UNIVERSIDAD AMERICANA
Facultad de Ingeniería y Arquitectura

Inteligencia de negocios

Análisis de fuentes diversas

Estudiante: Carlos Diego Toruño Espinales

Docente:
Arlene Jeannette López

Fecha:

18 de Septiembre 2024

El dataset utilizado en este análisis contiene información sobre reseñas de juegos y su actividad, estos sirven como una forma de retroalimentación directa de los jugadores, ayudando a los desarrolladores a identificar áreas de mejora y a medir la satisfacción general de los usuarios. Además, las reseñas influyen en la decisión de compra de nuevos jugadores, ya que un alto número de reseñas positivas puede generar confianza y atraer a más usuarios.

Se comenzó el proyecto con un análisis exploratorio inicial del conjunto de datos para donde se muestra el tipo de datos de las columnas para darnos una idea de los tipos de datos con los que estamos trabajando.

```
# Exploración inicial del dataset
print("Primeras filas del dataset:")
print(data.head())
print("\nInformación del dataset:")
print(data.info())
print("\nEstadísticas descriptivas del dataset:")
print(data.describe())
```

```
6   negative_reviews        67571 non-null  int64
7   total_reviews           67571 non-null  int64
8   rating                  67571 non-null  float64
9   primary_genre           67561 non-null  object
10  store_genres            67514 non-null  object
11  publisher               67110 non-null  object
12  developer               67443 non-null  object
13  detected_technologies   60265 non-null  object
14  store_asset_mod_time    67275 non-null  object
15  review_percentage       47767 non-null  float64
16  players_right_now       67565 non-null  object
17  24_hour_peak            67565 non-null  object
18  all_time_peak           67571 non-null  int64
19  all_time_peak_date      67565 non-null  object
```

Se calcularon las estadísticas básicas de las variables numéricas tales como media, mediana y la desviación estándar. Primero se separa cada columna del conjunto de datos en dos categorías, las cuales son numéricas y no numéricas.

```
# Estadística básica del dataset
numericas = data.select_dtypes(include=['float64', 'int64'])
no_numericas = data.select_dtypes(exclude=['float64', 'int64'])
```

Después de separar cada columna, procedemos a iterar las columnas numéricas en un ciclo el cual va calculando la media, mediana y moda de cada variables.

```
# Media, mediana y desviación estándar para variables numéricas
print("\nAnálisis descriptivo de variables numéricas:")
for col in numericas.columns:
    print(f"\nColumna: {col}")
    print(f"Media: {numericas[col].mean()}")
    print(f"Mediana: {numericas[col].median()}")
    print(f"Desviación Estándar: {numericas[col].std()}")
```

```
Columna: peak_players
Media: 952.8673691376478
Mediana: 7.0
Desviación Estándar: 19790.925578000897

Columna: positive_reviews
Media: 1273.527326811798
Mediana: 19.0
Desviación Estándar: 29551.634359398166

Columna: negative_reviews
Media: 216.8938005949298
Mediana: 6.0
Desviación Estándar: 5434.959528294874
```

Para las variables que no son numéricas hemos aplicado la moda, esto es para poder ver el valor que más aparece con frecuencia en las columnas.

```
# Moda para variables no numéricas
print("\nAnálisis descriptivo de variables no numéricas:")
for col in no_numericas.columns:
    print(f"\nColumna: {col}")
    print(f"Moda: {no_numericas[col].mode()[0]}")
```

De igual manera se iteran estas columnas y se imprime la salida de estos:

```
Columna: primary_genre
Moda: Indie (23)

Columna: store_genres
Moda: Casual (4), Indie (23)

Columna: publisher
Moda: Big Fish Games
```

Luego de analizar los datos del conjunto, utilizaremos la columna **positive review**, la cual dividiremos en conjuntos para poder realizar una tabla de frecuencia y ver la repetición de cada grupo:

```
# Agrupar los datos de 'positive_reviews' en rangos
bins = [0, 100, 500, 1000, 5000, 10000, 50000, 100000, data['positive_reviews'].max()]
labels = ['0-100', '101-500', '501-1000', '1001-5000', '5001-10000', '10001-50000', '50001-100000', '100001+']
data['positive_reviews_group'] = pd.cut(data['positive_reviews'], bins=bins, labels=labels, include_lowest=True)
```

