c("SiB_Ameritrash","SiB_Cooperativo","SiB_Eurogame","SiB_Experiencia"
,"SiB_Familiares","SiB_Fiesta","SiB_Infantil","SiB_Narrativo","SiB_Pa
raDos","SiB_Rápido","SiB_Solitario","SiB_Viaje")]

sumdata=data.frame(value=apply(sibuscas,2,sum))
sumdata$key=rownames(sumdata)
ggplot(data=sumdata, aes(x=key, y=value,
fill=key))+geom_bar(colour="black",
stat="identity")+theme(legend.position="none",axis.text.x =
element_text(angle = 90))+ylab("Recompte")+xlab("Si buscas...")
```



Edat (Edad): Per obtenir les dades de l'edat adequada pels jocs, mantindrem tota la informació si la guardem en dos atributs, un amb l'edat mínima i un altre amb la màxima (suposem que no existeixen "forats" en els rangs d'edat recomanada). Aquests dos nous atributs seran de tipus categòric ordinal.

```
zacatrus_data <- transform(zacatrus_data, EdadMinima= ifelse(grepl("de 0 a 3 años"  ,Edad), 0,
                                                     ifelse(grepl("de 3 a 6 años"  ,Edad), 3,
                                                     ifelse(grepl("de 6 a 8 años"  ,Edad), 6,
                                                     ifelse(grepl("de 8 a 10 años" ,Edad), 8,
                                                     ifelse(grepl("de 10 a 14 años",Edad), 10,
                                                     ifelse(grepl("de 14 a 18 años",Edad), 14,
                                                     ifelse(grepl("más de 18 años" ,Edad), 18,
                                                     NA)))))))

zacatrus_data <- transform(zacatrus_data, EdadMaxima= ifelse(grepl("más de 18 años" ,Edad), 99,
                                                     ifelse(grepl("de 14 a 18 años",Edad), 18,
                                                     ifelse(grepl("de 10 a 14 años",Edad), 14,
                                                     ifelse(grepl("de 8 a 10 años" ,Edad), 10,
                                                     ifelse(grepl("de 6 a 8 años"  ,Edad), 8,
                                                     ifelse(grepl("de 3 a 6 años"  ,Edad), 6,
                                                     ifelse(grepl("de 0 a 3 años"  ,Edad), 3,
                                                     NA)))))))


zacatrus_data <- select(zacatrus_data,-Edad)
```

En aquest cas sí que s'han de gestionar els valors buits. Per fer-ho, assignem per als jocs que no proporcionen la informació la mitjana d'edat mínima en els jocs que sí que ho fan, i la mitjana d'edat màxima.

```
aux <- dplyr::filter(zacatrus_data, !is.na(EdadMaxima))
mitjana_edat_maxima<-mean(aux$EdadMaxima)
zacatrus_data$EdadMaxima[is.na(zacatrus_data$EdadMaxima)] <-
mitjana_edat_maxima

mitjana_edat_minima<-mean(aux$EdadMinima)
zacatrus_data$EdadMinima[is.na(zacatrus_data$EdadMinima)] <-
mitjana_edat_minima
```

En aquest atribut tampoc cal gestionar els valors extrems.

Núm. jugadores: Per aquest atribut, les decisions preses són les mateixes que per l'edat. El substituïm per dos atributs nous, amb el nombre mínim i màxim de jugadors, i substituïm els valors buits per la mitjana de cadascun dels atributs creats.

```
zacatrus_data <- transform(zacatrus_data, Núm..jugadores = str_replace_all(Núm..jugadores, " ", ""))
zacatrus_data <- transform(zacatrus_data, NumJugadoresMinimo = ifelse(stri_length(Núm..jugadores) > 0, substr(Núm..jugadores,1,1) , NA))
zacatrus_data <- transform(zacatrus_data, NumJugadoresMaximo = ifelse(stri_length(Núm..jugadores) > 0, ifelse(grepl("+8",Núm..jugadores) , "8",
substr(Núm..jugadores,stri_length(Núm..jugadores),stri_length(Núm..jugadores))), NA))

zacatrus_data$NumJugadoresMinimo <-  as.numeric(zacatrus_data$NumJugadoresMinimo)
zacatrus_data$NumJugadoresMaximo <-  as.numeric(zacatrus_data$NumJugadoresMaximo)

zacatrus_data <- select(zacatrus_data,-Núm..jugadores)
```

```
aux <- dplyr::filter(zacatrus_data, !is.na(NumJugadoresMinimo))
mitjana_num_jug_min<-mean(aux$NumJugadoresMinimo)
zacatrus_data$NumJugadoresMinimo[is.na(zacatrus_data$NumJugadoresMinimo)] <- mitjana_num_jug_min

aux <- dplyr::filter(zacatrus_data, !is.na(NumJugadoresMaximo))
mitjana_num_jug_max<-mean(aux$NumJugadoresMaximo)
zacatrus_data$NumJugadoresMaximo[is.na(zacatrus_data$NumJugadoresMaximo)] <- mitjana_num_jug_max
```

Tiempo de juego: Pel que fa a la transformació, per aquest atribut fem exactament el mateix que en els casos anteriors. També la gestió de valors buits és la mateixa. Abans de calcular la mitjana, però, en aquest cas sí que s'han de controlar els valors extrems.

```
zacatrus_data <- transform(zacatrus_data, Tiempo.de.juego.mod = str_replace_all(Tiempo.de.juego, regex("[/|x]"), "-"))
zacatrus_data <- transform(zacatrus_data, Tiempo.de.juego.mod = str_replace_all(Tiempo.de.juego, regex("[^0-9|-]+"), ""))
zacatrus_data <- transform(zacatrus_data, TiempoDeJuegoMin = as.numeric(str_replace_all(Tiempo.de.juego.mod, regex("-[0-9]*"), "")))
zacatrus_data <- transform(zacatrus_data, TiempoDeJuegoMax = as.numeric(str_replace_all(Tiempo.de.juego.mod, regex("[0-9]*-"), "")))


zacatrus_data <- select(zacatrus_data,-Tiempo.de.juego)
zacatrus_data <- select(zacatrus_data,-Tiempo.de.juego.mod)
```

Comprovem que hi ha valors que són erronis amb molta probabilitat. Els posem a nul perquè siguin substituïts pel valor mitjà de l'atribut:

```
select(dplyr::filter(zacatrus_data, TiempoDeJuegoMin > 300), Nombre, TiempoDeJuegoMin)
```

Nombre <chr>	TiempoDeJuegoMin <dbl>
Adventure	390
Adventure	390
Escape	450
Bugs	90120
Horizons	6075

5 rows

```
zacatrus_data$TiempoDeJuegoMin[zacatrus_data$TiempoDeJuegoMin > 300] <- NA

zacatrus_data$TiempoDeJuegoMax[zacatrus_data$TiempoDeJuegoMax > 300] <- NA
```

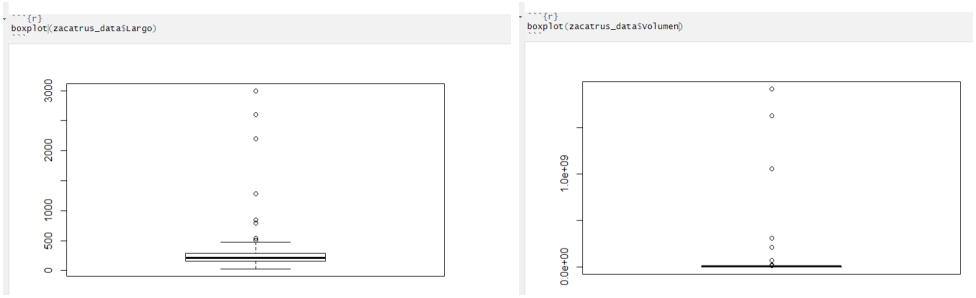
Finalment, establim la mitjana com a valor dels valors nuls.

```
aux <- dplyr::filter(zacatrus_data, !is.na(TiempoDeJuegoMin))
mitjana_temps_joc_min<-mean(aux$TiempoDeJuegoMin)
zacatrus_data$TiempoDeJuegoMin[is.na(zacatrus_data$TiempoDeJuegoMin)] <- mitjana_temps_joc_min

aux <- dplyr::filter(zacatrus_data, !is.na(TiempoDeJuegoMax))
mitjana_temps_joc_max<-mean(aux$TiempoDeJuegoMax)
zacatrus_data$TiempoDeJuegoMax[is.na(zacatrus_data$TiempoDeJuegoMax)] <- mitjana_temps_joc_max
```

Mides (Medidas): Crearem dos atributs amb aquest camp. D'una banda, el llarg (el valor màxim dels tres proporcionats), i d'altra, el volum de la caixa del joc.

```
zacatrus_data <- transform(zacatrus_data, Medidas = str_replace_all(Medidas," ", ""))

r <- regexpr("([0-9]+x[0-9]+x[0-9]+)", zacatrus_data$Medidas)
zacatrus_data$MedidasModif[which(r != -1)] <- regmatches(zacatrus_data$Medidas, r)

zacatrus_data <- transform(zacatrus_data, Medida1 = ifelse(grepl("cm",Medidas), 10,1) * as.numeric(str_replace_all(MedidasModif, regex("x[0-9|x]*"), "")))
zacatrus_data <- transform(zacatrus_data, Medida2 = str_replace_all(MedidasModif, regex("x[0-9]*$"), ""))
zacatrus_data <- transform(zacatrus_data, Medida2 = ifelse(grepl("cm",Medidas), 10,1) * as.numeric(str_replace_all(Medida2, regex("[0-9|x]*x"), "")))
zacatrus_data <- transform(zacatrus_data, Medida3 = ifelse(grepl("cm",Medidas), 10,1) * as.numeric(str_replace_all(MedidasModif, regex("[0-9|x]*x"), "")))

zacatrus_data <- transform(zacatrus_data, Largo = pmax(Medida1, Medida2, Medida3))
zacatrus_data <- transform(zacatrus_data, Volumen = Medida1 * Medida2 * Medida3)

zacatrus_data <- select(zacatrus_data,-Medidas, -MedidasModif)
```

Comprovem valors extrems:



Resulta molt evident que en la transformació, hi ha valors que no s'han obtingut de forma correcta, o que no estan ben definits a les dades. Com a exemple, trobem aquests jocs amb volums en centímetres, i que hauria d'indicar mil·límetres (a no ser que la capsa del joc mesuri dos metres amb 20 de llarg, fet poc probable):

2697		Edge Entertainment	
2698	130 x 220 x 37 cm.	Edge Entertainment	
2699	130 x 220 x 37 cm.	Edge Entertainment	
2700	130 x 220 x 37 cm.	Edge Entertainment	
2701	130 x 220 x 37 cm.	Edge Entertainment	
2702	300 x 165 x 40 mm.	Edge Entertainment	
2703	130 x 220 x 37 cm.	Edge Entertainment	
2704	140 x 155 x 80 mm	Edge Entertainment	
2705	311 x 203 x 102 mm.	Edge Entertainment	
2706		Maldito Games	Media

Posem a nul tots els llargs i els volums dels jocs que tenen una mida per sobre del tercer quartil.

