SOTC ASSESSMENT

Submitted By: Hasnain Iftikhar

Assessment Requirements (Any 2)

- Participation Analysis
- Results Analysis
- Prediction
- Athlete Profile Analysis
- Age Analysis

Prediction (Download Code's HTML File) | (IPYNB File)

- Exploratory Data Analytics
  - Summary
  - Columns Types
  - Data Filtering
  - Statistical Summary
  - Missing Values Check

In [7]:
```
print("\033[1mThe sum of null values in each column: \033[0m")
print(df.isnull().sum())
print("\n")
print("\033[1mHere is the statistical summary of provided data:\033[0m")
df.describe().T
```

```
The sum of null values in each column:
Competition              0
Sport                    0
Gender                   0
Rank                     0
TeamMembers              0
Nationality          15926
Age                   9012
HostContinent          780
Participants             0
Countries                0
Continents               0
RankWithinCountry        0
dtype: int64
```

Here is the statistical summary of provided data:

Out[7]:

	count	mean	std	min	25%	50%	75%	max
Rank	180597.0	13.751070	14.105322	1.0	4.0	9.0	17.0	70.0
Age	171585.0	24.162369	5.487487	9.0	20.0	23.0	27.0	70.0
Participants	180597.0	30.934478	29.531468	1.0	12.0	20.0	40.0	462.0
Countries	180597.0	21.867384	16.639597	1.0	11.0	17.0	27.0	132.0
Continents	180597.0	2.747787	2.144004	1.0	1.0	1.0	5.0	6.0
RankWithinCountry	180597.0	1.239428	0.628577	0.0	1.0	1.0	1.0	17.0

Prediction (Download Code's HTML File) | (IPYNB File)

- Missing Values Treatment
  - Tried first by filling missing values with statistics (numerical columns) and "unknown" string (categorical variables)
  - Best model accuracy was obtained by dropping them, so decided to go with dropna() function

In [8]:
```
# Treating Missing Values
df.dropna(inplace=True)
df.head()
```

Out[8]:

	Competition	Sport	Gender	Rank	TeamMembers	Nationality	Age	HostContinent	Participants	Countries	Continents	RankWithinCountry
0	Olympic Qualification Tournament - Asia	Archery	Men	4	No	IRI	18.0	Asia	49	21	1	1
1	Olympic Qualification Tournament - Asia	Archery	Men	5	No	BAN	21.0	Asia	49	21	1	1
2	Olympic Qualification Tournament - Asia	Archery	Men	6	No	KSA	25.0	Asia	49	21	1	1
3	Olympic Qualification Tournament - Asia	Archery	Men	7	No	IND	39.0	Asia	49	21	1	2
4	Olympic Qualification Tournament - Asia	Archery	Men	8	No	IRI	28.0	Asia	49	21	1	2

In [9]:
```
print("\033[1mThe sum of null values in each column: \033[0m")
print(df.isnull().sum())
print("\n")
print("The number of rows in provided set are: " +
      "\033[1m" + str(df.shape[0]) + "\033[0m" +
      "\nThe number of columns in provided dataset are: \033[1m" + str(df.shape[1]) + "\033[0m" + "\n\n")
print("\033[1mHere is the statistical summary of provided data:\033[0m")
print(df.describe().T)
```

```
The sum of null values in each column:
Competition          0
Sport                0
Gender               0
Rank                 0
TeamMembers          0
Nationality          0
Age                  0
HostContinent        0
Participants         0
Countries            0
Continents           0
RankWithinCountry    0
dtype: int64
```

Prediction (Download Code's HTML File) | (IPYNB File)

- Categorical to Numerical
  - Converted all the categorical variables to numerical variables using one-hot-encoding to use them as features during model training

In [10]:
```
# One-Hot-Encoding to change categorical variables to numericals

df = df.join(pd.get_dummies(df.Competition))
df = df.drop("Competition", axis=1)

df = df.join(pd.get_dummies(df.Sport))
df = df.drop("Sport", axis=1)

df = df.join(pd.get_dummies(df.Gender))
df = df.drop("Gender", axis=1)

df = df.join(pd.get_dummies(df.TeamMembers))
df = df.drop("TeamMembers", axis=1)

df = df.join(pd.get_dummies(df.HostContinent))
df = df.drop("HostContinent", axis=1)

df = df.join(pd.get_dummies(df.Nationality))
df = df.drop("Nationality", axis=1)
```

In [11]:
```
print("The number of rows after one-hot encoding are: " +
      "\033[1m" + str(df.shape[0]) + "\033[0m" +
      "\nThe number of columns after one-hot encoding are: \033[1m" + str(df.shape[1]) + "\033[0m" + "\n\n")
df.head()
```

```
The number of rows after one-hot encoding are: 156110
The number of columns after one-hot encoding are: 152
```

