Proyecto de consultas e implementación de predicción precios, sobre dataset de características inherentes a lanzamientos de videojuegos hasta el año 2020.

Problema.

Se tiene un datase con información relevante inherente a la descripción de videojuegos y sus respectivos lanzamientos. Finalmente, se tiene entre los registros, los precios con los cuales han sido lanzados. En tal sentido se plantean dos objetivos:

1. Desarrollar seis consultas sobre el dataset, tal que pueda ser extraída información útil para posteriores análisis. Dichas consultas (todas a partir del dato "año" como insumo) se enfocan en los siguiente:
    1. "genero(año)". Devuelve los primeros 5 generos de videojuegos mas lanzados en el año.
    2. "juegos(año)". Devuelve los títulos de todos los juegos lanzados en el año.
    3. "specs(año)". Devuelve las 5 especificaciones de juego mas lanzadas en el año.
    4. "earlyaccess(año)". Devuelve la cantidad de juegos que fueron lanzados bajo esta modalidad (lanzamiento temprano) en el año.
    5. "sentiment(año)". Devuelve en orden descendente la calificación perceptiva destacadas por consumidores en el año.
    6. "metascore(año)". Devuelve los 5 juegos con mejor calificación cuantitativa en el año.
2. Desarrollar una función que permita, basado en la compilación de datos del dataset provisto, suponer o predecir el precio que tendría un videojuego en su lanzamiento, en función a las características del mismo. Dicha función se denomino "pediccion()". Y se diseño en función de 4 atributos: especificaciones, genero, lanzamiento, sentimiento.

En lo siguiente del documento se desarrolla el detalle del trabajo realizado y entendimiento de los datos.

Que tenemos?.

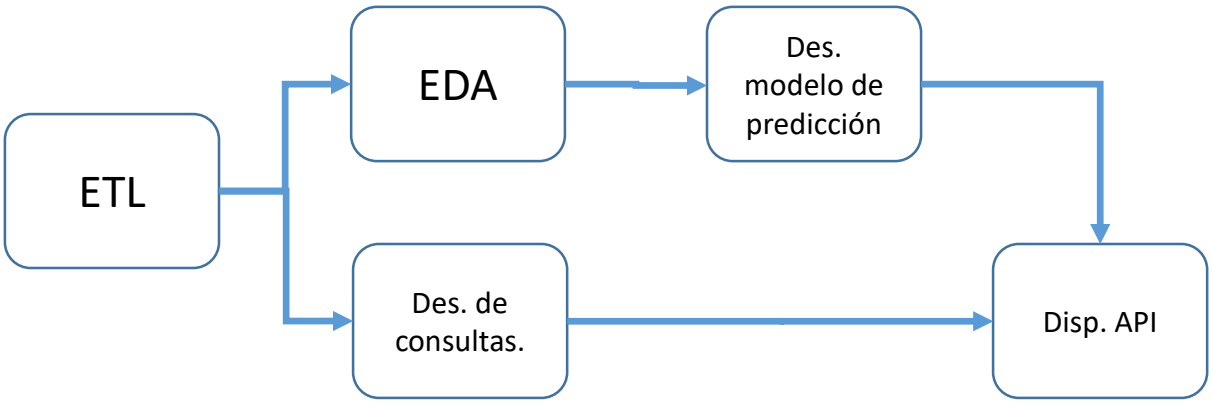
Tenemos un dataset “steam_games.json”, en el cual se hayan disponible los siguientes datos:

Columna	Descripcion
publisher	Nombre de la empresa que publica el juego.
genres	Generos a los cuales esta asociado el juego según sus creadores.
app_name	Nombre de la aplicación, si el juego no es un producto fisico.
title	Titulo comercial del juego.
url	Direccion del sitio web
release_date	Fecha oficial de lanzamiento del juego.
tags	Etiquetas que permiten identificar genero o cracteristicas de juego.
discount_price	Descuento sobre el precio de lanzamiento.
reviews_url	Vistas del jeugo en el sitio web.
specs	Especificaciones funcionales del juego.
price	Precio
early_access	Caracteristica de lanzamiento anticipado.
id	Id del producto.
developer	Desarrollador oficial del juego.
sentiment	Calificacion perceptiva del juego.
metascore	calificacion objetiva del juego.

El archivo con un tamaño de 20.747 KB, es decir, aproximadamente 20 MB. Algunos de sus atributos contienen datos variados, tal es el caso de :

1. “genres”, genero.
2. “specs”, especificaciones.
3. “tags”, etiquetas.

Metodología de trabajo.



Metodología de trabajo.

ETL del dataset.

Dataset fuente:
steam_games.json

Archivo de trabajo:
Juegos_ETL.ipynb

Datasets resultante:
stems_games_expandido.csv
func_genero.csv
func_specs.csv
func_juego.csv
func_sentiment.csv
func_early.csv
func_metascore.csv

- Apertura del archivo (expansión de datos agrupados).
- Análisis estructural.
  - Cantidad de registros no nulos por columna.
  - Revisión de estructura por dato (contenido y validez).
- Adecuación y/o transformaciones de datos y/o columnas.

Criterios básicos de transformación:

1. No se borraron registros, se crearon nuevas columnas: “año”, “mes”, ”dia” y “precio_1”.
2. Los nuevos datasets se generaron a partir de filtros aplicados sobre datos con contenido erróneo, ejemplos:
   1. Valores alfanumérico en datos numéricos.
   2. Valores faltantes sin afectación en datos laterales.