```
zacatrus_data$Largo[zacatrus_data$Largo > 288] <- NA
zacatrus_data$Volumen[zacatrus_data$Volumen > 4092000] <- NA
```

I seguim la mateixa estratègia que amb l'edat recomanada i el nombre de jugadors per emplenar els valors perduts. Assignem la mitjana dels atributs:

```
aux <- dplyr::filter(zacatrus_data, !is.na(Largo))
mitjana_largo<-mean(aux$Largo)
zacatrus_data$Largo[is.na(zacatrus_data$Largo)] <- mitjana_largo

aux <- dplyr::filter(zacatrus_data, !is.na(Volumen))
mitjana_volumen<-mean(aux$Volumen)
zacatrus_data$Volumen[is.na(zacatrus_data$Volumen)] <-
mitjana_volumen
```

Complexitat (Complejidad): És un atribut a priori vàlid. Com que és un valor de tipus ordinal, transformem els seus tres possibles valors a numèric.

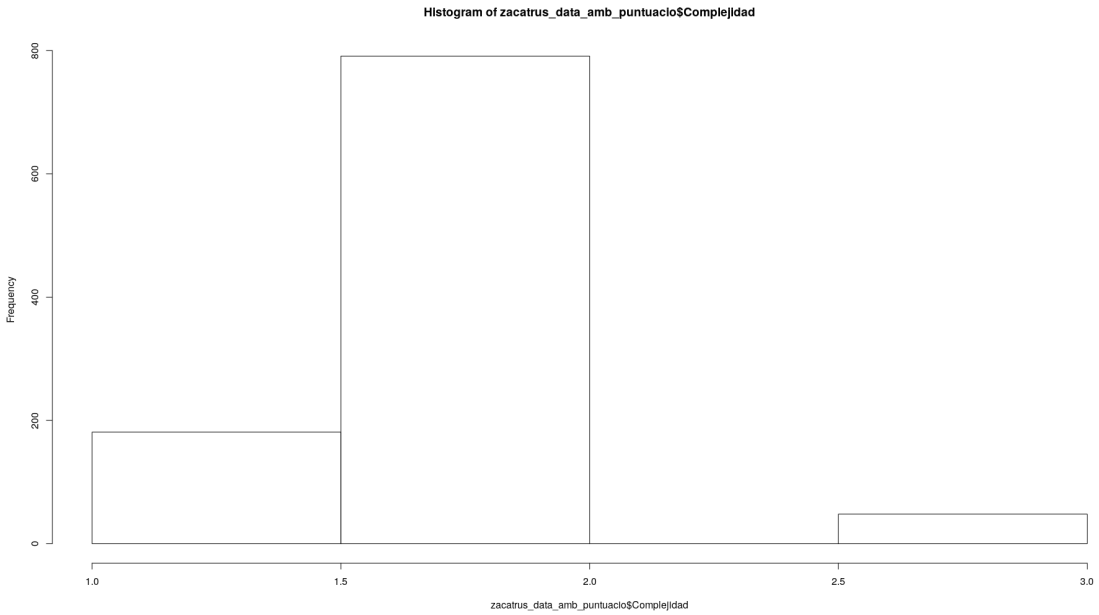
```
zacatrus_data$ComplejidadOrd[zacatrus_data$Complejidad == "Fácil"] <-
1
zacatrus_data$ComplejidadOrd[zacatrus_data$Complejidad == "Medio"] <-
2
zacatrus_data$ComplejidadOrd[zacatrus_data$Complejidad == "Difícil"]
<- 3
```

Per als valors buits, prenem la decisió d'assignar el valor mitjà, que és el que afecta menys a les anàlisis posteriors. Per tant, establim directament el valor 2:

```
zacatrus_data$ComplejidadOrd[is.na(zacatrus_data$ComplejidadOrd)] <-
2
```

Observem la distribució dels valors mitjançant un histograma:

```
hist(zacatrus_data_amb_puntuacio$Complejidad)
```



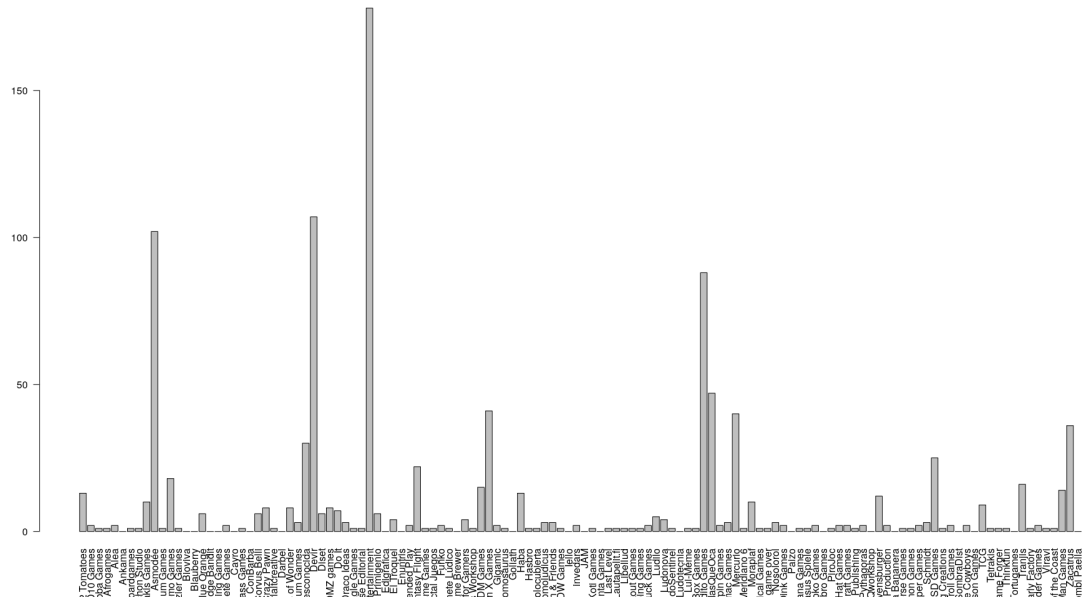
Editorial: És un camp categòric que pot proporcionar informació, i que es pot utilitzar a les anàlisis tal com l'hem obtingut, sense cap transformació. No pot tenir valors extrems, i els valors buits els substituïm per "Desconegut".

```
zacatrus_data$EditorialAux <- as.character(zacatrus_data$Editorial)
zacatrus_data$EditorialAux[zacatrus_data$Editorial == ""] <-
"Desconocida"
```

```
zacatrus_data$Editorial <- as.factor(zacatrus_data$EditorialAux )
```

Observem les dates en un histograma:

```
plot(zacatrus_data_amb_puntuacio$Editorial, las=2)
```

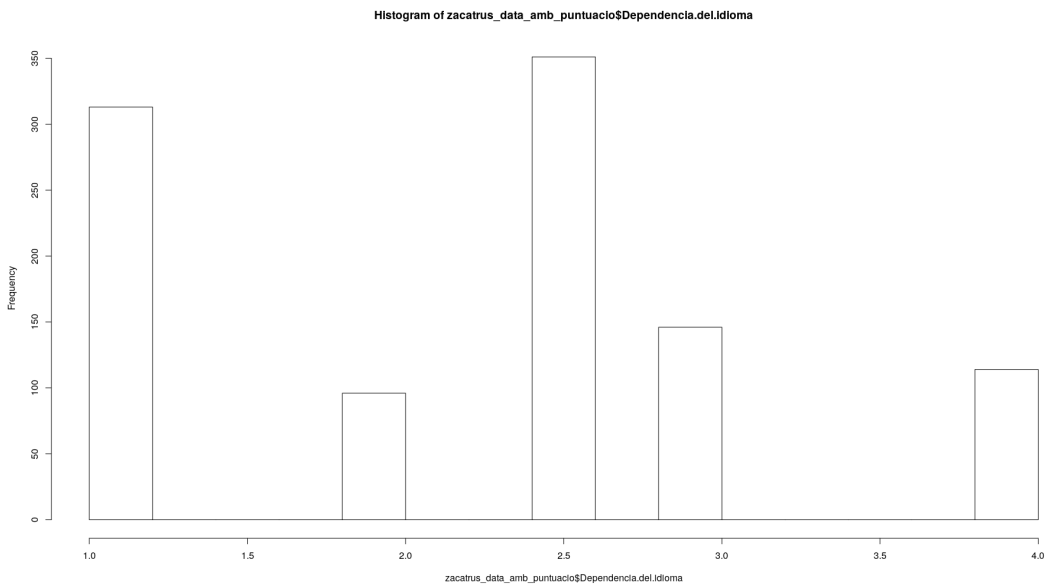


Dependencia del idioma: Apliquem la mateixa transformació que a la complexitat, i gestionem els valors buits de la mateixa manera, assignant el valor mitjà entre el valor més gran i el més petit:

```
zacatrus_data$DependenciaOrd[zacatrus_data$Dependencia.del.idioma ==
"Nula (Sólo instrucciones)"] <- 1
zacatrus_data$DependenciaOrd[zacatrus_data$Dependencia.del.idioma ==
"Poca"] <- 2
zacatrus_data$DependenciaOrd[zacatrus_data$Dependencia.del.idioma ==
"Media"] <- 3
zacatrus_data$DependenciaOrd[zacatrus_data$Dependencia.del.idioma ==
"Alta"] <- 4
zacatrus_data$DependenciaOrd[zacatrus_data$Dependencia.del.idioma ==
"Devir"] <- NA

zacatrus_data$DependenciaOrd[is.na(zacatrus_data$DependenciaOrd)] <-
2.5
```

Observem la distribució dels valors mitjançant un histograma:

```
hist(zacatrus_data_amb_puntuacio$Dependencia.del.idioma)
```



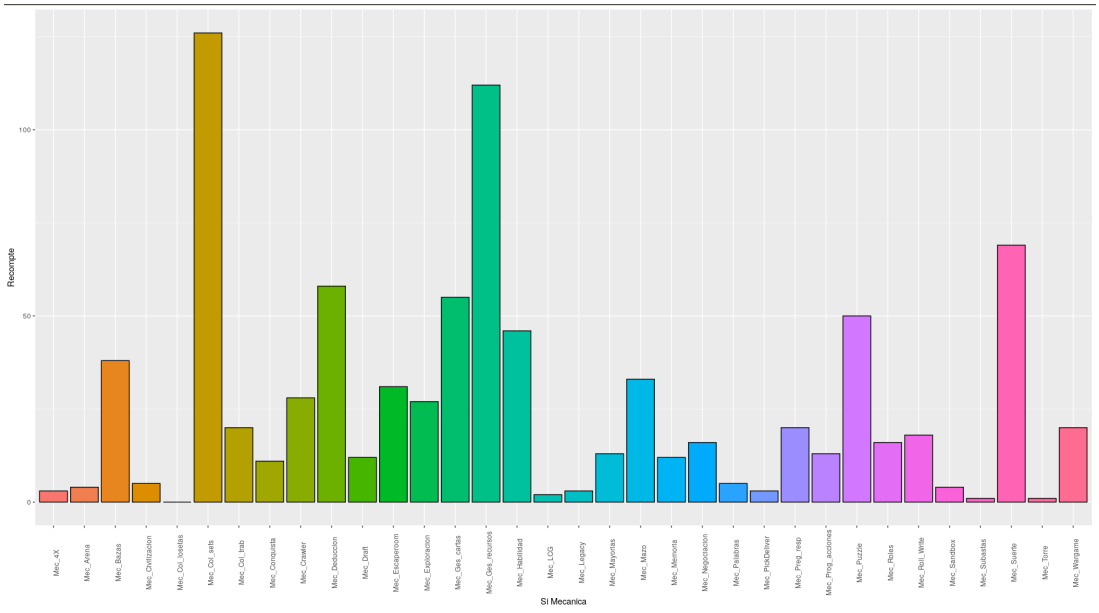
Mecànica: Aquest camp conté una llista amb les mecàniques que s'apliquen al joc. Les característiques d'aquest atribut són exactament les mateixes que les de la temàtica i "Si busques", i per tant apliquem la mateixa transformació:

```
zacatrus_data <- transform(zacatrus_data, Mec_4X =
grepl("4X",Mecánica))
zacatrus_data <- transform(zacatrus_data, Mec_Arena =
grepl("Arena",Mecánica))
...
zacatrus_data <- transform(zacatrus_data, Mec_Suerte =  grepl("Tienta
la suerte",Mecánica))
zacatrus_data <- transform(zacatrus_data, Mec_Wargame =
grepl("Wargame",Mecánica))

zacatrus_data <- select(zacatrus_data,-Mecánica)
```

A continuació visualitzem amb un gràfic de barres el nombre de jocs de cada grup "Si buscas...".

```
mecanica<-zacatrus_data_amb_puntuacio[,c("Mec_4x","Mec_Arena","Mec_Ba
zas","Mec_Col_sets","Mec_Col_losetas","Mec_Col_trab","Mec_Conquista",
"Mec_Crawler","Mec_Mazo","Mec_Deduccion","Mec_Torre","Mec_Draft","Mec
_Escaperoom","Mec_Civilizacion","Mec_Exploracion","Mec_Ges_cartas","M
ec_Ges_recursos","Mec_Habilidad","Mec_Palabras","Mec_LCG","Mec_Legacy
","Mec_Mayorias","Mec_Memoria","Mec_Negociacion","Mec_PickDeliver","M
```

```
ec_Preg_resp","Mec_Prog_acciones","Mec_Puzzle","Mec_Roles","Mec_Roll_
Write","Mec_Sandbox","Mec_Subastas","Mec_Suerte","Mec_Wargame")]

sumdata=data.frame(value=apply(mecanica,2,sum))
sumdata$key=rownames(sumdata)
ggplot(data=sumdata, aes(x=key, y=value,
fill=key))+geom_bar(colour="black",
stat="identity")+theme(legend.position="none",axis.text.x =
element_text(angle = 90))+ylab("Recompte")+xlab("Si Mecanica")
```



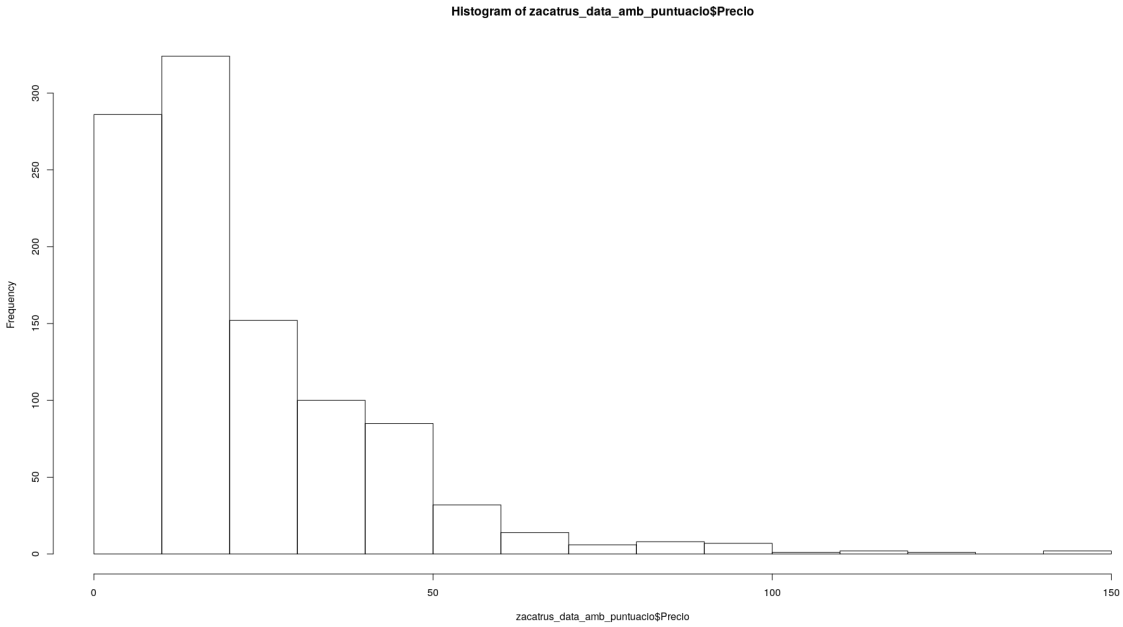
Preu(Precio): És una de les dades objectiu de la nostra anàlisi, i per tant òbviament forma part de les dades que s'utilitzen. L'única transformació necessària és transformar el valor a numèric.

```
zacatrus_data$Precio <- as.numeric(as.character(
zacatrus_data$Precio))
```

Comprovem que aquesta columna sí que té valors nuls i valors amb valor zero. Com que és una variable objectiu, no és convenient executar tasques d'estimació del valor, com per exemple assignar el valor de la mitjana o de la moda, perquè desvirtuaria molt el resultat del model. Per les anàlisis que no tinguin com a objectiu el preu, sí que el mantenim.

```
zacatrus_data_preus_amb_valor <- dplyr::filter(zacatrus_data, Precio
!= 0)

zacatrus_data_preus_amb_valor <- dplyr::filter(zacatrus_data_
preus_amb_valor, !is.na(Precio))
```

```
hist(zacatrus_data_amb_puntuacio$Precio)
```



Analitzem ara si apareixen valors extrems. Mirem primer els preus més baixos:

```{r}
zacatrus_data_preus_amb_valor[order(zacatrus_data_preus_amb_valor$Precio),]
```

	Nombre <fctr>	Precio <dbl>	Autor <fctr>
95	Wingspan: Guía de inicio rápido	1.35	Elizabeth Hargrave
850	Contadores Jet Black	1.35	
851	Contadores verde Jade	1.35	
901	Contadores Verde Esmeralda	1.35	
902	Contadores Marble Blue	1.35	
903	Contadores Bloodstone Red	1.35	
924	Contadores Blanco Perla	1.35	
942	Contadores Crystal Clear	1.35	
943	Contadores Sapphire Blue	1.35	
900	Contadores Ruby Red	1.40	

1-10 of 2,332 rows | 1-4 of 14 columns

Veiem que hi ha articles que amb tota seguretat són complements de jocs de taula, però no jocs de taula com a tals. Fent una inspecció visual, considerem que tots o gairebé tots els productes amb un preu menor de tres euros i mig estan mal categoritzats com a jocs de taula. Els eliminem:

```
zacatrus_data_preus_amb_valor <- dplyr::filter(zacatrus_data, Precio
>= 3.5)
```

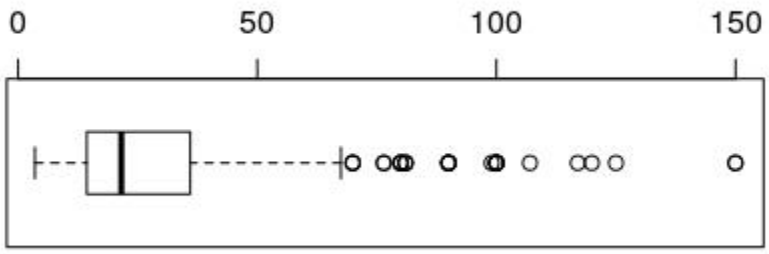
La variable “Precio” té una mitjana de 27.95.

```
summary(zacatrus_data_preus_amb_valor[zacatrus_data_preus_amb_valor[,
"ranking"] != "",]["Precio"]])
```

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.50   14.36   21.59   27.95   36.00  149.99
```

Creem un Boxplot per veure la distribució dels valors:

```
Precio.bp <-
boxplot(zacatrus_data_preus_amb_valor[zacatrus_data_preus_amb_valor[,
"ranking"] != "",][["Precio"]])
```




```
> Precio.bp$out
 [1] 107.00 149.95  99.99  79.99  79.99  79.99  90.00  76.45  89.95  99.99  90.00  80.99 117.00
[14]  99.95  99.95  81.00  69.95  81.00  99.99  79.95  99.99 149.99 124.99 119.95  69.95  89.99
[27]  89.95  99.00  69.95  76.45
```

A la base de dades trobem valors que superen en dues vegades la desviació estàndard, però una comprovació manual d'aquestes dades ens permet veure que efectivament són valors vàlids, i per tant no s'han d'eliminar.

