WUUPI
AVANCE 2

EQUIPO 4 | Máquinas de chambeo

INICIO DE CÓDIGO

- Importamos las bibliotecas esenciales para el análisis
- Realizamos la carga de los datos por el archivo csv
- Realizamos la exploración de datos, y la búsqueda de valores nulos por columna

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import matplotlib.pyplot as plt
```

```
#Cargamos los datos
data = pd.read_csv('DataAnalytics.csv')
```

NULOS Y OUTLIERS

```
valores_nulos = data.isnull().sum()
print(valores_nulos)
```

```
Administrador                0
Usuario                      0
botón correcto             762
tiempo de interacción      762
mini juego                 156
número de interacción      762
color presionado           762
dificultad                   0
fecha                        0
Juego                        0
auto push                  762
tiempo de lección          177
tiempo de sesión           606
dtype: int64
```

RESULTADO:

Varias columnas tienen **valores nulos**, especialmente:

"botón correcto", "tiempo de interacción", "número de interacción", "color presionado" y "auto push" con 762 valores nulos cada una.

SOLUCIÓN:

- Se **verifican** los tipos de **datos** de cada columna

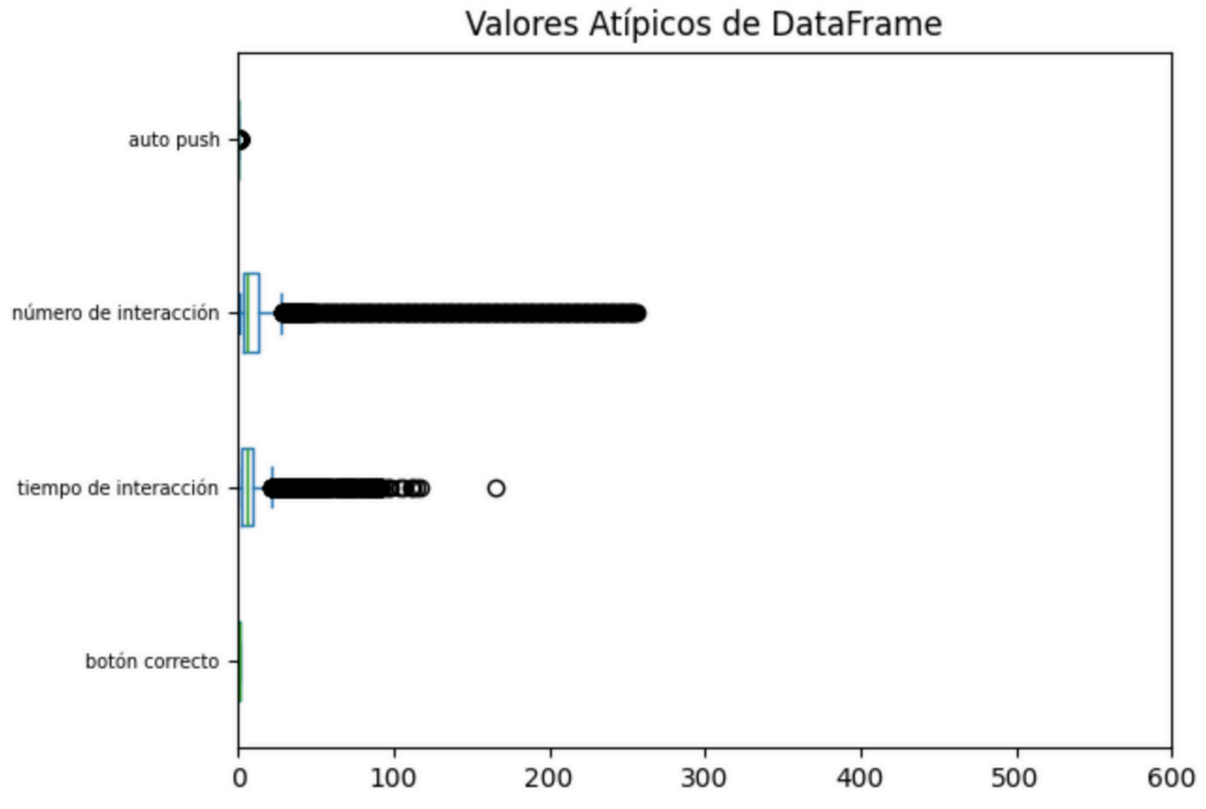
- **Separamos** las variables por **numéricas y cuantitativas**:

  - Cuantitativas (‘int64’ & ‘float64’): media

