Contents

Introducción

En este trabajo, llevaremos a cabo un análisis más profundo del csv generado en la PRA1 mediante el uso de webscraping. Este dataset contiene información variada sobre los juegos de la plataforma Steam listados en https://www.instant-gaming.com/.

La información acabarca precio, porcentaje de descuento, valoración y reseñas, asi como el titulo del juego, desarrollador y publisher. También una columna que contiene información acerca del juego en si, como puede ser el género, si es multijugador, online etc.

Dado que en los recientes años, y especialmente con el Covid, la industria del videojuego ha experimentado un crecimiento significativo. Con la aparición de plataformas en linea para la adiquisición de videojuegos, se genera una gran cantidad de información que puede ser aprovechada para estudiar las tendencias del mercado y comportamiento de consumidores.

Metodologia y Objetivos

Mediante técnicas de análisis exploratorio y modelado predictivo, buscaremos identificar patrones, tendencias y relaciones entre las variables presentes en el dataset, intentando obtener conocimientos y patrones relevantes en el ámbito de los videojuegos.

Algunas de las preguntas que nos planteamos abordar son: ¿Cuáles son los géneros de juegos más populares entre los usuarios? ¿Existe alguna relación entre el precio de un juego y su nivel de descuento? ¿Qué características influyen en la valoración de los usuarios sobre un juego?

Para conseguir esto, utilizaremos tecnicas de representación visual (graficos, tablas..) asi como el analisis de una serie de indicadores que nos proporcionarán información sobre el rendimiento y resultados de nuestros modelos.

Limpieza y procesado de datos

Presentación de los datos

El dataset consiste exclusivamente de juegos de la plataforma Steam. Consiste de aproximadamente 8000 observaciones y un total de 12 variables.

A continuación, echaremos un primer vistazo a nuestros datos

```
## Rows: 8,194
## Columns: 12
## $ title        <chr> "Assetto Corsa Competizione - 2023 GT World Challenge Pac~
## $ price        <chr> "9.99€", "13.89€", "19.09€", "14.49€", "36.87€", "18.99€"~
## $ discount     <chr> "-23%", "-31%", "-24%", "-28%", "-39%", "-24%", "-31%", "~
## $ developer    <chr> "Kunos Simulazioni", "Passtech Games", "Black Salt Games"~
## $ publisher    <chr> "505 Games", "Nacon", "Team17", "Paradox Interactive", "P~
## $ tags         <chr> "['Un solo jugador', 'Indies', 'Carreras', 'Simulación', ~
## $ release_date <chr> "19 abril 2023", "6 abril 2023", "30 marzo 2023", "18 abr~
## $ valoration   <int> 10, NA, 10, NA, 6, 10, 10, 10, 10, 7, 10, NA, 10, 10, 10,~
## $ reviews      <int> 1, 0, 16, 0, 211, 2, 517, 3, 517, 8, 59, 0, 1, 150, 150, ~
## $ stock        <chr> "En stock", "En stock", "En stock", "En stock", "En stock~
## $ descarga     <chr> "Descarga digital", "Descarga digital", "Descarga digital~
## $ f2p          <chr> "False", "False", "False", "False", "False", "False", "Fa~
```

Explicación de las variables

“Title”: Nombre del videojuego

“Price”: Precio del videojuego

“Discount”: Descuento frente al mercado físico.

“Developer”: Estudio que ha desarrollado el juego.

“Publisher”: Quien es el publicador del videojuego.

“tags”: Atributos del juego, como, por ejemplo, de que genero es, si es multijugador, etc.

“release_date”: Fecha de lanzamiento “valoration”: valoración media del videojuego

“reviews”: número de reseñas que tiene el videojuego

“stock”: Indica si hay disponibilidad/stock

“download”: Tipo de descarga.

“f2p”: Indica si el juego es gratis

Analisis de datos perdidos

Primero, vamos a ver si realmente encontramos elementos en blanco o perdidos en nuestro dataset

```
## [1] "Blancos"
```

```
##        title        price     discount    developer    publisher         tags
##            0          625          683          352          362            0
## release_date   valoration      reviews        stock     descarga          f2p
##           35           NA           NA          433         2803            0
```

Tras comprobar si hay elementos en blanco vemos que:

- Precio tiene 625 elementos,
- Discount 683,
- Developer 352,
- Publisher 362,
- Release_Date: 35
- Stock 433,
- Descarga 2803

```
## [1] "NA"
```

```
##        title        price     discount    developer    publisher         tags
##            0            0            0            0            0            0
## release_date   valoration      reviews        stock     descarga          f2p
##            0         4130          321            0            0            0
```

Respecto a los NA vemos que hay:

- 4130 en Valoration
- 321 en Reviews

Se observa que muchas valoraciones nulas tienen relación a cuando el valor de Review es 0, vamos a comprobarlo.

```
## [1] 321
```

```
## [1] 3809
```

Con esto podemos ver que efectivamente todos los valores nulos de valoration son debido a que hay reviews en 0 o reviews nulas.

Transformaciones

Transformar la columna precio a numérica

Transformar descuento a numérica

Transformar columna stock a dictómica dependiendo si hay stock o no

Vemos que la columna Tags son listas de caracteres, por lo que modificarlo para quitar los caracteres "[" y "]" para mejor visualización

Transformar F2p a dictómica con un 0 si es False y un 1 si es True

Creación columna precio inicial

Creamos una nueva columna para ver si es DLC o no

Tratado de blancos

Como parte del tratado de blancos, vamos a sustituir las casillas en blanco de las variables developer y publisher por "Unknown" y los blancos de descarga digital como "No" descarga digital.

Extracción de info de la columna tags

Como hemos podido observar, la columna tags contiene información variada sobre los juegos. Para facilitar el trabajo con los datos y para que el dataframe quede más limpio, desglosaremos esta variable en varias.

Por un lado, en los que respecta al genero, vamos a considerar los generos más comunes: "Acción", "Aventura", "Carreras", "Aventura", "Estrategia", "Deporte", "Simulación", "RPG", "Indies", "Gestión". Se planteará de la siguiente forma: El primer elemento que aparezca en el string, será el genero principal del juego, y el segundo elemento que aparezca será el genero secundario. De no tener más de un genero, la columna genre2 tendrá el mismo valor que genre 1

También crearemos una columna nueva que indicará si el juego es indie, y si tiene multijugador, a modo de true/ false

Como podemos ver, este es el aspecto de nuestro nuevo dataframe, pero el problema de los NA persiste.

La columna que mas valores perdidos presenta es la columna de valoración. Para no perder la mitad de las observaciones, imputaremos los valores perdidos de la columna valoration utilizando el paquete "mice". Este paquete utiliza un algoritmo random forest para imputar los valores perdidos de la columna valoration.