Puntuació (Ranking): Com el preu, és un dels valors objectius, i per tant formarà part de les dades que s'utilitzen a les anàlisis. Fem el mateix procés que amb els preus, començant transformant el tipus de l'atribut a double:

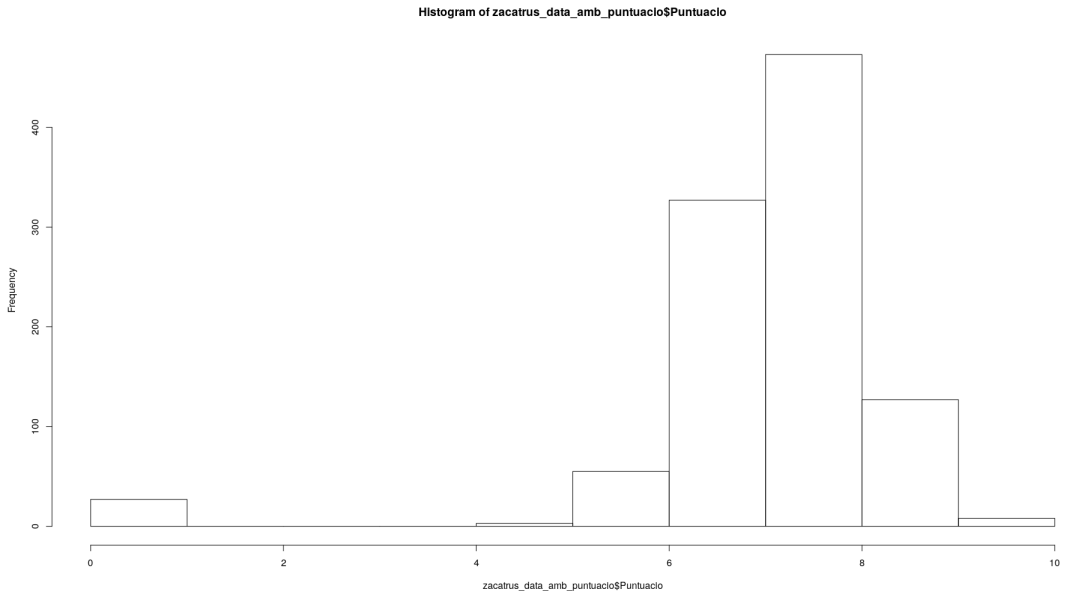
```
zacatrus_data$Puntuacio <- as.double(str_replace_all(as.character
(zacatrus_data$ranking), ",", "."))
```

Continuem eliminant tots els jocs que no tenen puntuació en un nou dataset:

```
zacatrus_data_amb_puntuacio <- dplyr::filter(zacatrus_data, Puntuacio
> 0)
zacatrus_data_amb_puntuacio <-
dplyr::filter(zacatrus_data_amb_puntuacio, !is.na(Puntuacio))
```

Si analitzem els valors extrems, veiem que no hi ha valors fora del rang 0-10 de les puntuacions, així que tots ells poden ser possibles, no hi ha outliers.

```
hist(zacatrus_data_amb_puntuacio$Puntuacio)
```



4. Anàlisi de les dades.

a. Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).

Tot i que en una anàlisi real utilitzaríem tots els atributs creats, per simplificar la realització d'aquesta pràctica excloem els atributs que s'han creat a partir d'altres atributs multivalor (Autor, Temàtica, "Si buscas…" i Mecànica).

```
zacatrus_data <- select(zacatrus_data, Nombre, Editorial,  EdadMinima, EdadMaxima, NumJugadoresMinimo, NumJugadoresMaximo,
TiempoDeJuegoMin, TiempoDeJuegoMax, Largo, Volumen, Complejidad, Dependencia.del.idioma, Puntuacio, Precio)

zacatrus_data_amb_preu <- select(zacatrus_data_amb_preu, Nombre, Editorial,  EdadMinima, EdadMaxima, NumJugadoresMinimo,
NumJugadoresMaximo, TiempoDeJuegoMin, TiempoDeJuegoMax, Largo, Volumen, Complejidad, Dependencia.del.idioma, Puntuacio,
Precio)

zacatrus_data_amb_puntuacio <- select(zacatrus_data_amb_puntuacio, Nombre, Editorial,  EdadMinima, EdadMaxima,
NumJugadoresMinimo, NumJugadoresMaximo, TiempoDeJuegoMin, TiempoDeJuegoMax, Largo, Volumen, Complejidad,
Dependencia.del.idioma, Puntuacio, Precio)

zacatrus_data_amb_preu_i_puntuacio <- select(zacatrus_data_amb_preu_i_puntuacio, Nombre, Editorial,  EdadMinima, EdadMaxima,
NumJugadoresMinimo, NumJugadoresMaximo, TiempoDeJuegoMin, TiempoDeJuegoMax, Largo, Volumen, Complejidad,
Dependencia.del.idioma, Puntuacio, Precio)
```

Fem un summary de les dades, que ens doni una pista de quins grups de dades podria ser interessant analitzar.

```
       Nombre                  Editorial      EdadMinima      EdadMaxima    NumJugadoresMinimo NumJugadoresMaximo
Star      : 172   Edge Entertainment: 651   Min.   : 0.00   Min.   : 3.00   Min.   :1.000      Min.   :1.000
Pathfinder:  56   Devir             : 255   1st Qu.:10.00   1st Qu.:18.00   1st Qu.:2.873      1st Qu.:4.000
Canción   :  36   Nosolorol         : 253   Median :10.87   Median :47.73   Median :2.873      Median :4.612
Arkham    :  34   Desconocida       : 230   Mean   :10.87   Mean   :47.74   Mean   :2.873      Mean   :4.617
Ninja     :  34   Fantasy Flight    : 208   3rd Qu.:14.00   3rd Qu.:99.00   3rd Qu.:3.000      3rd Qu.:5.000
Señor     :  28   Asmodée           : 177   Max.   :18.00   Max.   :99.00   Max.   :8.000      Max.   :8.000
(Other)   :2548   (Other)           :1134
TiempoDeJuegoMin TiempoDeJuegoMax     Largo          Volumen         Complejidad    Dependencia.del.idioma
Min.   :  1.00   Min.   :  2.00   Min.   : 29.0   Min.   :   5887   Min.   :1.000   Min.   :1.000
1st Qu.: 30.00   1st Qu.: 45.00   1st Qu.:187.7   1st Qu.:1455144   1st Qu.:2.000   1st Qu.:2.500
Median : 50.24   Median : 59.70   Median :187.7   Median :1455144   Median :2.000   Median :2.500
Mean   : 50.30   Mean   : 59.78   Mean   :187.7   Mean   :1455990   Mean   :1.952   Mean   :2.477
3rd Qu.: 50.24   3rd Qu.: 60.00   3rd Qu.:187.7   3rd Qu.:1455144   3rd Qu.:2.000   3rd Qu.:2.500
Max.   :240.00   Max.   :240.00   Max.   :286.0   Max.   :4092000   Max.   :3.000   Max.   :4.000

  Puntuacio         Precio
Min.   : 0.000   Min.   :  3.50
1st Qu.: 6.668   1st Qu.: 13.45
Median : 7.183   Median : 19.95
Mean   : 7.008   Mean   : 24.81
3rd Qu.: 7.688   3rd Qu.: 31.46
Max.   :10.000   Max.   :190.00
NA's   :1894     NA's   :626
```

Tenim dos atributs categòrics, i és possible que sigui interessant fer una comparació entre els jocs que pertanyen a una categoria o a una altra. Per exemple, podem comparar els jocs de les tres sagues més comunes, Star (Star Wars), Pathfinder i Canción (Canción de hielo y fuego). També podem comparar els preus i les valoracions de les quatre principals editorials (Edge, Devir, Nosolorol i Fantasy Flight).

```
zacatrus_data.star <- zacatrus_data[zacatrus_data$Nombre == "Star",]
zacatrus_data.pathfinder <- zacatrus_data[zacatrus_data$Nombre ==
"Pathfinder",]
zacatrus_data.cancion <- zacatrus_data[zacatrus_data$Nombre ==
"Canción",]

zacatrus_data.edge <- zacatrus_data[zacatrus_data$Editorial == "Edge
Entertainment",]
zacatrus_data.devir <- zacatrus_data[zacatrus_data$Editorial ==
"Devir",]
zacatrus_data.nosolorol <- zacatrus_data[zacatrus_data$Editorial ==
"Nosolorol",]
zacatrus_data.fantasyflight <- zacatrus_data[zacatrus_data$Editorial
== "Fantasy Flight",]
```

Aquestes agrupacions permetran més endavant analitzar quin efecte tenen les sagues i les editorials en el preu i en la puntuació dels jocs.

b. Comprovació de la normalitat i homogeneïtat de la variància.

Comprovarem per a tots els atributs numèrics, quins segueixen una distribució normal, i quins no. Per fer-ho, partim del codi que apareix a la pràctica de mostra, i el modifiquem per aplicar els tests de Shapiro-Wilk, el de Kolmogorov-Smirnov i el de Anderson-Darling.

```r
alpha = 0.05
col.names = colnames(zacatrus_data)
for (i in 1:ncol(zacatrus_data) - 2) {
  if (i == 1) cat("La variable segueix una distribució normal segons el test de Shapiro?\n")
  if (is.integer(zacatrus_data[,i]) | is.numeric(zacatrus_data[,i])) {
    cat(col.names[i])
    cat(": \n")
    cat("\tShapiro-Wilk ")
    p_val = shapiro.test(zacatrus_data[,i])$p.value
    cat(" p-value: ")
    cat(p_val)
    if (p_val < alpha) cat(" (NO)")
    else cat(" (SI)")
    cat("\n\tAnderson-Darling ")
    p_val = ad.test(zacatrus_data[,i])$p.value
    cat(" p-value: ")
    cat(p_val)
    if (p_val < alpha) cat(" (NO)")
    else cat(" (SI)")
    cat("\n\tKolmogorov-Smirnov ")
    p_val = ks.test(zacatrus_data[,i], pnorm, mean(zacatrus_data[,i]), sd(zacatrus_data[,i]))$p.value
    cat(" p-value: ")
    cat(p_val)
    if (p_val < alpha) cat(" (NO)\n")
    else cat(" (SI)\n")
  }
}
```

Per als atributs preu i puntuació, l'adaptem per utilitzar els datasets específics a on no hi ha valors buits. Obtenim els següents resultats:

```
La variable segueix una distribució normal segons el test de Shapiro?
EdadMinima:
        Shapiro-Wilk  p-value: 7.912501e-42 (NO)
        Anderson-Darling  p-value: 3.7e-24 (NO)
        Kolmogorov-Smirnov ties should not be present for the Kolmogorov-Smirnov test p-value: 0 (NO)
EdadMaxima:
        Shapiro-Wilk  p-value: 2.853711e-55 (NO)
        Anderson-Darling  p-value: 3.7e-24 (NO)
        Kolmogorov-Smirnov ties should not be present for the Kolmogorov-Smirnov test p-value: 0 (NO)
NumJugadoresMinimo:
        Shapiro-Wilk  p-value: 5.508778e-60 (NO)
        Anderson-Darling  p-value: 3.7e-24 (NO)
        Kolmogorov-Smirnov ties should not be present for the Kolmogorov-Smirnov test p-value: 0 (NO)
NumJugadoresMaximo:
        Shapiro-Wilk  p-value: 5.360623e-42 (NO)
        Anderson-Darling  p-value: 3.7e-24 (NO)
        Kolmogorov-Smirnov ties should not be present for the Kolmogorov-Smirnov test p-value: 0 (NO)
TiempoDeJuegoMin:
        Shapiro-Wilk  p-value: 4.524977e-51 (NO)
        Anderson-Darling  p-value: 3.7e-24 (NO)
        Kolmogorov-Smirnov ties should not be present for the Kolmogorov-Smirnov test p-value: 0 (NO)
TiempoDeJuegoMax:
        Shapiro-Wilk  p-value: 1.405226e-49 (NO)
        Anderson-Darling  p-value: 3.7e-24 (NO)
        Kolmogorov-Smirnov ties should not be present for the Kolmogorov-Smirnov test p-value: 0 (NO)
Largo:
        Shapiro-Wilk  p-value: 5.524663e-62 (NO)
        Anderson-Darling  p-value: 3.7e-24 (NO)
        Kolmogorov-Smirnov ties should not be present for the Kolmogorov-Smirnov test p-value: 0 (NO)
Volumen:
        Shapiro-Wilk  p-value: 1.191037e-62 (NO)
        Anderson-Darling  p-value: 3.7e-24 (NO)
        Kolmogorov-Smirnov ties should not be present for the Kolmogorov-Smirnov test p-value: 0 (NO)
Complejidad:
        Shapiro-Wilk  p-value: 1.118596e-68 (NO)
        Anderson-Darling  p-value: 3.7e-24 (NO)
        Kolmogorov-Smirnov ties should not be present for the Kolmogorov-Smirnov test p-value: 0 (NO)
Dependencia.del.idioma:
        Shapiro-Wilk  p-value: 1.424927e-53 (NO)
        Anderson-Darling  p-value: 3.7e-24 (NO)
        Kolmogorov-Smirnov ties should not be present for the Kolmogorov-Smirnov test p-value: 0 (NO)
Precio:
        Shapiro-Wilk  p-value: 1.211606e-46 (NO)
        Anderson-Darling  p-value: 3.7e-24 (NO)
        Kolmogorov-Smirnov ties should not be present for the Kolmogorov-Smirnov test p-value: 0 (NO)
Puntuacio:
        Shapiro-Wilk  p-value: 0.000306014 (NO)
        Anderson-Darling  p-value: 0.09836671 (SI)
        Kolmogorov-Smirnov ties should not be present for the Kolmogorov-Smirnov test p-value: 0.8734734 (SI)
```

