SEMANA 2

Wuupi

Reporte de hallazgos

Carlos Fonseca Olivares- A01734538

Emily Bueno Romero - A01736939

Juan José Zamorano Balderas - A01642396

Juan Sebastián Mejía De Gortari - A01722536

02

Descripción general

Introducción:

Este codigo contiene un análisis centrado en la limpieza de datos nulos y la exploración de un conjunto de datos usando herramientas comunes en ciencia de datos. A lo largo del archivo se realizan tareas para mejorar la calidad de los datos y prepararlos para un análisis más profundo.

	Administrador	Usuario	botón correcto	tiempo de interacción	mini juego	número de interacción	color presionado	dificultad	fecha	Juego	auto push	tiempo de lección	tiempo de sesión
0	nicolas	nicolas	1.0	5.399169	Despegue	1.0	blue	Episodio 1	25/01/2024 09:26:42 a. m.	Astro	0.0	0.0	0.0
1	nicolas	nicolas	0.0	1.283400	Despegue	2.0	violet	Episodio 1	25/01/2024 09:26:46 a. m.	Astro	0.0	0.0	0.0
2	nicolas	nicolas	1.0	2.700226	Despegue	3.0	green	Episodio 1	25/01/2024 09:26:48 a. m.	Astro	0.0	0.0	0.0
3	nicolas	nicolas	0.0	3.050262	Despegue	4.0	green	Episodio 1	25/01/2024 09:26:57 a. m.	Astro	0.0	0.0	0.0
4	nicolas	nicolas	0.0	4.750256	Despegue	5.0	green	Episodio 1	25/01/2024 09:26:58 a. m.	Astro	0.0	0.0	0.0

Limpieza de datos

Base limpia, resultados limpios

Se dividieron las variables del conjunto de datos en dos tipos:

- Cuantitativas: números.
- Cualitativas: textos o fechas.

Después, se identificaron las columnas con datos faltantes (valores nulos) y se aplicaron técnicas para eliminarlos o reemplazarlos según el caso.

```
data_cuanti=data.select_dtypes(include=["float64","int64","float","int"])
data_cuali=data.select_dtypes(include=["object","datetime","category","datetime64[ns]"])

cuantias_mean=data_cuanti.fillna(round(data_cuanti.mean(),1))
cualis_bfill=data_cuali.fillna(method="bfill")
cualis_ffill=data_cuanti.fillna(method="ffill")

data_sin_nulos = pd.concat([cuantias_mean,cualis_ffill],axis=1)

data_sin_nulos
```

Código utilizado para la limpieza

	botón correcto	tiempo de interacción	número de interacción	auto push	tiempo de lección	tiempo de sesión	botón correcto	tiempo de interacción	número de interacción	auto push	tiempo de lección	tiempo de sesión
0	1.0	5.399169	1.0	0.0	0.000000	0.000000	1.0	5.399169	1.0	0.0	0.000000	0.000000
1	0.0	1.283400	2.0	0.0	0.000000	0.000000	0.0	1.283400	2.0	0.0	0.000000	0.000000
2	1.0	2.700226	3.0	0.0	0.000000	0.000000	1.0	2.700226	3.0	0.0	0.000000	0.000000
3	0.0	3.050262	4.0	0.0	0.000000	0.000000	0.0	3.050262	4.0	0.0	0.000000	0.000000
4	0.0	4.750256	5.0	0.0	0.000000	0.000000	0.0	4.750256	5.0	0.0	0.000000	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...
5860	0.5	10.000000	13.3	0.0	6.300000	332.240000	1.0	7.099429	4.0	0.0	7.461668	332.240000
5861	0.0	2.135419	1.0	0.0	0.000000	0.000000	0.0	2.135419	1.0	0.0	0.000000	0.000000
5862	0.5	10.000000	13.3	0.0	2.271806	12.400000	0.0	2.135419	1.0	0.0	2.271806	0.000000
5863	0.5	10.000000	13.3	0.0	6.300000	6.478299	0.0	2.135419	1.0	0.0	2.271806	6.478299
5864	0.5	10.000000	13.3	0.0	6.300000	12.400000	0.0	2.135419	1.0	0.0	2.271806	6.478299

Resultados

Limpieza de Outliers

Se identificaron valores que se salían de lo normal (outliers) usando el rango intercuartílico, que es una forma de ver qué tan dispersos están los datos. Estos valores extremos fueron eliminados o ajustados.