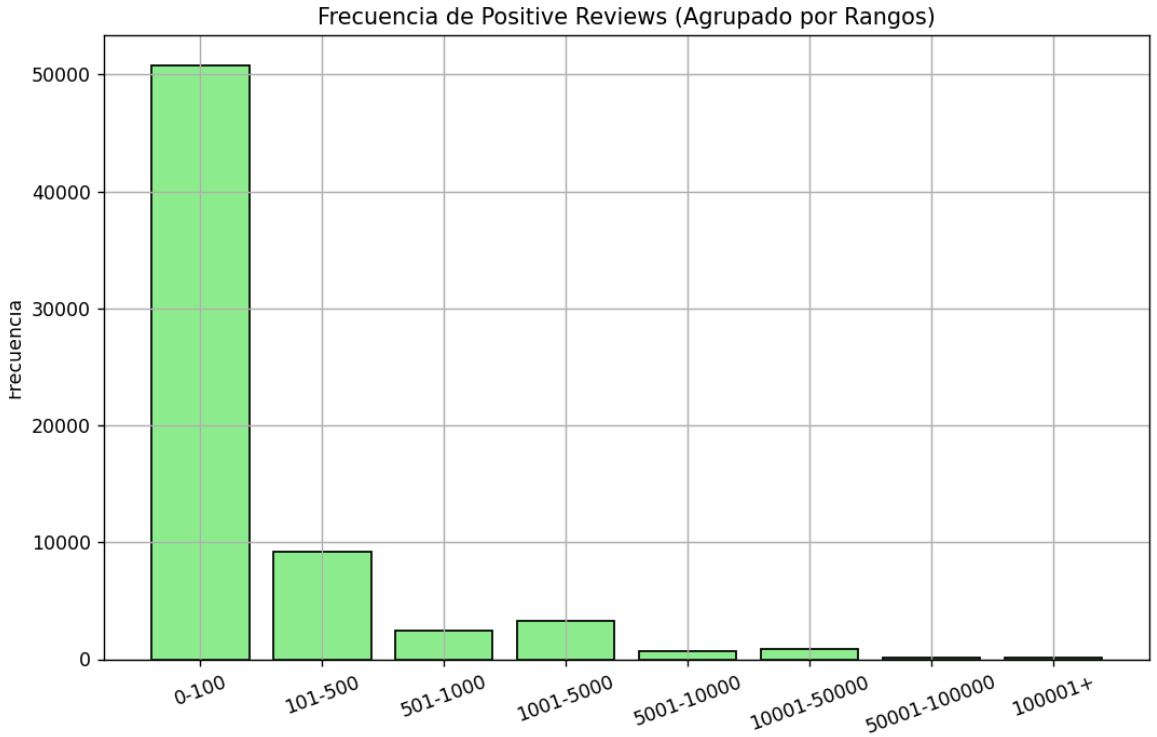
```
Tabla de Frecuencia de Positive Reviews

positive_reviews_group
0-100           50819
101-500          9155
501-1000         2439
1001-5000        3304
5001-10000        739
10001-50000       840
50001-100000      154
100001+           121
```

Luego de realizar las estadísticas básicas, se realizó un gráfico de barras para mostrar la frecuencia de la columna.

```
# Gráfico de barras para los grupos de 'positive_reviews'
plt.figure(figsize=(10, 6))
plt.bar(tabla_frecuencia_positive_review_grouped.index.astype(str),
        tabla_frecuencia_positive_review_grouped.values, color='lightgreen', edgecolor='black')
plt.title('Frecuencia de Positive Reviews (Agrupado por Rangos)')
plt.xlabel('Rango de Positive Reviews')
plt.ylabel('Frecuencia')
plt.xticks(rotation=20)
plt.grid(True)
plt.show()
```

En este código se utiliza la librería matplotlib para crear un gráfico de barras y mostrar la distribución de las reseñas positivas agrupadas por grupos.



El gráfico de barras muestra la frecuencia de las reseñas positivas agrupadas en conjuntos. La mayoría de juegos tiene entre 0 y 100 reseñas positivas con una frecuencia de más de 50,000 jugadores en ese rango. A medida que el número de reseñas positivas aumenta, la frecuencia disminuye de manera notable.