Assumint com a hipòtesi nul·la que la població està distribuïda normalment, si el p-valor és més petit que el nivell de significació (α=0,05), llavors la hipòtesi nul·la és rebutjada i es conclou que les dades no compten amb una distribució normal. Si, per contra, el p-valor és major a α, es conclou que no es pot rebutjar aquesta hipòtesi i s'assumeix que les dades segueixen una

distribució normal. En aquest cas, el p-value de tots els atributs numèrics tret de la puntuació és pràcticament 0 o 0 en tots els tests, i podem dir que cap d'ells segueix una distribució normal. Pel que fa al cas concret de la puntuació, els dos tests menys restrictius sí que detecten una normalitat en les seves dades.

Pel que fa a l'homogeneïtat de les variàncies, ho comprovarem per als grups de dades definits a l'apartat anterior. Farem quatre comprovacions, dues per l'atribut preu, i dues per l'atribut puntuació. Per cadascun dels atributs, es farà fent una divisió per les 3 sagues més comunes, i per les quatre editorials amb més jocs. Obtenim els següents resultats:

```
zacatrus_data_per_saga_preu <- dplyr::filter(zacatrus_data_amb_preu, Nombre == "Star" | Nombre ==
"Pathfinder" | Nombre == "Canción")

fligner.test(Precio ~ Nombre, data = zacatrus_data_per_saga_preu)

zacatrus_data_per_saga_puntuacio <- dplyr::filter(zacatrus_data_amb_puntuacio, Nombre == "Star" | Nombre ==
"Pathfinder" | Nombre == "Canción")

fligner.test(Puntuacio ~ Editorial, data = zacatrus_data_per_saga_puntuacio)

zacatrus_data_per_editorial_preu <- dplyr::filter(zacatrus_data_amb_preu, Editorial == "Edge Entertainment"
| Editorial == "Devir" | Editorial == "Nosolorol" | Editorial == "Fantasy Flight")

fligner.test(Precio ~ Nombre, data = zacatrus_data_per_editorial_preu)

zacatrus_data_per_editorial_puntuacio <- dplyr::filter(zacatrus_data_amb_puntuacio, Editorial == "Edge
Entertainment" | Editorial == "Devir" | Editorial == "Nosolorol" | Editorial == "Fantasy Flight")

fligner.test(Puntuacio ~ Editorial, data = zacatrus_data_per_editorial_puntuacio)
```

```
        Fligner-Killeen test of homogeneity of variances

data:  Precio by Nombre
Fligner-Killeen:med chi-squared = 16.416, df = 2, p-value = 0.0002725


        Fligner-Killeen test of homogeneity of variances

data:  Puntuacio by Editorial
Fligner-Killeen:med chi-squared = 7.0601, df = 4, p-value = 0.1327


        Fligner-Killeen test of homogeneity of variances

data:  Precio by Nombre
Fligner-Killeen:med chi-squared = 417.3, df = 351, p-value = 0.008539


        Fligner-Killeen test of homogeneity of variances

data:  Puntuacio by Editorial
Fligner-Killeen:med chi-squared = 11.846, df = 3, p-value = 0.007931
```

Els dos primers resultats corresponen a l'anàlisi de variàncies amb les dades agrupades per nom de la saga. S'estudia en primer lloc l'atribut Precio i en segon lloc la Puntuació. A continuació estudiem l'homogeneïtat de les variàncies amb les dades agrupades per editorial. S'estudia en primer lloc l'atribut Preu i en segon lloc l'atribut Puntuació.

Veiem que només per l'atribut puntuació, quan separem les dades per editorial, el p-value és més gran que 0.05 i per tant podem considerar les variàncies són homogènies. En la resta de casos, les variàncies als diferents grups es poden considerar estadísticament diferents.

c. Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents.

Anàlisi 1. Quins atributs quantitatius afecten més al preu i a la puntuació? Són els mateixos?

Per poder saber quin dels atributs quantitatius es troba més fortament relacionat amb el preu dels productes, i per tant ens dóna una informació més important per poder realitzar una predicció sobre el preu dels jocs. Com que cap dels atributs segueix una distribució normal, descartem el coeficient de correlació de Pearson, i apliquem directament el coeficient de Spearman. Per fer el càlcul, ens basem en el codi de la pràctica d'exemple:

```
corr_matrix <- matrix(nc = 2, nr = 0)
colnames(corr_matrix) <- c("estimate", "p-value")
# Calcular el coeficiente de correlación para cada variable cuantitativa
# con respecto al campo "precio"
for (i in 1:(ncol(zacatrus_data_amb_preu) - 1)) {
  if (is.integer(zacatrus_data_amb_preu[,i]) | is.numeric(zacatrus_data_amb_preu[,i])) {
    spearman_test = cor.test(zacatrus_data_amb_preu[,i], zacatrus_data_amb_preu[,length(zacatrus_data_amb_preu)], method = "spearman")
    corr_coef = spearman_test$estimate
    p_val = spearman_test$p.value
    # Add row to matrix
    pair = matrix(ncol = 2, nrow = 1)
    pair[1][1] = corr_coef
    pair[2][1] = p_val
    corr_matrix <- rbind(corr_matrix, pair)
    rownames(corr_matrix)[nrow(corr_matrix)] <- colnames(zacatrus_data_amb_preu)[i]
  }
}
```

	estimate	p-value
EdadMinima	0.25596357	1.833544e-35
EdadMaxima	0.09870673	2.310170e-06
NumJugadoresMinimo	-0.04783177	2.231323e-02
NumJugadoresMaximo	-0.09188263	1.101847e-05
TiempoDeJuegoMin	0.26065748	9.312390e-37
TiempoDeJuegoMax	0.30752681	3.518056e-51
Largo	0.25792419	5.320217e-36
Volumen	0.25734281	7.686991e-36
Complejidad	0.15767865	3.573002e-14
Dependencia.del.idioma	0.06663312	1.448147e-03
Puntuacio	0.22786557	2.908726e-28

Veiem com les variables que presenta una correlació més gran amb el preu són les relacionades amb el temps de joc, seguides per la grandària de la capsa. Tot i ser les que presenten una correlació més gran, aquesta no és gaire significativa.

Fem la mateixa anàlisi per la puntuació:

	estimate	p-value
EdadMinima	0.40693327	1.179238e-40
EdadMaxima	0.09438221	2.997109e-03
NumJugadoresMinimo	-0.22013979	2.686232e-12
NumJugadoresMaximo	-0.21710092	5.412682e-12
TiempoDeJuegoMin	0.46048671	5.839900e-53
TiempoDeJuegoMax	0.50111348	7.135198e-64
Largo	0.08650251	6.542707e-03
Volumen	0.16190563	3.156626e-07
Complejidad	0.33678431	1.332237e-27
Dependencia.del.idioma	0.20625966	6.060898e-11
Precio	0.42343933	4.142959e-35

En aquest cas, també el temps de joc són les variables que tenen una influència més gran sobre la puntuació del joc. El següent factor a destacar és el preu.

Anàlisi 2. Com afecta la saga del joc al seu preu? Com afecta l'editorial a la puntuació dels jocs?

Per a la realització d'aquestes anàlisis farem proves de contrast d'hipòtesi. Per a la variable objectiu preu teníem que no segueix una distribució normal, ni tampoc no hi havia homogeneïtat en la variància en la partició per sagues. D'altra banda, com per les tres sagues que volem analitzar tenim un nombre de mostres més gran de 30, podem utilitzar igualment el contrast d'anàlisi. Compararem les tres sagues dos a dos, utilitzant un valor de significació de 0.05 per determinar si estadísticament es pot dir que una saga té un preu més gran que una altra o no.

```
zacatrus_data_amb_preu.star.precios <- zacatrus_data_amb_preu.star$Precio
zacatrus_data_amb_preu.pathfinder.precios <- zacatrus_data_amb_preu.pathfinder$Precio
zacatrus_data_amb_preu.cancion.precios <- zacatrus_data_amb_preu.cancion$Precio


t.test(zacatrus_data_amb_preu.pathfinder.precios, zacatrus_data_amb_preu.star.precios, alternative = "less")
t.test(zacatrus_data_amb_preu.pathfinder.precios, zacatrus_data_amb_preu.cancion.precios, alternative = "less")
t.test(zacatrus_data_amb_preu.star.precios, zacatrus_data_amb_preu.cancion.precios, alternative = "less")
```

```
[1] 56

        Welch Two Sample t-test

data:  zacatrus_data_amb_preu.pathfinder.precios and zacatrus_data_amb_preu.star.precios
t = -2.4604, df = 158.61, p-value = 0.007475
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -1.877648
sample estimates:
mean of x mean of y
 20.47893  26.21148


        Welch Two Sample t-test

data:  zacatrus_data_amb_preu.pathfinder.precios and zacatrus_data_amb_preu.cancion.precios
t = -1.8302, df = 46.841, p-value = 0.03679
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.6701484
sample estimates:
mean of x mean of y
 20.47893  28.53806


        Welch Two Sample t-test

data:  zacatrus_data_amb_preu.star.precios and zacatrus_data_amb_preu.cancion.precios
t = -0.53005, df = 46.798, p-value = 0.2993
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 5.039009
sample estimates:
mean of x mean of y
 26.21148  28.53806
```

Quan comparem els preus de la saga "Pathfinder" amb els de la saga "Star Wars", el valor de significació és menor del límit que ens hem marcat. Per tant, es pot determinar que els jocs de "Pathfinder" són més barats que el d'"Star Wars". Pel que fa a la comparació entre "Pathfinder" i "Canción de hielo y fuego", també es pot determinar que els primers són, de mitjana, més barats, tot i que amb un nivell de confiança menor. Finalment, pel que fa a la comparació entre "Star Wars" i "Canción de hielo y fuego", no es poden treure conclusions: el p-value és més gran que el límit del valor de significació marcat.