La primera etapa de trabajo genero un conjunto de “datasets” transformados desde el punto de vista estructural de los datos y adecuados según el objeto de uso.

func_genero.csv, solo con tres columnas: 'id','año','genres‘.
func_specs.csv, con tres columnas: 'id','año','specs‘.
func_juego.csv, con tres columnas: 'id','año','title‘.
func_sentiment.csv, tres columnas: 'id','año','sentiment‘.
func_early.csv, tres columnas: 'id','año','early_access‘.
func_metascore.csv, con cuatro columnas: 'id','año','metascore',"title“.

Finalmente “stems_games_expandido.csv
” un archivo total con todas las adecuaciones generales implementadas:

1. Se completo valores en la columna “title” con valores de la columna “app_name”.
2. Todos los datos de columnas agrupadas, se expandieron.
3. Se agregaron algunas columnas con fines e facilitar el trabajo posterior y no eliminar datos.
4. Se filtro la cantidad de columnas a: 'id', 'publisher', 'genres', 'tags', 'title', 'specs', 'early_access', 'sentiment', 'metascore', 'año', 'precio_1‘.

Metodología de trabajo.

Desarrollo de consultas.

Dataset fuente:
func_genero.csv
func_specs.csv
func_juego.csv
func_sentiment.csv
func_early.csv
func_metascore.csv

Archivo de trabajo:
Juegos_Prep_Func

Formatos de salida de cada función con ejemplo:

Función:
genero(año)

```
genero(2017,df_generos)
```

```
{'Indie': {'counts': 5929},
 'Action': {'counts': 3525},
 'Casual': {'counts': 3147},
 'Adventure': {'counts': 2798},
 'Strategy': {'counts': 2257}}
```

Ejemplo: año = 2018

Salida:
Formato: diccionario.
Orden descendente por frecuencia.

Función:
juegos(año)

```
juegos(2019,df_titulos)
```

```
"{'Juegos': ['Raji: An Ancient Epic',
'The Legendary Player - Make Your Reputation - OPEN BETA',
'The End of an Age: Fading Remnants']}"
```

Ejemplo: año = 2019

Salida:
Formato: diccionario.
Sin orden. Listado de juegos.

Metodología de trabajo.

Función:
specs(año)

```
espec(2018,df_espec)
```

```
{'Single-player': {'counts': 87},
 'Steam Achievements': {'counts': 42},
 'Steam Cloud': {'counts': 22},
 'Full controller support': {'counts': 20},
 'Steam Trading Cards': {'counts': 18},
 'Partial Controller Support': {'counts': 18},
 'Downloadable Content': {'counts': 15},
```

Ejemplo: año = 2018

Salida:
Formato: diccionario.
Orden descendente por frecuencia.

Función:
earlyacces(año)

```
earlyacces(2018,df_early)
```

```
21
```

Ejemplo: año = 2018

Salida:
Único valor. Cantidad de juegos.

Función:
sentimiento(año)

```
sentimiento(2018,df_sentiment)
```

```
{'1 user reviews': {'counts': 8},
 '3 user reviews': {'counts': 6},
 'Mixed': {'counts': 6},
 '2 user reviews': {'counts': 4},
 'Mostly Positive': {'counts': 3},
 'Very Positive': {'counts': 3},
 '4 user reviews': {'counts': 2},
```

Ejemplo: año = 2018

Salida:
Formato: diccionario
Orden descendente por cantidad de revisiones.

Función:
sentimiento(año)

```
metascore(2016,df_metascore)
```

```
{'metascore': {'Out of the Park Baseball 17': 92.0,
  'Tumblestone': 91.0,
  "Stephen's Sausage Roll": 90.0,
  'NBA 2K17': 90.0,
  'Tadpole Treble': 90.0}}
```

Ejemplo: año = 2016

Salida:
Formato: diccionario
Orden descendente por cantidad de revisiones.

Metodología de trabajo.

EDA

Dataset fuente:
stems_games_expandido.csv

Archivo de trabajo:
EDA_ML_1.ipynb
EDA_ML_2.ipynb

Datasets producidos:
DatasetML_General.csv
DatasetML_Efectivo.csv

A continuación exposición breve del trabajo realizado y documentado en el archivo EDA_ML_1,ipynb.

Aplicación de los siguientes criterios de preparación:

1. Uso de la mayor cantidad de datos posible. Esto implicara la minimización de acciones de eliminación y exclusión posible.
2. Al mismo tiempo, la minimización de atributos. Esto implicara el descarte de atributos que no aportan información de interés al modelo o que no se encuentren en condiciones de aportar dicha información (ejemplo, muy pocos datos).

En tal sentido, se descartaron las siguientes columnas (con su respectivo razonamiento):

1. “id”, la identificación de cada producto (videojuego) no es relevante con respecto al modelo de predicción. Pues solo aporta información administrativa, pero no aporta al comportamiento de la variable “precio”.
2. “title”. A efectos del modelo de predicción tampoco este atributo aporta valor, pues el mismo solo esta orientado a su identificación comercial.
3. Se filtraron los registros con valores nulos, pues, a efectos del modelo de predicción si es necesario trabajar con todos los datos validos. El filtrado de valores nulos se realizo solo después de identificar a priori los primeros atributos descartables.

Metodología de trabajo.

4. Se descarto el atributo “tags”, esto debido a dos argumentos:
    1. Los valores contenidos de este atributo son extremadamente similares a los contenidos en el atributo “genres”.
    2. Adicionalmente, este atributo contiene menor cantidad da datos que el atributo “genres”.
5. Se descarto el atributo “metascore” pues la cantidad de datos contenida en el dataset es extremadamente bajo, lo cual, o restringe el uso de gran cantidad de información en el modelo, o lo imputamos y producimos sesgo. Dadas estas razones, no se considero para el modelo de predicción.
6. Se restringieron todos los registros (filas) cuyos valores en precio, eran irregulares, es decir, con contenido errado (No numérico, nulos, con símbolos).
7. Finalmente se descartaron después de las acciones anteriores, todos los registros (filas) repetidas.