Recordemos que no eliminamos ciertos espacios en blanco que llamamos unknown. Esto representa que desconocemos el desarrollador o publicador. Estas columnas podemos guardarlas o no (en este caso las eliminaremos). Para no reducir demasiado el dataset, imputaremos los nas de descuento y precio también

Imputación knn de valoration y reviews (multiple options) y tratado de NAs

Tras imputar ciertos valores perdidos numericos, vamos a deshacernos de los desconocidos, asi como de la columna tags.

De esta forma el dataset final tendria esta estructura

PCA Analysis

Vamos a continuar ralizando un analisis PCA (Princiapl Component Analysis). Este analisis es muy utilizado a la hora de trabajar con datos de alta dimensionalidad, permitiendo reducir el tamaño del conjunto de datos reteniendo una cantidad razonable de información. En este sentido, nuestro dataframe no es excesivamente grande, pero aun asi, realizar la PCA es interesante, ya que nos puede aportar información acerca de variables nos aportan mas información, y que variables pueden ser prescindibles.

No obstante, para aplicar el analisis, tendremos que sometes nuestro dataframe a otra serie de transformacione.

Para esto, crearemos un nuevo dataset para no alterar el que ya tenemos

Scaling numerics

Dado que el algoritmo PCA solo puede procesar datos numericos (y aunque no sea la mejor opción), transformaremos las columnas factor a numericas. Además, este metodo también es muy sensible a la relatividad de los datos. Para evitar asignar más peso del debido a ciertas variables, procederemos a escalar los datos numericos. Como defecto el procedimiento que utilizamos utiliza el proceso Z-score normalization, que centra

las variables con media 0 y desviación estandar 1. También para este analisis se desglosa la columna que indica la fecha de lanzamiento.

```
## [1] "Spanish_Spain.1252"
```

Convert factors and bool to numeric

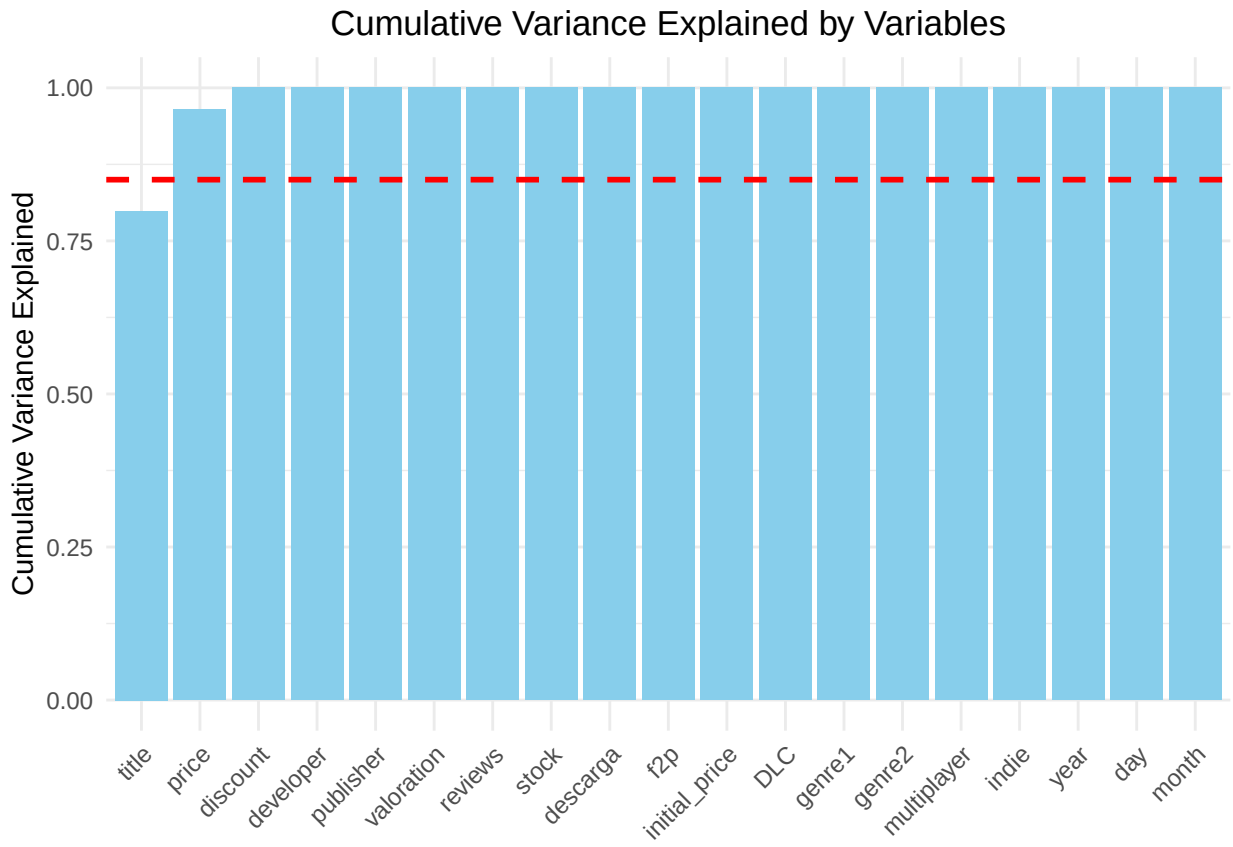
Tras todas las conversiones pertinentes, podemos tomar un vistazo rapido a nuestro dataset

PCA

Variable	VarianceExplained	CumulativeVariance
title	0.7989245	0.7989245
price	0.1658591	0.9647835
discount	0.0352094	0.9999930
developer	0.0000042	0.9999972
publisher	0.0000010	0.9999981
valoration	0.0000004	0.9999985
reviews	0.0000002	0.9999987
stock	0.0000002	0.9999989
descarga	0.0000002	0.9999991
f2p	0.0000002	0.9999993

Aqui podemos observar los resultados del análisis PCA, donde se muestra la varianza explicada de cada variable y la varianza acumulada explicada. Sorprende/salta a la vista que 2 variables explican practicamente la totalidad de la varianza. Representemos esto y analicemos en mas profundidad los resultados

Representación gráfica resultados PCA



Aqui vemos los resultados con más claridad. Como ya vimos en los resultados, 2 variables explican casi la totalidad del dataset. Esto quiere decir que la mayoria de información relevante se concentra en estas dos variables, y potencialmente tienen una alta contribución en los patrones y estructura de nuestros datos. Este resultado sorprende por varias razones: Factores como el genero del juego, la valoración o la cantidad de reseñas no tiene mucha relevancia en este contexto. También sorprende porque en una industria tan grande como la del videojuego no es normal que gran cantidad de la información se resuma en 2 variables.

No obstante, al haber aplicado este test a un dataframe en condiciones subopticas, es mejor coger "con pinzas" este resultado.

##dataframe for Modelization

Analisis Exploratorio de los datos

Tras realizar todos los cambios y transformaciones pertinentes a nuestro conjunto de dato, podemos proceder a un primer analisis exploratorio de nuestros datos.