A continuació, analitzarem si l'editorial a la qual pertany un joc té un efecte en la puntuació que rep. En aquest cas, el nombre de mostres amb puntuació per editorial no sempre arriben a les 30 mostres. No obstant això, sí que es va poder determinar la normalitat de les dades de puntuació amb els dos tests menys restrictius, i també es va determinar l'homogeneïtat de la variància de la puntuació per editorials, i per tant també podem aplicar el contrast d'hipòtesis. Com en l'anàlisi anterior, fem el contrast dos a dos i obtenim els següents resultats:

```
zacatrus_data_amb_puntuacio.edge.puntuacions <- zacatrus_data_amb_puntuacio.edge$Puntuacio
zacatrus_data_amb_puntuacio.nosolorol.puntuacions <- zacatrus_data_amb_puntuacio.nosolorol$Puntuacio
zacatrus_data_amb_puntuacio.fantasyflight.puntuacions <- zacatrus_data_amb_puntuacio.fantasyflight$Puntuacio
zacatrus_data_amb_puntuacio.devir.puntuacions <- zacatrus_data_amb_puntuacio.devir$Puntuacio

t.test(zacatrus_data_amb_puntuacio.nosolorol.puntuacions, zacatrus_data_amb_puntuacio.devir.puntuacions, alternative = "less")
t.test(zacatrus_data_amb_puntuacio.nosolorol.puntuacions, zacatrus_data_amb_puntuacio.edge.puntuacions, alternative = "less")
t.test(zacatrus_data_amb_puntuacio.devir.puntuacions, zacatrus_data_amb_puntuacio.edge.puntuacions, alternative = "less")
t.test(zacatrus_data_amb_puntuacio.nosolorol.puntuacions, zacatrus_data_amb_puntuacio.fantasyflight.puntuacions, alternative = "less")
t.test(zacatrus_data_amb_puntuacio.devir.puntuacions, zacatrus_data_amb_puntuacio.fantasyflight.puntuacions, alternative = "less")
t.test(zacatrus_data_amb_puntuacio.edge.puntuacions, zacatrus_data_amb_puntuacio.fantasyflight.puntuacions, alternative = "less")
```

```
        Welch Two Sample t-test

data:  zacatrus_data_amb_puntuacio.nosolorol.puntuacions and zacatrus_data_amb_puntuacio.devir.puntuacions
t = -13.104, df = 16.076, p-value = 2.677e-10
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -0.926921
sample estimates:
mean of x mean of y
 6.195623  7.264975


        Welch Two Sample t-test

data:  zacatrus_data_amb_puntuacio.nosolorol.puntuacions and zacatrus_data_amb_puntuacio.edge.puntuacions
t = -14.841, df = 12.939, p-value = 8.373e-10
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.9998293
sample estimates:
mean of x mean of y
 6.195623  7.330983


        Welch Two Sample t-test

data:  zacatrus_data_amb_puntuacio.devir.puntuacions and zacatrus_data_amb_puntuacio.edge.puntuacions
t = -0.73884, df = 231.61, p-value = 0.2304
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf 0.0815334
sample estimates:
mean of x mean of y
 7.264975  7.330983
```

```
        Welch Two Sample t-test

data:  zacatrus_data_amb_puntuacio.nosolorol.puntuacions and zacatrus_data_amb_puntuacio.fantasyflight.puntuacions
t = -12.955, df = 22.693, p-value = 2.874e-12
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf -1.54589
sample estimates:
mean of x mean of y
 6.195623  7.977358


        Welch Two Sample t-test

data:  zacatrus_data_amb_puntuacio.devir.puntuacions and zacatrus_data_amb_puntuacio.fantasyflight.puntuacions
t = -4.9108, df = 33.064, p-value = 1.194e-05
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -0.466895
sample estimates:
mean of x mean of y
 7.264975  7.977358


        Welch Two Sample t-test

data:  zacatrus_data_amb_puntuacio.edge.puntuacions and zacatrus_data_amb_puntuacio.fantasyflight.puntuacions
t = -4.5437, df = 30.833, p-value = 3.994e-05
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.4051339
sample estimates:
mean of x mean of y
 7.330983  7.977358
```

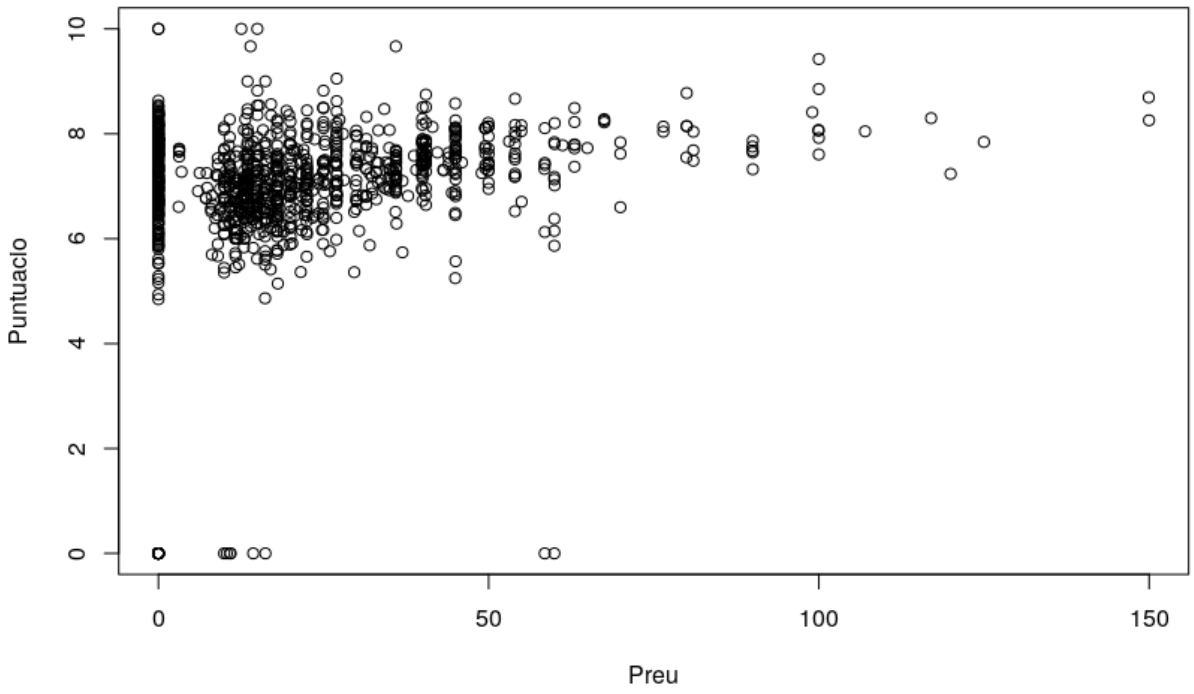
Amb aquestes dades, podem extreure les següents conclusions:

- Es pot dir que els jocs de Fantasy Flight són els més valorats quan els comparem amb qualsevol altra de les editorials.
- Els jocs de Nosolorol són els menys valorats
- Entre els jocs de Devir i d'Edge no es pot determinar estadísticament quins estan millor valorats.

Anàlisi 3. Model de regressió lineal. Predicció del preu.

En primer lloc inspeccionem visualment la relació entre preu i puntuació. Aparentment no hi ha una relació forta entre elles. Sí que els valors més alts de preu es troben en els jocs més ben valorats, però com hem vist en l'anàlisi de la variable "Precio", aquests es poden vendre a preus més alts, però per l'interval de preus fins a 70 euros no s'aprecia cap diferencia en el preu respecte de la puntuació del joc.

```
plot(zacatrus_data_amb_puntuacio$Precio,zacatrus_data_amb_puntuacio$P
untuacio,xlab="Preu",ylab="Puntuacio")
```



Per demostrar estadísticament aquesta no-correlació que veiem en el gràfic utilitzarem el coeficient de correlació de Spearman, ja que no podem dir que les variables segueixin una distribució normal i per tant no podem utilitzar el coeficient de Pearson.

```
cor.test(zacatrus_data_amb_puntuacio$Precio,zacatrus_data_amb_puntuacio$Puntuacio, method = "spearman")
```

```
        Spearman's rank correlation rho

data:  zacatrus_data_amb_puntuacio$Precio and zacatrus_data_amb_puntuacio$Puntuacio
S = 135870000, p-value = 6.549e-14
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.2318076

Warning message:
In cor.test.default(zacatrus_data_amb_puntuacio$Precio, zacatrus_data_amb_puntuacio$Puntuacio,  :
  Cannot compute exact p-value with ties
```

El coeficient de correlació (rho) pren valors entre -1 i +1, essent 1 una correlació forta (negativa o positiva, respectivament) i 0 una no-correlació entre les variables.
Així doncs podem dir que la correlació entre preu i puntuació és lleugerament positiva, de manera que els preus amb millor puntuació tendeixen a tenir preus més cars.
Acompanyant-nos amb el resultat del gràfic, veiem que els jocs més barats tenen un rang de puntuacions gran, mentre que en els jocs més cars, en general tenen bones puntuacions.

Intentem predir el Preu del joc en funció de la Puntuació mitjançant una anàlisi de regressió lineal i aconseguim una bondat de l'ajust d'un 0.05038 (essent 1 un ajust perfecte). Clarament l'ajust aconseguit és insuficient per explicar la variable 'preu':

```
regr_punt<-lm(Precio~Puntuacio, data=zacatrus_data_amb_puntuacio)
```

```
> summary(regr_punt)

Call:
lm(formula = Precio ~ Puntuacio, data = zacatrus_data_amb_puntuacio)

Residuals:
    Min      1Q  Median      3Q     Max
-31.734 -11.984  -3.520   9.439 124.222

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.4866     3.2944  -0.755    0.451
Puntuacio     3.4220     0.4612   7.420 2.46e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.21 on 1018 degrees of freedom
Multiple R-squared:  0.05131,   Adjusted R-squared:  0.05038
F-statistic: 55.06 on 1 and 1018 DF,  p-value: 2.458e-13
```

Afegim la variable Editorial i la bondat de l'ajust puja fins a 0.1245:

```
regr_punt_edit<-lm(Precio~Puntuacio+Editorial,
data=zacatrus_data_amb_puntuacio)
```

```
EditorialWizards of the Coast  24.48326   20.13672   1.216  0.22435
EditorialZ-Man Games           20.62686    7.47858   2.758  0.00593 **
EditorialZacatrus              -8.37487    6.28262  -1.333  0.18286
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.4 on 917 degrees of freedom
Multiple R-squared:  0.2122,    Adjusted R-squared:  0.1245
F-statistic: 2.421 on 102 and 917 DF,  p-value: 6.578e-12
```