Finalmente y en este punto, el dataset insumo mantiene 190.768 filas, y 9 columnas (8 atributos, 1 variable objetivo).

A continuación exposición breve del trabajo realizado y documentado en el archivo EDA_ML_2.ipynb.

Como pudo observarse, en la primera etapa del EDA se tomaron decisiones de tipo estructural de la información con base al uso en el modelo de predicción a implementar.

En tal sentido, procedemos entonces a describir brevemente las actividades de análisis y situaciones halladas en la información.

1. Análisis sobre la variable objetivo “precio”.
    1. Outliers
        1. En este respecto se identifico importantes desviaciones que modifican su distribución, sin embargo, esto es un fenómeno común. Constantemente se dan lanzamientos de ediciones especiales o limitadas que impactan en el precio. Por tanto tal comportamiento extremo debe considerarse.
    2. Distribución.
        1. Se observo que la variable precio aun cuando presenta una clara tendencia, su dispersión es sumamente alta.

Metodología de trabajo.

MODELO ML

Dataset fuente:
DatasetML_Efectivo.csv
DatasetML_General.csv

Archivo de trabajo:
Mod_Prediccion_ML.ipynb

A continuación los criterios básicos implementados en el modelo y en la función de predicción:

En el modelo:

1. Se utilizo regresin lineal y se aplico dicho algoritmo sobre dos conjuntos de datos:
    1. DatasetML_Efectivo.csv

        Este Dataset contiene todos los datos preprocesados para ser consumidos por el modelo pero, con aquellos cuyo precio no excede de los 19.96$ (el razonamiento de este seccionamiento se explica posteriormente en el documento). Este filtro es consecuencia de la extracción del 85% de la población de precio que se comporta sin "outliers", es decir es el rango de valores que mantiene un comportamiento normal y que genera una desv estándar de 4.37.
    2. DatasetML_General.csv

        Este Dataset, también preprocesado, contiene la totalidad de los datos, es decir, considera a la "población outlier" del precio. Esta población total genera unos estadísticos del precio bastante dispersos, desv estándar de 13.44. Esta variabilidad en el precio tiene sentido en el rango de la población outlier, pero no en el rango mas concentrado. Sin embargo, es de tener en cuenta que este compotamiento "outlier" no es consecuencia de errores, es también un fenómeno normal en el precio de videojuegos, por tanto, hay que considerarlo en el modelo de predicción.

Metodología de trabajo.

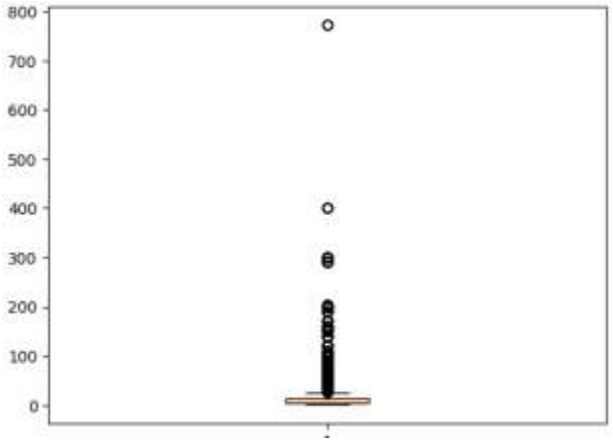
Modelo de predicción:

A continuación describiremos en modelo secuencial como opera la función de prediccion:

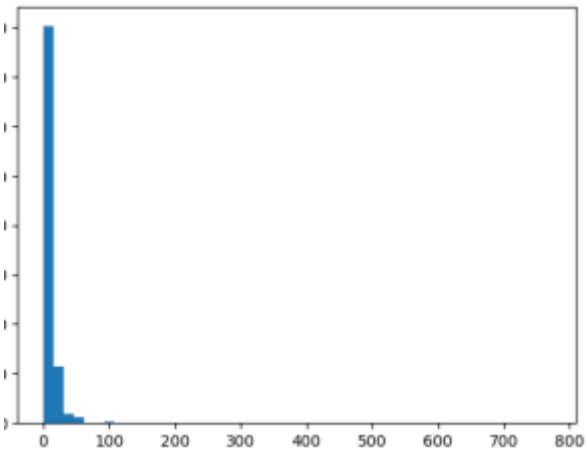
1. Se hace un requerimiento a la función, suministrando valores de “especificación”, “genero”, “lanzamiento temprano” (earlyaccess) y “sentimiento”. Todos estos parámetros en su formato origen, es decir, en cadenas de texto, con la excepción de “lanzamiento que es un parámetro booleano, se imputa 0 o 1.
2. La función consume estos parámetros y calcula el precio con dos modelos de regresión bienal.
    1. Regresión lineal sobre la población completa de precios (población general). Una población en la cual el precio tiene una desviación estándar de 13.44.
    2. Regresión lineal sobre la población de precio excluyendo a la población outlier (población efectiva). Una población con desviación estándar de 4.34. Lo cual ocurre en el 85% de los datos.
3. Una vez con los precios proyectados para ambas poblaciones del precio. Se aplica una distribución de probabilidad uniforme (puesto que la condición del precio “outlier” parece estar condicionado por el azar, no hay atributos en el dataset utilizado que explique tales cambios).
    - P(que se den precios outlier) = 0.15
      Es 0.15 puesto que esta es la probabilidad de caer en precios cuyo comportamiento es “outlier”.
    - P(que se den precios dentro de la población densamente agrupada) = 0.85.
4. Según sea la probabilidad obtenida, la función entoncs toma el valor de precio predicho de un modelo o del otro (pob gneral o de la pob efectiva).
5. Finalmente, entrega dicho precio proyectado pero con la desviación estadard que le corresponde a la población efectiva o general según el modelo de regresión seleccionado.
6. Finalmente, entrega un tercer y ultimo dato, “precio probable”. Este precio probable obedece a que aun cuando la variable precio es conceptualmente continua, existe algunos precios que concentran las ocurrencias. La idea es entonces luego de entregar el precio proyectado por el modelo y la desviación estándar que corresponde, entregar también, el precio frecuente mas cercano al proyectado dentro del rango de variabilidad.