Developers y publisher

Empecemos por los desarrolladores y publicadores.

	Var1	Freq	Mean_Valoration
2143	Paradox Development Studio	123	7.00
844	Dovetail Games	86	10.00
1013	Feral Interactive (Linux)	79	9.00
1596	KOEI TECMO GAMES CO., LTD.	73	9.00
615	Colossal Order Ltd.	65	2.00
497	CAPCOM Co., Ltd.	62	10.00
237	Aspyr (Linux)	54	8.00
1098	Frontier Developments	44	10.00
2517	SCS Software	41	9.17
2690	Square Enix	39	8.29
1848	Milestone S.r.l.	37	10.00
2871	Techland	37	5.00
196	Arc System Works	35	9.00
1280	Haemimont Games	34	8.00
336	Bethesda Game Studios	33	3.00

Aqui podemos observar los Publicadores que más juegos tienen publicados. Encabeza Paradox y Dovetail games (y algunos nombres conocidos dentro de la industria). También podemos observar en la tabla la valoracion media de los juegos de cada desarrollador

Hagamos ahora lo mismo para los publicadores de videojuegos

	Var1	Freq	Mean_Valoration
1112	Paradox Interactive	340	10.00
1309	SEGA	182	10.00
178	BANDAI NAMCO Entertainment	160	9.00
22	2K	136	10.00
1387	Square Enix	127	10.00
1505	THQ Nordic	114	8.00
209	Bethesda Softworks	101	10.00
1017	Nacon	98	10.00
1352	Slitherine Ltd.	90	9.17
528	Feral Interactive (Linux)	79	8.55
557	Focus Entertainment	78	10.00
375	Daedalic Entertainment	70	9.00
1460	Team17 Digital Ltd	69	6.00
399	Devolver Digital	65	5.00
587	Fulqrum Publishing	62	6.00

Tras analizar quienes son las empresas que mas aparecen, vamos a ver que tipos de juegos hace cada empresa.

publisher	Acción	Aventura	Carreras	Deporte	Estrategia	RPG	Simulación
2K	48	4	6	2	17	10	49
BANDAI NAMCO Entertainment	128	10	6	0	1	12	3
Bethesda Softworks	65	0	0	0	0	36	0
Daedalic Entertainment	17	34	0	0	6	5	8
Devolver Digital	48	11	0	1	1	2	2
Feral Interactive (Linux)	42	0	5	0	32	0	0
Focus Entertainment	41	15	0	0	10	4	8

publisher	Acción	Aventura	Carreras	Deporte	Estrategia	RPG	Simulación
Fulqrum Publishing	24	5	0	0	12	14	7
Nacon	26	5	26	20	1	4	16
Paradox Interactive	19	10	0	0	49	43	219
SEGA	77	6	7	1	49	20	22
Slitherine Ltd.	0	1	0	0	58	2	29
Square Enix	69	10	0	0	6	41	1
Team17 Digital Ltd	33	4	0	1	22	1	8
THQ Nordic	56	16	7	0	5	23	7

```
##     Acción   Aventura   Carreras    Deporte Estrategia        RPG Simulación
##        693        131         57         25        269        217        379
```

Esto nos puede proporcinar información relevante sobre las empresas mas presentes en el mercado y a que tipo de juegos desarrollan con más frecuencia: Vemos, por ejemplo, que Paradox se dedica principalmente a juegos de Simulación/Estrategia, mientras que Bandai se especializa en juegos de acción. Por otro lado, vemos que generos como Deporte y Carreras son mas nicho, y solo una de los desarrolladores mas frecuentes se centran en este tipo de juegos.

Asimismo, podemos observar que el genero de acción es con diferencia el mas producido, seguido de simulación y estrategia

Multiplayer, Indie y Dlc

DLC

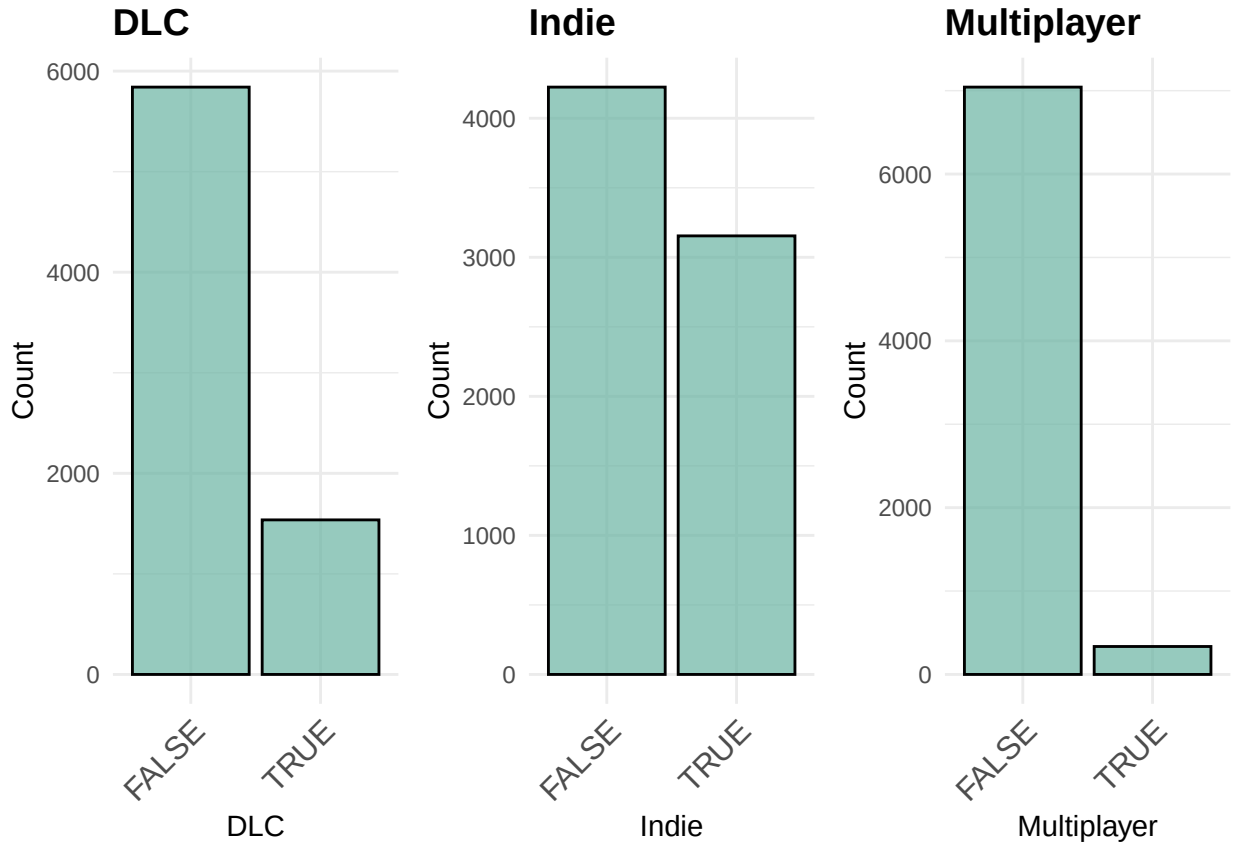
Vamos a comprobar que proporción de juegos tienen multijugador, son indies, y cuantos de ellos son dlcs (pack de expansión)

