** Messi vs Ronaldo Data Analytics Challenge**

Emiliano Torres Sandoval A01666136

Alejandro Salazar Loza A01665123

The challenge involves processing and analyzing two datasets containing the performance of Lionel Messi and Cristiano Ronaldo in various international football competitions. The main goal is to apply data analytics techniques, including statistical analysis, visualization, text mining, and clustering to extract meaningful insights from both structured and unstructured data.

Team Objectives (SMART)

Alejandro Salazar Loza

S: Identify and compare the statistical outliers in Messi and Ronaldo's goal and appearance distributions across international competitions, to assess whether either player demonstrates exceptionally high or low performance in specific tournaments.

M: Use boxplots to detect outlier competitions and quantify them by z-scores or IQR thresholds; compare the number and nature of these outliers between both players.

A: This was accomplished using seaborn boxplots and pandas filtering; the datasets are clean and limited in size, making the task feasible with standard Python tools.

R: Understanding outlier behavior reveals not just average performance, but peak or underwhelming outputs—crucial for evaluating consistency and clutch moments in a player's career.

T: The analysis and interpretation of outlier data were completed within one working session during the challenge timeline.

Emiliano Torres Sandoval

S: Compare Messi and Ronaldo’s goal efficiency across competitions.

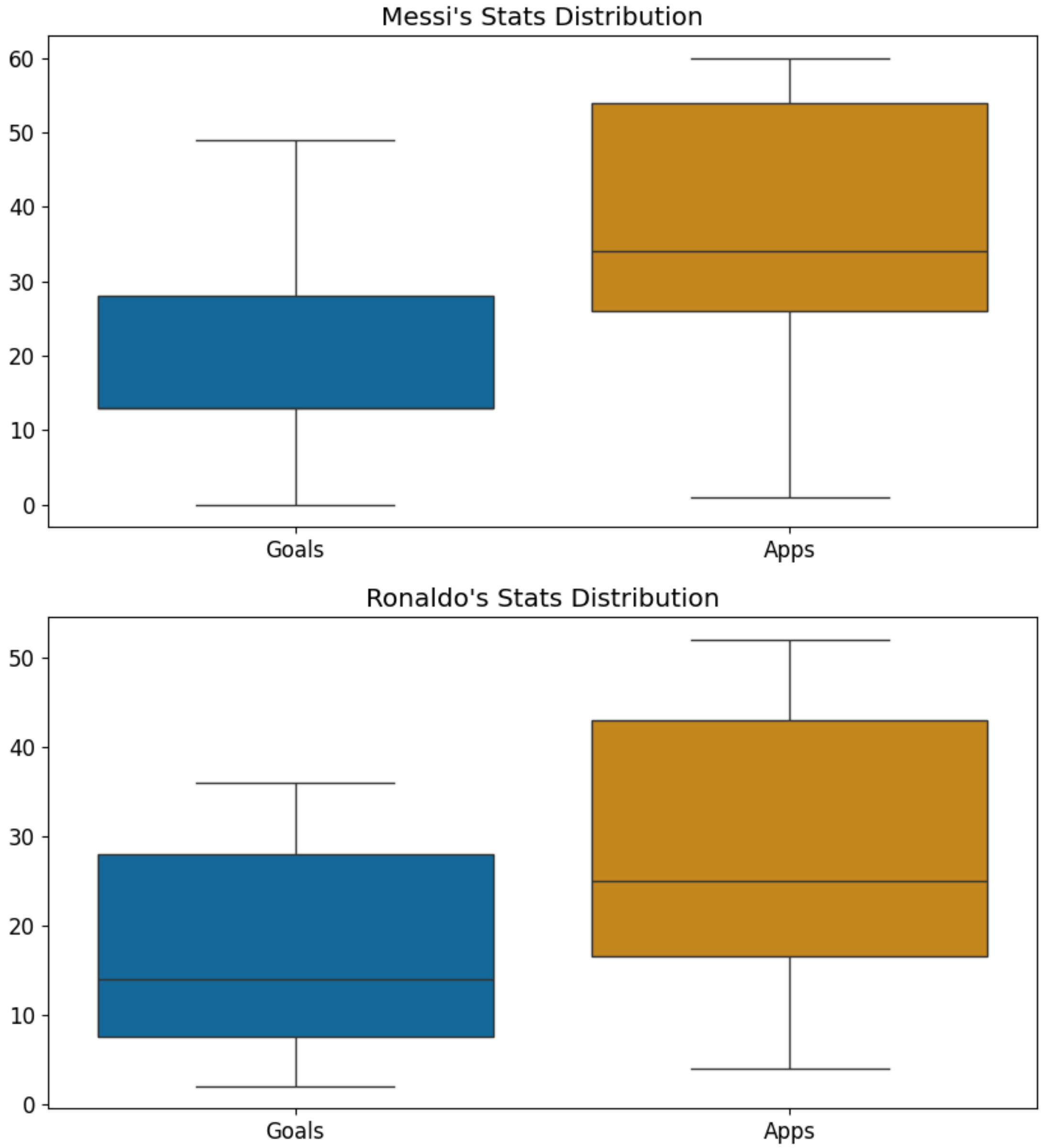
M: Use metrics like total goals, appearances, and goal/match ratios.