Prediction (Download Code's HTML File) | (IPYNB File)

- Model Building
  - Train Test Splitting of Data
  - Model Training
  - Model Accuracy Check (~80%)

In [13]:
```
# Model Training

from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor

X, y = df.iloc[:, :-1], df.iloc[:, -1]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20, random_state=42)

scaler = StandardScaler()

X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

rf = RandomForestRegressor()
rf.fit(X_train_scaled, y_train)
```

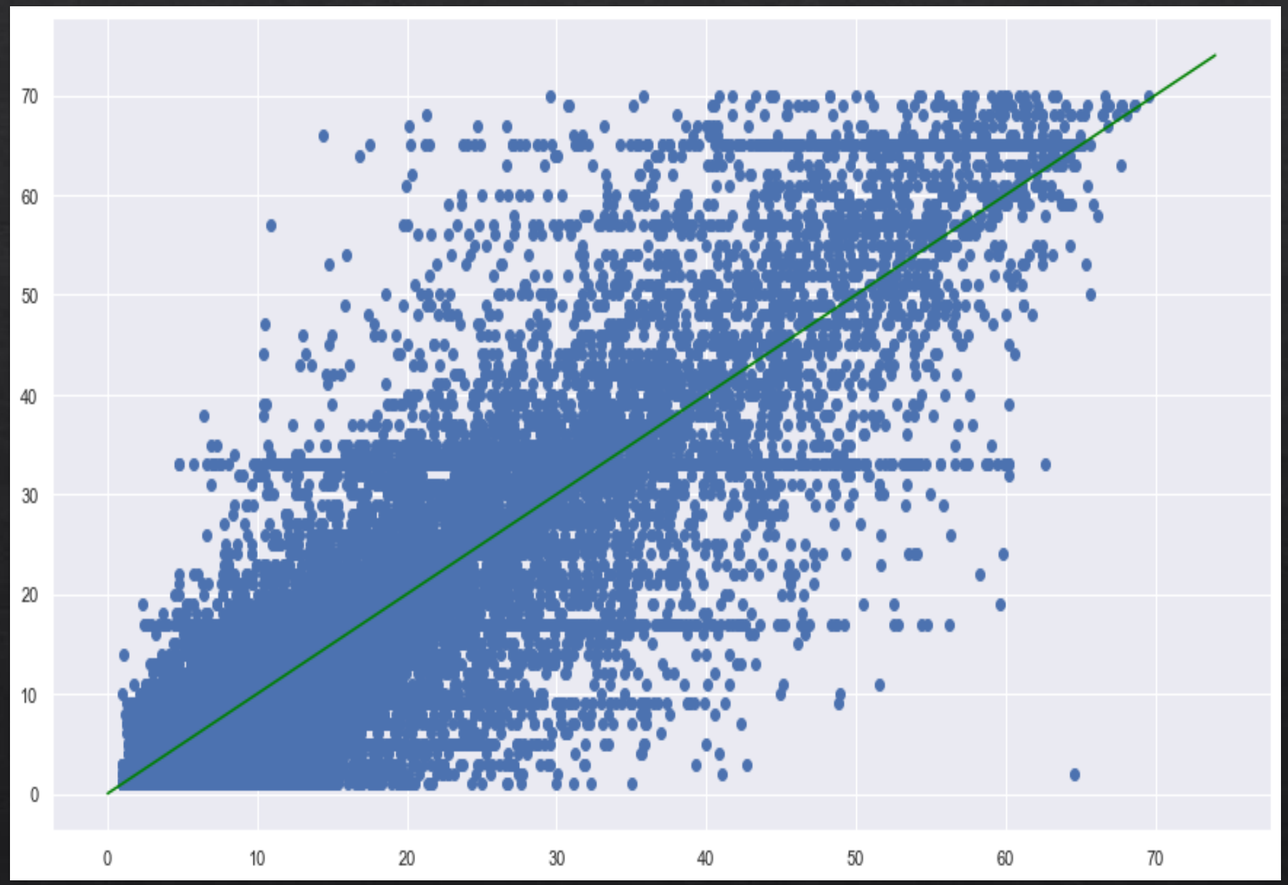
Out[13]:
```
RandomForestRegressor()
```

In [14]:
```
y_pred = rf.predict(X_test_scaled)
print("The accuracy of built model is: " +
      "\033[1m" + str(round(rf.score(X_test_scaled, y_test)*100,2)) + "%\033[0m")
```

```
The accuracy of built model is: 79.56%
```

Prediction (Download Code's HTML File) | (IPYNB File)

- Predicted Ranks Plotting



Prediction (Download Code's HTML File) | (IPYNB File)