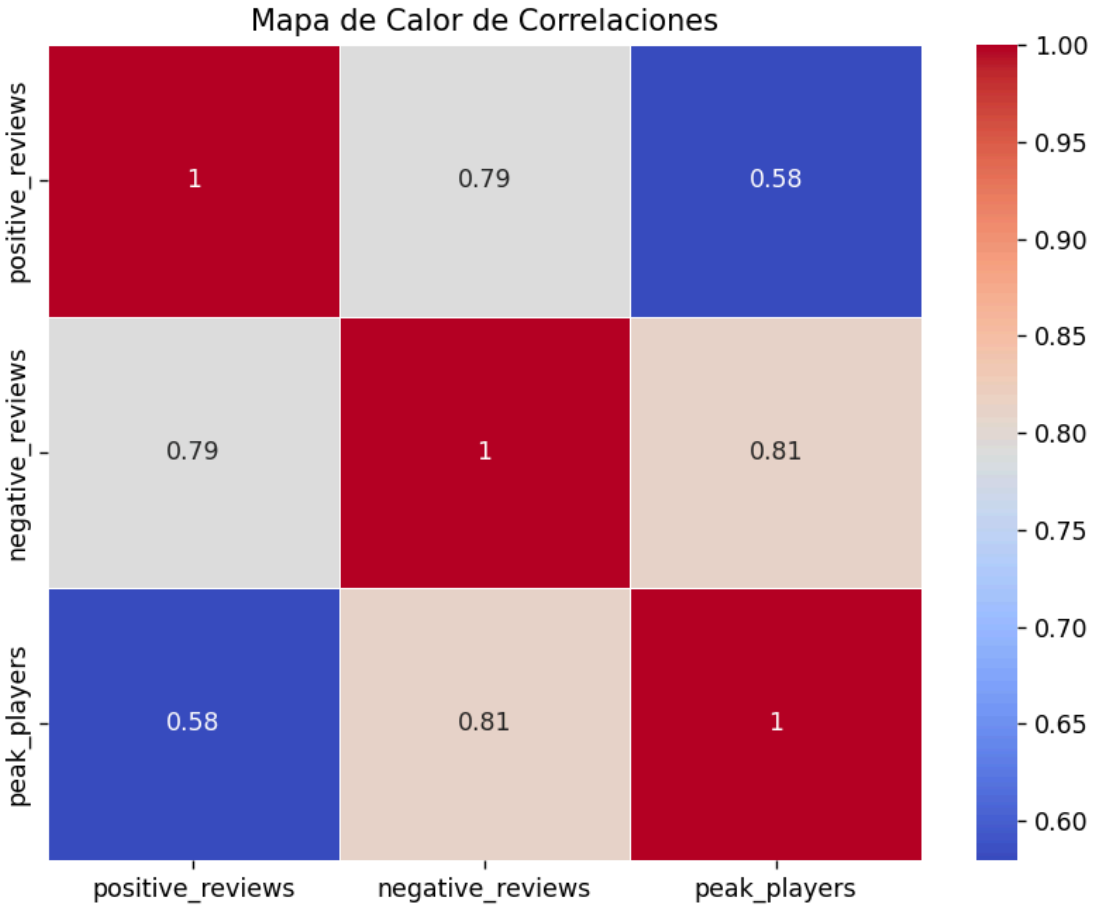
- **Análisis y descripción de dos patrones de variables**

Para el análisis de estos patrones en las variables se creó un mapa de calor para comprender de mejor manera los hallazgos. El siguiente mapa de calor muestra las relaciones entre 3 variables claves del conjunto de datos, cada valor dentro de una celda representa los coeficientes de relación.

