Tecnológico de Monterrey
Escuela de Ingeniería y Ciencias

**Campus CCM**

# Activity 4: Heatmaps and boxplots

CCS Tools


Estudiante:

Alejandro Salazar Loza A01665123

Emiliano Torres Sandoval


Link Repo: https://github.com/RA-Rauw/Activity-1-Repository

Fecha de entrega:

7 de mayo del 2025

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from google.colab import files

# Código para subir archivos en Google Colab
print("Por favor, sube el archivo 'messi_competition_goals.csv'")
uploaded_messi = files.upload()
messi_filename = list(uploaded_messi.keys())[0]

print("Por favor, sube el archivo 'ronaldo_competition_goals.csv'")
uploaded_ronaldo = files.upload()
ronaldo_filename = list(uploaded_ronaldo.keys())[0]

# Cargar los datos de Messi y Ronaldo desde los archivos subidos
messi_data = pd.read_csv(messi_filename)
ronaldo_data = pd.read_csv(ronaldo_filename)

# Renombrar la columna 'Caps' en los datos de Ronaldo para que coincida con 'Apps' de Messi
ronaldo_data.rename(columns={'Caps': 'Apps'}, inplace=True)

# Añadir una columna para identificar a cada jugador
messi_data['Player'] = 'Messi'
ronaldo_data['Player'] = 'Ronaldo'

# Combinar los datos de ambos jugadores
combined_data = pd.concat([messi_data, ronaldo_data], ignore_index=True)

# Configuración de estilo para mejorar la visualización
sns.set(style="whitegrid")
plt.figure(figsize=(15, 10))
```

```
# 1. Box and Whisker Plot (Diagrama de cajas y bigotes)
plt.subplot(2, 2, 1)
sns.boxplot(data=combined_data, x='Player', y='Goals')
plt.title('Distribución de Goles por Jugador')
plt.xlabel('Jugador')
plt.ylabel('Goles')

# Box plot para Apps (apariciones)
plt.subplot(2, 2, 2)
sns.boxplot(data=combined_data, x='Player', y='Apps')
plt.title('Distribución de Apariciones por Jugador')
plt.xlabel('Jugador')
plt.ylabel('Apariciones')

# 2. Histograma de goles
plt.subplot(2, 2, 3)
sns.histplot(data=combined_data, x='Goals', hue='Player', kde=True, bins=10, element="step")
plt.title('Histograma de Goles')
plt.xlabel('Goles')
plt.ylabel('Frecuencia')

# 3. Crear un dataframe pivote para el análisis de correlación
pivot_df = combined_data.pivot_table(index='Competition', columns='Player', values=['Goals', 'Apps'], aggfunc='sum')
pivot_df.columns = [f"{col[1]}_{col[0]}" for col in pivot_df.columns]
pivot_df = pivot_df.reset_index()
pivot_df.fillna(0, inplace=True)

# Obtener solo las columnas numéricas para la correlación
numeric_columns = pivot_df.select_dtypes(include=[np.number]).columns
correlation_matrix = pivot_df[numeric_columns].corr()
```

```python
# 4. Mapa de calor para la correlación
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', linewidths=0.5)
plt.title('Mapa de Calor de Correlación')
plt.tight_layout()

# Mostrar estadísticas descriptivas
print("Estadísticas descriptivas para Messi:")
print(messi_data.describe())
print("\nEstadísticas descriptivas para Ronaldo:")
print(ronaldo_data.describe())

# Comparar eficiencia (goles por aparición)
messi_data['Eficiencia'] = messi_data['Goals'] / messi_data['Apps']
ronaldo_data['Eficiencia'] = ronaldo_data['Goals'] / ronaldo_data['Apps']

print("\nEficiencia de goles por aparición para Messi:")
print(messi_data[['Competition', 'Eficiencia']].sort_values(by='Eficiencia', ascending=False))
print("\nEficiencia de goles por aparición para Ronaldo:")
print(ronaldo_data[['Competition', 'Eficiencia']].sort_values(by='Eficiencia', ascending=False))

# Crear una figura adicional para una comparación detallada por competición
plt.figure(figsize=(14, 8))
competitions = set(messi_data['Competition'].tolist() + ronaldo_data['Competition'].tolist())