- Results on Testing Data

In [16]:
```
print("\033[1mTesting Dataset \033[0m")
X_new_scaled = scaler.transform([X_test.iloc[3]])
print("\nTest_A\nThe actual Rank is: " + str(y_test.iloc[3]))
print("The predicted Rank is: " + str(rf.predict(X_new_scaled)))
X_new_scaled = scaler.transform([X_test.iloc[0]])
print("\nTest_B\nThe actual Rank is: " + str(y_test.iloc[0]))
print("The predicted Rank is: " + str(rf.predict(X_new_scaled)))
X_new_scaled = scaler.transform([X_test.iloc[7]])
print("\nTest_C\nThe actual Rank is: " + str(y_test.iloc[7]))
print("The predicted Rank is: " + str(rf.predict(X_new_scaled)))
```

Testing Dataset

```
Test_A
The actual Rank is: 3
The predicted Rank is: [3.87666667]

Test_B
The actual Rank is: 3
The predicted Rank is: [4.81]

Test_C
The actual Rank is: 9
The predicted Rank is: [10.89]
```

Power BI Dashboard (Download PowerBI File)

- Participation Analysis
- Resulted Ranks Analysis
- Players Age Analysis

Note: Kindly download the attached PowerBI file to check the analysis. In next slides, you can find the exported report as well, but to experience the full analysis like tooltips and sliders, use PowerBI file as some components do not export in same manner in pdf.

SPORTS MAJOR EVENTS ANALYSIS

2001 2023

Summary:

Analysis is performed to explore National and International trends across multiple sports regarding players participation in world-renowned competitions, their resulted ranks across different sports disciplines, their peak performances in different ages and a lot more. Dived into the details and understood the various factors that influence the exciting world of competitive athletics.

165K Total Participants	2619 KSA Participants	124 KSA - Women Participants	17.24 KSA - Average Rank
26 KSA Participants - Average Age	24 Globally - Medalists Average Age	286 KSA Medals Count	11% KSA - Medalists Percentage

PAGES

- Participation Anaysis
- Rank Analysis
- Age Analysis

Designed & Developed by **Hasnain Iftikhar**

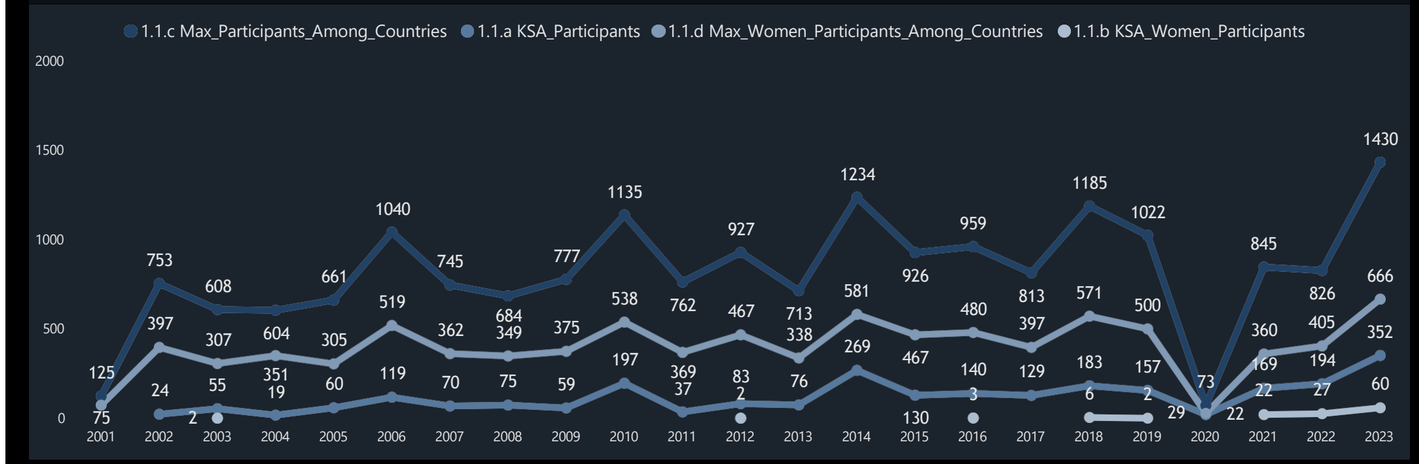
SPORTS MAJOR EVENTS ANALYSIS

CompetitionSet: All

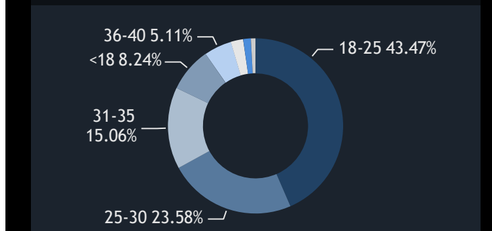
Sports: All

Year: 2023

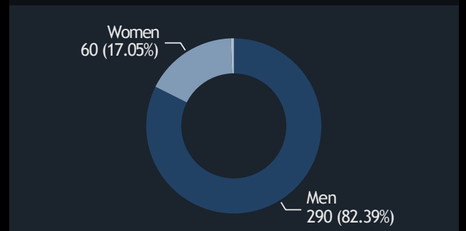
International Sports Landscape: KSA Participation, Global Trends, and Women's Involvement Over the Years



