Exploratory Data Analysis

T20 Internationals Dataset (2005-2023)

using pandas, matplotlib and seaborn

Dataset is available on **kaggle**

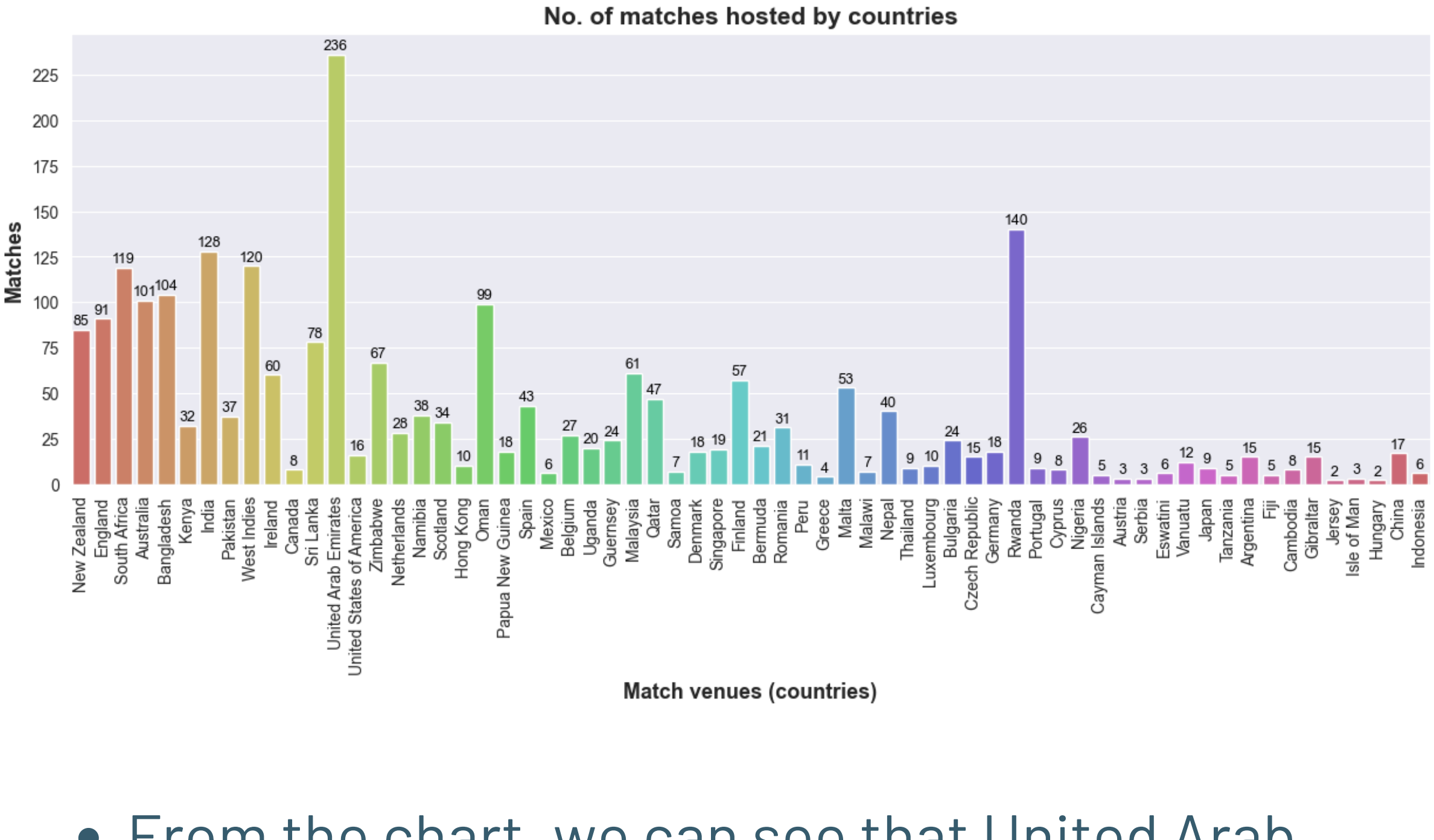
ali-bin-kashif

1

Which country has hosted the most matches?

```
# Setting figure size
plt.figure( figsize = (15,5) )

# Setting figure style or theme
sns.set_style('darkgrid')

# Plotting chart
ax = sns.countplot(data=matches_data, x='match_venue_(country)', hue='match_venue_(country)', palette='hls')

# Setting labels with each bar count
for container in ax.containers:
    ax.bar_label(container, label_type="edge", padding=1, size=9, color="black")

# Customizations
ax.tick_params('x', rotation=90)
ax.set_xlabel("Match venues (countries)", fontweight = 'bold', fontsize = 13)
ax.set_ylabel("Matches", fontweight = 'bold', fontsize = 13)
ax.set_title("No. of matches hosted by countries", fontweight = 'bold', fontsize = 15)
ax.set_yticks([0,25,50,75,100,125,150,175,200,225])

# Show chart
plt.show()
```

1



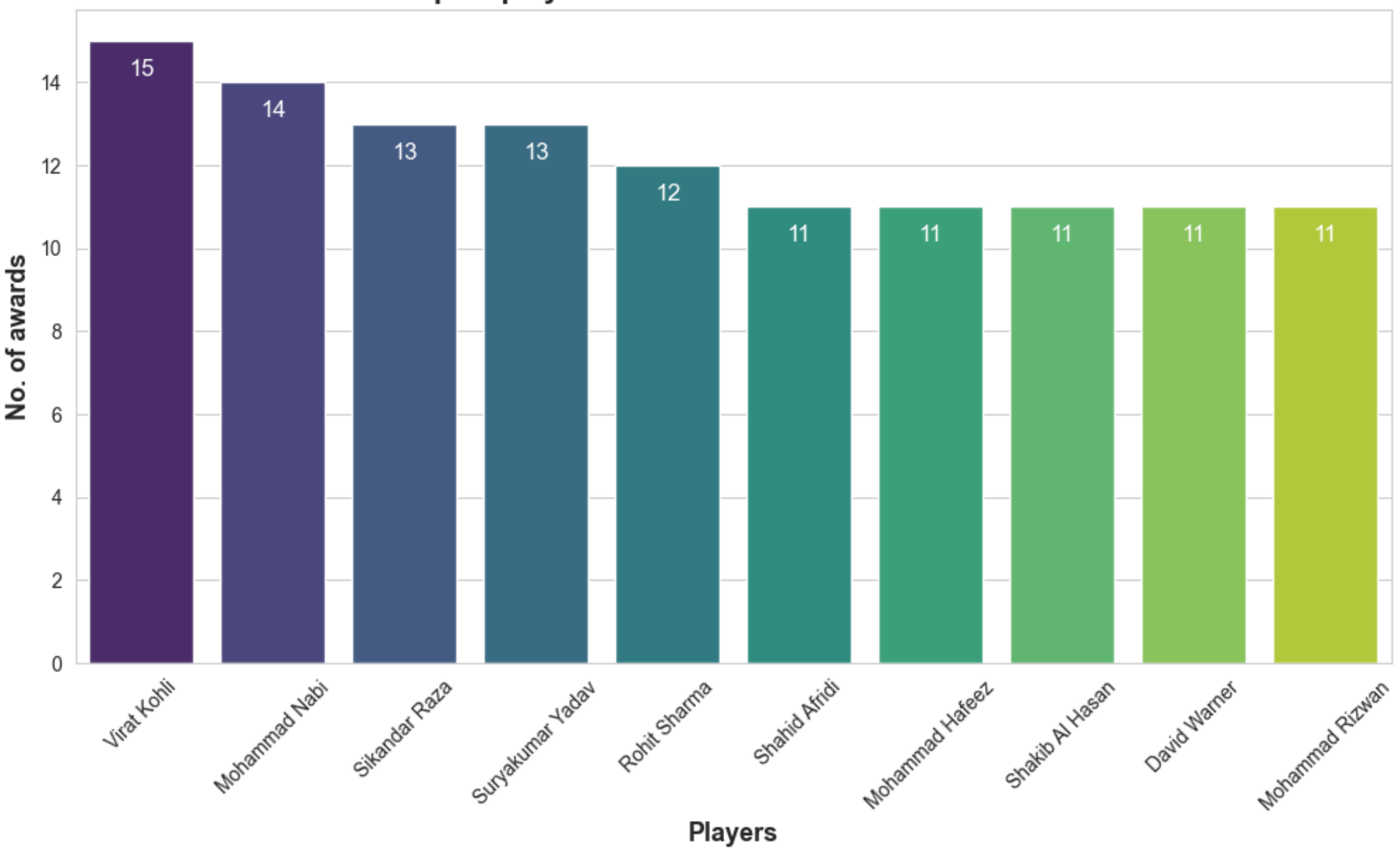
- From the chart, we can see that United Arab Emirates has hosted the most number of T20I matches till now i.e 236