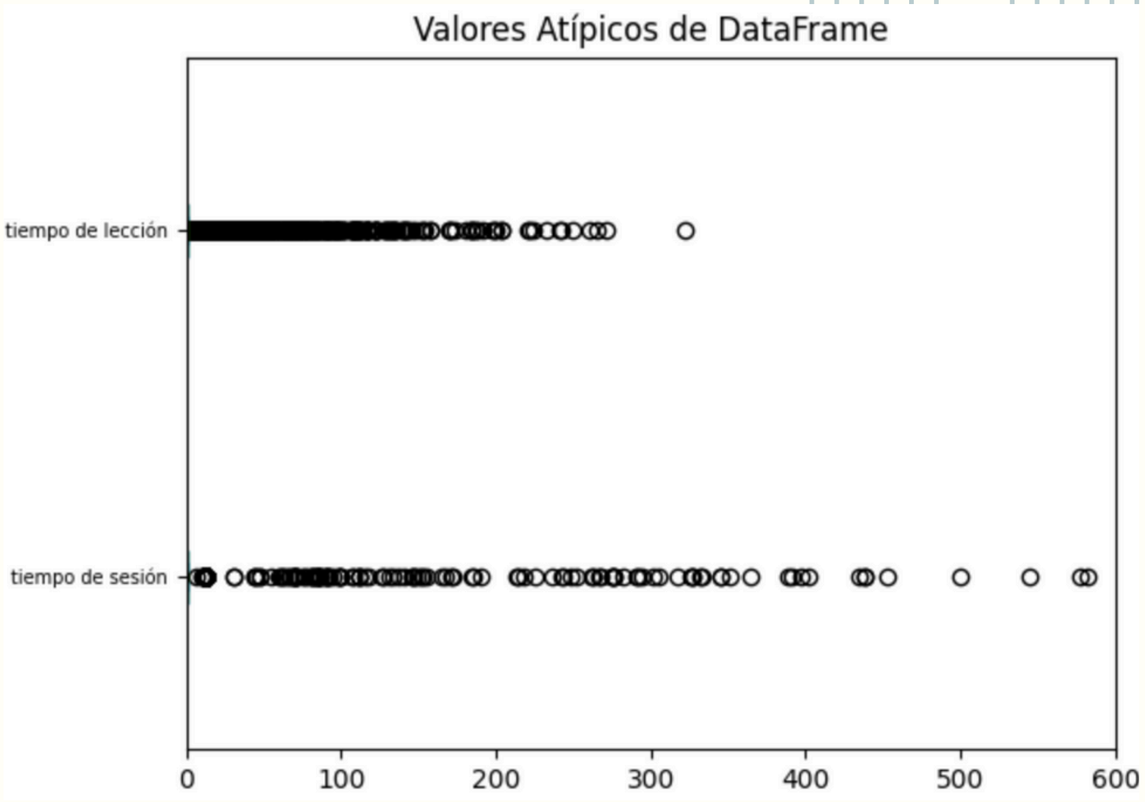
  - Cualitativas (‘object’): “Sin dato”

```
numericas = data.select_dtypes(include=['int64', 'float64'])
cualitativas = data.select_dtypes(include=['object'])
```

Valores Atípicos del DataFrame



Variables: auto push, número de interacción, tiempo de interacción y botón correcto
Método: rango intercuartílico (IQR)



Variables: tiempo de lección y tiempo de sesión (por la gran cantidad de registros “0”)
Método: percentil 1 y 99

CONVERSIÓN DE VARIABLES

categóricas ⟶ numéricas

```
cat1 = data_final.groupby(['Administrador'])['Administrador'].count().sort_values(ascending=False)
cat1
```

✓ 0.0s

```
Administrador
ALEIDA            3260