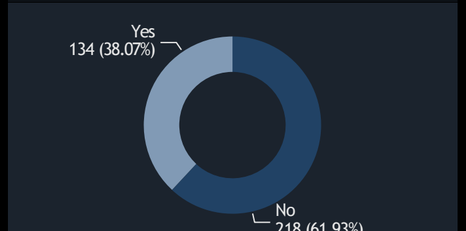
KSA - Age Group



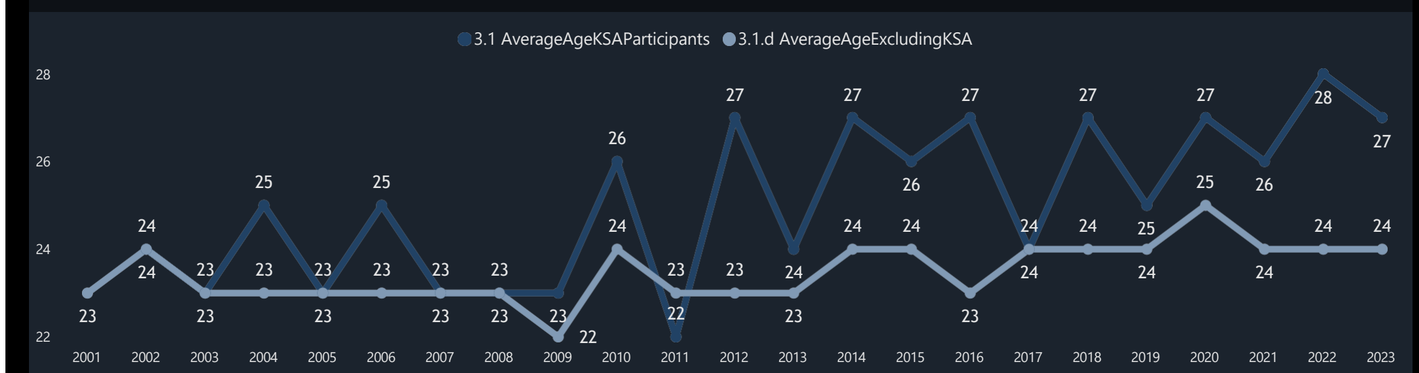
KSA - Gender Distribution



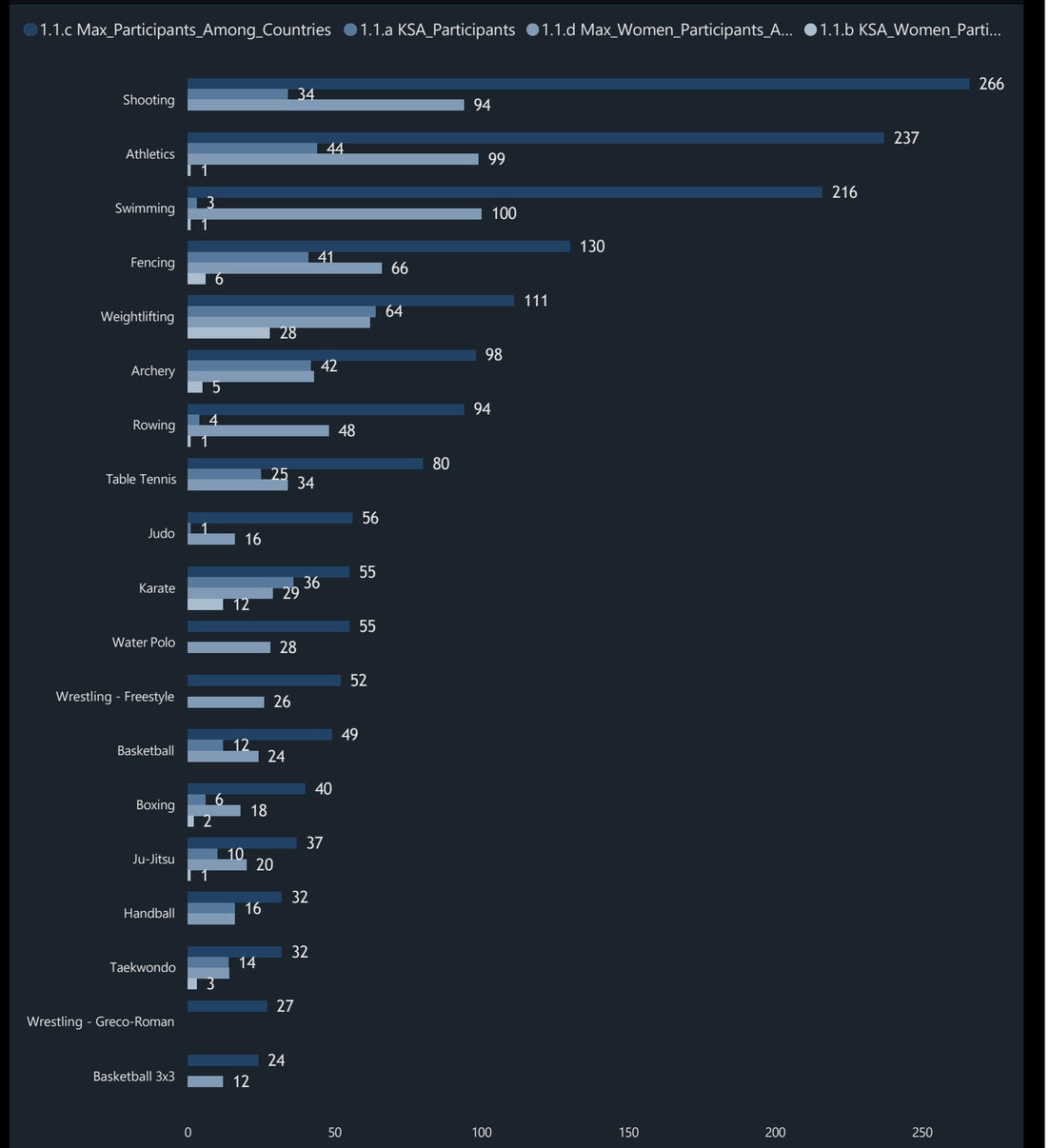
KSA - Team (Yes/No) Distribution