Metodología de trabajo.

La variable objetivo “PRECIO”. Véanse las siguientes graficas:



En el diagrama de caja se pueda observar el impacto importante de los valores “outliers” en el precio, conjunto que en adelante le llamaremos “población outlier”. Población que como ya se menciono es parte del comportamiento legitimo de la variable.



En el diagrama de frecuencia se evidencia que la gran acumulación de valores se encuentra en un rango menor, aproximadamente en un precio menor o igual a 60$.

```
Promedio:  10.55
desviacion:  13.44
Cuartil 20%:  2.99
Cuartil 40%:  4.99
Mediana:  6.99
Cuartil 60%:  9.99
Cuartil 80%:  14.99
```

Cuando consideramos toda la población de valores de “precio” obtenemos estadísticos que nos permiten aseverar las siguientes conclusiones:

1. La dispersión es sumamente alta, teniendo en cuenta el rango de valores de la variable en el 80% de los datos. 13.44 y una media de 10.55, bastante alejada de su mediana
2. Es una distribución cuyo 20% de los datos genera un sesgo importante que impacta en la dispersión. Sin embargo, se trata de datos legítimos.

Metodología de trabajo.

La variable “PRECIO” y la “población efectiva”,

Siendo el precio, una variable sesgada por el impacto de una “población outlier”, es de interés analizar el comportamiento de la variable si extraemos a esta población.

Para ello nos definimos una meta de desviación estándar, es decir, nos proponemos a identificar el porcentaje de la población que permitiría obtener una desviación estándar de 5 o menor. Este fue el resultado:

```
Poblacion efectiva:  0.85
Precio de inflexion:  19.96
Desviacion Standar:  4.37
```

Es decir, cuando la desviación estándar se obtiene por debajo de 5, el porcentaje de la población (“precio”) implícita es de 85% y la cota de precio (punto de inflexión del precio) es de 19.96 $.

El calculo se corresponde con la observación y además 4.37 $ como desviación estándar es un parámetro de dispersión que concuerda con la lógica de negocio.

Es decir, estos datos nos permiten suponer lo siguiente:

1. Al restringir la “población outlier” la variable objetivo adquiere valores estadísticos mejor ajustados. Sin embargo, dicha “población outlier” es parte necesaria en el modelo.

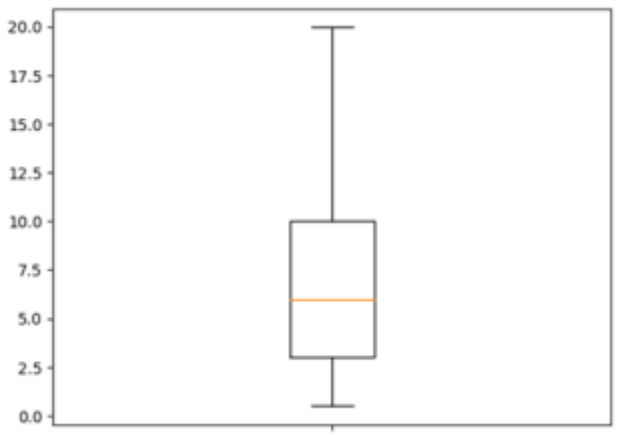
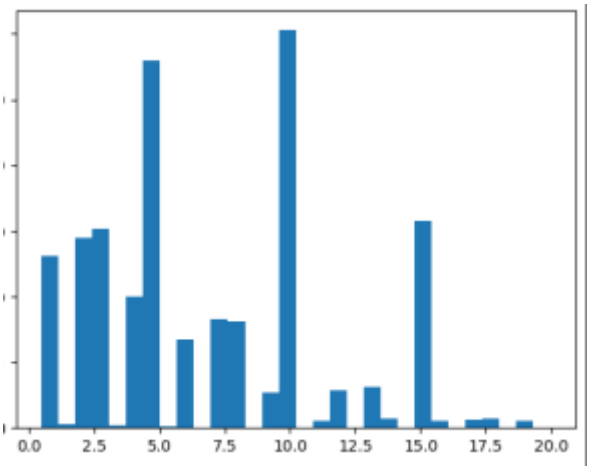


Diagrama de caja para la variable precio excluyendo a la “población outlier”. Le daremos el nombre a este conjunto de “población efectiva”.

Diagrama de frecuencia de la “población efectiva”. Es decir del 85% de los datos desde precio = 0 $ hasta precio = 19.96 $.



Cuando al diagrama aumentamos la cantidad de “bins” es posible ver la tendencia a un comportamiento normal, sin embargo es útil observar que, aunque la variable precio no es una variable categoría, sino mas bien continua, hay ciertos valore de precio que tienden a ser comunes. Se entiende que tal agrupación se corresponde a razonamientos de mercadotecnia.