Vemos que, como es de esperar, la gran mayoria de registros son juegos, y no packs de expansión

Indie

Se puede apreciar que buena parte de los registros son juegos indie (llamados independientes, dado que son desarrollados por estudios mas pequeños/modestos). Puede ser indicio de un mercado bastante competitivo por el numero de competidores existentes

Multiplayer



Sorprende este resultado ya que en esta epoca se espera que los juegos tengan servicio en la nube, y multi-jugador a través de la nube. Esto puede indicar que probablemente hayamos interpretado la columna tags de forma erronea, y que "un solo jugador" no necesariamente implique ser SOLO de un jugador.

Relación entre variables

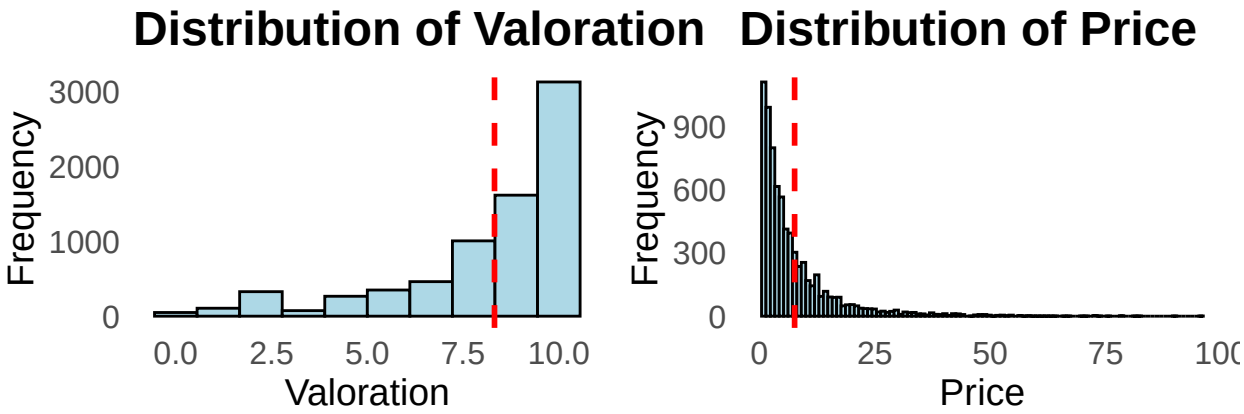
Tras analizar individualmente, veamos como se relacionan nuestras variables entre si

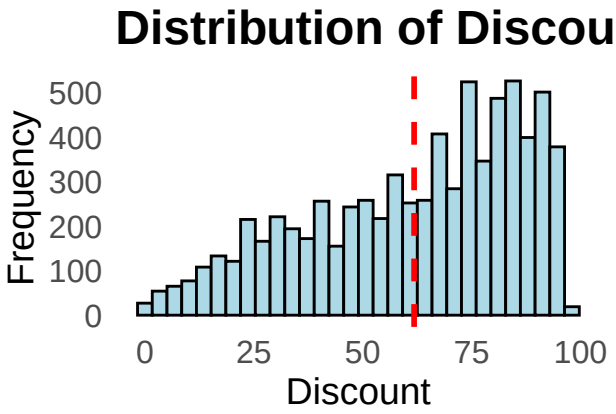
Histogramas/distribuciones

Vamos a comenzar representando los histogramas de algunas de nuestras variables numericas, como pueden ser precio, descuento o valoración. Estas gráficas nos darán indicios sobre la distribución que siguen nuestras variables, y si esta se asemeja a la normal.

A primera vista se ve que Valoration no sigue una distribución normal. Salta a la vista que hay un alto numero de valoraciones muy altas/maximas. La linea vertical roja indica la media.

(ver si dejar y normal o en escala logaritmica). Podemos observar que hay algunos juegos que tienen un coste mas elevado de lo normal. Esto se podria considerar outliers, pero dado que existen packs de juegos + expansión, estos altos precios son outliers legitimos. Aun asi, price tampoco sigue una distribución normal, y la gran mayoria de juegos tienen un precio inferior a 25 euros (lo que cabe esperar de un sitio que oferta juegos de una forma mas barata)