KSA vs. World Participants Average Age Over the Years



Global Sports Trends: KSA vs. World Participation Across Multiple Sports



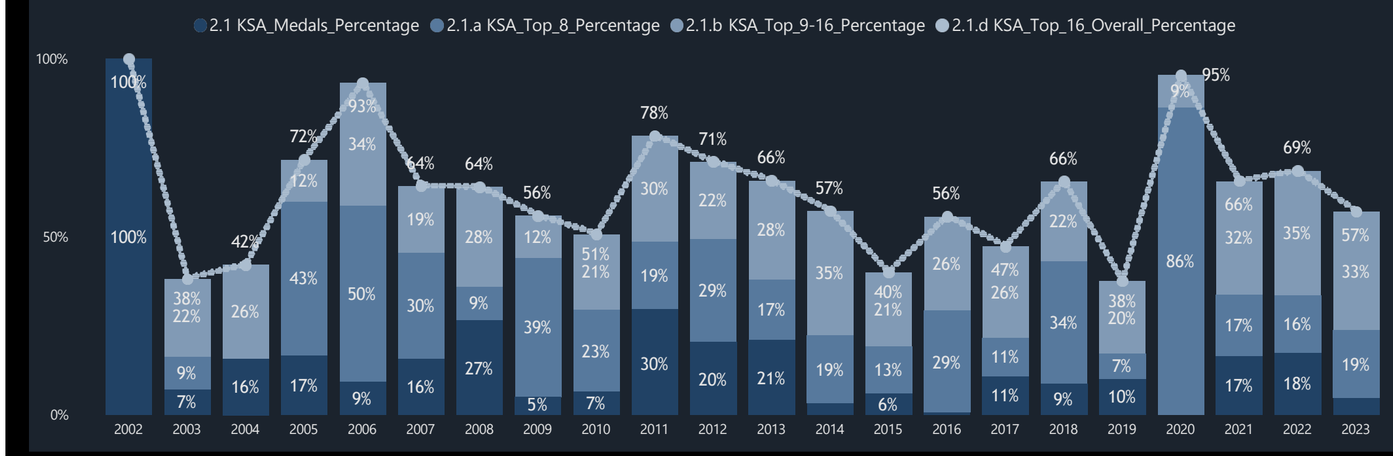
SPORTS MAJOR EVENTS ANALYSIS

