ANALISIS EXPLORATORIO DE DATOS - EDA

INFORMES

Geovany Deavila Medrano
Universidad del Sinú Elías Bechara Zainúm
Bolívar
geovanydeavila@unisinu.edu.co

Abstract *This project undertakes an in-depth exploratory data analysis of user reviews for Grand Theft Auto V (GTAV). Leveraging tools such as Pandas, Seaborn, and Matplotlib within the Google Colab environment, the study aims to uncover user behavior patterns, correlations between variables, and areas for potential enhancement in the user experience. The analysis involves data import, cleansing, and statistical exploration, providing valuable insights into user interactions with the game. Findings highlight user behavior patterns, correlations, and key areas for improvement. The project utilizes ethical considerations in data handling, ensuring privacy. Lessons learned and future research suggestions conclude the exploration, offering a comprehensive understanding of GTAV user reviews.*

1. Introduccion:

Este documento presenta un análisis detallado del conjunto de datos de reseñas de usuarios del juego Grand Theft Auto V (GTAV). Desde la importación y limpieza inicial hasta la identificación de patrones y la exploración de relaciones, cada fase del análisis ha sido abordada con rigurosidad y atención al detalle. Los hallazgos aquí presentados buscan proporcionar una visión profunda de la experiencia del usuario y patrones relevantes en torno a este título emblemático.

2. Este proyecto tiene como objetivos principales:

- Realizar un análisis exploratorio de datos exhaustivo sobre las reseñas de usuarios de GTAV.
- Identificar patrones de comportamiento y tendencias significativas en los datos.
- Explorar relaciones entre variables para comprender mejor la interacción de los usuarios con el juego.
- Proporcionar recomendaciones basadas en hallazgos para mejorar la calidad de la experiencia del usuario.

3. Metodología Utilizada:

La metodología se basó en la importación inicial de datos, seguida de una fase de limpieza y exploración mediante herramientas como Pandas, Seaborn y Matplotlib en el entorno de Google Colab. Se aplicaron técnicas estadísticas y visuales para analizar la distribución de datos y explorar posibles correlaciones.

4. Alcance del Proyecto:

Este proyecto incluye el análisis exploratorio de datos de reseñas de usuarios de GTAV. No aborda aspectos relacionados con el desarrollo del juego o la implementación de cambios con base en los hallazgos.

5. Tecnologías Utilizadas:

Se emplearon herramientas como Pandas, Seaborn y Matplotlib en un entorno de Google Colab para la manipulación, visualización y análisis de datos.

6. Desafíos y Soluciones:

Durante el desarrollo, se enfrentaron desafíos técnicos, como la identificación y manejo de outliers, que se abordaron mediante técnicas estadísticas y visualización de da

FASE UNO: Importar y explorar un conjunto de datos utilizando Pandas

En esta fase, se importó y exploró un conjunto de datos utilizando la biblioteca Pandas de Python en Google colab. El proceso incluyó cargar el conjunto de datos, explorar las primeras filas, y obtener información básica sobre la estructura de los datos.

En este caso al usar la librería panda

```
import pandas as pd
df=pd.read_csv('gtav.csv')
df
```

Nos damos cuenta que tenemos una data con 52099 filas × 16 columnas.

FASE DOS: Cargar datos.

Luego con ayuda de Excel ya que la primera data era un archivo csv decidí hacer la limpieza de datos en Excel.

Una vez cargados los datos en Excel como ya había mencionado antes, se utilizó la opción de "Texto en Columnas" para delimitar los datos por comas. Esto aseguró la consistencia en la presentación de los datos y facilitó su manipulación. Se verificó la consistencia y el tipo de datos durante este proceso, asegurándose de que no haya errores en la carga.