2

Players with most man-of-the-match awards?

```
# Grouping the dataframe by 'mom_player' columns and aggregating the rows by co
unting the no. of awards
motm_players = matches_data.groupby('mom_player')[['match_id']].count().rename
(columns={'match_id' : 'awards'})

# Merging the result dataframe with player info on each unique player id
motm_with_names = motm_players.merge(players_info, left_index=True, right_on='p
layer_id')

# Setting size and theme
plt.figure(figsize=(12,6))
sns.set_style('whitegrid')

# Plotting by 10 largest no. of awards
ax = sns.barplot(x='player_name', y='awards', data=motm_with_names.nlargest(1
0,'awards'), hue='player_name', palette='viridis')

# Customizations
ax.tick_params('x', rotation=45)
ax.set_title("Top 10 players with most MotM awards in T20Is", fontweight = 'bol
d', fontsize = 15)
ax.set_xlabel('Players', fontweight = 'bold', fontsize = 13)
ax.set_ylabel('No. of awards', fontweight = 'bold', fontsize = 13)

# Setting bar label on each container
for container in ax.containers:
    ax.bar_label(container, padding=-20, size=12, color='white')
```

2



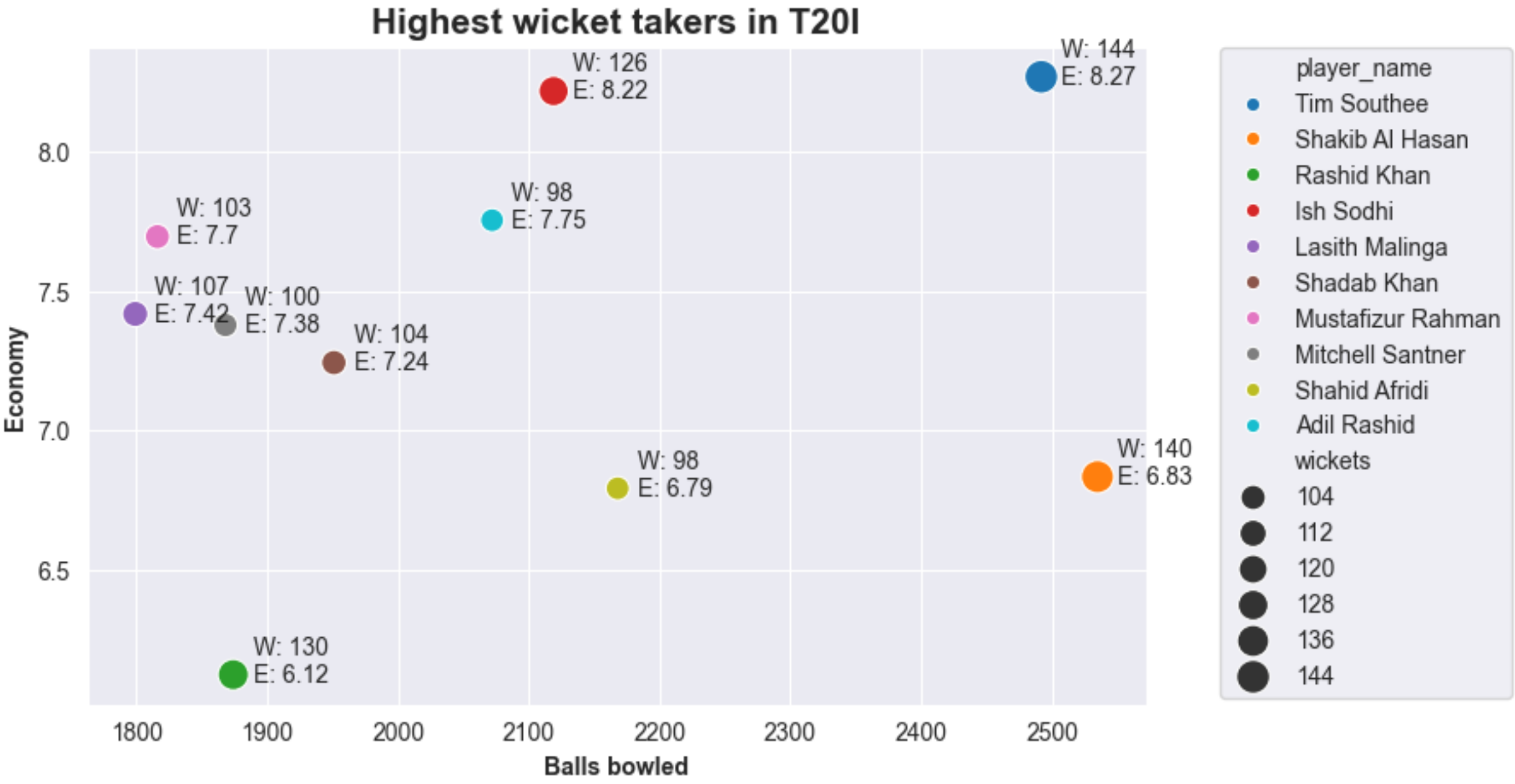
- Virat Kholi has won the most Man-of-the-Match awards followed by Muhammad Nabi and Sikandar Raza.

3

Highest wicket takers in T20 Internationals?

```python