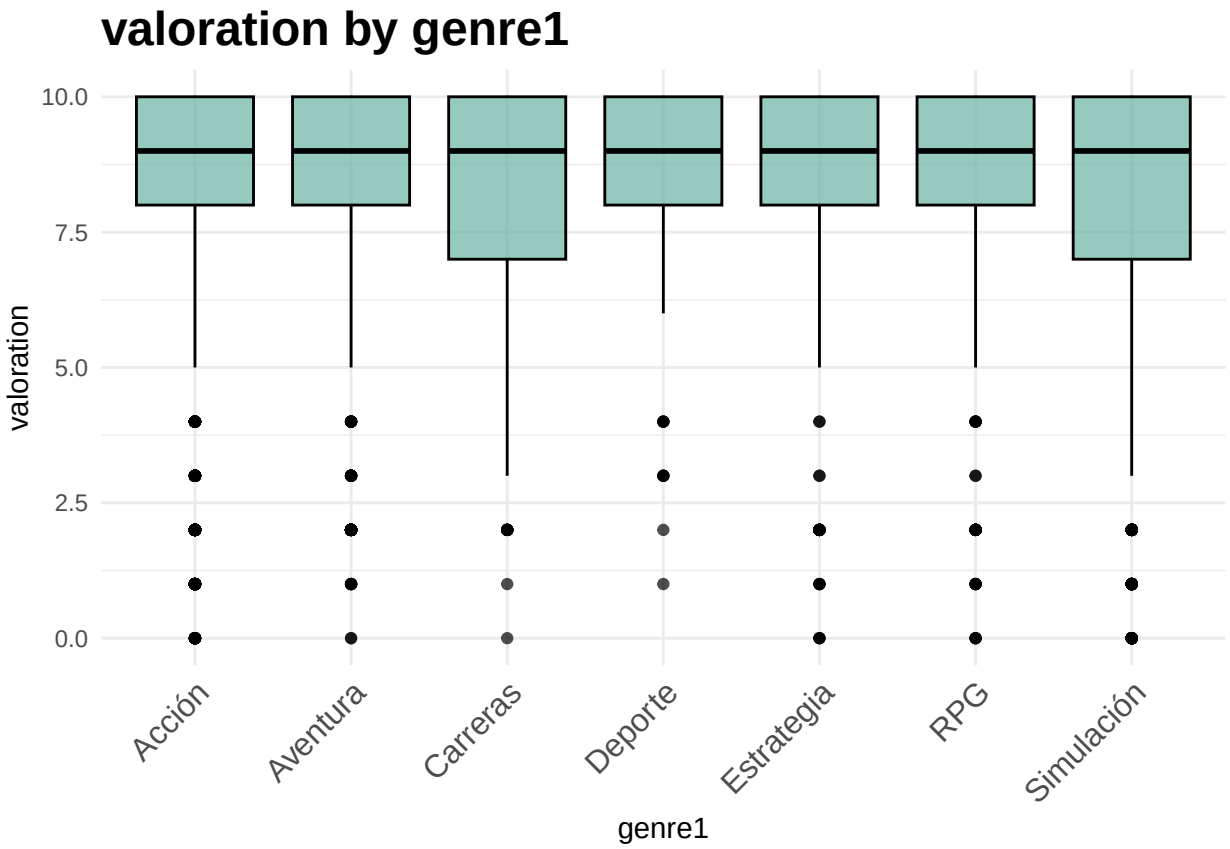


De nuevo, nos encontramos una distribución normal, y vemos que el descuento medio que ofrece la pagina es un 60%

Relación entre variables numericas-categoricas

Mostraremos a continuación analizando las relaciones entre variables numericas y categoricas. Por ejemplo, podemos apreciar diferencias en las valoraciones de juegos de distintos juegos? O en su precio? Veamoslo

Valoración y género



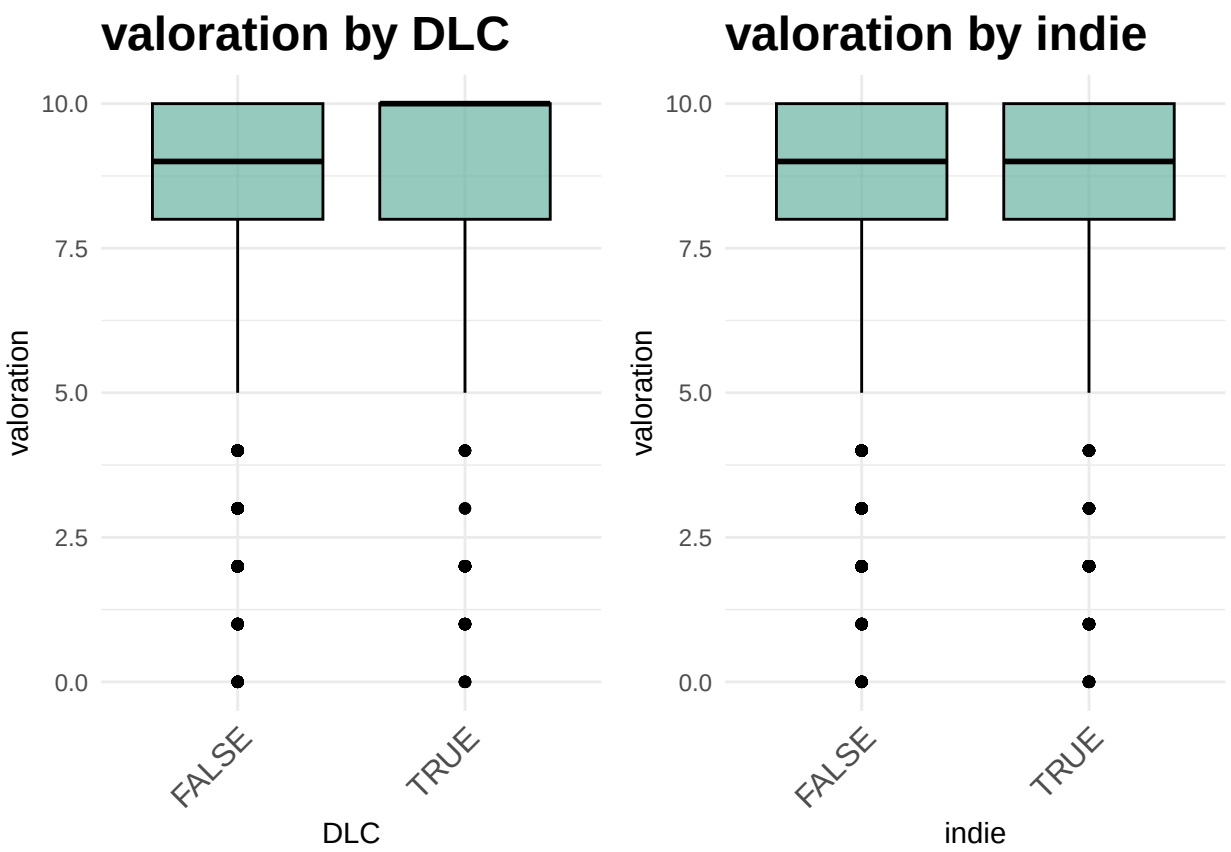
A primera vista, no se aprecian grandes diferencias (vemos que hay valoraciones malas en todos los generos), salvo en el genero de deportes, que aparentemente tiene una valoración media más baja.

Valoración y DLC

También podemos ver si las expansiones tienen en media mejor nota que los juegos (quizá la nota del juego base influye en la del DLC)

Puede confirmarse nuestras sospechas, aunque solo con esta información no se puede confirmar