CompetitionSet: All

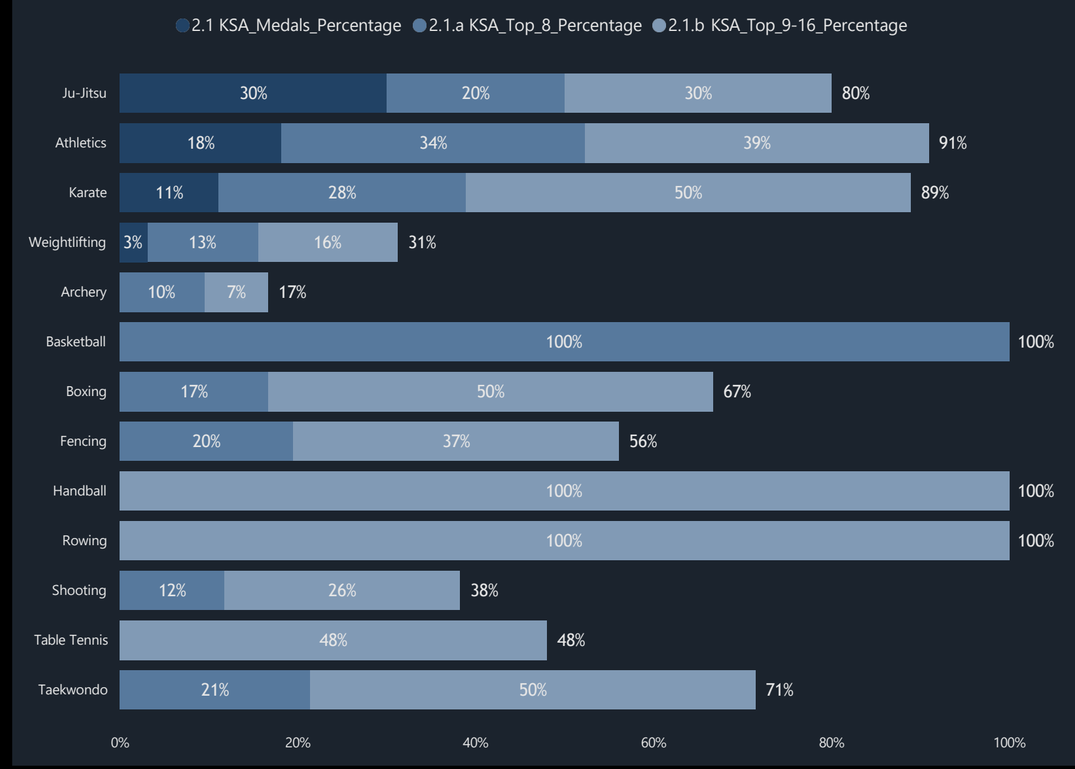
Sports: All

Year: 2023

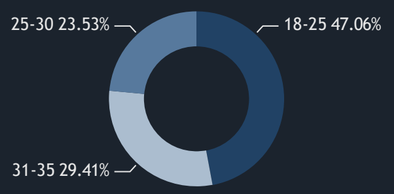
Performance Metrics: KSA Medals, Top 8, & Top 16 Percentage Over the Years