# Crear DataFrames para la comparación por competición
comparison_data = []
for comp in competitions:
    messi_row = messi_data[messi_data['Competition'] == comp]
    ronaldo_row = ronaldo_data[ronaldo_data['Competition'] == comp]

    messi_goals = messi_row['Goals'].sum() if not messi_row.empty else 0
    messi_apps = messi_row['Apps'].sum() if not messi_row.empty else 0

    ronaldo_goals = ronaldo_row['Goals'].sum() if not ronaldo_row.empty else 0
    ronaldo_apps = ronaldo_row['Apps'].sum() if not ronaldo_row.empty else 0
```

```
    comparison_data.append({
        'Competition': comp,
        'Messi_Goals': messi_goals,
        'Messi_Apps': messi_apps,
        'Ronaldo_Goals': ronaldo_goals,
        'Ronaldo_Apps': ronaldo_apps
    })

comparison_df = pd.DataFrame(comparison_data)

# Gráfico de barras para comparar goles por competición
plt.subplot(1, 2, 1)
comparison_df_melted = pd.melt(comparison_df, id_vars=['Competition'],
                               value_vars=['Messi_Goals', 'Ronaldo_Goals'],
                               var_name='Player', value_name='Goals')
comparison_df_melted['Player'] = comparison_df_melted['Player'].str.replace('_Goals', '')

sns.barplot(data=comparison_df_melted, x='Competition', y='Goals', hue='Player')
plt.title('Comparación de Goles por Competición')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()

# Gráfico de barras para comparar apariciones por competición
plt.subplot(1, 2, 2)
comparison_df_melted = pd.melt(comparison_df, id_vars=['Competition'],
                               value_vars=['Messi_Apps', 'Ronaldo_Apps'],
                               var_name='Player', value_name='Appearances')
comparison_df_melted['Player'] = comparison_df_melted['Player'].str.replace('_Apps', '')

sns.barplot(data=comparison_df_melted, x='Competition', y='Appearances', hue='Player')
plt.title('Comparación de Apariciones por Competición')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()

plt.show()
```

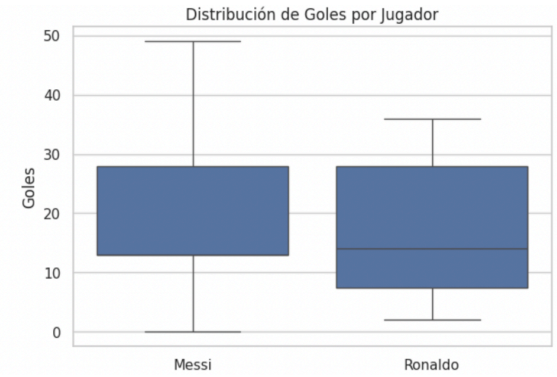
```
Estadísticas descriptivas para Messi:
              Apps      Goals
count    5.000000   5.000000
mean    35.000000  20.600000
std     23.579652  18.716303
min      1.000000   0.000000
25%     26.000000  13.000000
50%     34.000000  13.000000
75%     54.000000  28.000000
max     60.000000  49.000000