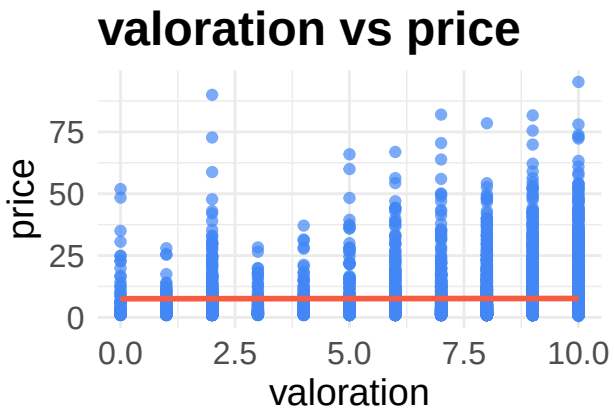
5.4.3: Valoración y Indie

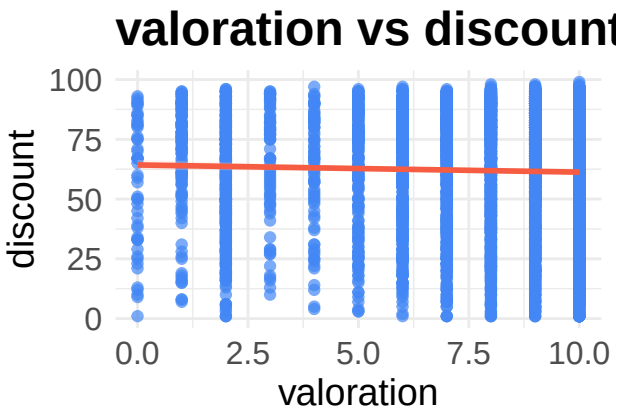


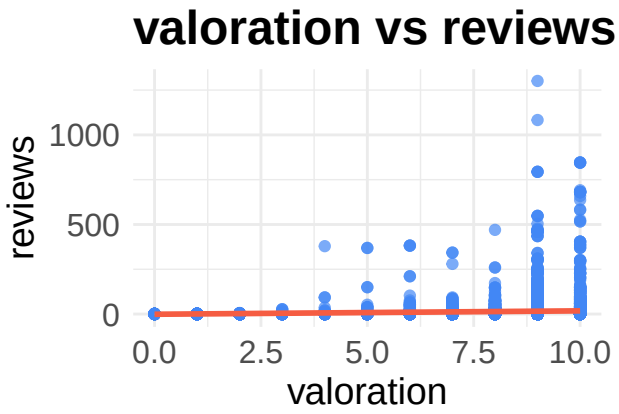
No se aprecian diferencias visibles entre valoracion media de juegos indie o no. Esto habla bien de estudios mas pequeños e independientes, que a pesar de tener menos recursos crean juegos con valoraciones a la altura de grandes empresas.

Relación entre variables numéricas

Algo que no puede ser tan obvio a primera vista como lo puede ser la relación entre el precio y el descuento, es como reacciona ante cambios de variables como precio, descuento o numero de reseñas.



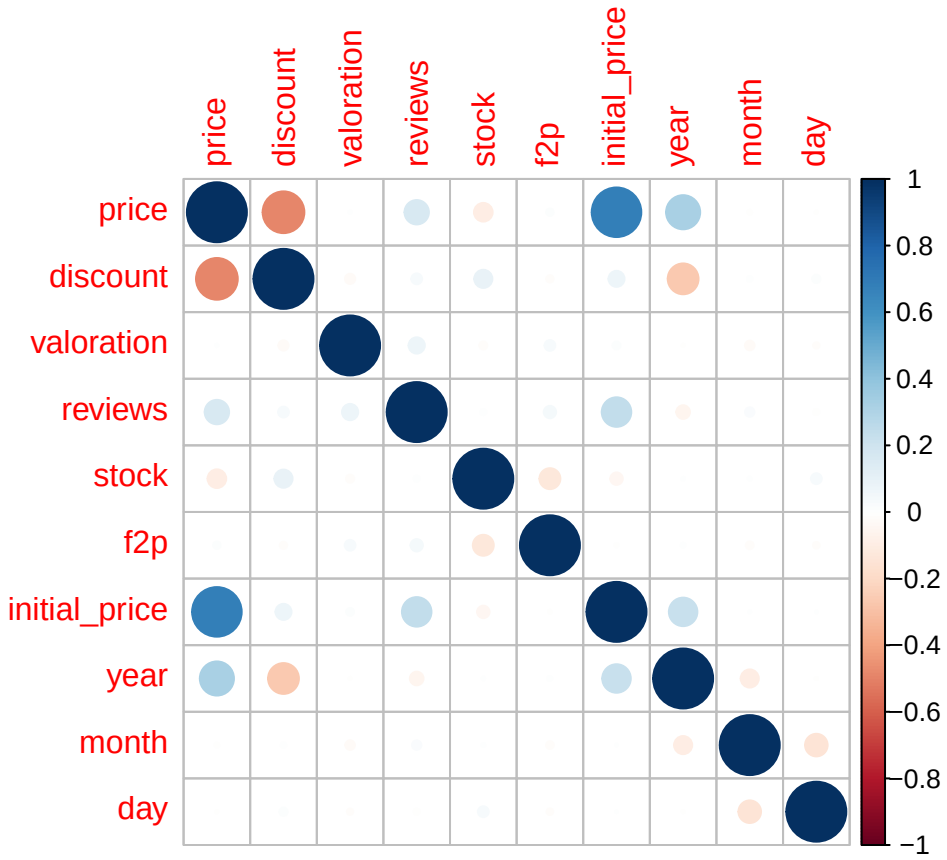




Aparentemente, la valoración no influye en el precio de forma notable. Tampoco se aprecia una gran relacion entre descuento y valoración, al igual que la cantidad de reseñas no influye en la valoración

Para confirmar lo que se intuye en los graficos mediante el proceso de contraste de hipótesis

Correlación de variables numéricas



No se aprecia correlacion, salvo con el precio/porecio inicial, lo cual es de esperar. Por el resto, el dataframe no presenta problemas de colinealidad

Contrastes de hipótesis

Contraste de medias para acción y deporte

Selección de muestras Queremos ver si los generos impactan en la media de la valoración. Para ello, vamos a ver si hay una diferencia significativa entre generos de juegos.Vamos a comparar el genero de acción con deporte, para confirmar o desmentir la intución que nos han dado los gráficos

Para aplicar el test correcto y correspondiente, primero hemos de comprobar una serie de cosas: Que distribución siguen los datos, y si las muestras tienen varianzas iguales o distintas.

Test de shapiro (normalidad) Ya vimos en los gráficos anteriores que la mayoria de variables no siguen una distribución normal. De todas maneras, lo vamos a comprobar aplicando el test de shapiro, que tiene como hipótesis nula la normalidad de los datos, e hipótesis alternativa la asunción de no normalidad.

```
##
##  Shapiro-Wilk normality test
##
## data:  accion_data_valoration
## W = 0.72825, p-value < 2.2e-16
```

```
##
##  Shapiro-Wilk normality test
##
## data:  deporte_data_valoration
## W = 0.78056, p-value = 3.789e-08
```

Vemos que ambos p-valores son muy muy pequeñas, esto quiere decir que rechazamos la hipotesis nula. Por lo tanto, la distribución de nuestros datos no es normal y se confirma lo que hemos intuido graficamente.

Test de Levene (igualdad de varianzas) El test de Levene es una prueba estadística utilizada para comprobar si dos muestras tienen igualdad de varianza.

El procedimiento del test consiste en calcular una estadística de prueba basada en las diferencias absolutas entre los valores de las muestras y su media. Se puede utilizar tanto para muestras independientes como pareadas.

A contianuación, vamos a analizar la varianza de ambos grupos, y ver si son iguales o no. Dado que nuestros datos no estan distribuidos de forma normal, utilizaremos el test Levene.El test Levene tiene como hipotesis nula la asunción de igualdad de varianza, y como hipótesis alternativa el rechazo de igualdad de varianzas.