Estadísticos para la “población efectiva”.

```
Promedio:  6.93
desviacion:  4.36
Cuartil 20%:  2.99
Cuartil 40%:  4.99
Mediana:  5.99
Cuartil 60%:  7.99
Cuartil 80%:  9.99
```

Estadísticos para la “población total”, es decir, efectiva + outlier..

```
Promedio:  10.55
desviacion:  13.44
Cuartil 20%:  2.99
Cuartil 40%:  4.99
Mediana:  6.99
Cuartil 60%:  9.99
Cuartil 80%:  14.99
```

precio_1	valor	cantidad	prob
19.99	39.99	13285	0.48
29.99	19.99	3884	0.14
24.99	29.99	2953	0.11
39.99	24.99	2384	0.09
49.99	61.99	1334	0.05

Observemos los valores de precio mas frecuentes en la –”población outlier”. Los tres precios mas frecuentes: 39.99 $, 19,99 $ y 29.99 $ respectivamente.

precio_1	valor	cantidad	prob
9.99	3.99	30191	0.19
4.99	1.59	27844	0.17
14.99	0.99	15665	0.10
2.99	9.99	14534	0.09
1.99	6.99	14453	0.09

Observemos los valores de precio mas frecuentes en la –”población efectiva”. Los tres precios mas frecuentes: 3.99 $, 4,99 $ y 14.99 $ respectivamente.

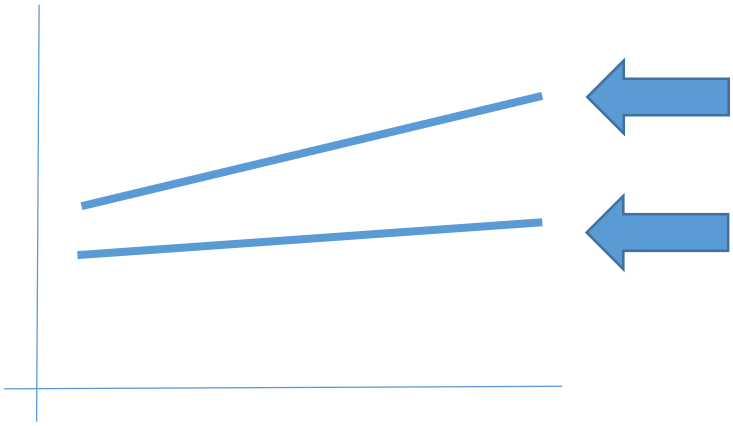
En función de las observación expuestas en la variable “precio”, la variable objetivo, se destacan los siguientes inconvenientes:

1. La población total de valores en la variable “precio” muestra un comportamiento extremadamente sesgado aunque legitimo. Sin embargo,
    1. La sección de los datos (85%) inherentes a la “población efectiva” si presenta un comportamiento normal.
    2. La sección de los datos (15%) inherentes a la “población outlier” también presenta una distribución normal.

Decisiones:

En vista de la particularidad ya expuesta se consideran los siguientes criterios:

1. Ambas secciones de la población de datos del precio serán consideradas, la “población efectiva” y la “población total”. Es decir. Se trabajara en lo siguiente considerando dos comportamiento posibles en el precio, un comportamiento que no considera a la “población outlier” y otro comportamiento que si la considera.



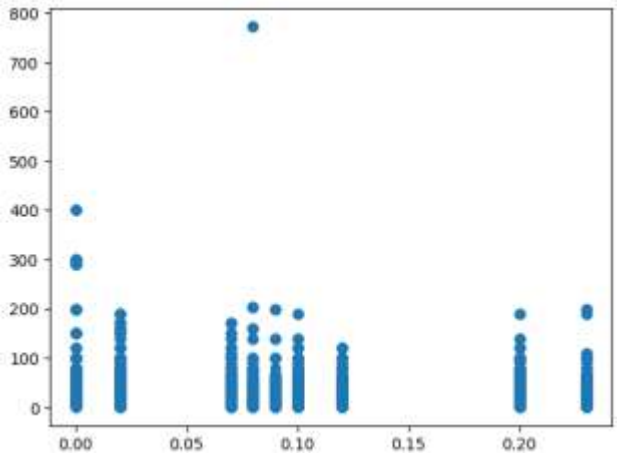
Existen dos comportamiento ligeramente distintos en el precio.

Pero cual comportamiento considerar a efectos del modelo de predicción ?, la respuesta es AMBOS.

En este punto, el objetivo será entonces verificar cual(es) de los atributos disponibles describe o explica mejor el cambio del precio (el punto de inflexión), en tal sentido ese será(an) el mejor atributo a utilizar.

Analicemos entonces la dispersión de cada uno de los atributos disponibles con la variable objetivo “precio”.