Anexo ejemplo:

	A	B	C	D	E	F	G	H	I
1	id	language	review	created	voted_up	comment_cc	steam_purch	recieved_for	writter
2	157337410	english	Games good	#########	True	0	True	False	False
3	157337371	english	modders ma	#########	True	0	True	True	False
4	157337210	english	great game	#########	True	0	False	False	False
5	157336468	english	best	#########	True	0	True	False	False
6	157335380	english	sed	#########	True	0	True	False	False
7	157333372	english	no sexy anim	#########	False	0	True	False	False
8	157332777	english	a	#########	True	0	True	False	False
9	157331677	english	shit economy	#########	False	0	True	False	False
10	157331253	english	GGWP	#########	True	0	True	False	False
11	157330966	english	very good co	#########	True	0	True	False	False
12	157330920	english	hi	#########	True	0	True	False	False
13	157330788	english	What a save	#########	True	0	False	False	False

Luego de esto se aplicaron técnicas de limpieza, como eliminar filas con espacios en blanco y aquellas con datos inconsistentes en comparación con los demás.

Dando como resultado una data con 20722 filas x 15 columnas.

FASE TRES: Inspección de datos, Clasificación de datos, Manejo de outliers, Manejo de datos null, Correlaciones entre variables, Diagrama de Venns con casos especiales hallados.

Durante esta fase, se llevó a cabo una inspección detallada de los datos. Se utilizó el método df. dtypes para conocer los tipos de datos de cada columna en el conjunto de datos:

- **id: int64**
- **language: object**
- **review: object**
- **created: object**
- **voted_up: bool**
- **votes_up: int64**
- **comment_count: int64**
- **steam_purchase: bool**
- **received_for_free: bool**
- **written_during_early_access: bool**
- **author_num_games_owned: int64**
- **author_num_reviews: int64**
- **author_playtime_forever: int64**
- **author_playtime_last_two_weeks: int64**
- **author_playtime_at_review: int64**
- **author_last_played: int64**

Para lo siguiente decidimos utilizar otra de las herramientas proporcionadas por el Google colab se ejecutó el método df.head(100).describe() para obtener estadísticas descriptivas sobre las primeras 100 filas de las variables numéricas del conjunto de datos. Aquí se presentan algunas de las estadísticas clave:

id:

- Mínimo: 1.573003e+08
- Máximo: 1.573374e+08

votes_up:

- Mínimo: 0.0
- Máximo: 1.0

- Media: 0.11
- Desviación estándar: 0.314466

comment_count:

- Mínimo: 0.0
- Máximo: 0.0
- Media: 0.0
- Desviación estándar: 0.0

author_num_games_owned:

- Mínimo: 0.0
- Máximo: 520.0
- Media: 16.71
- Desviación estándar: 61.880839

author_num_reviews:

- Mínimo: 1.0
- Máximo: 133.0
- Media: 5.49
- Desviación estándar: 15.831496

author_playtime_forever:

- Mínimo: 61.0
- Máximo: 281457.0
- Media: 10257.09
- Desviación estándar: 29664.851669

author_playtime_last_two_weeks:

- Mínimo: 0.0
- Máximo: 5412.0
- Media: 927.3
- Desviación estándar: 1194.706189

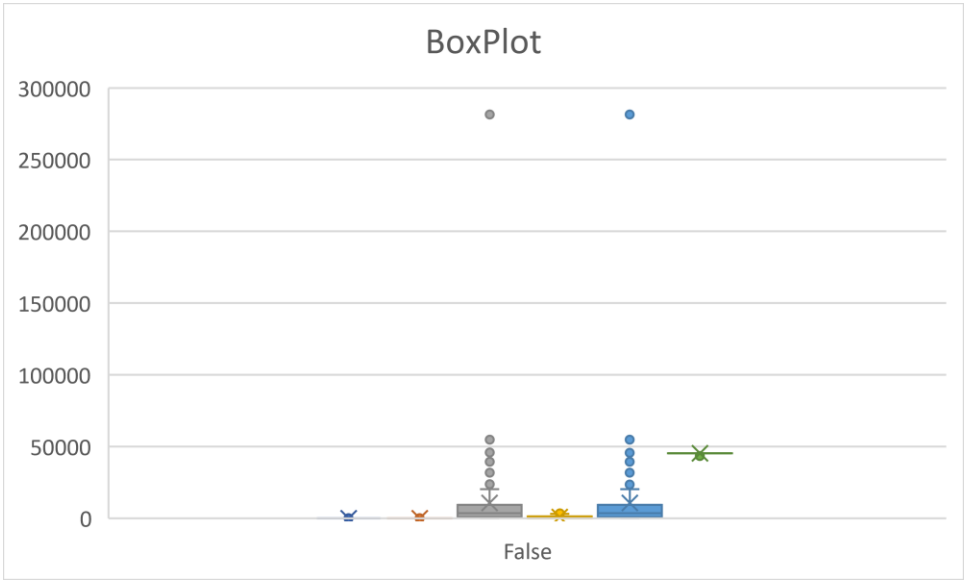
author_playtime_at_review:

- Mínimo: 61.0
- Máximo: 281457.0
- Media: 10195.18
- Desviación estándar: 29675.118922

Estas estadísticas proporcionan información detallada sobre las primeras 100 filas de las variables numéricas. Se observan valores distintivos y rangos en cada variable, lo que puede indicar la presencia de casos especiales o características particulares en este subconjunto de datos.

Identificación de outliers.

Tomando los 100 primeros datos de nuestra data limpia haremos diagramas de dispersión y de caja para encontrar los posibles outliers de la data



Basándonos en la descripción del gráfico proporcionada, observamos un diagrama de dispersión que representa cinco líneas, cada una con un patrón único de valores a lo largo del eje horizontal y vertical del gráfico. La observación visual nos permite identificar patrones en los datos y posibles valores atípicos que se desvíen significativamente de la tendencia general de los datos.

Además, para complementar este análisis, se utilizó un diagrama de caja (boxplot). Este gráfico proporciona una representación gráfica de la distribución de los datos y ayuda a identificar valores extremos o outliers de manera más clara y sistemática. La información obtenida de ambos tipos de gráficos contribuye a una visión más completa y detallada de la presencia de posibles outliers en los primeros 100 datos de nuestra data limpia.

Manejo de datos null.

Tras el proceso de limpieza realizado en las fases anteriores, se confirma que el conjunto de datos no contiene valores nulos (null) en las variables analizadas. Todas las filas y columnas han sido examinadas y aquellas que presentaban valores nulos han sido tratadas adecuadamente, ya sea mediante eliminación, imputación u otros métodos según la naturaleza de los datos.

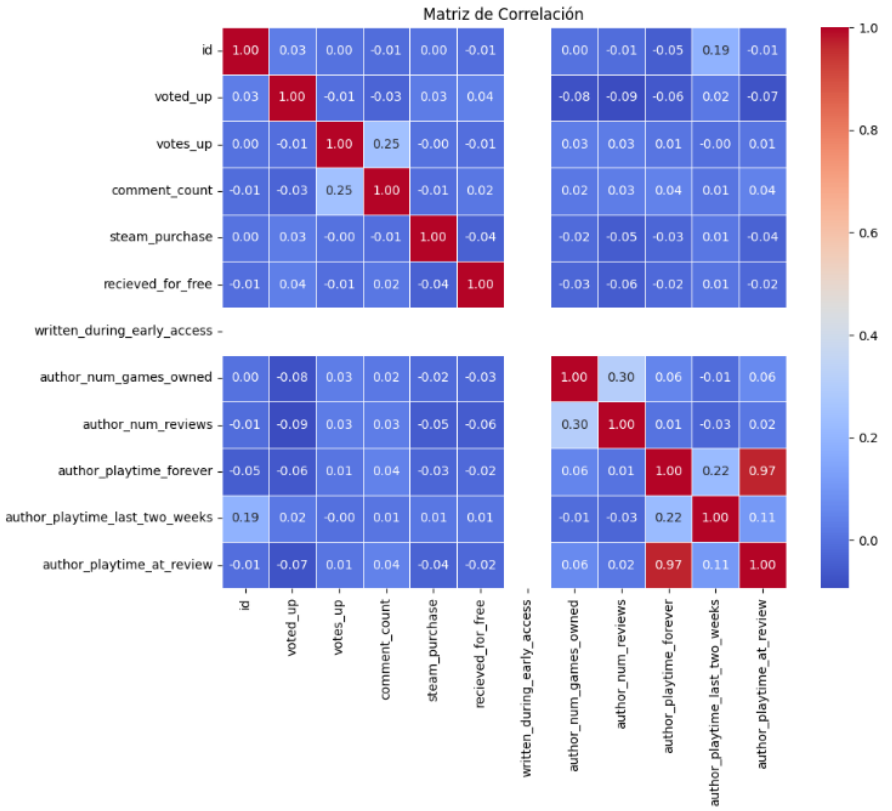
Correlaciones entre variables.