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
## group    6  6.5648 6.267e-07 ***
##       7371
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P valor extremadamente pequeño. Rechazamos la hipotesis nula de igualdad de varianzas, y por tanto tenemos varianzas desiguales. El test levene analiza la variación de varianzas entre grupos de una variable discreta. Esto quiere decir que las varianzas entre todos los niveles de genre1 son desiguales.

Test a Realizar Dado que nos encontramos con dos muestras de varianza poblacional desconocida y distinta, podemos aplicar el test de Welch (t de welch), como alternativa a t de Student

El test tendrá como hipotesis nula la igualdad de medias x1=x2. Hipotesis alternativa x1!=x2

Aplicación Apliquemos ahora el test correspondiente:

```
##
##  Welch Two Sample t-test
##
## data:  accion_data_valoration and deporte_data_valoration
## t = 1.1737, df = 61.751, p-value = 0.245
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2596627  0.9980417
## sample estimates:
## mean of x mean of y
##  8.352796  7.983607
```

P valor inferior a 0.05. Esto significa que rechazamos la hipotesis nula, y podemos afirmar que existen diferencias significativas entre medias Comprobamos y afirmamos lo que se podia intuir en el grafico: el genero influye en la nota final, almenos en el caso de accion/deporte.

Podemos, a través de otro analisis de contraste, ver si esto se refleja en el precio

Contraste medias deporte/acción en cuanto a precio

Como ya hemos presentado el procedimiento para justificar el test, mostraremos directamente los resultados del test

```
##
##  Shapiro-Wilk normality test
##
## data:  accion_data_precio
## W = 0.70722, p-value < 2.2e-16
```

```
##
##  Shapiro-Wilk normality test
##
## data:  deporte_data_precio
## W = 0.74938, p-value = 7.414e-09
```

```
##
##  Welch Two Sample t-test
##
## data:  accion_data_precio and deporte_data_precio
## t = 1.0563, df = 63.261, p-value = 0.2949
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.8876094  2.8783366
## sample estimates:
## mean of x mean of y
##  8.043396  7.048033
```

A pesar de la diferencia de medias en cuanto a valoraciones, no se puede afirmar que exista una diferencia en la media de precios.

Modelos de predicción

En este apartado vamos a construir dos modelos predictivos: regresión lineal y decision tree.

Antes de nada (y aunque no se muestre), se recomienda partir los datos en dos grupos, 1 de entrenamiento y otro de test, con el que podemos medir el rendimiento de nuestro modelo.

El objetivo de este apartado es ver si podemos predecir el precio de un juego basado en información contenida en el conjunto de datos:

Regresión lineal

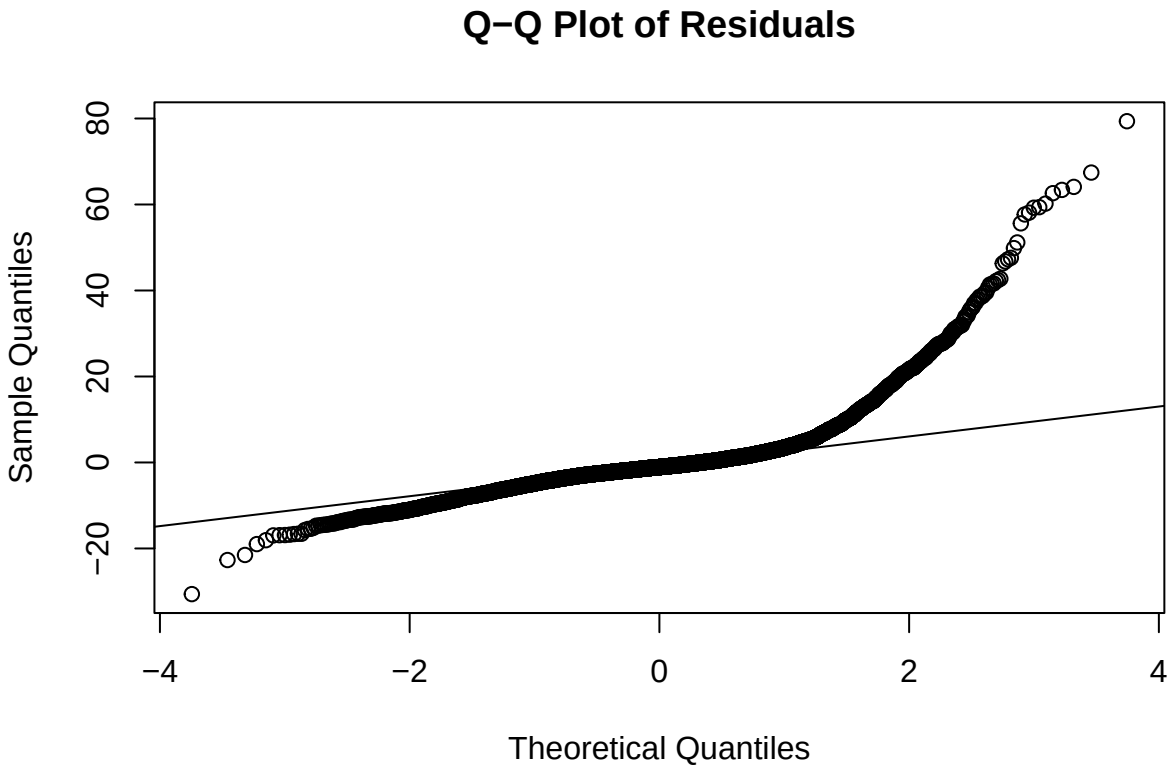
Predecimos precio.

Vamos a hacer una regresión sobre los precios para ver qué variables influyen en gran medida. Vamos a empezar analizando Reviews, Discount y Valoration.

```
## # A tibble: 5 x 5
##   term         estimate std.error statistic   p.value
##   <chr>           <dbl>     <dbl>     <dbl>     <dbl>
```

```
## 1 (Intercept) -858.      46.1        -18.6  8.20e- 75
## 2 reviews        0.0285   0.00160     17.8  4.02e- 69
## 3 discount      -0.156    0.00422    -37.1  4.91e-269
## 4 valoration    -0.0965   0.0428      -2.25 2.43e-  2
## 5 year           0.434    0.0228      19.0  5.62e- 78
```

```
## [1] "El R cuadrado del modelo es 0.312904"
```




```
## [1] 102.5865
```

Análisis:

- E R-cuadrado es de 0.25, esto significa que el modelo explica un 25% de la variabilidad de la variable precio. Esto es bastante poco.
- Las variables reviews y discount son significativas debido a que su p-valor es menor a 0.05, aunque sus coeficientes son muy bajos.
- La variable valoration no es significativa debido a que p-valor es mayor a 0.05, por lo que la quitaremos del modelo.