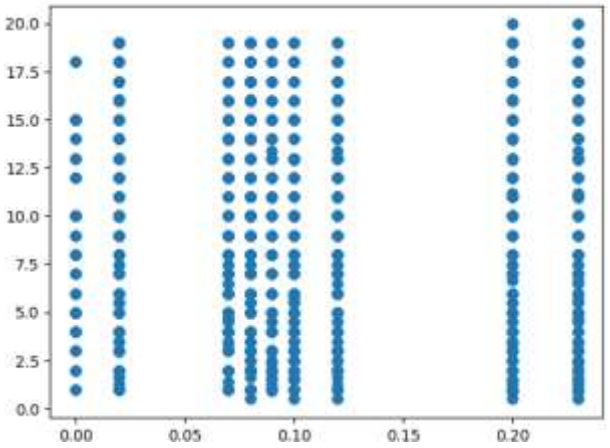
Genero vs precio
Población total



Correlación

	escala_genres	precio_1
escala_genres	1.000000	-0.229938
precio_1	-0.229938	1.000000

Genero vs precio
Población efectiva (sin outlier



Correlación

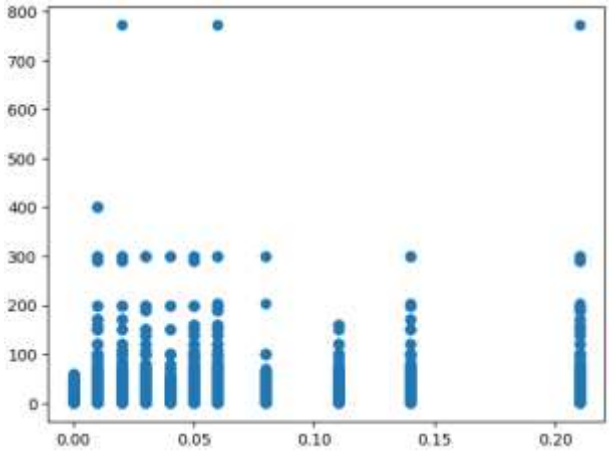
	escala_genres	precio_1
escala_genres	1.000000	-0.082279
precio_1	-0.082279	1.000000

Observaciones:

1. Aunque con respecto a la población total es claro que existe una tendencia y concentración de datos, cuando disminuimos a la población y observamos la dispersión con la población efectiva del precio, vemos que la “UNIFORMIDAD” aumenta, es decir, la dispersión sigue siendo alta.

Analicemos entonces la dispersión de cada uno de los atributos disponibles con la variable objetivo “precio”.

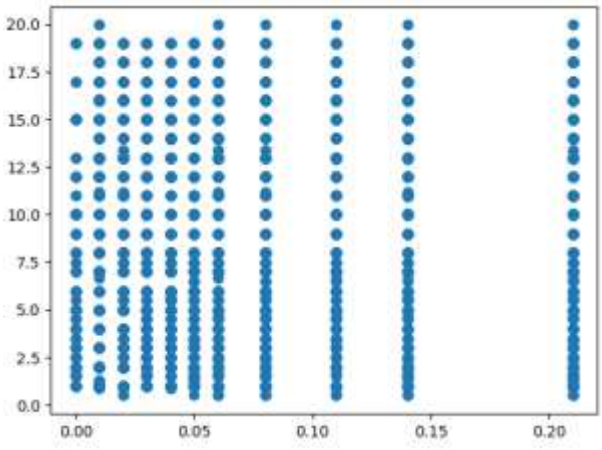
Especificaciones vs precio
Población total



Correlación

	escala_specs	precio_1
escala_specs	1.000000	-0.104843
precio_1	-0.104843	1.000000

Especificaciones vs precio
Población efectiva (sin outlier)



Correlación

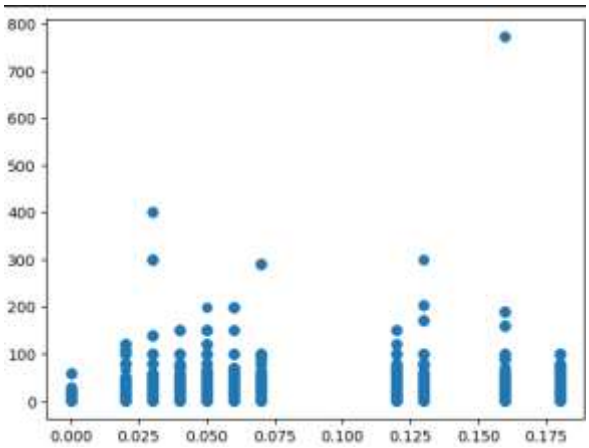
	escala_specs	precio_1
escala_specs	1.000000	-0.069913
precio_1	-0.069913	1.000000

Observaciones:

1. La observación es similar a la expuesta con el atributo generos. Tendencia a concentrarse hacia los precios bajos pero aun allí, hay alta dispersión.

Analicemos entonces la dispersión de cada uno de los atributos disponibles con la variable objetivo “precio”.

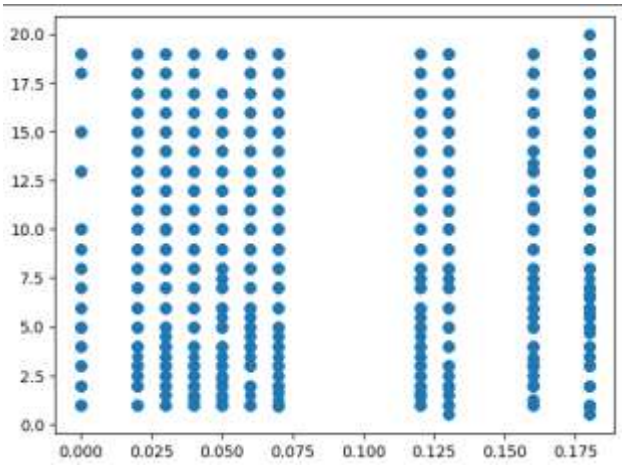
Sentimiento vs precio
Población total



Correlación

	escala_sentiment	precio_1
escala_sentiment	1.00000	0.05266
precio_1	0.05266	1.00000

Sentimiento vs precio
Población efectiva (sin outlier)