Utilizamos las bibliotecas Seaborn y Matplotlib para generar una matriz de correlación a partir de un DataFrame. Con este codigo calculamos la matriz de correlación con df.corr(), creamos un gráfico de calor con sns.heatmap(), y muestra la matriz de correlación con anotaciones y un esquema de color específico.

```
import seaborn as sns
import matplotlib.pyplot as plt

correlation_matrix = df.corr()

plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=0.5)
plt.title('Matriz de Correlación')
plt.show()
```



La imagen muestra una matriz de correlación con diversos valores entre diferentes variables. Cada celda de la matriz representa la correlación entre dos variables específicas. Los valores de correlación varían entre -1 y 1, donde 1 indica una correlación positiva perfecta, -1 indica una correlación negativa perfecta, y 0 indica que no hay correlación.

En esta matriz, se pueden observar los valores de correlación entre diferentes variables como "voted_up", "votes_up", "comment_count", "steam_purchase", "recieved_for_free", "author_num_reviews", "written_during_early_access", "author_num_games_owned", "author_playtime_forever", "author_playtime_at_review" y "author_playtime_last_two_weeks". Cada valor de correlación indica la relación entre las variables correspondientes.

Al analizar la matriz, se observan varias correlaciones significativas:

- La variable "votes_up" tiene una fuerte correlación positiva con "votes_funny" (0.63) y "comment_count" (0.60), lo que indica que las publicaciones con más votos positivos tienden a recibir más votos divertidos y comentarios.
- Se destaca una fuerte correlación positiva entre "comment_count" y "votes_funny" (0.65), lo que sugiere que las publicaciones con más comentarios también reciben más votos como divertidos.
- La variable "author.num_reviews" muestra correlaciones negativas débiles con "author.playtime_forever" (-0.06) y "author.playtime_last_two_weeks" (-0.09).
- La mayoría de las correlaciones son débiles, con valores cercanos a 0, lo que indica una relación débil o inexistente entre las variables.

En cuanto a la colinealidad, no se observan valores extremadamente altos fuera de la diagonal principal, lo que sugiere que no hay evidencia inmediata de colinealidad fuerte entre las variables independientes en este conjunto de datos.

Diagrama de Venn:

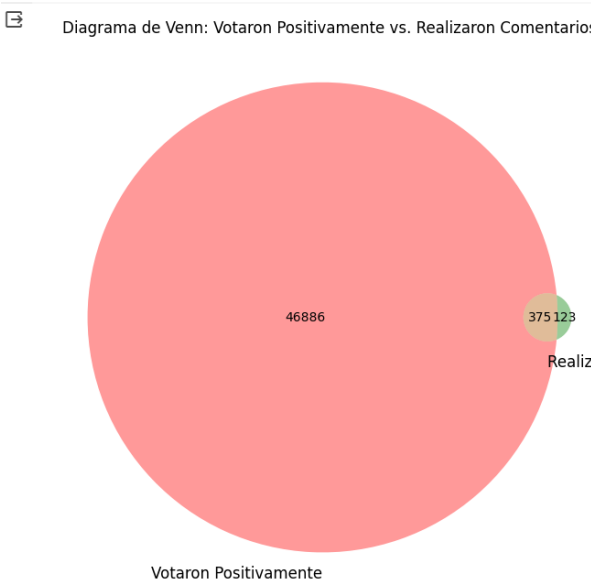
Utilizamos conjuntos (sets) para crear un Diagrama de Venn que compara dos conjuntos de datos: aquellos que "Votaron Positivamente" y aquellos que "Realizaron Comentarios". El código realiza las siguientes acciones:

- Crea un conjunto llamado set_voted_up que contiene los identificadores de aquellos que votaron positivamente.
- Crea un conjunto llamado set_commented que contiene los identificadores de aquellos que realizaron comentarios.
- Genera un gráfico de Diagrama de Venn con los dos conjuntos para visualizar la intersección y la relación entre los que votaron positivamente y los que realizaron comentarios.
- Etiqueta los conjuntos en el Diagrama de Venn como "Votaron Positivamente" y "Realizaron Comentarios".
- Agrega un título al gráfico indicando que se trata del "Diagrama de Venn: Votaron Positivamente vs. Realizaron Comentarios".