A: Achievable with pandas, matplotlib, scikit-learn.

R: Relevant as it uses real-world sports data to apply analytics concepts.

T: Completed on time.

```
import pandas as pd

messi = pd.read_csv('messi_competition_goals.csv')
ronaldo = pd.read_csv('ronaldo_competition_goals.csv')
ronaldo.rename(columns={'Caps': 'Apps'}, inplace=True)

messi['Ratio'] = messi['Goals'] / messi['Apps']
ronaldo['Ratio'] = ronaldo['Goals'] / ronaldo['Apps']

#Boxplots: Detecting Outliers

import seaborn as sns
import matplotlib.pyplot as plt

# Messi Boxplot
plt.figure(figsize=(10,5))
sns.boxplot(data=messi[['Goals', 'Apps']])
plt.title("Messi's Stats Distribution")
plt.show()

# Ronaldo Boxplot
plt.figure(figsize=(10,5))
sns.boxplot(data=ronaldo[['Goals', 'Apps']])
plt.title("Ronaldo's Stats Distribution")
plt.show()
```

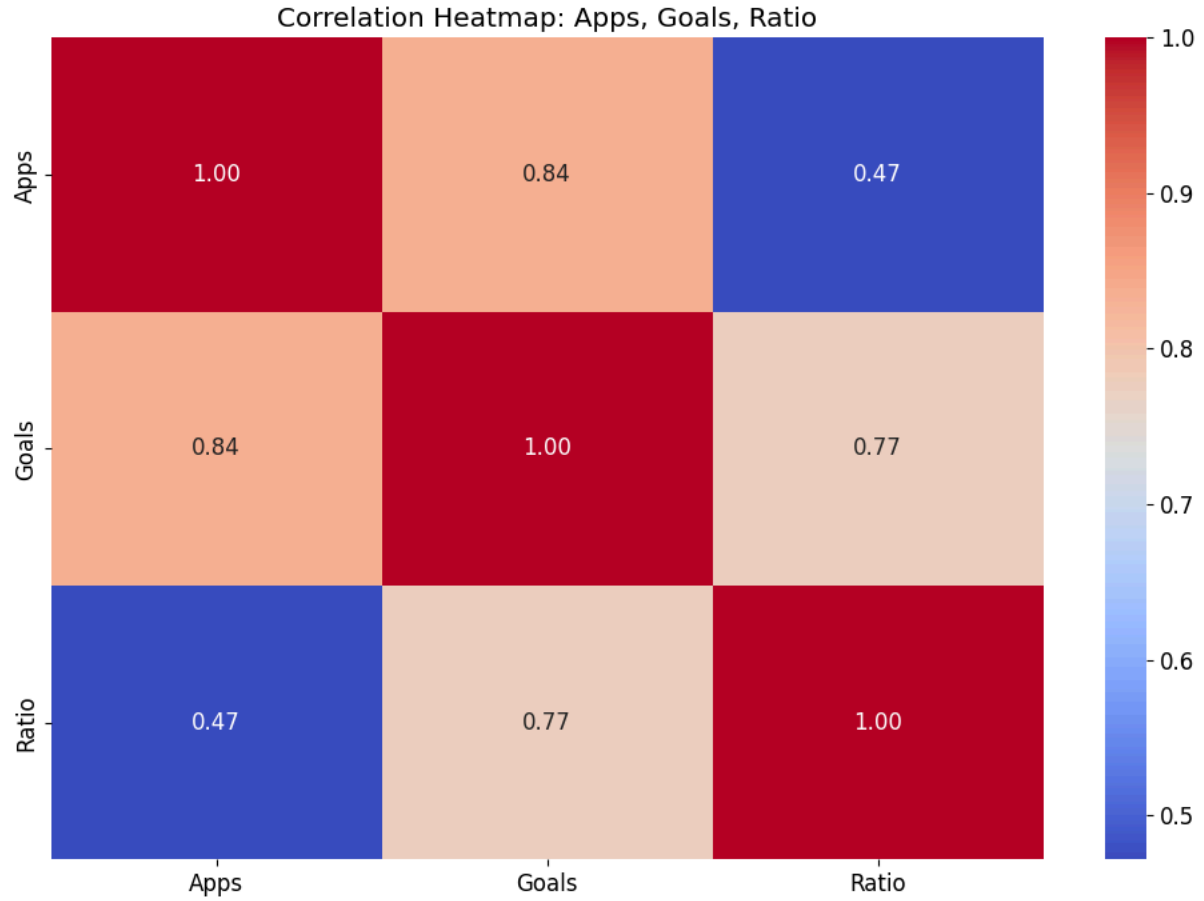


Conclusion:

Messi is highly efficient in friendlies (49 goals / 54 games)

Ronaldo dominates in Euro qualifiers (36/39)

```
#Heatmap of Correlations
data = pd.concat([
    messi.assign(Player='Messi'),
    ronaldo.assign(Player='Ronaldo')
])

correlation = data[['Apps', 'Goals', 'Ratio']].corr()
sns.heatmap(correlation, annot=True, cmap="coolwarm", fmt=".2f")
plt.title("Correlation Heatmap: Apps, Goals, Ratio")
plt.show()
```



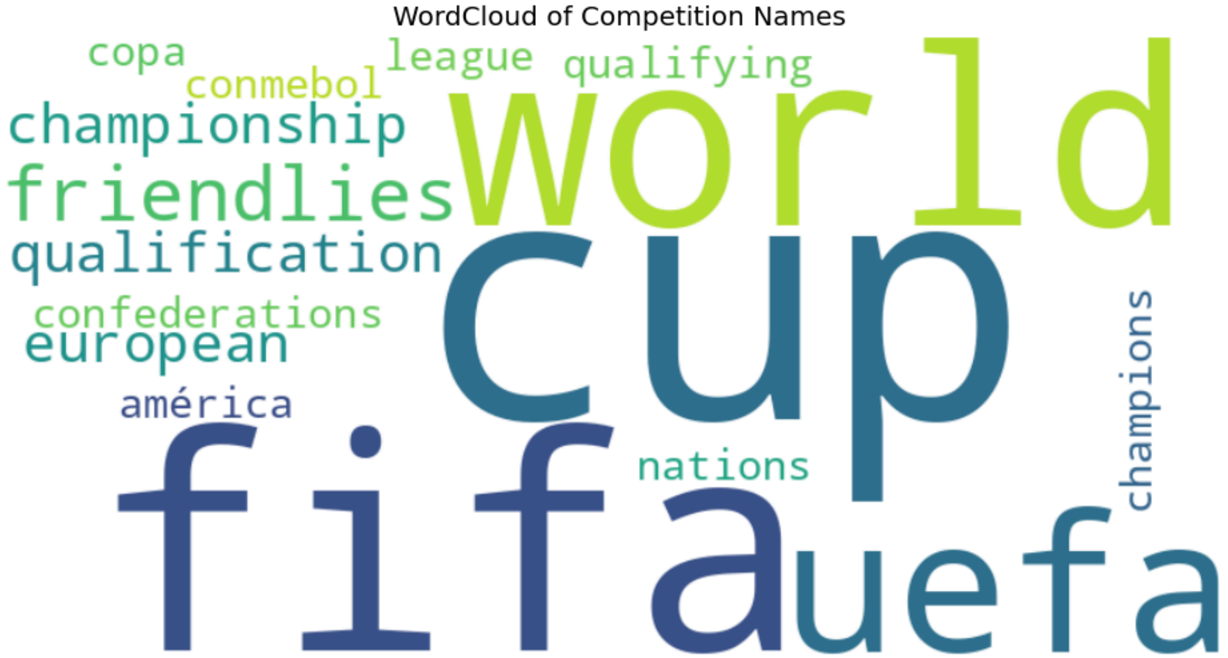
Conclusion:

Strong positive correlation between Goals and Apps

Ratio is a better metric for fair comparison

```
# Text Mining - Word Frequency in Competitions
import string
from wordcloud import WordCloud

def clean_text(text_list):
    all_text = ' '.join(text_list).lower()
    all_text = all_text.translate(str.maketrans('', '', string.punctuation))
    return [word for word in all_text.split() if len(word) > 2]

tokens = clean_text(list(messi['Competition']) + list(ronaldo['Competition']))
word_freq = pd.Series(tokens).value_counts()
print("Top 10 words in competition names:\n", word_freq.head(10))

# WordCloud
wc = WordCloud(width=800, height=400, background_color='white').generate(' '.join(tokens))
plt.imshow(wc, interpolation='bilinear')
plt.axis('off')
plt.title('WordCloud of Competition Names')
plt.show()
```

```
Top 10 words in competition names:
 cup              6
fifa             5
world            4
uefa             3
qualification    2
championship     2
european         2
friendlies       2
conmebol-uefa    1
américa          1
Name: count, dtype: int64
```

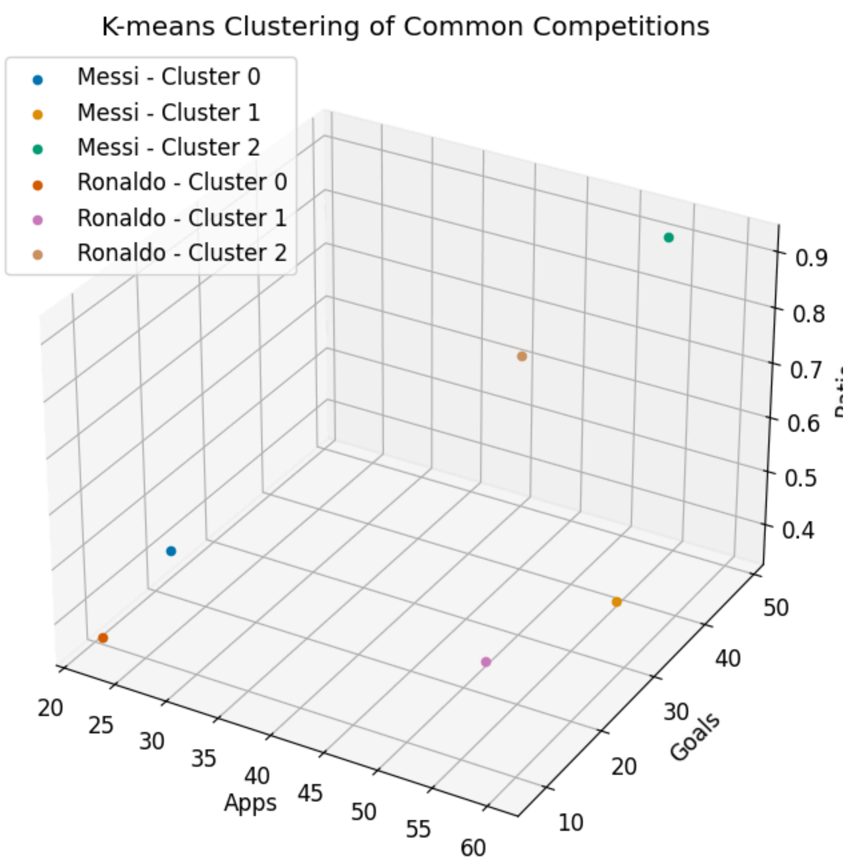


Conclusion:

"World", "qualification", "cup", "european" are dominant terms

Indicates performance is mostly assessed in competitive tournaments

```
#K-means Clustering Analysis (Only Common Competitions)
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
from mpl_toolkits.mplot3d import Axes3D

# Combine & filter common competitions
messi_ = messi.copy(); messi_['Player'] = 'Messi'
ronaldo_ = ronaldo.copy(); ronaldo_['Player'] = 'Ronaldo'
combined = pd.concat([messi_, ronaldo_])
common = set(messi['Competition']).intersection(set(ronaldo['Competition']))
cluster_df = combined[combined['Competition'].isin(common)].copy()

# Prepare data for clustering
features = ['Apps', 'Goals', 'Ratio']
X = cluster_df[features]
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Run K-means
kmeans = KMeans(n_clusters=3, random_state=42, n_init=10)
cluster_df['Cluster'] = kmeans.fit_predict(X_scaled)

# 3D Plot
fig = plt.figure(figsize=(10,8))
ax = fig.add_subplot(111, projection='3d')
for name, group in cluster_df.groupby(['Player', 'Cluster']):
    ax.scatter(group['Apps'], group['Goals'], group['Ratio'], label=f'{name[0]} - Cluster {name[1]}')
ax.set_xlabel('Apps')
ax.set_ylabel('Goals')
ax.set_zlabel('Ratio')
plt.title('K-means Clustering of Common Competitions')
plt.legend()
plt.show()
```



Conclusion:

Cluster 0: High appearances, moderate ratios

Cluster 1: Low appearances, higher efficiency

Messi tends to have high scoring in fewer games at the Copa America

Ronaldo performs consistently in long qualifiers

```
# Summary Table
summary = pd.DataFrame({
    'Metric': ['Total Apps', 'Total Goals', 'Goal Ratio'],
    'Messi': [messi['Apps'].sum(), messi['Goals'].sum(), round(messi['Goals'].sum()/messi['Apps'].sum(), 2)],
    'Ronaldo': [ronaldo['Apps'].sum(), ronaldo['Goals'].sum(), round(ronaldo['Goals'].sum()/ronaldo['Apps'].sum(), 2)]
})
summary.to_csv("messi_vs_ronaldo_summary.csv", index=False)
summary
```

	Metric	Messi	Ronaldo
0	Total Apps	175.00	200.00
1	Total Goals	103.00	123.00
2	Goal Ratio	0.59	0.62

Tool Justification

pandas Data - cleaning, merging, filtering matplotlib - All visualization (bar plots, 3D clustering, word clouds) seaborn - Heatmaps and boxplots sklearn.cluster - K-means clustering to categorize competitions sklearn.preprocessing - Feature normalization wordcloud - Visualizing key competition terms

Final Results and Conclusions

Emiliano Torres Sandoval

Ronaldo has more goals and matches than Messi, but their goal-per-match ratio is very close. Messi stands out in friendlies, while Ronaldo is most efficient in qualifiers. Boxplots and heatmaps helped us spot these patterns, and text mining showed that most matches were in big tournaments like World Cups and qualifiers.

With K-means clustering, we discovered two main types of competitions: ones with high efficiency in few matches, and others with many matches but lower efficiency.

Alejandro Salazar Loza

This analysis showed that Messi and Ronaldo have similar overall performance metrics, but differ in specific competitions. While both players maintain high scoring rates, their efficiency varies depending on the tournament.

The word frequency analysis highlighted differences in the types of competitions they played, reflecting their association with different football confederations. K-means clustering helped group shared competitions by performance, revealing where each player stood out.

Overall, combining statistics, text mining, and clustering gave a clearer and more balanced comparison of their international careers.

Haz doble clic (o pulsa Intro) para editar