KSA Performance: Resulted Ranks Percentage Across Multiple Sports



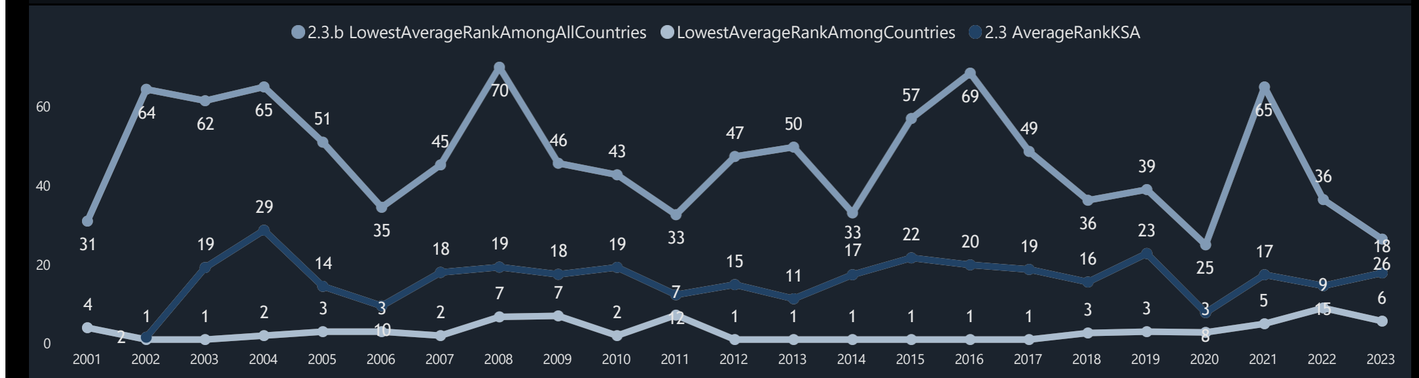
KSA Medalists - Age Group



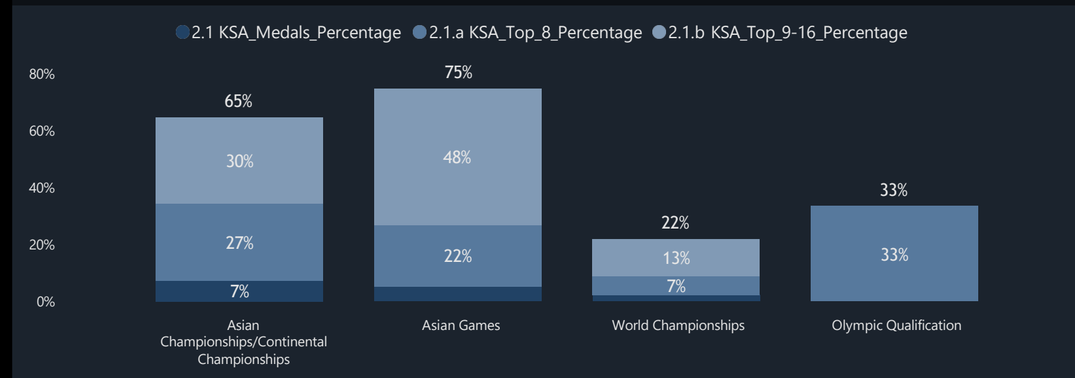
KSA Medalists - Gender Distribution

KSA Medalists - Team (Yes/No)

Performance Metrics: Highest & Lowest Average Rank vs KSA Average Rank Over the Years



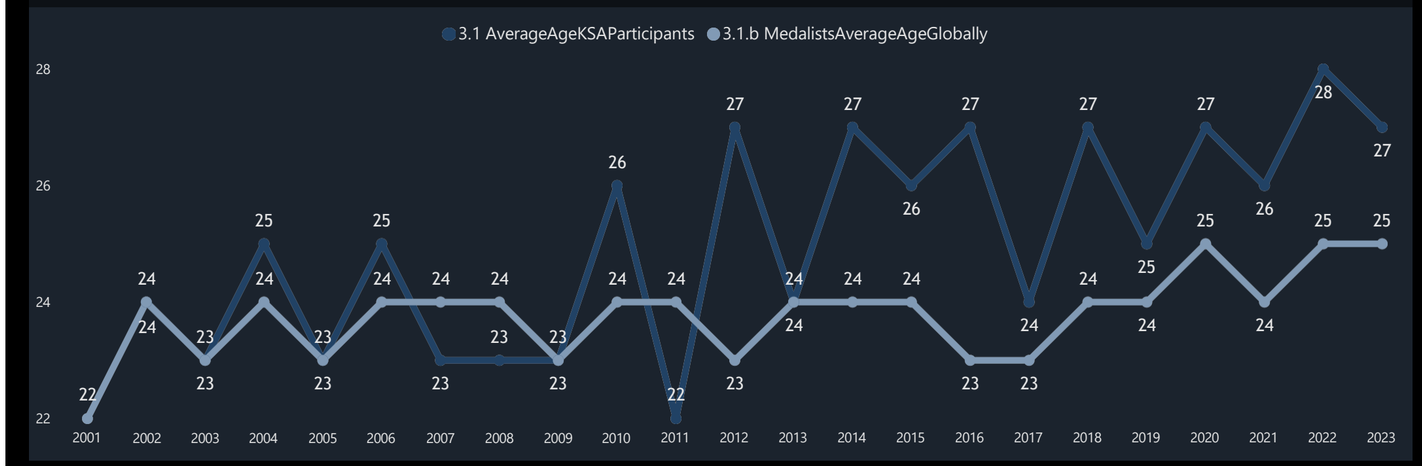
KSA Resulted Ranks Percentage Across Multiple Competitions