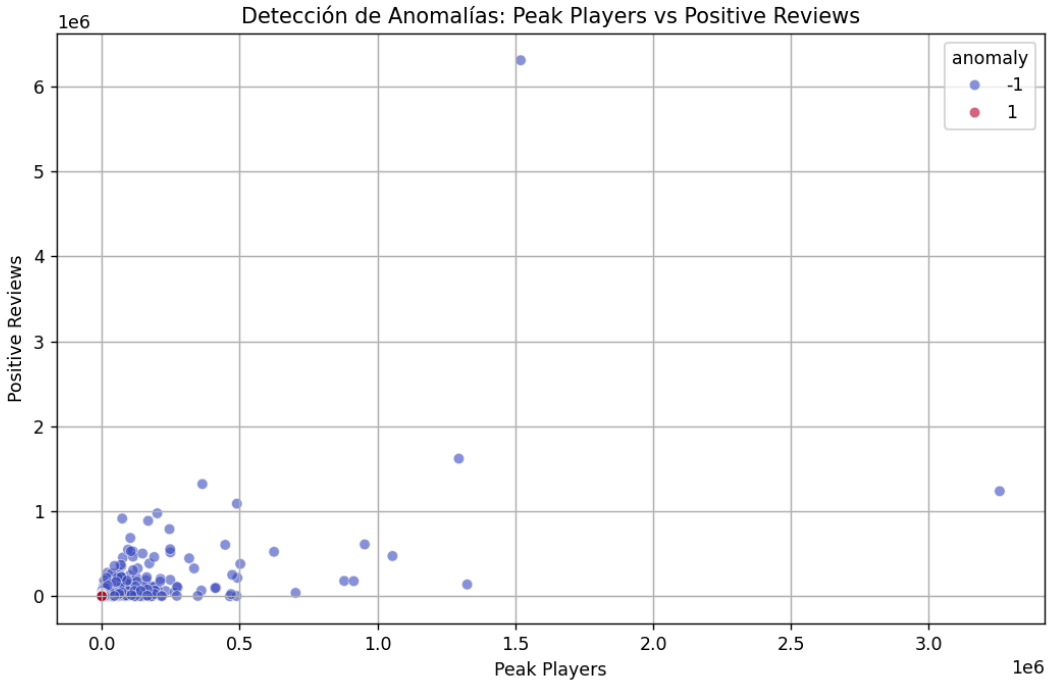
Positive Reviews y Negatives Reviews: Indica que los juegos con muchas reseñas positivas tienden también a recibir un número alto de reseñas negativas.

Positive Reviews y Peak Players: Indica que hay una relación moderada entre el número de jugadores máximos y las reseñas positivas. Esto sugiere que los juegos con más jugadores suelen recibir más reseñas positivas, pero no es una relación tan fuerte como entre las reseñas positivas y negativas.

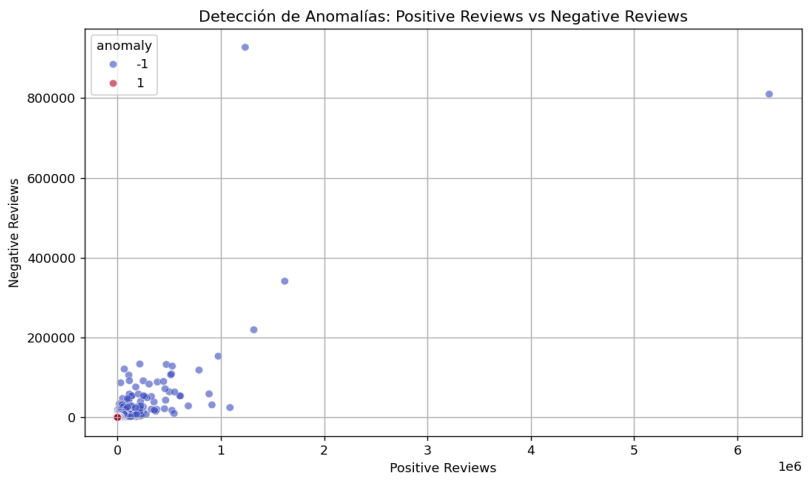
Negative Review y Peak Players: Muestran una relación relativamente fuerte. Esto sugiere que los juegos con un mayor número de jugadores también tienden a tener más reseñas negativas, posiblemente debido a que la mayor visibilidad puede generar más críticas.



Otro gráfico generado fue el diagrama de detección de anomalías, el cual es es una visualización gráfica que permite identificar datos atípicos o inusuales en un conjunto de datos. Estos datos, conocidos como anomalías u outliers, son puntos que se desvían significativamente del patrón general de los datos y pueden ser indicadores de comportamientos inesperados o errores en el sistema.