	Administrador	Usuario	mini juego	color presionado	dificultad	fecha	Juego	botón correcto	tiempo de interacción	número de interacción	auto push	tiempo de lección	tiempo de sesión
0	nicolas	nicolas	Despegue	blue	Episodio 1	25/01/2024 09:26:42 a. m.	Astro	1.0	5.399169	1.0	0.0	0.000000	0.000000
1	nicolas	nicolas	Despegue	violet	Episodio 1	25/01/2024 09:26:46 a. m.	Astro	0.0	1.283400	2.0	0.0	0.000000	0.000000
2	nicolas	nicolas	Despegue	green	Episodio 1	25/01/2024 09:26:48 a. m.	Astro	1.0	2.700226	3.0	0.0	0.000000	0.000000
3	nicolas	nicolas	Despegue	green	Episodio 1	25/01/2024 09:26:57 a. m.	Astro	0.0	3.050262	4.0	0.0	0.000000	0.000000
4	nicolas	nicolas	Despegue	green	Episodio 1	25/01/2024 09:26:58 a. m.	Astro	0.0	4.750256	5.0	0.0	0.000000	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...
5860	ALEIDA	ESMERALDA	NaN	NaN	Episodio 1	28/05/2024 04:15:49 p. m.	Astro	0.5	10.000000	13.3	0.0	6.300000	332.240000
5861	ALEIDA	JOSE JAVIER	Asteroides	green	Episodio 3	04/06/2024 11:09:54 a. m.	Astro	0.0	2.135419	1.0	0.0	0.000000	0.000000
5862	ALEIDA	JOSE JAVIER	Asteroides	NaN	Episodio 3	04/06/2024 11:09:58 a. m.	Astro	0.5	10.000000	13.3	0.0	2.271806	12.400000
5863	ALEIDA	JOSE JAVIER	NaN	NaN	Episodio 3	04/06/2024 11:09:58 a. m.	Astro	0.5	10.000000	13.3	0.0	6.300000	6.478299
5864	ALEIDA	JOSE JAVIER	MiniGame_0	NaN	Episodio 3	04/06/2024 11:09:59 a. m.	Astro	0.5	10.000000	13.3	0.0	6.300000	12.400000

Conversion de variables categoricas

```python
data["Administrador"]=data["Administrador"].str.replace("ALEIDA","1")
data["Administrador"]=data["Administrador"].str.replace("nicolas","2")
data["Administrador"]=data["Administrador"].str.replace("LEONARDO","3")
data["Administrador"]=data["Administrador"].str.replace("DENISSE","4")
data["Administrador"]=data["Administrador"].str.replace("SERGIO ANGEL","5")
data["Administrador"]=data["Administrador"].str.replace("CARLOS ENRIQUE","6")
data["Administrador"]=data["Administrador"].str.replace("YAEL DAVID","7")
data["Administrador"]=data["Administrador"].str.replace("AUSTIN","8")
data["Administrador"]=data["Administrador"].str.replace("VALENTIN","9")
data["Administrador"]=data["Administrador"].str.replace("erick","10")
data["Administrador"]=data["Administrador"].str.replace("IKER BENJAMIN","11")
data["Administrador"]=data["Administrador"].str.replace("KYTZIA","12")
data["Administrador"]=data["Administrador"].str.replace("BENJAMIN","13")
data
```

Python

	Administrador	Usuario	botón correcto	tiempo de interacción	mini juego	número de interacción	color presionado	dificultad	fecha	Juego	auto push	tiempo de lección	tiempo de sesión
0	2	nicolas	1.0	5.399169	Despegue	1.0	blue	Episodio 1	25/01/2024 09:26:42 a. m.	Astro	0.0	0.000000	0.000000
1	2	nicolas	0.0	1.283400	Despegue	2.0	violet	Episodio 1	25/01/2024 09:26:46 a. m.	Astro	0.0	0.000000	0.000000
2	2	nicolas	1.0	2.700226	Despegue	3.0	green	Episodio 1	25/01/2024 09:26:48 a. m.	Astro	0.0	0.000000	0.000000
3	2	nicolas	0.0	3.050262	Despegue	4.0	green	Episodio 1	25/01/2024 09:26:57 a. m.	Astro	0.0	0.000000	0.000000
4	2	nicolas	0.0	4.750256	Despegue	5.0	green	Episodio 1	25/01/2024 09:26:58 a. m.	Astro	0.0	0.000000	0.000000

Estas son algunas de las gráficas presentadas en el código

¿Qué son las variables categóricas?
Son aquellas que contienen etiquetas o categorías en lugar de números. Por ejemplo:

- Nombres de personas ("ALEIDA", "LEONARDO", "nicolas").
- Tipos de juego ("Carrera", "Puzzle", "Aventura").
- Clasificaciones como "Bajo", "Medio", "Alto".

Estas variables no pueden ser usadas directamente en operaciones matemáticas, por eso es necesario transformarlas.

¿Por qué convertirlas a números?

- Los modelos de análisis estadístico y machine learning solo trabajan con números.
- Las etiquetas de texto no tienen valor matemático, por lo tanto, no se pueden comparar directamente.
- Convertirlas a valores numéricos permite graficarlas, analizarlas y procesarlas automáticamente.