nicolas            440
LEONARDO           371
DENISSE            302
SERGIO ANGEL       243
CARLOS ENRIQUE     228
YAEL DAVID         224
AUSTIN             199
VALENTIN           163
erick              158
IKER BENJAMIN      128
KYTZIA              98
BENJAMIN            51
Name: Administrador, dtype: int64
```

método

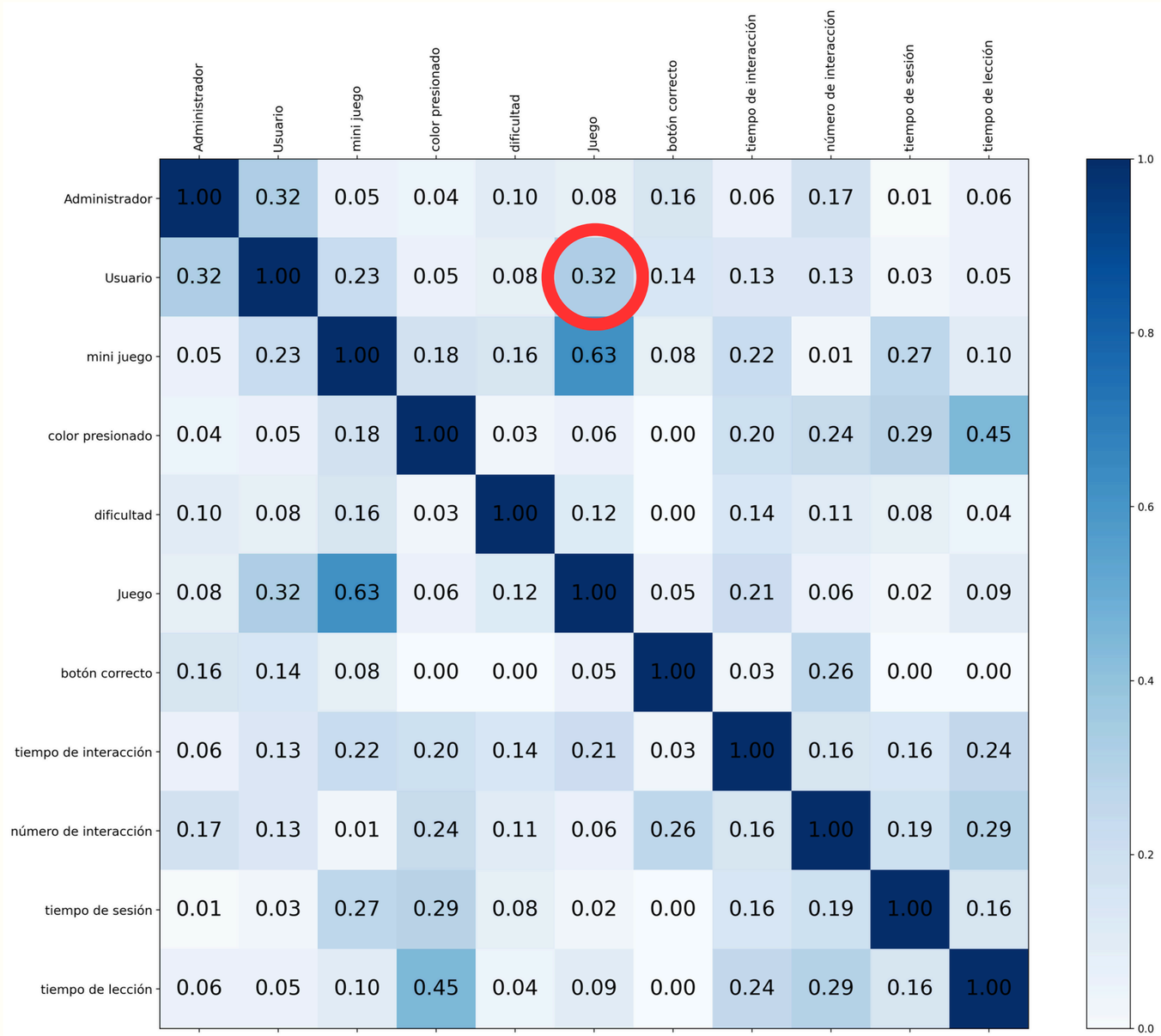
Frecuencia

```
data_final.Administrador = data_final.Administrador.replace({'ALEIDA':'1'}, regex=False)
data_final.Administrador = data_final.Administrador.replace({'nicolas':'2'}, regex=False)
data_final.Administrador = data_final.Administrador.replace({'LEONARDO':'3'}, regex=False)