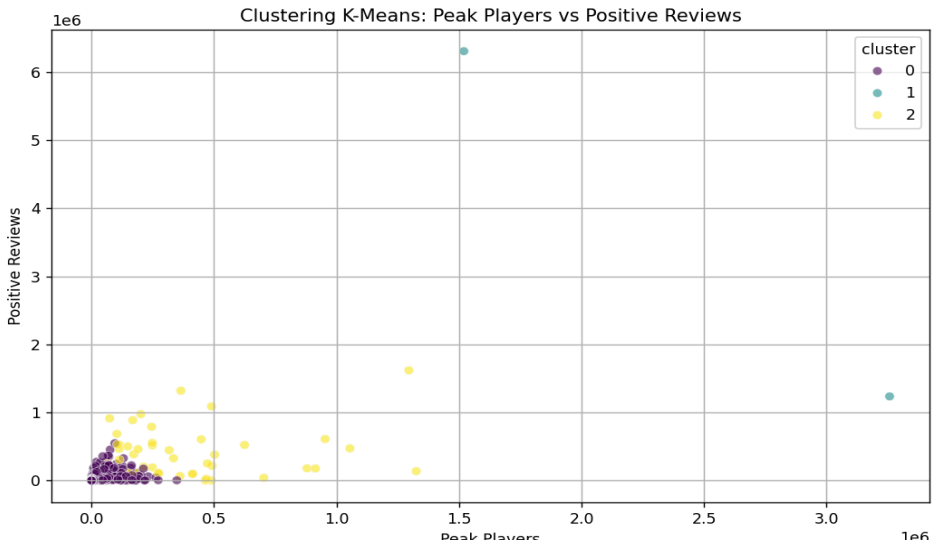


Este gráfico muestra la dispersión y detección de anomalías entre la cantidad máxima de jugadores y los reviews positivos de los juegos, en este gráfico los puntos azules representan los datos considerados normales, y los puntos rojos representan anomalías.
La mayoría de los juegos se concentran en la esquina inferior izquierda, lo que indica que la mayoría tienen pocos jugadores y reseñas positivas. Sin embargo, algunos puntos dispersos reflejan juegos con valores extraordinariamente altos en una o ambas variables, indicando títulos que destacan del resto del conjunto.
Otra relación relevante es la de **positive review y negative review.** El análisis revela que hay juegos con una alta cantidad de reseñas, lo cual es raro, lo que indica una alta interacción de los usuarios. Algunos de estos juegos, que tienen un balance inusual entre reseñas positivas y negativas, pueden reflejar títulos que generan fuertes opiniones tanto favorables como desfavorables.



- **Técnicas para descubrir información relevante**

Para la detección de anomalías utilizamos la técnica **clustering**, la cual es una técnica de machine learning no supervisado que se utiliza para agrupar datos en subconjuntos llamados clusters. Cada cluster contiene datos que son más similares entre sí que con los de otros clusters. El objetivo es encontrar patrones o estructuras ocultas en los datos sin que haya etiquetas predefinidas.



En el gráfico del clustering de Peak Players y Positive Reviews, este se agrupa en 3 conjuntos:

Color morado: Agrupa juegos con pocos jugadores máximos y pocas reseñas positivas. La mayoría de los juegos pertenecen a este cluster, lo que sugiere que estos juegos no alcanzan un alto nivel de popularidad.

Color Amarillo: Representa juegos que tienen una mayor dispersión en cuanto a jugadores máximos y reseñas positivas, incluyendo algunos títulos con valores extremadamente altos en ambas variables.

Color Cyan: Agrupa juegos con un número elevado de jugadores máximos y un número significativo de reseñas positivas. Estos juegos se destacan por tener un nivel de popularidad superior al promedio.

Conclusiones

Distribución de reseñas y jugadores: La mayoría de los juegos en el dataset presentan un bajo número de reseñas, tanto positivas como negativas, y cuentan con una cantidad reducida de jugadores máximos simultáneos. Esto indica que una gran parte de los títulos en Steam no alcanzan una alta popularidad, lo que es común en plataformas con una amplia oferta de productos, donde solo unos pocos títulos logran destacarse.

Los análisis de correlación muestran una relación positiva entre las reseñas positivas y negativas, así como entre el número de jugadores máximos y las reseñas. Esto sugiere que los juegos con más jugadores tienden a generar más reseñas, y que la popularidad de un título en términos de jugadores se refleja en la cantidad de interacciones de los usuarios