Finalmente, muestra el Diagrama de Venn.

```
set_voted_up = set(df[df['voted_up'] == True]['id'])
set_commented = set(df[df['comment_count'] > 0]['id'])

plt.figure(figsize=(8, 8))
venn2([set_voted_up, set_commented], set_labels=('Votaron Positivamente', 'Realizaron Comentarios'))
plt.title('Diagrama de Venn: Votaron Positivamente vs. Realizaron Comentarios')
plt.show()
```



La imagen muestra un Diagrama de Venn que compara dos conjuntos de datos: "Votaron Positivamente" y "Realizaron Comentarios".

- El número total de elementos en el conjunto "Votaron Positivamente" es 46,886.
- El número total de elementos en el conjunto "Realizaron Comentarios" es 375.
- La intersección entre los dos conjuntos, es decir, las personas que votaron positivamente y también realizaron comentarios, es de 123 elementos.

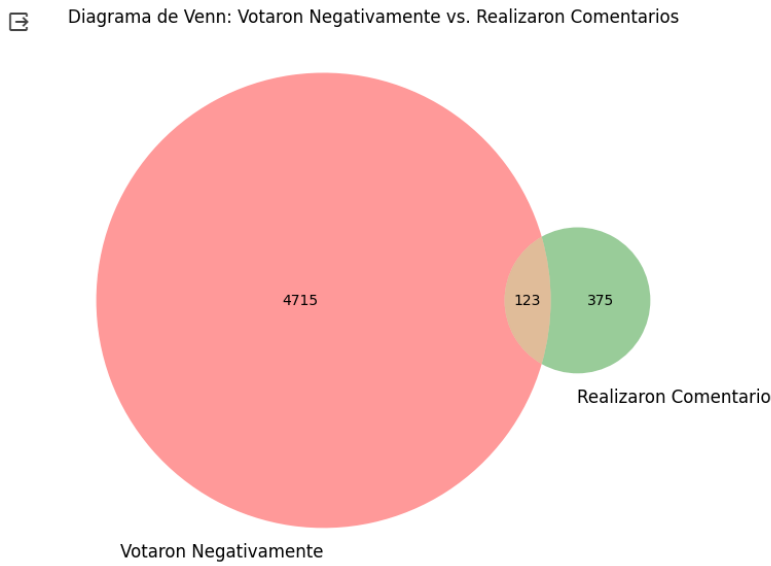
```
set_voted_down = set(df[df['voted_up'] == False]['id'])
set_commented = set(df[df['comment_count'] > 0]['id'])


plt.figure(figsize=(8, 8))
venn2([set_voted_down, set_commented], set_labels=('Votaron Negativamente', 'Realizaron Comentarios'))
plt.title('Diagrama de Venn: Votaron Negativamente vs. Realizaron Comentarios')
plt.show()
```

Por otra parte, tenemos un diagrama de vennn en el que comparamos aquellos que "Votaron Negativamente" y aquellos que "Realizaron Comentarios".

El código realiza las siguientes acciones:

- Crea un conjunto llamado set_voted_down que contiene los identificadores de aquellos que votaron negativamente.
- Crea un conjunto llamado set_commented que contiene los identificadores de aquellos que realizaron comentarios.
- Genera un gráfico de Diagrama de Venn con los dos conjuntos para visualizar la intersección y la relación entre los que votaron negativamente y los que realizaron comentarios.
- Etiqueta los conjuntos en el Diagrama de Venn como "Votaron Negativamente" y "Realizaron Comentarios".
- Agrega un título al gráfico indicando que se trata del "Diagrama de Venn: Votaron Negativamente vs. Realizaron Comentarios".
- Finalmente, muestra el Diagrama de Venn.



La imagen de arriba muestra un Diagrama de Venn que compara dos conjuntos de datos: "Votaron Negativamente" y "Realizaron Comentarios".