Estadísticas descriptivas para Ronaldo:
              Apps      Goals
count    7.000000   7.000000
mean    28.571429  17.571429
std     18.100250  13.806486
min      4.000000   2.000000
25%     16.500000   7.500000
50%     25.000000  14.000000
75%     43.000000  28.000000
max     52.000000  36.000000
```

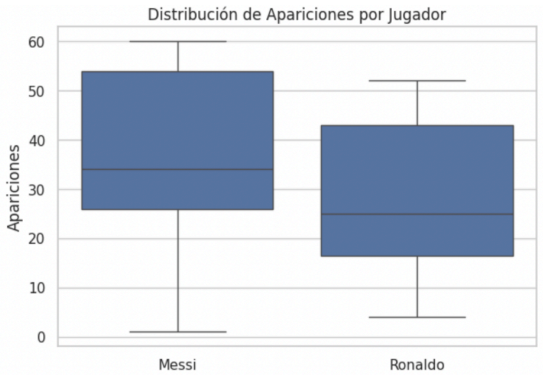
Eficiencia de goles por aparición para Messi:

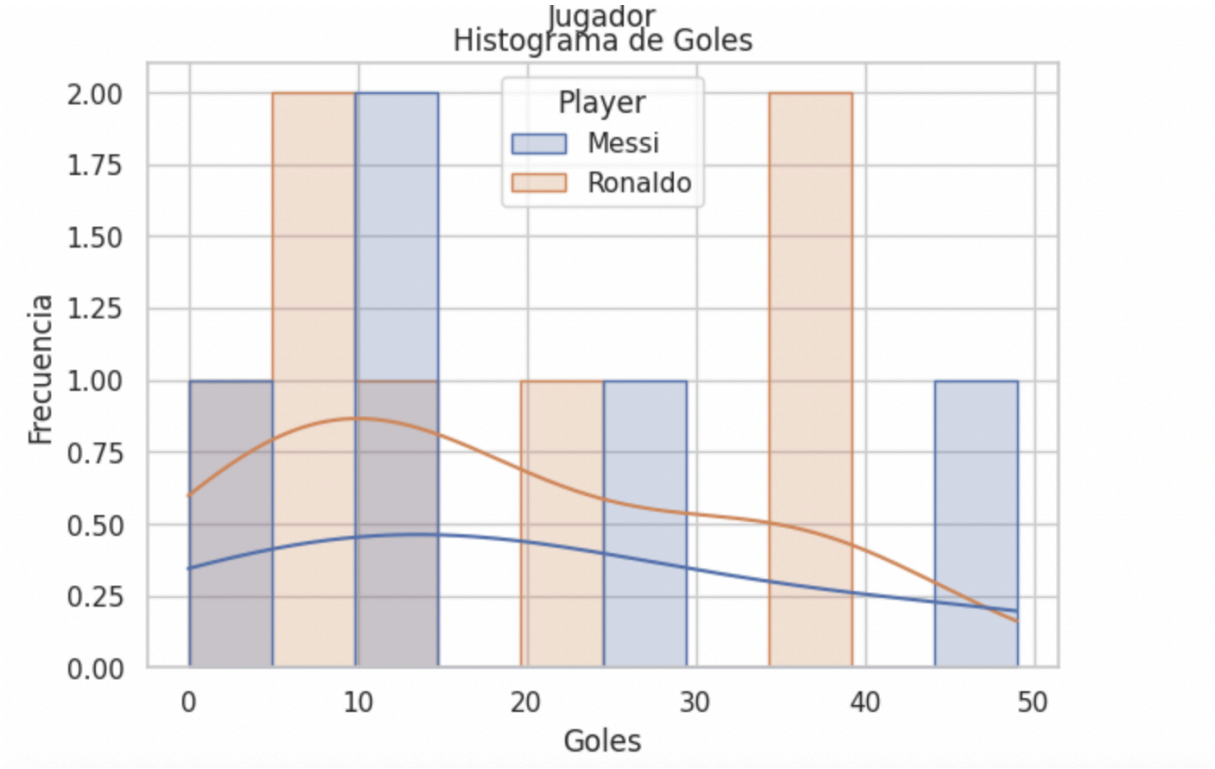
| | Competition | Eficiencia |
|---|---|---|
| 0 | Friendlies | 0.907407 |
| 3 | FIFA World Cup | 0.500000 |
| 2 | FIFA World Cup qualification | 0.466667 |
| 1 | Copa América | 0.382353 |
| 4 | CONMEBOL–UEFA Cup of Champions | 0.000000 |

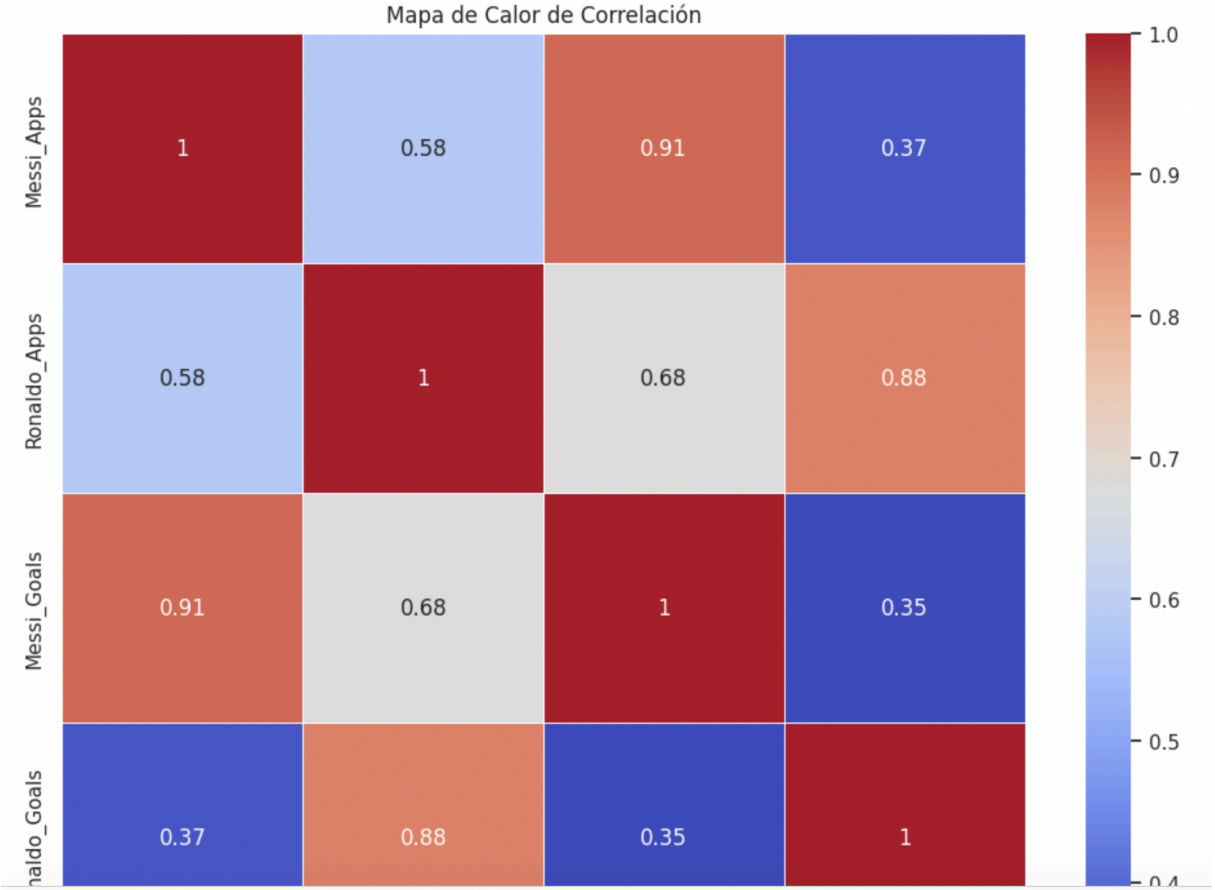
Eficiencia de goles por aparición para Ronaldo:

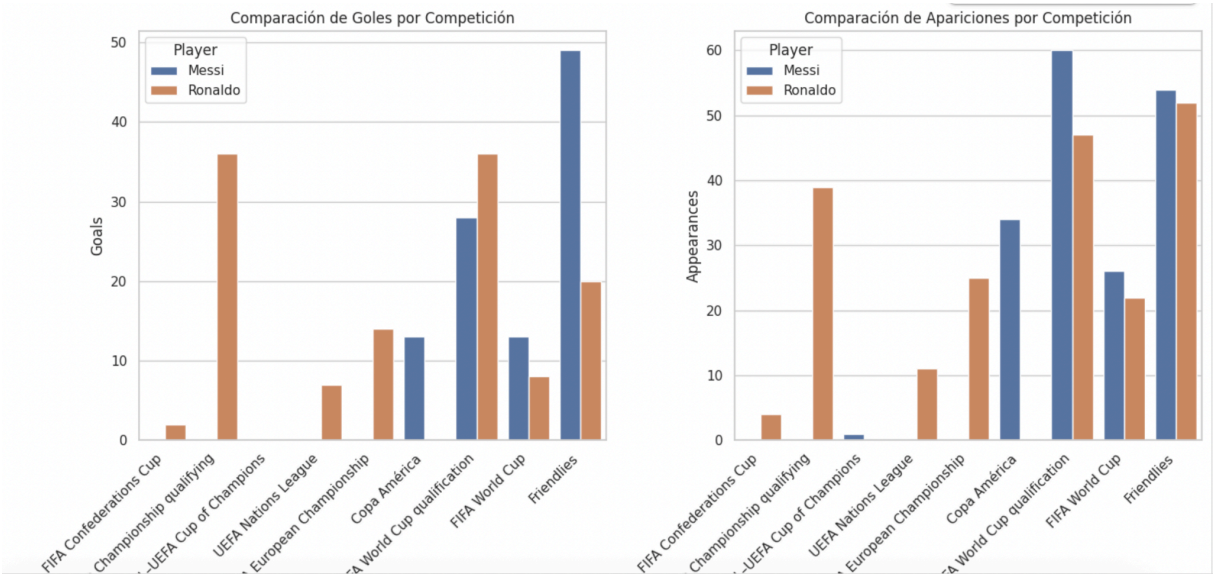
| | Competition | Eficiencia |
|---|---|---|
| 1 | UEFA European Championship qualifying | 0.923077 |
| 4 | FIFA World Cup qualification | 0.765957 |
| 3 | UEFA Nations League | 0.636364 |
| 2 | UEFA European Championship | 0.560000 |
| 6 | FIFA Confederations Cup | 0.500000 |
| 0 | Friendlies | 0.384615 |
| 5 | FIFA World Cup | 0.363636 |




Jugador
Histograma de Goles
Player
Messi
Ronaldo
Frecuencia
Goles
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
0
10
20
30
40
50


Mapa de Calor de Correlación
Messi_Apps
Ronaldo_Apps
Messi_Goals
1
0.58
0.91
0.37
0.58
1
0.68
0.88
0.91
0.68
1
0.35
0.37
0.88
0.35
1
1.0
0.9
0.8
0.7
0.6
0.5




Answers to the Statistical Analysis of Messi and Ronaldo Data