Vamos a probar ahora añadiéndole una variable cualitativa como es genero, como referencia pondremos Accion

```
## # A tibble: 10 x 5
##    term              estimate std.error statistic   p.value
```

```
##    <chr>               <dbl>     <dbl>     <dbl>     <dbl>
##  1 (Intercept)      -863.      46.3       -18.6   4.18e- 75
##  2 reviews             0.0273   0.00159    17.2   1.67e- 64
##  3 discount           -0.157    0.00420   -37.4   9.42e-273
##  4 year                0.436    0.0229     19.0   3.40e- 78
##  5 genre1Aventura     -1.54     0.292      -5.29  1.26e-  7
##  6 genre1Carreras      1.15     0.585       1.96  5.02e-  2
##  7 genre1Deporte      -0.777    1.11       -0.699 4.85e-  1
##  8 genre1Estrategia   -1.24     0.394      -3.15  1.66e-  3
##  9 genre1RPG           1.50     0.398       3.76  1.70e-  4
## 10 genre1Simulación   -1.35     0.282      -4.77  1.86e-  6
```

```
## [1] "El R cuadrado del modelo es 0.322317"
```

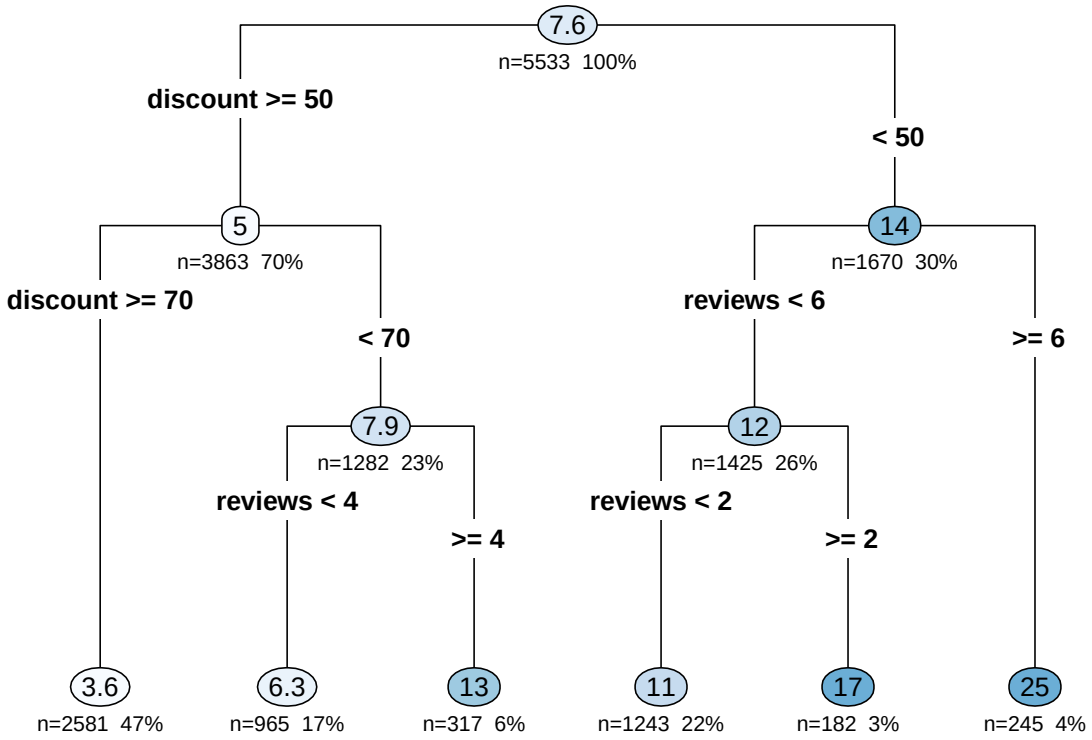
Análisis:

- El R-cuadrado sigue siendo muy bajo, casi un 27% este modelo.
- Las variables reviews y género siguen siendo significativas y mantienen sus coefcientes
- Los géneros: AVenturas, Carreras, Estrategie, Indies y RPG son significativos. Sus coeficientes son de: -1.29, 3.07, -1.84, -2.48 y de 2.30 respectivamente. Estos coeficientes son mayores a los del modelo anterior.
- Los géneros Deporte y Simulación no son significativos debido a que su p-valor es mayor a 0.05.

Sobre la regresión podemos decir que no son muy buenos modelos debido a su R-cuadrado que es muy pequeño, hay otros modelos con otras herramientas que serían mejores para este tipo de predicciones.

Probaremos si Decision Tree es un mejor modelo.

Decision Tree




```
## [1] "R-squared: 0.3089"
```

```
## [1] "MSE: 56.1216"
```

Podemos ver que las métricas mejoran un poco, aunque no de forma significativa, sin alcanzar un R cuadrado de 0.35. El error cuadrado total se ve reducido también, aunque de nuevo sin ser una mejora sustancial.

En cada nodo se puede observar el valor (medio) predicho para cada nodo. Además vemos el numero de observaciones en cada nodo, y el % de observaciones que cae en este nodo. También se indica la condición que ha de cumplir cada observación para avanzar por un lado u otro.

Conclusiones

Tras el analisis realizado sobre el dataset de partida, podemos extraer las siguientes conclusiones:

- La valoración media de los juegos en la plataforma es de un 8.5.
- Hemos descubierto quienes son las 15 compañias que mas juegos han publicado en la plataforma y la valoración media obtenida de cada uno.
- Tomando como referencia estas 15 compañias, hemos analizado cuales son los generos de juegos más populares entre ellas: Destaca con más del doble de juegos que cualquier otra categoria la de acción. En segundo lugar encontramos simulación y en tercer lugar estrategia
- Poco menos que la mitad de los juegos publicados en esta plataforma son indies.

- Graficamente hemos intuido que la valoración media de los DLC son mayores a los juegos no DLC.
- Hemos comprobado mediante contrastes de hipótesis que el género puede influir de forma significativa el la media de valoración (y hemos visto que deporte es el genero con peor valoración)
- Los datos no siguen una distribución normal
- La variabilidad total del dataset está explicada principalmente en 2 variables: title y price. Title en este caso actua como identificador único, lo que provoca un alto nivel de overfitting.
- Modelos predictivos con un R cuadrado en torno al 30%, un valor bastante bajo, indicando un modelo poco fiable.Esto puede darse a que nuestro dataset no captura variables relevantes en la industria, como puede ser el tamaño del juego(A, AA, AAA), tamaño de la empresa, costes de desarrollo etc. Otra posibilidad es que nuestros modelos no son capaces de capturar las relaciones no lineales existentes entre las variables.

Con todo esto, a pesar de no poder crear modelos predictivos muy buenos, si hemos obtenido información valiosa acerca de la industria del videojuego, sus empresas más grandes en cuanto a numero de juegos se refiere, que géneros son los más producidos y si este influye de alguna manera en la valoración.

Participaciones

Column1	Column2
Investigación previa	SMM, DS
Redacción de respuestas	SMM, DS
Desarrollo del código	SMM, DS
Video	-