Correlación

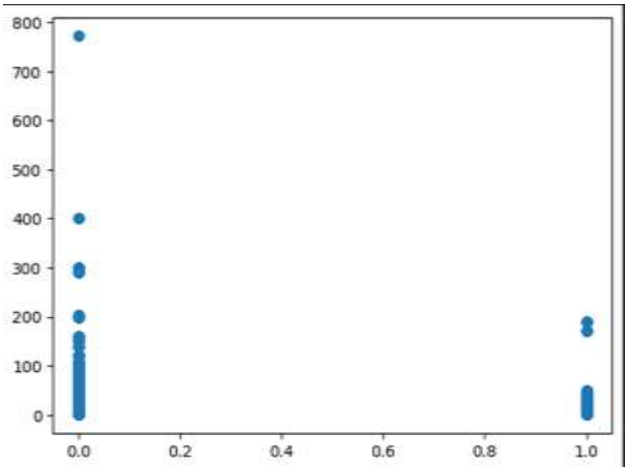
	escala_sentiment	precio_1
escala_sentiment	1.000000	0.122745
precio_1	0.122745	1.000000

Observaciones:

1. Igual a las anteriores.

Analicemos entonces la dispersión de cada uno de los atributos disponibles con la variable objetivo “precio”.

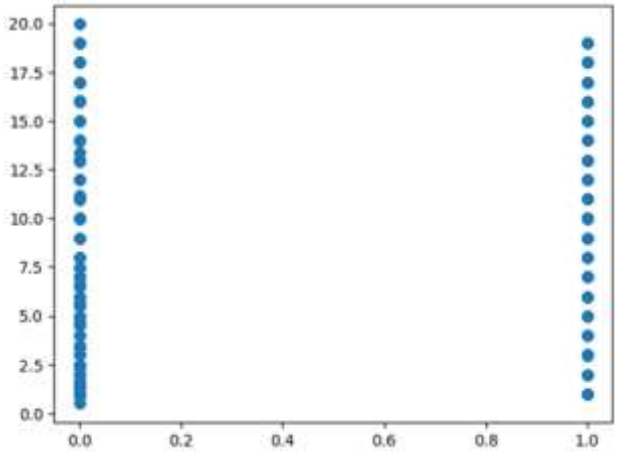
Lanzamiento vs precio
Población total



Correlación

	escala_early	precio_1
escala_early	1.000000	0.029848
precio_1	0.029848	1.000000

Lanzamiento vs precio
Población efectiva (sin outlier)



Correlación

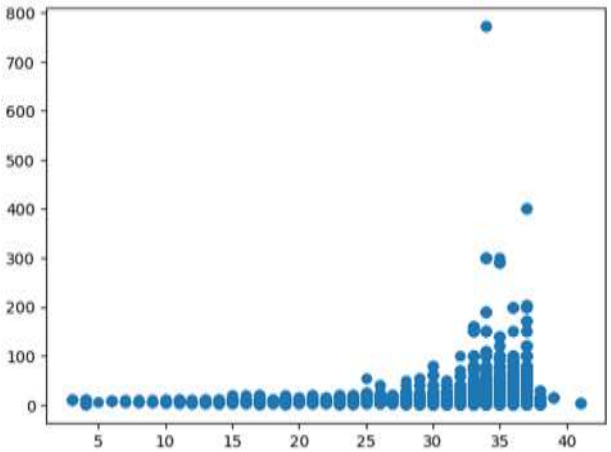
	escala_early	precio_1
escala_early	1.000000	0.171126
precio_1	0.171126	1.000000

Observaciones:

1. La principal observación aquí es sin lugar a duda es que en la medida que se implementan lanzamientos tempranos, el precio de lanzamiento tiende a ser menor. Sin embargo, dentro de la población efectiva, dicha conclusión se confirma, pero queda claro que se trata de una tendencia muy leve, pues el precio tiende a distribuirse con bastante uniformidad.

Analicemos entonces la dispersión de cada uno de los atributos disponibles con la variable objetivo “precio”.

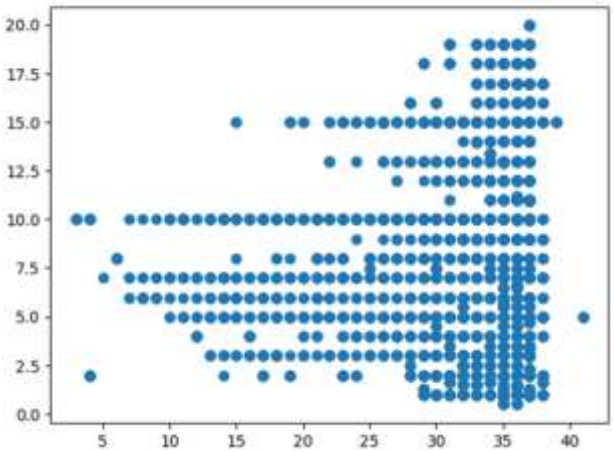
Año vs precio
Población total



Correlación

	año	precio_1
año	1.000000	0.018344
precio_1	0.018344	1.000000

Año vs precio
Población efectiva (sin outlier)



Correlación

	año	precio_1
año	1.000000	-0.030243
precio_1	-0.030243	1.000000

Observaciones:

1. Con respecto al año, queda claro que en la medida que transcurre el tiempo se acrecentar la producción de juegos, pero también desde el punto de vista de la observación, también mantiene la misma tendencia de comportamiento conjunto que el resto de los atributos con respecto al precio.

Matriz de correlación completa entre todas as variables consideradas según la población del conjunto precio.