data_final.Administrador = data_final.Administrador.replace({'DENISSE':'4'}, regex=False)
data_final.Administrador = data_final.Administrador.replace({'SERGIO ANGEL':'5'}, regex=False)
data_final.Administrador = data_final.Administrador.replace({'CARLOS ENRIQUE':'6'}, regex=False)
data_final.Administrador = data_final.Administrador.replace({'YAEL DAVID':'7'}, regex=False)
data_final.Administrador = data_final.Administrador.replace({'AUSTIN':'8'}, regex=False)
data_final.Administrador = data_final.Administrador.replace({'VALENTIN':'9'}, regex=False)
data_final.Administrador = data_final.Administrador.replace({'erick':'10'}, regex=False)
data_final.Administrador = data_final.Administrador.replace({'IKER BENJAMIN':'11'}, regex=False)
data_final.Administrador = data_final.Administrador.replace({'KYTZIA':'12'}, regex=False)
data_final.Administrador = data_final.Administrador.replace({'BENJAMIN':'13'}, regex=False)
```

RESULTADOS

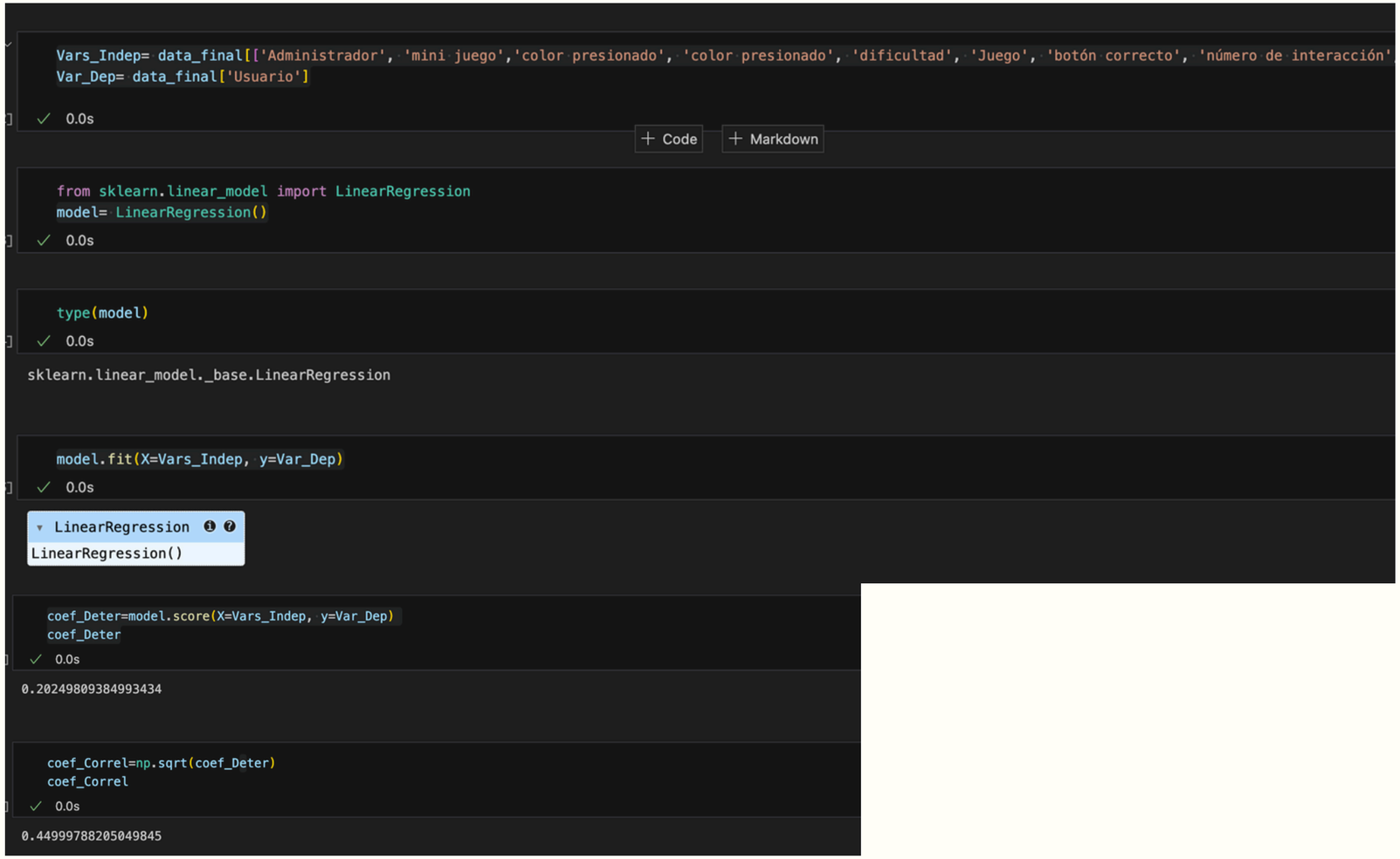
Primer heatmap

analizamos los coeficientes de la variable “Usuario” con las demás variables



Aplicamos un modelo de regresión linear múltiple

1) Identificamos variable dependiente (usuario) e independientes (las demás)
2) Generamos el modelo
3) Calculamos el coeficiente de determinacion
4) Calculamos coeficiente de correlación múltiple



COEFICIENTES

REGRESIÓN LINEAR SIMPLE

0.32

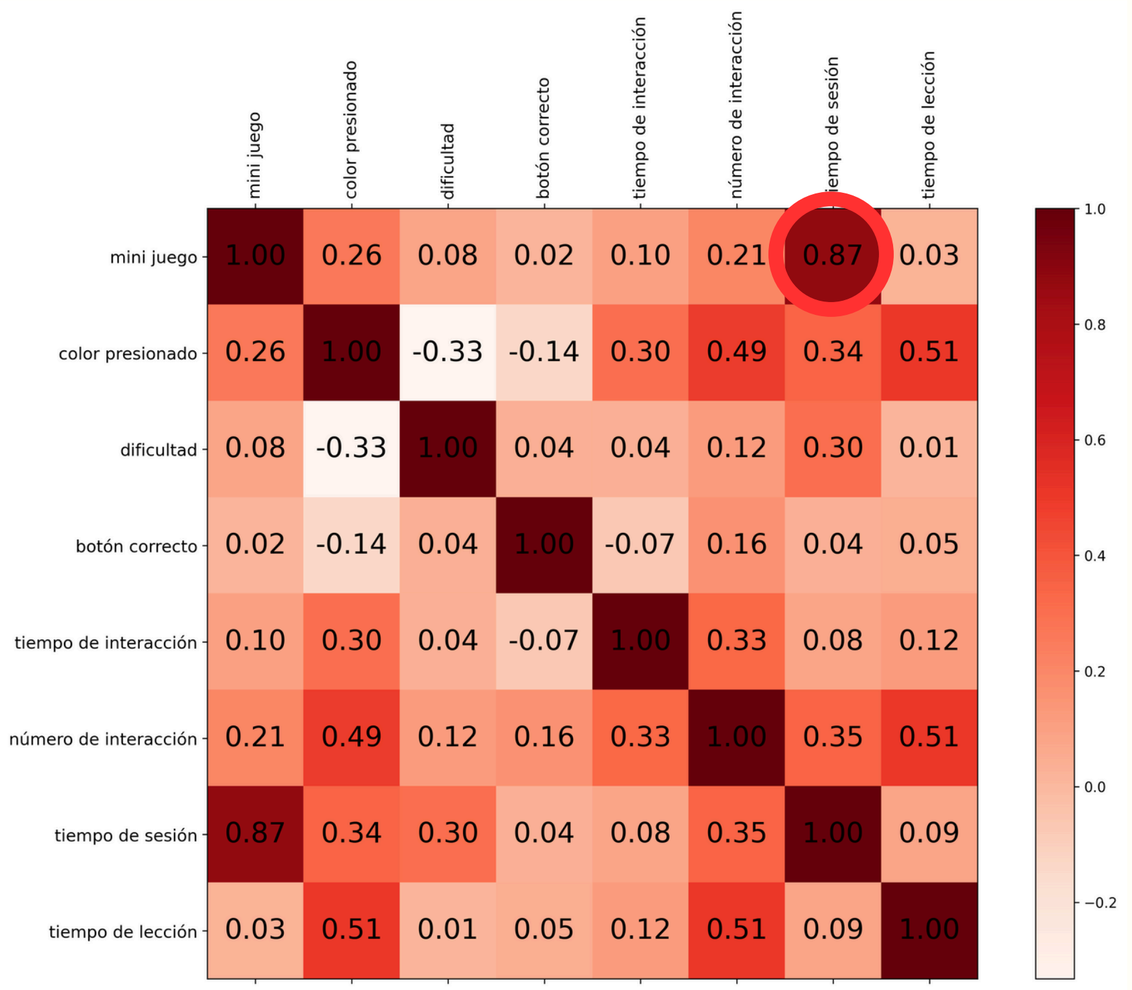