Afegint només les variables quantitatives de la base de dades obtenim un 0.1808:

```
regr_qtt<-lm(Precio~Puntuacio+EdadMinima+EdadMaxima+NumJugadoresMinim
o+NumJugadoresMaximo+Largo+Volumen+Complejidad+Dependencia.del.idioma
, data=zacatrus_data_amb_puntuacio)
summary(regr_qtt)
```

```
Call:
lm(formula = Precio ~ Puntuacio + EdadMinima + EdadMaxima + NumJugadoresMinimo +
    NumJugadoresMaximo + Largo + Volumen + Complejidad + Dependencia.del.idioma,
    data = zacatrus_data_amb_puntuacio)

Residuals:
    Min      1Q  Median      3Q     Max
-33.369 -12.134  -3.837   8.323 127.693

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           -2.555e+01  6.389e+00  -4.000 6.81e-05 ***
Puntuacio              2.300e+00  4.381e-01   5.251 1.85e-07 ***
EdadMinima             1.247e+00  2.343e-01   5.323 1.26e-07 ***
EdadMaxima             1.490e-01  1.510e-02   9.870  < 2e-16 ***
NumJugadoresMinimo    -1.999e+00  8.704e-01  -2.297   0.0218 *
NumJugadoresMaximo     2.972e-01  3.807e-01   0.781   0.4352
Largo                  2.547e-02  2.577e-02   0.988   0.3232
Volumen                1.562e-06  1.187e-06   1.315   0.1887
Complejidad            3.411e+00  1.439e+00   2.370   0.0180 *
Dependencia.del.idioma 8.104e-02  6.755e-01   0.120   0.9045
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.77 on 1010 degrees of freedom
Multiple R-squared:  0.188,     Adjusted R-squared:  0.1808
F-statistic: 25.99 on 9 and 1010 DF,  p-value: < 2.2e-16
```

Afegim la resta de variables excepte el nom del joc i l'autor i arribem a un ajust de 0.3496.

```
regr_tot_ex_nom_aut<-lm(Precio~Puntuacio+Editorial+Tem_Abstracto+Tem_
Comercio+Tem_Egipto+Tem_Medicina+Tem_Oriental+Tem_Trenes+Tem_Agricult
ura+Tem_Cómic+Tem_Electrónica+Tem_Medieval+Tem_Piratas+Tem_Audiovisua
```

```
l+Tem_Animales+Tem_Cuentos+Tem_EspadaBrujería+Tem_Misterio+Tem_Politi
ca+Tem_Urbano+Tem_ArteLiteratura+Tem_Cyberpunk+Tem_Fantasía+Tem_Mitol
ogía+Tem_Postapocalíptico+Tem_Vampiros+Tem_Carreras+Tem_Deportes+Tem_
Gastronómica+Tem_Música+Tem_Steampunk+Tem_Videojuegos+Tem_Ciencia+Tem
_Detectivesca+Tem_Historia+Tem_Naturaleza+Tem_Superhéroes+Tem_Vikingo
s+Tem_CienciaFicción+Tem_Dinosaurios+Tem_Maya+Tem_Oeste+Tem_Terror+Te
m_Zombies+SiB_Ameritrash+SiB_Cooperativo+SiB_Eurogame+SiB_Experiencia
+SiB_Familiares+SiB_Fiesta+SiB_Infantil+SiB_Narrativo+SiB_ParaDos+SiB
_Rápido+SiB_Solitario+SiB_Viaje+EdadMinima+EdadMaxima+NumJugadoresMin
imo+NumJugadoresMaximo+TiempoDeJuegoMin+TiempoDeJuegoMax+Largo+Volume
n+Complejidad+EditorialAux+Dependencia.del.idioma+Mec_4X+Mec_Arena+Me
c_Bazas+Mec_Col_sets+Mec_Col_losetas+Mec_Col_trab+Mec_Conquista+Mec_C
rawler+Mec_Mazo+Mec_Deduccion+Mec_Torre+Mec_Draft+Mec_Escaperoom+Mec_
Civilizacion+Mec_Exploracion+Mec_Ges_cartas+Mec_Ges_recursos+Mec_Habi
lidad+Mec_Palabras+Mec_LCG+Mec_Legacy+Mec_Mayorias+Mec_Memoria+Mec_Ne
gociacion+Mec_PickDeliver+Mec_Preg_resp+Mec_Prog_acciones+Mec_Puzzle+
Mec_Roles+Mec_Roll_Write+Mec_Sandbox+Mec_Subastas+Mec_Suerte+Mec_Warg
ame, data=zacatrus_data_amb_puntuacio)
```

```
EditorialAuxFresMe Games                NA         NA      NA       NA
EditorialAuxFractal Juegos              NA         NA      NA       NA
 [ reached getOption("max.print") -- omitted 102 rows ]
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.72 on 822 degrees of freedom
Multiple R-squared:  0.4754,    Adjusted R-squared:  0.3496
F-statistic: 3.781 on 197 and 822 DF,  p-value: < 2.2e-16
```

Aquestes anàlisis de regressió ens donen una idea de l'ajust que podem obtenir, però per tal que els resultats siguin fiables no podem deduir el model de la totalitat de la mostra, sinó que ens cal fer un sampling per separar les dades d'entrenament i de test.

Utilitzant el mètode de sampling de les dades de test i training leave-one-out:

```
> library(caret)
> ctrl <- trainControl(method="LOOCV")
> model <-
train(Precio~Puntuacio,data=zacatrus_data_amb_puntuacio,method="lm",t
rControl=ctrl)
> print(model)
```

```
> ctrl <- trainControl(method="LOOCV")
> model <- train(Precio~Puntuacio,data=zacatrus_data_amb_puntuacio,method="lm",trControl=ctrl)
> print(model)
Linear Regression

1020 samples
   1 predictor

No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 1019, 1019, 1019, 1019, 1019, 1019, ...
Resampling results:

  RMSE      Rsquared    MAE
  20.23082  0.04751768  14.6668

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Amb aquest model de regressió, el coeficient de determinació (R^2) és de 0.0475 (on 1 seria el millor ajust). Deduïm doncs que només amb la puntuació no n'hi ha prou per explicar el preu d'un joc.

Per intentar millorar l'ajust, incorporem les variables quantitatives del dataframe:

```
model <-
train(Precio~Puntuacio+EdadMinima+EdadMaxima+NumJugadoresMinimo+NumJugadoresMaximo+Largo+Volumen+Complejidad+Dependencia.del.idioma,data=zacatrus_data_amb_preu,method="lm",trControl=ctrl)
> print(model)
```

```
> model <- train(Precio~Puntuacio+EdadMinima+EdadMaxima+NumJugadoresMinimo+NumJugadoresMaximo+Largo+Volumen+Complejidad+Dependencia.del.idioma,data=zacatrus_data_amb_puntuacio,method="lm",trControl=ctrl)
> print(model)
Linear Regression

1020 samples
   9 predictor

No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 1019, 1019, 1019, 1019, 1019, 1019, ...
Resampling results:

  RMSE      Rsquared   MAE
  18.83606  0.1745038  13.7504

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Afegint les variables següents una a una la bondat de l'ajust va millorant fins a aconseguir un 0.17450:

Variables	R2
Puntuacio	0.04751
+EdadMinima	0.08787
+EdadMaxima	0.16291
+NumJugadoresMinimo	0.16599
+NumJugadoresMaximo	0.16474
+Largo	0.17138
+Volumen	0.17253
+Complejidad	0.17562

En afegir les variables Nom, Autor i Editorial ens trobem que són de tipus text i la funció les converteix a binàries. Afegir-les, així com les variables Tema, Sibuscas i Mecanica que ja són binàries, afegeixen un gran nombre de variables a l'anàlisi de regressió, i això fa que, junt amb el fet que la base de dades no és gaire gran, els resultats de l'anàlisi no siguin fiables.
Si comparem els resultats de la regressió amb totes les dades i fent el pas de sampling veiem que la bondat de l'ajust és lleugerament inferior però similar a l'obtinguda sense.

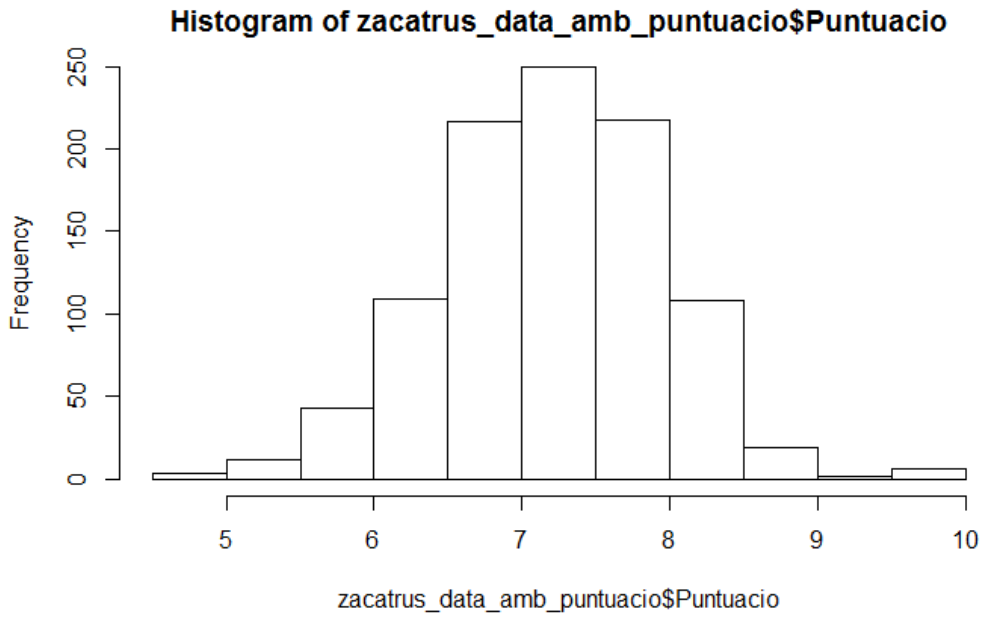
R2 ajustat

Variables independents	**Només regressió lineal**	**Sampling (leave-one-out)**
Puntuacio	0.0504	0.0475
+Quantitatives	0.1808	0.1745
+Resta	0.3496	

Per concloure l'anàlisi de regressió, podríem dir que les variables de les quals disposem no són suficients per explicar d'una manera satisfactòria el Preu.

5. Representació dels resultats a partir de taules i gràfiques.

Primer mostrem gràficament la normalitat de l'atribut puntuació, i comparem les variàncies de les puntuacions quan les agrupem a les principals editorials.