Análisis de correlaciones lineales con respecto a “Usuario” (Heatmap)

Se analizó cómo se relaciona la variable “Usuario” con las demás variables numéricas del conjunto de datos. Como "Usuario" es una categoría (por ejemplo, nombres o etiquetas), primero se convirtió a valores numéricos utilizando su frecuencia (cuántas veces aparece cada uno). Luego, se calculó una matriz de correlación para encontrar relaciones estadísticas entre todas las columnas numéricas, y finalmente se generó un heatmap (mapa de calor) para visualizar mejor dichas relaciones

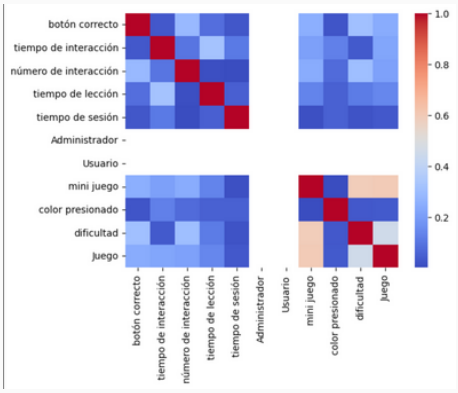


Imagen 1. Las mayores correlaciones fueron: mini juego - Juego | mini juego - dificultad

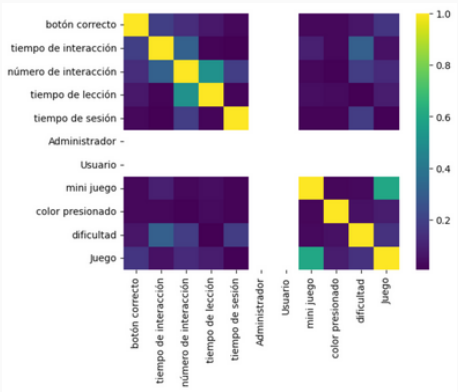


Imagen 2. Las mayores correlaciones fueron: mini juego - Juego. Seguido de cerca por: tiempo de interacción - dificultad | tiempo de interacción - número de interacción.

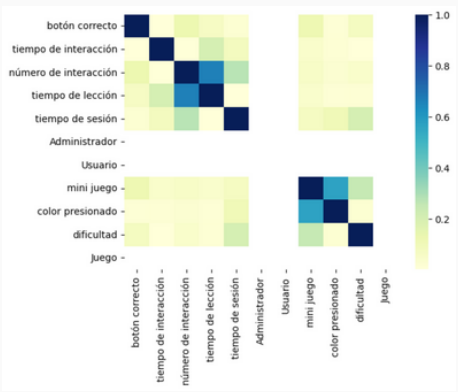


Imagen 3. La mayor correlación fue: mini juego - color presionado.

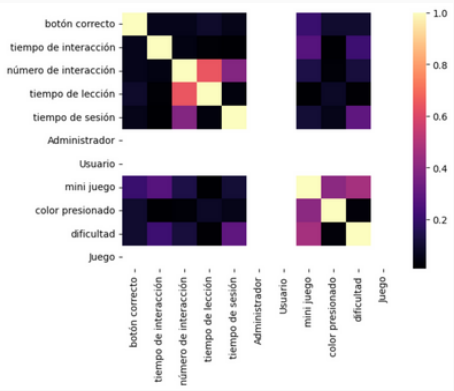


Imagen 4. La mayor correlación fue: mini juego - dificultad. Seguido de cerca por: mini juego - color presionado.

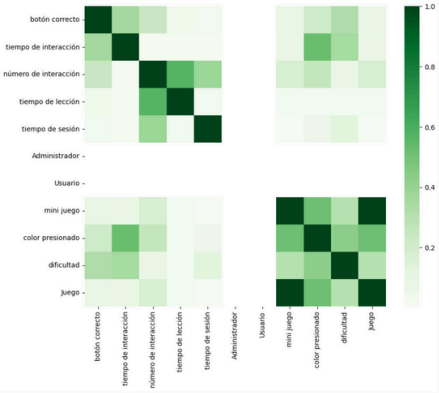


Imagen 5. La mayor correlación fue: mini juego - Juego.

¿Para que sirve?

Esto ayuda a identificar qué variables pueden tener influencia sobre el tipo de usuario, lo que es útil para clasificar, predecir o personalizar análisis

Modelo Lineal Múltiple

Luego de revisar qué variables tienen relación individual con "Usuario", se construyó un modelo de regresión lineal múltiple. Esto significa que se usan varias columnas como predictores para estimar o predecir otra columna (en este caso, la versión numérica de “Usuario”).