# Grouping data by bowler ids and aggregating balls, wickets and economy
bowlers_record = bowling_data.groupby('bowler_id')[['balls', 'wickets', 'economy']].agg(
    {'balls' : 'sum',
     'wickets' : 'sum',
     'economy' : 'mean'})

# Extracting top 10 bowlers by wickets
top_10_bowlers = bowlers_record.nlargest(10, 'wickets')

# Merging dataframe with players info
top_10_bowlers_names = top_10_bowlers.merge(players_info, left_index=True, right_on='player_id')

# Setting size and theme
plt.figure(figsize=(8,5))
sns.set_style('darkgrid')

# Plotting with balls bowled on x-axis and economy on y-axis
ax = sns.scatterplot(x='balls', y='economy', data=top_10_bowlers_names, size='wickets', hue='player_name', sizes=(100,200))

# Customizing legend
plt.legend(bbox_to_anchor=(1.07, 1), loc='upper left', borderaxespad=0)

# Annotating points on the plot with wickets and economy
for lab,row in top_10_bowlers_names.iterrows():
    ax.annotate(f"W: { int(row['wickets']) }\nE: { round(row['economy'],2) }", xy=(row['balls']+15, row['economy']-0.025))

# Customizing the chart
ax.set_title("Highest wicket takers in T20I", fontweight = 'bold', fontsize=15)
ax.set_xlabel("Balls bowled", fontweight = 'bold')
ax.set_ylabel("Economy", fontweight = 'bold')

plt.show()
```

3



- Rashid Khan is the most economical bowler with 130 wickets.
- Tim Southee has the highest wickets uptill now but also with a high economy.

4

Highest run scorers in T20 Internationals?