1. Are there any variables that do not provide information?

In this dataset, all variables provide relevant information for the analysis. The "Competition" variable allows us to categorize the data, while "Goals" and "Apps"/"Caps" are the fundamental quantitative metrics for evaluating player performance. However, we could consider that the "Player" column that we added for identification is merely organizational and does not provide direct statistical value to the analysis of each player separately.

2. If you had to eliminate variables, which ones would you remove and why?

It would not be advisable to eliminate any of the main variables (Competition, Goals, Apps/Caps) as they are all essential for analyzing the players' performance. These three variables are the minimum necessary to conduct a meaningful comparative analysis. If it were absolutely necessary to reduce variables, we could only dispense with the "Player" column if we keep each player's data in separate datasets.

3. Are there any variables with unusual data?

When examining the boxplots, some potentially unusual values can be identified:

- In Messi's case, his performance in friendly matches (49 goals in 54 appearances) shows notably high efficiency compared to other competitions.

- For Ronaldo, his goals in the European Championship qualifiers (36 goals in 39 appearances) represent particularly high efficiency.

These values are not necessarily errors, but rather distinctive characteristics of each player's performance in certain competitions.

4. If you compare the variables, are they all in similar ranges?

No, the "Goals" and "Apps"/"Caps" variables show different ranges:

- The "Apps"/"Caps" variable generally has higher values, ranging between 1 and 60.

- The "Goals" variable shows lower values, between 0 and 49.

This difference is logical and expected, as a player typically does not score a goal in every match they play.

5. Do you think this affects the data analysis?

Yes, the difference in scales can affect certain types of analysis. For example:

- In correlation analysis, scale differences can influence the interpretation of results.

- When creating visualizations, it may be necessary to adjust scales for clear representation.

- For fair comparisons between different competitions and players, it is more appropriate to use normalized metrics such as efficiency (goals per appearance) instead of absolute values.

6. Can you find any similar groups? What are these groups?

In the analysis, several natural groups of competitions can be identified:

1. Official international competitions:

   - FIFA World Cup
   - Copa América
   - UEFA European Championship

2. Qualifiers:

   - FIFA World Cup qualification
   - UEFA European Championship qualifying

3. Non-competitive matches:

   - Friendlies

4. Secondary competitions:

   - UEFA Nations League
   - FIFA Confederations Cup
   - CONMEBOL–UEFA Cup of Champions

These groups reflect different levels of importance and types of competition in international football, which could explain the differences in player performance.