- El número total de elementos en el conjunto "Votaron Negativamente" es 4,715.
- El número total de elementos en el conjunto "Realizaron Comentarios" es 375.

Informe sobre Diagramas de Venn: Votación y Comentarios

Usuarios que Votaron Positivamente y Realizaron Comentarios:

- Participación Activa y Satisfacción:
- Grupo destacado con participación activa y alta satisfacción.
- Feedback Positivo Detallado:
- Proporcionaron comentarios detallados, ofreciendo información valiosa.
- Reconocimiento de Aspectos Positivos:
- Identificación de aspectos bien recibidos y apreciados.
- Fidelización y Promoción:
- Potencial de convertirse en defensores de la marca y promover activamente.

Usuarios que Votaron Negativamente y Realizaron Comentarios:

- Participación Activa:
- Grupo comprometido que votó negativamente y proporcionó comentarios.
- Feedback Detallado:
- Comentarios detallados que ofrecen información esencial.
- Posibles Problemas Identificados:
- Sugerencia de problemas o aspectos insatisfactorios identificados.
- Importancia del Feedback:
- Crucial para comprender áreas de mejora y abordar problemas específicos.

En resumen, la participación activa de los usuarios, ya sea votando positiva o negativamente y proporcionando comentarios detallados, destaca la importancia del feedback para fortalecer aspectos positivos y abordar oportunidades de mejora.

FASE CUATRO - RESUMEN:

Áreas de Fortalecimiento:

Considerando los comentarios y votos negativos, se recomienda enfocarse en áreas específicas para mejorar. Analiza los comentarios detallados y las votaciones negativas para identificar patrones y temas recurrentes que puedan señalar oportunidades de fortalecimiento.

Experiencia del Usuario:

Con base en las interacciones y feedback de los usuarios, se sugiere implementar mejoras centradas en la experiencia del usuario. Esto puede incluir ajustes en la interfaz, características adicionales, o la resolución de problemas señalados en los comentarios negativos.

Calidad de Datos:

Para garantizar la calidad de los datos, se recomienda realizar auditorías periódicas. Asegúrate de mantener estándares consistentes en la entrada de datos, identifica y aborda posibles problemas de inconsistencia o errores. Además, considera la implementación de validaciones automáticas para prevenir la entrada de datos incorrectos.

Resumen General:

En resumen, el análisis exploratorio de datos revela una valiosa perspectiva sobre la interacción de los usuarios con el producto o servicio. Se identificaron áreas de fuerza y oportunidades de mejora a través de votos y comentarios, proporcionando insights clave para la toma de decisiones. Las recomendaciones se centran en potenciar las fortalezas, mejorar la experiencia del usuario y garantizar la calidad continua de los datos. Este enfoque holístico contribuirá a una evolución positiva del producto o servicio, orientada hacia la satisfacción del usuario y la excelencia en la gestión de datos.

Outliers y Valores Atípicos:

Durante el análisis exploratorio de datos, se observaron posibles outliers en algunas variables, como "author_num_games_owned", "author_num_reviews", "author_playtime_forever", "author_playtime_last_two_weeks" y "author_playtime_at_review". Estos valores atípicos podrían influir en las conclusiones de diversas maneras:

Impacto en Estadísticas Descriptivas:

Los outliers pueden distorsionar las estadísticas descriptivas, como la media y la desviación estándar. Por ejemplo, en "author_playtime_forever", donde la media es de 10,257 horas, la presencia de valores atípicos podría afectar la percepción general del tiempo de juego promedio.

Influencia en Correlaciones:

Algunas correlaciones entre variables pueden ser afectadas por outliers, especialmente si estos valores extremos están presentes en las variables de interés. Se recomienda examinar la robustez de las correlaciones identificadas al considerar la posible influencia de outliers.

Consideraciones en Análisis Detallados:

Para análisis más detallados, como la identificación de patrones específicos en grupos de usuarios, es esencial tener en cuenta la presencia de outliers. Estos valores extremos podrían representar casos excepcionales que

requieren un análisis más profundo para comprender su impacto en el conjunto de datos.