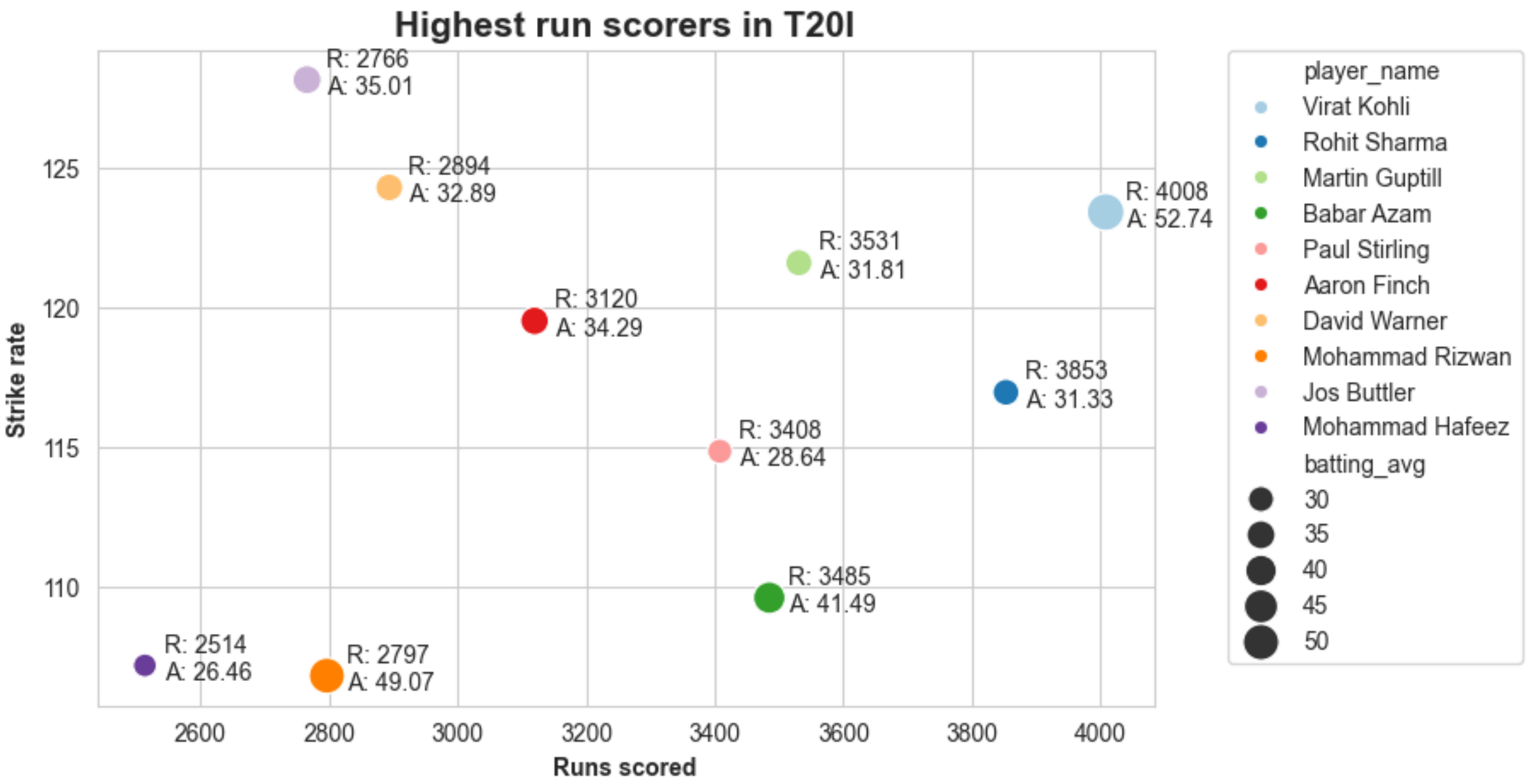
```
# Grouping each batting record by the batsmen and aggregating their performances metrics i.e runs, balls, fours, sixes, strikeRate
batsman_records = batting_data.pivot_table(values=['runs', 'balls', 'fours', 'sixes', 'strikeRate', 'isOut'], index='batsman', aggfunc={'runs': 'sum', 'balls' : 'sum', 'fours' : 'sum', 'sixes' : 'sum', 'strikeRate' : 'mean', 'isOut' : 'sum'})

# Filtering top 10 by runs
top_10_batsman = batsman_records.nlargest(10,'runs').reset_index()

# Calculating batting average i.e dividing runs by no. of times got out
top_10_batsman['batting_avg'] = top_10_batsman['runs'] / top_10_batsman['isOut']

top_10_batsman_names = top_10_batsman.merge(players_info, left_on='batsman', right_on='player_id')
```