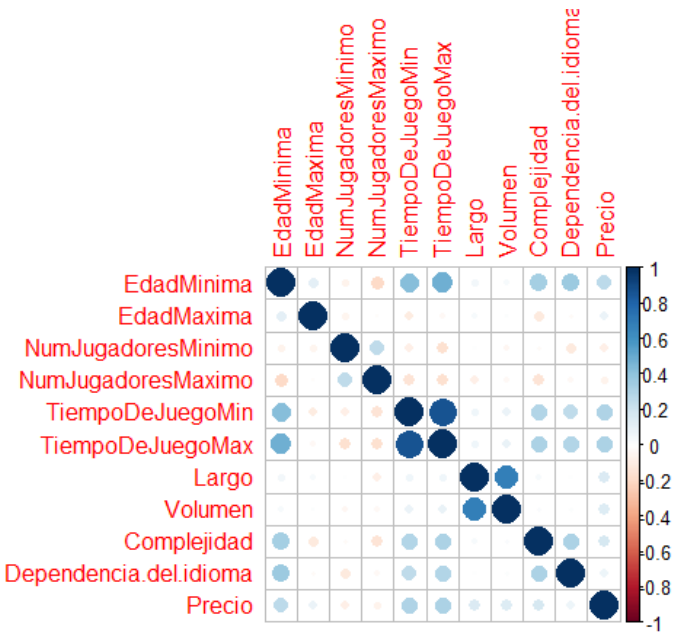


	Devir	**Edge**	**FantasyFlight**	**Nosolorol**
Std Dev	0.6661887	0.7528862	0.6052209	0.08243407

Es pot comprovar que les desviacions són similars tret de la de Nosolorol, aquesta a causa que només disposa de 3 jocs amb puntuació.

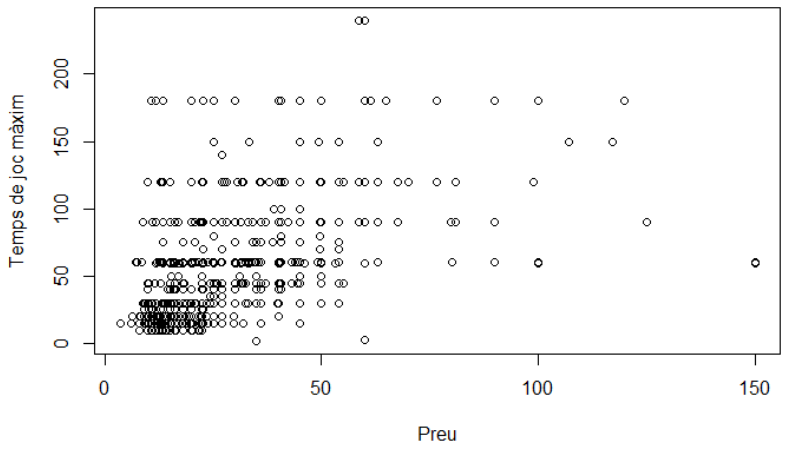
També es pot veure gràficament les correlacions de les variables numèriques amb el preu i amb la puntuació:

```
corr.res<-cor(select(zacatrus_data_amb_preu, EdadMinima, EdadMaxima,
NumJugadoresMinimo, NumJugadoresMaximo, TiempoDeJuegoMin,
TiempoDeJuegoMax, Largo, Volumen, Complejidad,
Dependencia.del.idioma, Precio))

corrplot(corr.res,method="circle")
```



Es pot veure com els atributs amb una correlació més forta amb el preu són el temps de joc, i l'edat mínima, tot i que aquesta correlació tampoc és gaire forta. Veiem aquesta relació en una gràfica:

```
plot(zacatrus_data_amb_puntuacio$Precio,zacatrus_data_amb_puntuacio$TiempoDeJuegoMax,xlab="Preu",ylab="Temps de joc màxim")
```

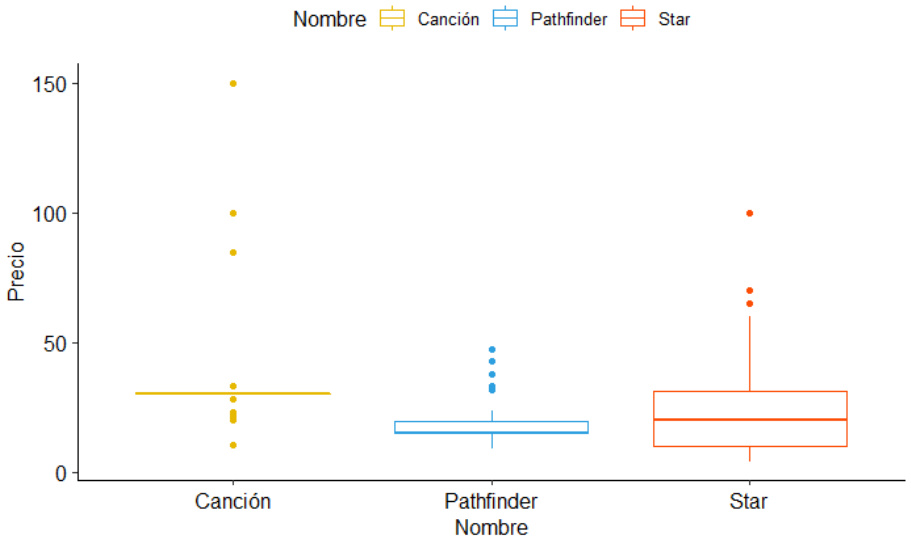


Es pot veure una lleugera correlació, però no massa forta.

A continuació, veurem visualment com afecta la saga del joc al seu preu

```
p <- ggplot(zacatrus_data_per_saga_preu, aes(x = Nombre, y = Precio))
bxp <- p + geom_boxplot(aes(color = Nombre)) +
```

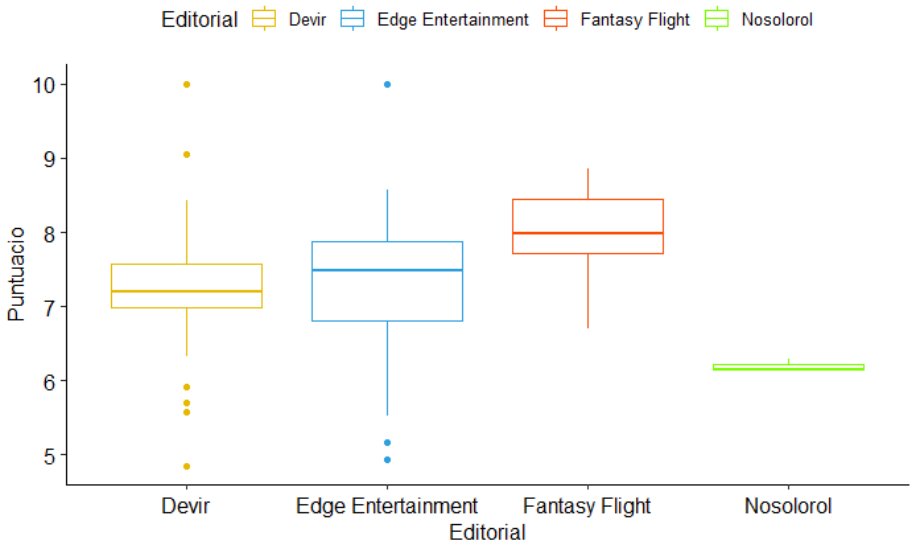
```
  scale_color_manual(values = my3cols)
```



Es pot veure com els preus de la saga "Star" són més dispersos, i per tant no permet determinar que els seus preus són menors que els de "Canción". Els de la saga "Pathfinder" en canvi tenen menys variabilitat i es pot determinar que són els més barats de les tres sagues.

Veiem ara l'efecte de l'editorial en la puntuació:

```
p <- ggplot(zacatrus_data_per_editorial_puntuacio, aes(x = Editorial,
y = Puntuacio))
bxp <- p + geom_boxplot(aes(color = Editorial)) +
  scale_color_manual(values = my3cols)
```



Gràficament resulten evidents les conclusions extretes de forma estadística: els jocs de Nosolorol són els menys valorats, els de Fantasy Flight els més, i els d'Edge i Devir tenen puntuacions similars.

6. Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

Hem començat la pràctica plantejant-nos uns objectius determinats sobre les dades. Per poder complir aquests objectius, el primer pas ha estat realitzar un procés de neteja sobre cadascun dels atributs que s'han obtingut al dataset creat a partir de les dades a la botiga online de Zacatrus. Aquest procés de neteja ha implicat l'extracció d'atributs a partir d'uns altres (saga a partir del nom), la transformació d'atributs multivalor en conjunts d'atributs binaris, i altres processos de transformació i neteja de les dades.

Posteriorment, s'han realitzat tres proves estadístiques sobre les dades, que ens havien de permetre trobar respostes a les preguntes que ens havíem plantejat originalment. L'anàlisi de correlació ens ha permès determinar quins dels atributs numèrics tenia una relació més forta amb els atributs que eren objectiu, i per tant haurien de ser més útils per realitzar prediccions. Hem vist com els jocs amb un temps de joc més gran indica, lleugerament, que el joc tendirà a tenir també un preu més gran. També es pot veure que, tal com ens podia indicar la lògica, els jocs amb més volum tendeixen a ser més cars. Pel que fa a la puntuació, també es veu una relació directa entre el temps de joc i la valoració dels usuaris. En canvi, en aquest cas una capsa més gran no suposa una aprovació major per part dels jugadors.

A continuació, amb les proves de contrast d'anàlisis, hem comprovat per dos dels atributs categòrics com afecten el preu i a la puntuació dels jocs. Hem vist com la saga a la qual pertany el joc pot tenir efecte en el preu (a vegades, el nom comercial també "es paga"). S'ha pogut determinar estadísticament com les sagues "Star Wars" i "Canción de hielo y fuego" tenen de mitjana un preu més gran que la saga "Pathfinder" que no té una gran superproducció audiovisual al darrere. Amb el mateix tipus d'anàlisi, s'ha pogut determinar que també algunes editorials de jocs parteixen d'una valoració més gran per part dels jugadors que altres. En concret, els jocs de Fantasy Flight són millor valorats que no pas la de la resta d'editorials analitzades. Tot i que els de Nosolorol hem determinat estadísticament que són els menys valorats, el nombre tan petit de mostres no ens permet realitzar aquesta afirmació.

Finalment, s'ha generat un model de regressió que, tot i utilitzar tots els atributs numèrics disponibles, no ha estat capaç de proporcionar un nivell de bondat de l'ajust suficient que faci possible utilitzar el model en un sistema de predicció de preus dels jocs. És evident que els factors que poden afectar a un joc de taula són molts més dels que hem pogut obtenir de la pàgina web de Zacatrus, i molts d'ells són atributs "no mesurables" com per exemple l'art de la portada del joc. De tota manera, es podria analitzar si nous models amb altres mètodes d'aprenentatge supervisat més enllà de la regressió lineal poden generar prediccions millors dels preus a partir dels atributs disponibles.

7. Codi: Cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i representació de les dades. Si ho preferiu, també podeu treballar en Python.

El codi de la pràctica es pot obtenir al següent repositori de GitHub:
https://github.com/jmontufo/zacatrusAnalyzer

Contribucions

Contribuciones	Firma
Investigació prèvia	JM, MM
Redacció de les respostes	JM, MM
Desenvolupament codi	JM, MM