REGRESIÓN LINEAR MÚLTIPLE

0.44

RESULTADOS POR USUARIO

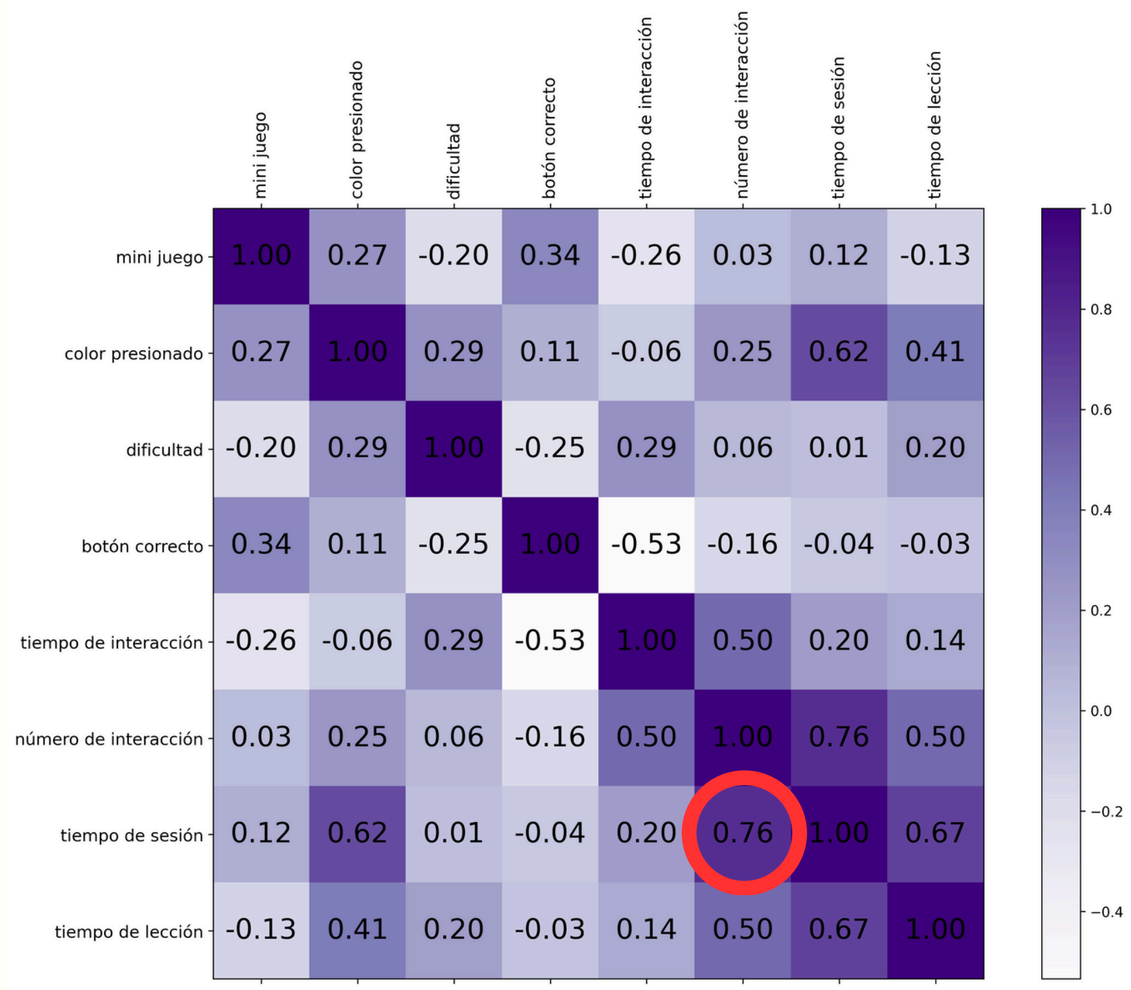
Benjamín

Variable: tiempo de sesión



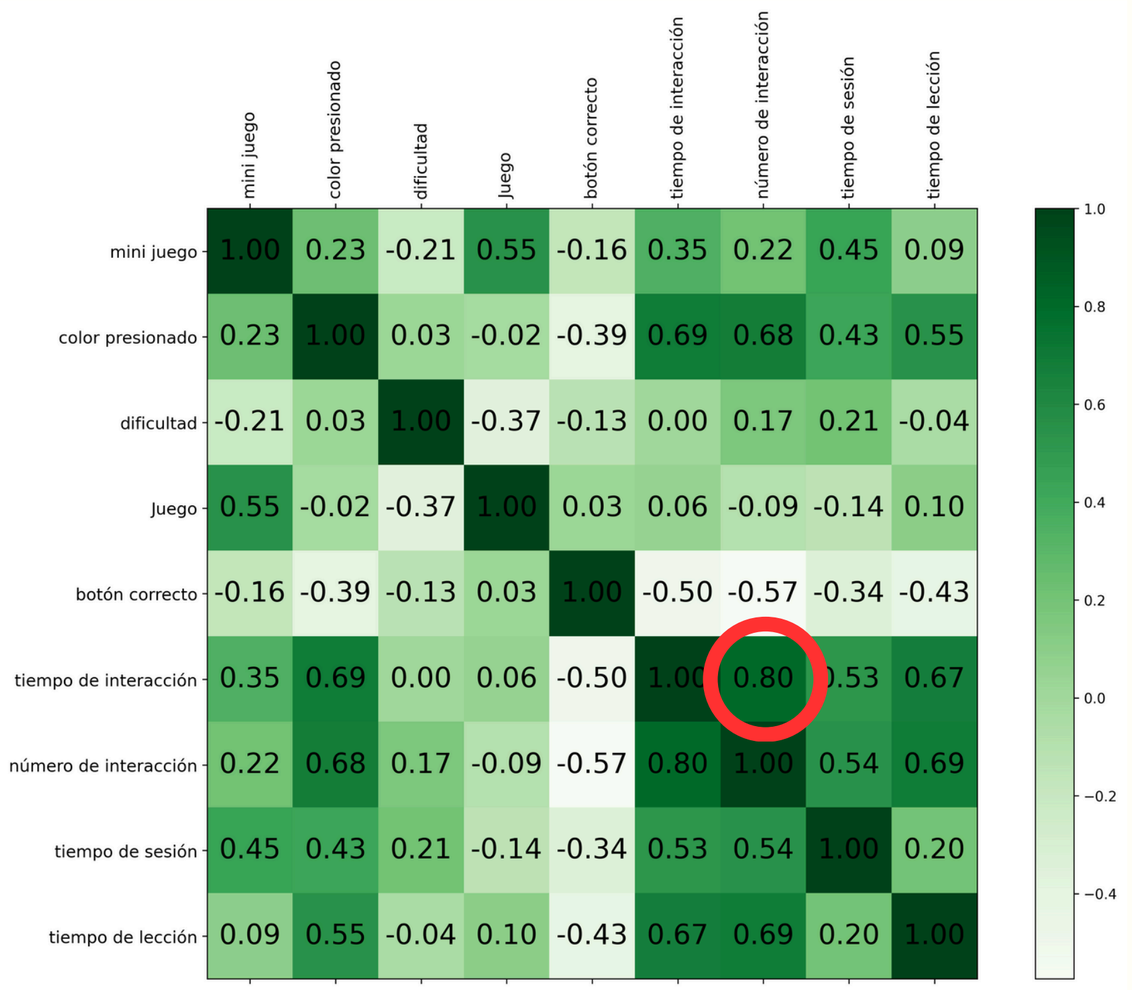
Carlos Abel

Variable: tiempo de sesión



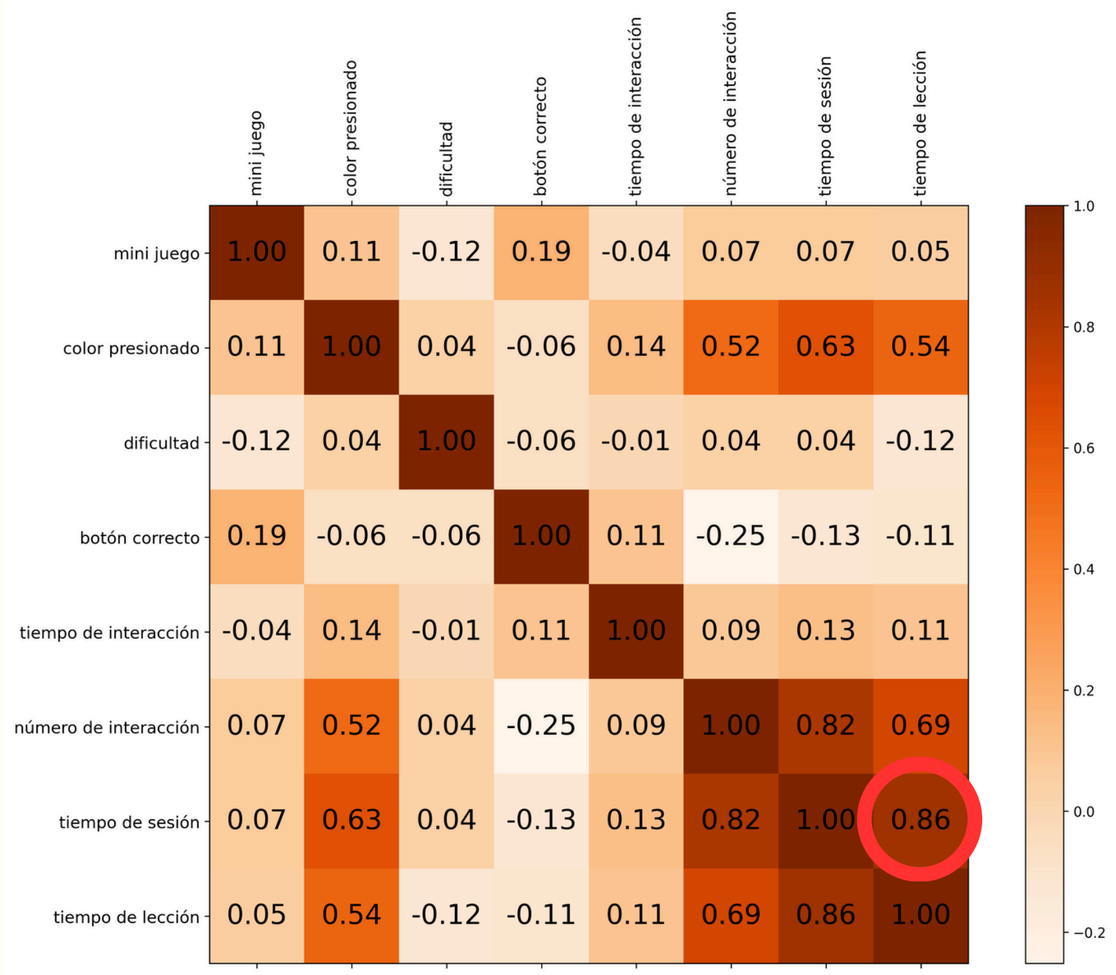
Carlos Enrique

Variable: tiempo de interacción



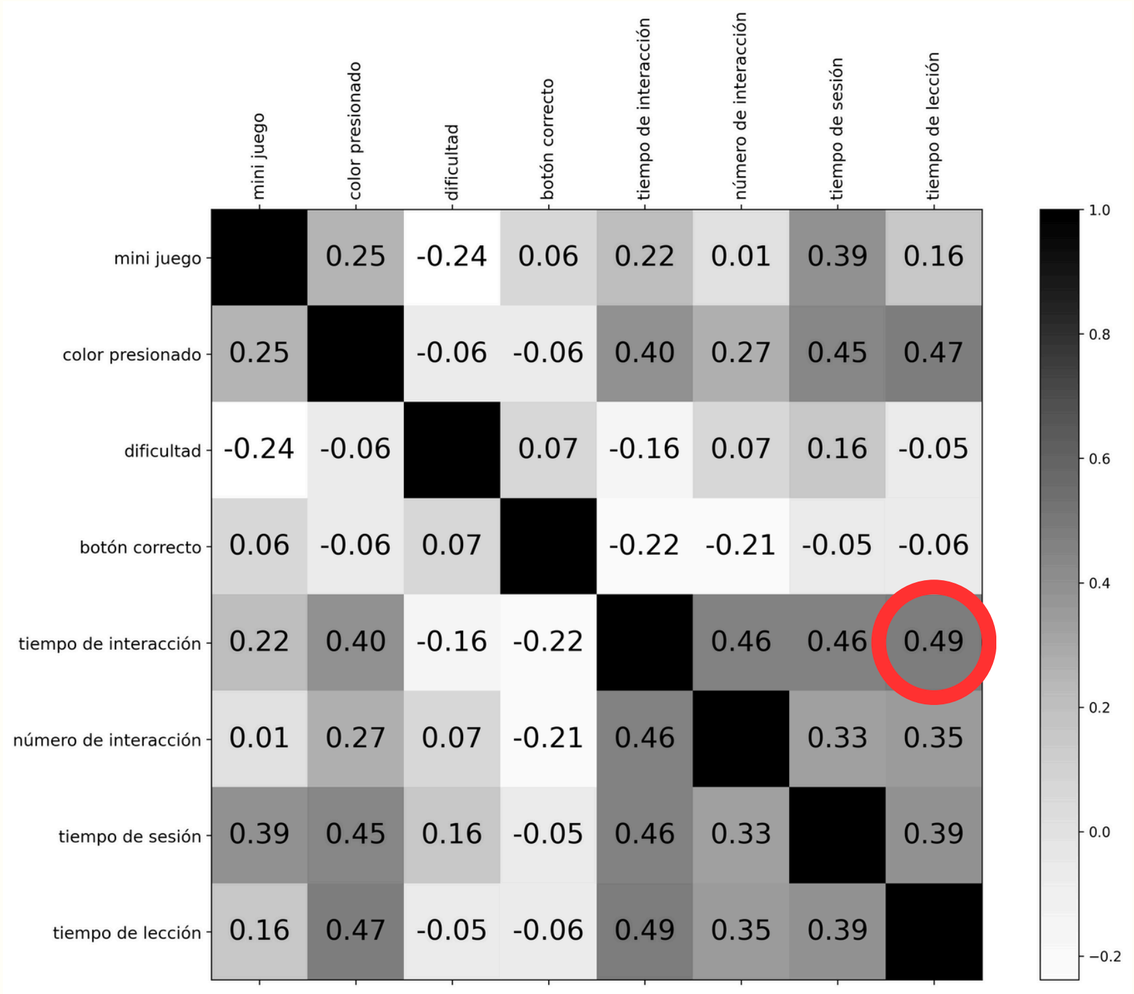
Concepción

Variable: tiempo de lección



Denisse

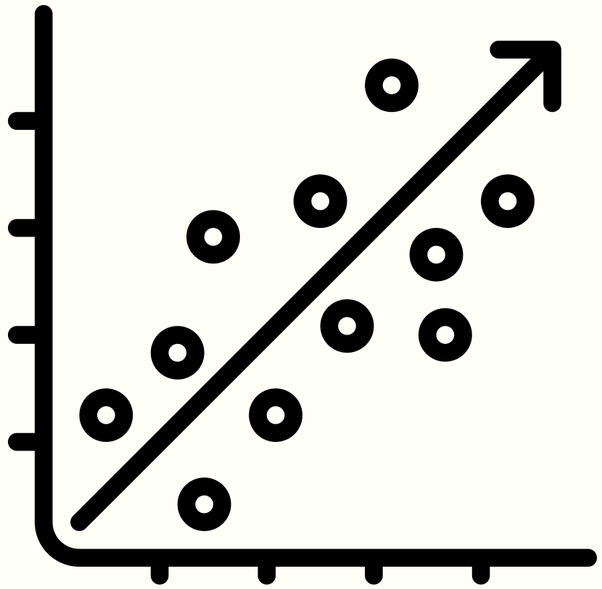
Variable: tiempo de lección



Comparativo

Usuario	Variable seleccionada (dependiente)	Coeficiente correlación linear	Coeficiente correlación linear múltiple
Benjamín	tiempo de sesión	**0.87**	**0.94**
Carlos Abel	tiempo de sesión	**0.76**	**0.92**
Carlos Enrique	tiempo de interacción	**0.80**	**0.85**
Concepción	tiempo de lección	**0.86**	**0.88**
Denisse	tiempo de lección	**0.49**	**0.59**

Conclusiones



- **Correlaciones variables por usuario:** Los coeficientes de correlación lineal y múltiple varían notablemente entre usuarios, destacando que variables como "tiempo de sesión" y "tiempo de interacción" tienen impactos diferenciados en cada caso.
- **Mejora en modelos múltiples:** En todos los usuarios analizados, el coeficiente de correlación lineal múltiple supera al lineal simple, lo que sugiere que incluir más variables en el modelo mejora la precisión de las predicciones.

Gracias