4

Highest run scorers in T20 Internationals?

```
# Setting size and theme
plt.figure(figsize=(8,5))
sns.set_style('whitegrid')

# Plotting with balls bowled on x-axis and economy on y-axis
ax = sns.scatterplot(x='runs', y='strikeRate', data=top_10_batsman_names, size='batting_avg',
                     hue='player_name', sizes=(100,250), palette='Paired')

# Customizing legend
plt.legend(bbox_to_anchor=(1.07, 1), loc='upper left', borderaxespad=0)

# Annotating points on the plot with wickets and economy
for lab,row in top_10_batsman_names.iterrows():
    ax.annotate(f"R: { int(row['runs']) }\nA: {round(row['batting_avg'], 2)}", xy=(row['runs']+30, row['strikeRate']-0.5))

# Customizing the chart
ax.set_title("Highest run scorers in T20I", fontweight = 'bold', fontsize=15)
ax.set_xlabel("Runs scored", fontweight = 'bold')
ax.set_ylabel("Strike rate", fontweight = 'bold')

plt.show()
```

4



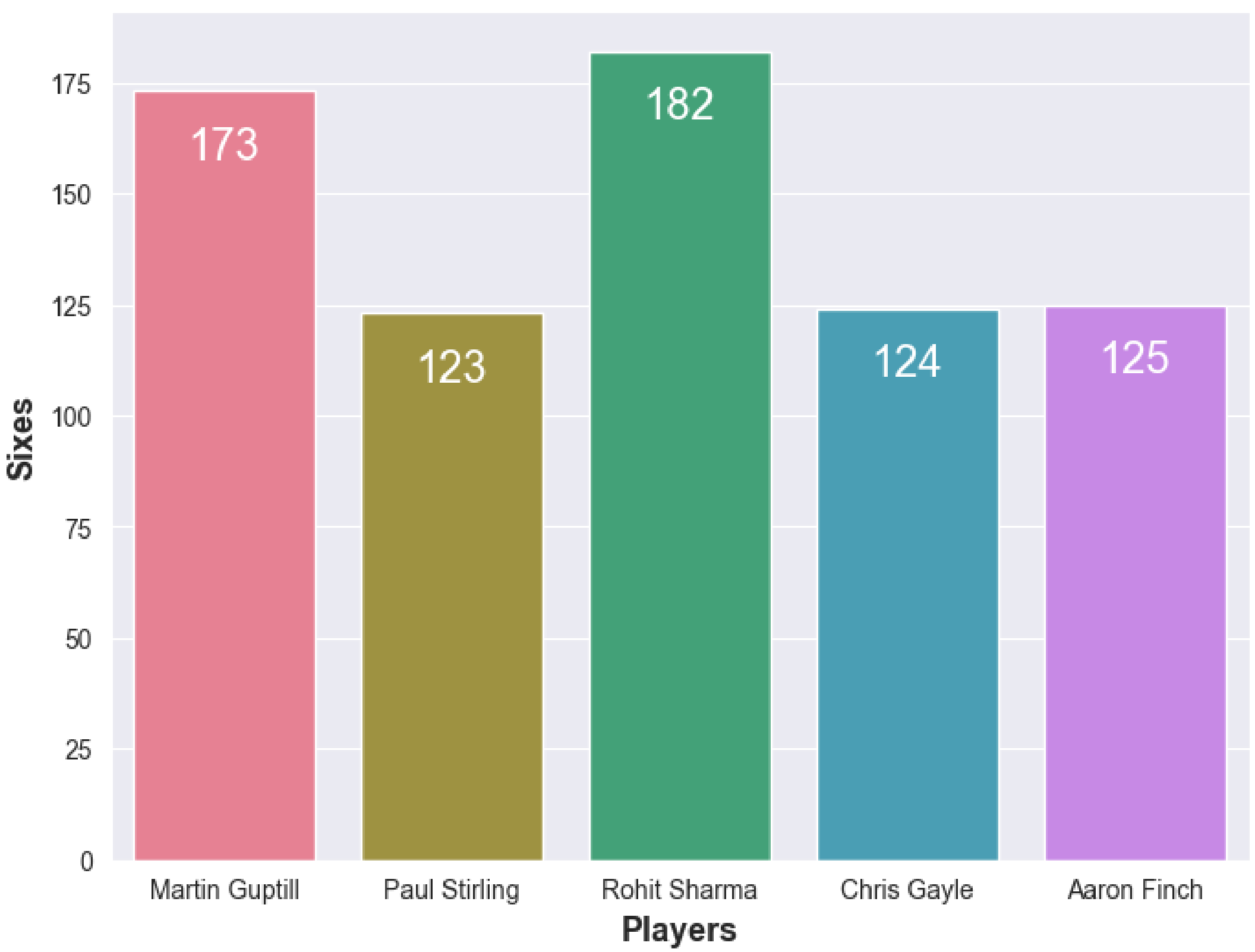
- Highest Run scorer is Virat Kohli with 4008 runs with average of 52.74
- Jos Buttler has the highest strike rate.
- Babar Azam has scored 3485 runs with average of 41.49.