Utilizando la población total del precio

	escala_genres	escala_specs	escala_sentiment	escala_early	año	precio_1
escala_genres	1.000000	0.091471	0.047994	-0.077073	-0.029648	-0.230710
escala_specs	0.091471	1.000000	0.010575	-0.069934	-0.054496	-0.104871
escala_sentiment	0.047994	0.010575	1.000000	0.019294	-0.080002	0.051868
escala_early	-0.077073	-0.069934	0.019294	1.000000	0.128040	0.029228
año	-0.029648	-0.054496	-0.080002	0.128040	1.000000	0.018344
precio_1	-0.230710	-0.104871	0.051868	0.029228	0.018344	1.000000

Utilizando la población efectiva del precio.

	escala_genres	escala_specs	escala_sentiment	escala_early	año	precio_1
escala_genres	1.000000	0.065481	0.038458	-0.101308	-0.022785	-0.081536
escala_specs	0.065481	1.000000	0.010600	-0.074150	-0.048328	-0.069551
escala_sentiment	0.038458	0.010600	1.000000	0.006488	-0.092055	0.121911
escala_early	-0.101308	-0.074150	0.006488	1.000000	0.128130	0.168980
año	-0.022785	-0.048328	-0.092055	0.128130	1.000000	-0.030243
precio_1	-0.081536	-0.069551	0.121911	0.168980	-0.030243	1.000000

En general, los valores de correlación son bajos. Evidenciando la existencia de alta dispersión. Sin embargo, dicha l dispersión existe en la variable precio en si misma.

CONCLUSIONES:

1. La variable objetivo es dispersa pero tiene tendencia.
2. El conjunto de atributos disponibles, no explican a la población outlier del precio. Pareciera que dicho comportamiento es inherente a otras condiciones no expuestas en el dataset fuente.
3. Todas las variables independientes tienen comportamiento homologo cuando se contrastan con la variable objetivo precio. Ello podría permitir una reducción importante en el numero de variables a considerar para el modelo.
4. El atributo “Publisher” se descarto dado que evidenciamos que tiene mucho mas de 1000 valores categóricos distintos cuyo impacto en el modelo podría mas que aportar información, diluir la información ya existente.
5. El atributo año se descartara dado que:
    1. Otras variables explican al objetivo de la misma manera.
    2. Es poco relevante generar dependencia del precio de un juego al tiempo, pues la idea es generar pronostico en función de las características del producto.
6. El precio en si mismo describe un patrón de frecuencia suficientemente normal en du población efectiva, y la existencia de una población outlier no esta suficientemente explicada por los atributos disponibles, al mismo tiempo, es una población legitima a considerar. En este sentido se implementara un criterio de probabilidad al resultado de la proyección del precio. Esto es:
    1. Dado los atributos insumo, se proyectara un precio (una predicción), tanto en un modelo que considere a la población total del precio, como en un modelo que solo considere la población efectiva del precio.
    2. Se calculara un valor aleatorio del 0 al 100. Es decir, se simulara los escenarios de población efectiva (85%) y población total (15%).
    3. Dependiendo del resultado aleatorio se entrega:
        1. Precio proyectado.
        2. Variabilidad del precio. Si el precio proyectado se corresponde al modelo de la población total, la variabilidad se corresponderá a la desviación calculada desv = 13.44. De lo contrario se entregara desviación calculada para la población efectiva desv = 4.37.
        3. Se entregara un precio probable. Este precio dependerá de la cercanía entre el precio proyectado y los precios mas frecuentes identificados en la población efectiva y en la población outlier.

MODELO ML REGRESION LINEAL

POR QUE REGRESION LINEAL?

1. El problema requiere predicción de una variable numérica cuantitativa continua.
2. La implementación de este modelo provee la posibilidad de reducir mucho el requerimiento de procesamiento, aspecto este que adquiere relevancia cuando dicho modelo pretende ser disponibilizado en una API.
3. Las variables regresoras muestran tendencia con la variable regresada.
4. Facilidad en la implementación y actualización.

CONSIDERACIONES IMPLICITAS EN EL MODELO

1. La dispersión es alta, sin embargo, se hace uso de características implícitas de la variable precio para ajustar el resultado del modelo (distribución de probabilidad entre la implementación en una población con outliar y sin outlier).

2. Se utilizaron las cuatro variables “specs”, “genres”, “sentiment” y “earlyaccess” dado que la combinación de las cuatro genero el menor error y factor de determinación.

Aplicado en la población total.

```
Error en datos de train: 171.3570883230106
Error en datos de test: 156.67723716292363
Coeficiente de determinacion:  0.06376457240280053
```

Aplicado a la población efectiva (sin población outlier)

```
Error en datos de train: 18.01028744673435
Error en datos de test: 17.976131500189236
Coeficiente de determinacion:  0.05024305725342737
```

MODELO ML REGRESION LINEAL

ALGUNAS CORRIDAS.

```
prediccion("Online Multi-Player","Video Production", 1,"Overwhelmingly Positive")
✓ 0.0s
"{'Precio_proyectado': 7.18, 'Variacion': 4.37, 'Precio_Probable': 4.99}"
```

```
prediccion("Cross-Platform Multiplayer","Video Production", 1,"Overwhelmingly Positive")
✓ 0.0s
"{'Precio_proyectado': 7.23, 'Variacion': 4.37, 'Precio_Probable': 4.99}"
```

```
prediccion("Cross-Platform Multiplayer","Audio Production", 0,"Overwhelmingly Positive")
✓ 0.0s
"{'Precio_proyectado': 7.33, 'Variacion': 4.37, 'Precio_Probable': 4.99}"
```