Se dividieron los datos en dos grupos: uno para entrenar el modelo y otro para probar qué tan bien funciona. Se usó la librería scikit-learn

```
from sklearn.linear_model import LinearRegression
model= LinearRegression()
type(model)
```

```
[ ]  #Corroboramos cual es el coeficiente de Determinación de nuestro modelo
     coef_Deter=model.score(X=Vars_Indep, y=Var_Dep)
     print("Coeficiente de Determinación",coef_Deter)
     #Corroboramos cual es el coeficiente de Correlación de nuestro modelo
     coef_Correl=np.sqrt(coef_Deter)
     print("Coeficiente de Correlación",coef_Correl)

     Coeficiente de Determinación 0.04013733821604415
     Coeficiente de Correlación 0.20034305132957356
```

Imagen 1: Al parecer los coeficientes no salieron muy altos como para superar a las correlaciones simples. Esto puede indicar que las variables realmente no aportan un valor real para predecir respecto al usuario.

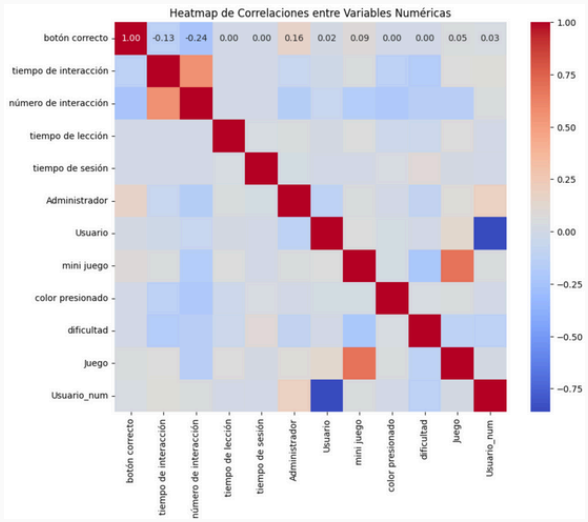
¿Para que sirve?
Un modelo lineal múltiple puede capturar relaciones combinadas entre muchas variables, lo que permite predecir con más precisión que si solo se usara una a la vez.

Comparación de Heatmaps para Cada Usuario

Una vez segmentados los datos, se calculó la matriz de correlación para cada subconjunto, considerando únicamente las variables numéricas. A continuación, se representaron estas correlaciones mediante mapas de calor (heatmaps), lo cual permitió visualizar gráficamente la intensidad y dirección de las relaciones entre variables para cada usuario de forma individual.
Este enfoque facilitó la comparación entre usuarios, permitiendo identificar si existen diferencias significativas en los patrones de comportamiento de las variables. Por ejemplo, se pudo observar si ciertas variables mantienen una relación fuerte en algunos usuarios mientras que en otros no, o si la dirección de la correlación cambia completamente.

```python
#Convertir la columna 'Usuario'
usuario_freq = data['Usuario'].value_counts().to_dict()
data['Usuario_num'] = data['Usuario'].map(usuario_freq)

#Filtra columnas numéricas
df_numericas = data.select_dtypes(include=["number"])

#matriz de correlación
correlation_matrix = df_numericas.corr()

#Mostrar correlación de todas las variables con 'Usuario_num'
correlaciones_usuario = correlation_matrix[['Usuario_num']].sort_values(by='Usuario_num', ascending=False)
print("Correlaciones con 'Usuario':\n")
print(correlaciones_usuario)

#heatmap
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title("Heatmap de Correlaciones entre Variables Numéricas")
plt.show()
```

```
Correlaciones con 'Usuario':

                     Usuario_num
Usuario_num             1.000000
Administrador           0.197205
tiempo de interacción   0.078805
número de interacción   0.053552
mini juego              0.045528
botón correcto          0.033926
Juego                   0.014637
tiempo de lección       0.010216
tiempo de sesión        0.003584
color presionado       -0.002673
dificultad             -0.134204
Usuario                -0.861846
```



¿Para que sirve?
Este tipo de análisis es útil para identificar patrones únicos por grupo, detectar comportamientos diferenciados y considerar la posibilidad de implementar modelos personalizados en lugar de generalizados.

Conclusión

Después de limpiar la base de datos, eliminar los espacios vacíos y quitar los valores que no seguían el comportamiento normal, se logró tener un conjunto de datos más confiable y ordenado. Esto significa que ahora sí se pueden hacer análisis y sacar conclusiones sin preocuparse por errores o información incompleta.

Por ejemplo, ahora se pueden comparar los resultados entre diferentes niveles de dificultad, analizar qué juegos son más efectivos o ver cuánto tiempo dedican en promedio los usuarios. También se pueden hacer gráficas, resúmenes o modelos que nos digan quién está aprendiendo más, qué partes del sistema son más atractivas o dónde podrían estar los errores. En resumen, este proceso de preparación hizo que los datos sean mucho más útiles y valiosos para tomar decisiones.