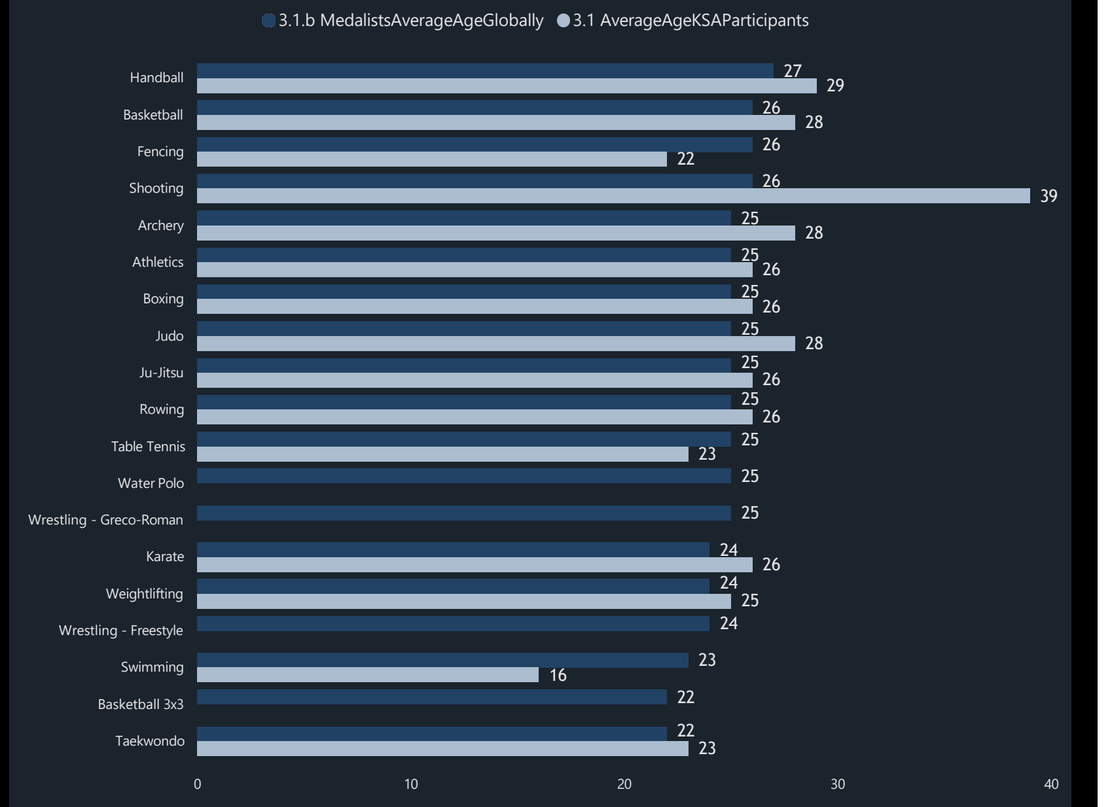
SPORTS MAJOR EVENTS ANALYSIS

CompetitionSet: All | Sports: All | Year: 2023

Global Medalists Average Age vs. KSA Participants Age Over the Years



Global Medalists Average Age vs. KSA Participants Age Across Multiple Sports



Insights:

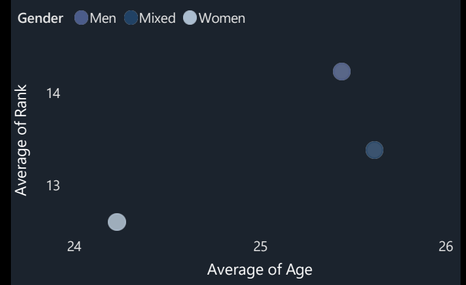
The proportion of participants ranking below 16 has been decreasing since last three years.

KSA's top 16 percentage is notably low in sports like table tennis, archery, and swimming. This trend is consistent across world championships and summer games competitions, suggesting a potential area for improvement.

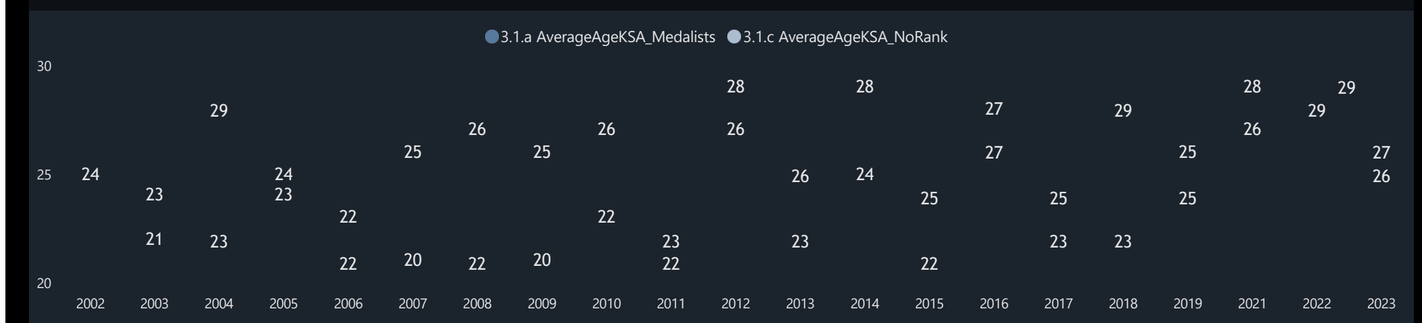
The average age of participants from KSA, approximately 26, surpasses the global average age of medalists, which stands at around 24.

Enhancing KSA's women participation is crucial as it currently lags significantly behind the global average for women's involvement in sports.

Age-Rank Trend Among Gender



KSA Age Trends: Medalists vs. the Rest



Age to Rank Analysis Across Multiple Sports