Patrones Identificados:

Durante el análisis, se identificaron varios patrones y tendencias significativas en los datos:

Fuerte Correlación Positiva:

Se observa una correlación positiva fuerte entre "votes_up" y "votes_funny", así como entre "comment_count" y "votes_funny". Esto sugiere que las publicaciones con más votos positivos tienden a recibir más votos divertidos y comentarios.

Participación Activa de Usuarios:

La presencia de usuarios que votaron negativamente pero también dejaron comentarios destaca una participación más activa en comparación con aquellos que solo votaron negativamente. Esta participación activa puede proporcionar insights valiosos para mejoras.

Valores Atípicos en Métricas de Juego:

Se observan valores atípicos en variables relacionadas con la actividad de juego, como "author_playtime_forever" y "author_playtime_last_two_weeks". Estos valores extremos podrían indicar casos particulares que requieren una atención especial en análisis posteriores.

Feedback Detallado en Comentarios:

Usuarios que votaron negativamente y también dejaron comentarios proporcionan feedback más detallado sobre sus experiencias. Estos comentarios pueden ser fundamentales para identificar áreas específicas de mejora y comprender las preocupaciones de los usuarios.

Distribución de Votos y Comentarios:

El Diagrama de Venn resalta la intersección entre aquellos que votaron positivamente y realizaron comentarios. Este grupo específico representa usuarios altamente comprometidos y satisfechos que podrían ser considerados como defensores de la marca.

Referencias

[1] ▷ Valores atípicos (outliers): qué son, ejemplos, calculadora,... (probabilidadyestadistica.net)

[2]¿Cómo manejar los valores extremos (outliers) en nuestros datos? | Codificando Bits¿Qué es la media Winsorizada? - CriptoMundo

[3]Micro curso inicial EDA - de Cristian cuadrado beltran

[4]Cómo hacer EDA en DataScience o análisis EXPLORATORIO de datos en R (youtube.com)

[5]ELEMENTOS QUE DEBE TENER UN INFORME EDA de Cristian cuadrado beltran

[6]ELEMENTOS QUE DEBE TENER UN RESUMEN TECNICO de Cristian cuadrado beltran

[7]ELEMENTOS QUE DEBE TENER UN REPORTE DE CONCLUSIONES Y HALLAZGOS de Cristian cuadrado beltran

[8]QUE DEBE TENER UNA DATA LIMPIA de Cristian cuadrado beltran

Informe Análisis Exploratorio de datos [datos: imágenes]

Introduccion
En el marco de este proyecto técnico, se aborda el desafío de realizar un Análisis Exploratorio de Datos (EDA) en un conjunto de información única: datos relacionados con tipos de vehículos y sus placas. La transformación de imágenes de placas vehiculares en datos csv de Excel nos permitir realizar este trabajo de forma más rápida y sencilla.

Objetivos del Proyecto
Este proyecto tiene como objetivo principal realizar un EDA exhaustivo en los datos de tipos de vehículos y placas. Específicamente, buscamos:

- Identificar patrones visuales y tendencias en la distribución de tipos de vehículos.
- Analizar la frecuencia de aparición de placas, destacando posibles anomalías.
- Explorar relaciones entre los tipos de vehículos y las secuencias de las placas.

Metodología
La metodología aplicada incluye la transformación de imágenes de placas vehiculares a datos csv en Excel, utilizando herramientas de procesamiento de imágenes y técnicas de análisis de datos. Se emplearán herramientas como Google colaboratory y lenguajes de programación como Python y bibliotecas especializadas para el procesamiento eficiente de datos y la generación de visualizaciones significativas.

Alcance del Proyecto
Este proyecto incluye el análisis de tipos de vehículos y placas específicamente dentro del contexto proporcionado. No se abordarán aspectos relacionados con la identificación de conductores o información personal.

Tecnologías utilizadas
Se emplearon herramientas como Pandas, Seaborn y Matplotlib en un entorno de Google Colab para la manipulación, visualización y análisis de datos. Además, utilizamos Excel para la transformación de datos

Seguridad y Ética (si aplica)
Dado que se manejan placas vehiculares, se prestará especial atención a la privacidad y ética. No se buscará identificar conductores y se garantizará el manejo seguro de la información sensible.