5

Who has hitted most sixes?

```
# Filter top 5 players with most sixes and merging it with player info dataframe
top_5_sixes = batsman_records.nlargest(5,'sixes')
top_5_sixes_names = top_5_sixes.merge(players_info, left_index=True, right_on='player_id')

# Setting size and theme
plt.figure(figsize=(8,6))
sns.set_style('darkgrid')

# Plotting values and setting title and labels
ax = sns.barplot(x='player_name', y='sixes', data=top_5_sixes_names.sample(5), hue='player_name', palette='husl')
ax.set_title("Most sixes in T20Is", fontweight = 'bold', fontsize=15)
ax.set_xlabel("Players", fontweight = 'bold', fontsize=13)
ax.set_ylabel("Sixes", fontweight = 'bold', fontsize=13)

# Setting label on each bar
for container in ax.containers:
    ax.bar_label(container, padding=-30, fontsize = 17, color='white')

```

5



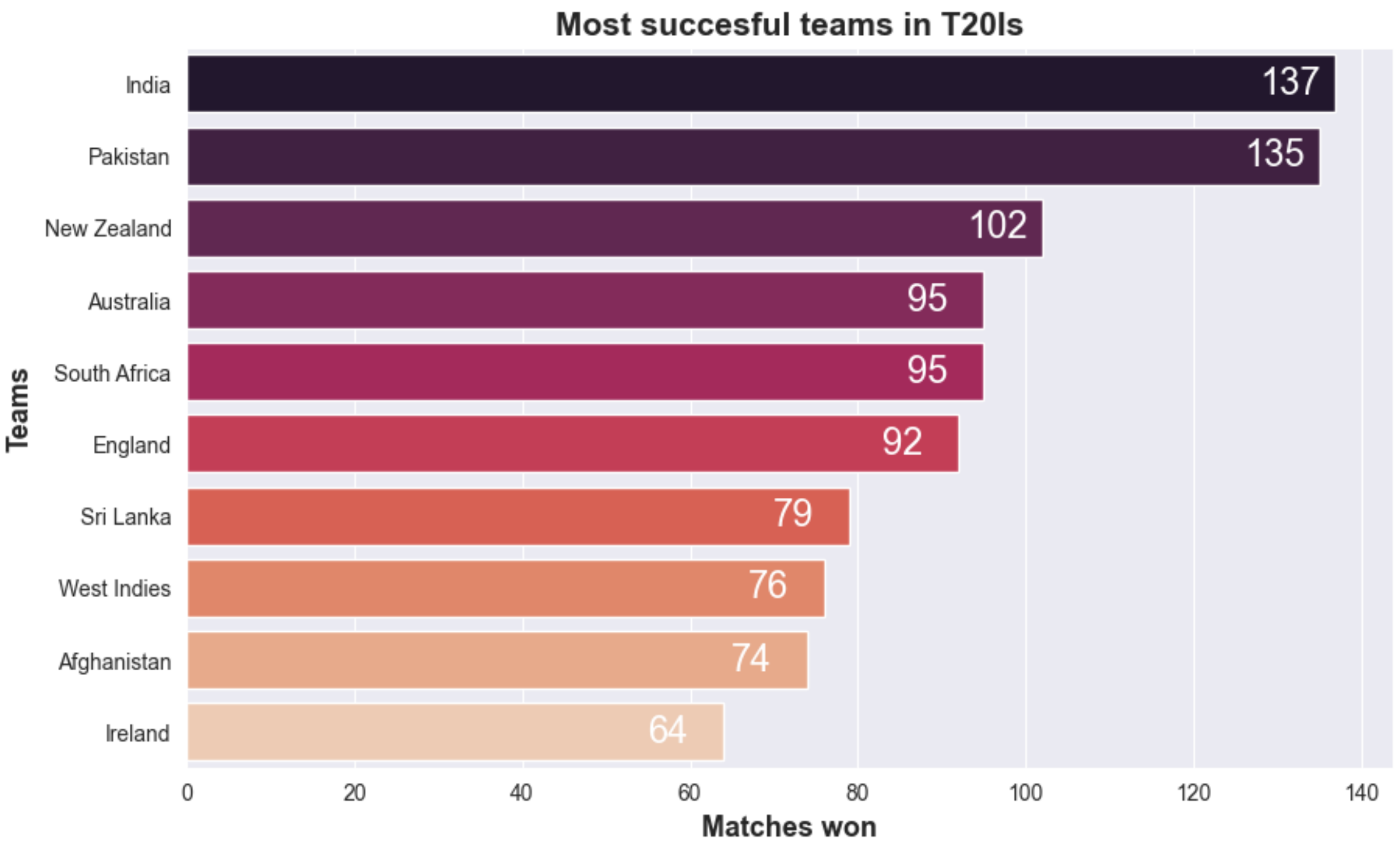
- Rohit Sharma has hitted most sixes till now i.e 182 in T20Is.

6

Most successful teams in T20Is?

```python
# Grouping by match winners and counting them
matches_win_countries = matches_data.pivot_table(values='match_id', index='match_winner', aggfunc='count').reset_index()

# Renaming and extracting 10 largest
top_10_countries = matches_win_countries.rename(columns={'match_id':'matches_won', 'match_winner' : 'country'}).nlargest(10,'matches_won')

plt.figure(figsize=(10,6))
sns.set_style('darkgrid')

# Plotting
ax =sns.barplot(x='matches_won', y='country', data=top_10_countries, label="Won", hue='country', palette='rocket')

# Customizaitons
ax.set_title("Most succesful teams in T20Is", fontweight = 'bold', fontsize=15)
ax.set_xlabel("Matches won", fontweight = 'bold', fontsize=13)
ax.set_ylabel("Teams", fontweight = 'bold', fontsize=13)

# Setting label on each bar
for container in ax.containers:
    ax.bar_label(container, padding=-35, fontsize = 17, color='white')

plt.show()
```

6



- Pakistan and India are amongst the most successful teams in terms of matches won.

Thank you, thats all uptill now!