OLYMPICS DATA INSIGHTS

Presented by: Tatev Gomtsyan

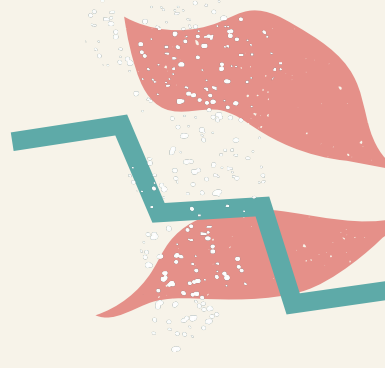# Introduction: Key Assumptions

WHO (client)

A sports apparel company

WHAT (objective)

Identify 2 countries with emerging female medalists and their respective sports. Recommend a strategic campaign to move into new markets, and expanding brand visibility and customer base.

WHY

The Olympics attract billions of viewers worldwide. Success stories uplift and empower younger generations.

The Data

Open-source Olympics data (1896 to 2016)

Cleaning

```
[405]:  # imputing NaNs
        athletes['Height'].fillna(athletes['Height'].median(), inplace=True)
        athletes['Weight'].fillna(athletes['Weight'].median(), inplace=True)
        athletes['Age'].fillna(athletes['Age'].median(), inplace=True)
        #athletes['Medal'].fillna('No Medal', inplace=True)

Games column doesn't give us additional information we don't already have. The year and season va

[406]:  athletes = athletes.drop(columns=['Games'])

[407]:  # Sort the df by 'Year' in ascending order
        athletes = athletes.sort_values(by='Year', ascending=True).reset_index(drop=True)
```

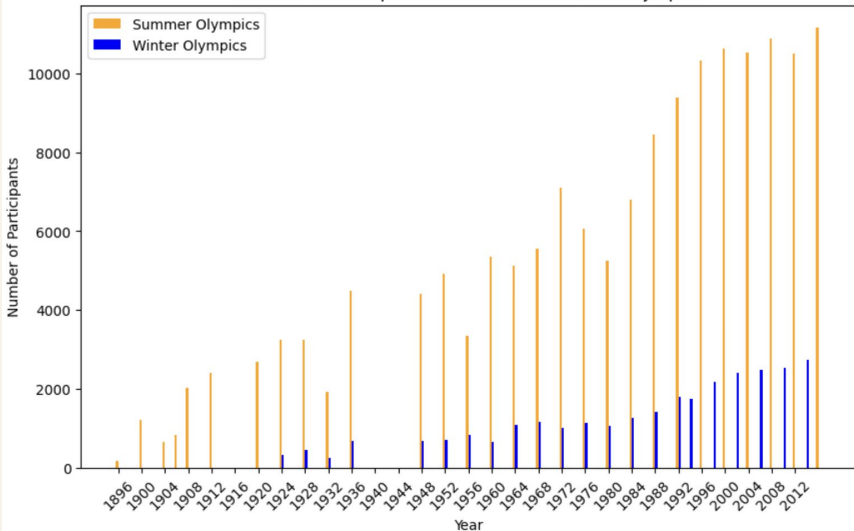|	Age	Height	Weight
count	271116.000000	271116.00000	271116.000000
mean	25.502493	175.26374	70.539500
std	6.287361	9.27917	12.578184
min	10.000000	127.00000	25.000000
25%	22.000000	170.00000	63.000000
50%	24.000000	175.00000	70.000000
75%	28.000000	180.00000	75.000000
max	97.000000	226.00000	214.000000

Statistical Summaries

```
[522]: print(winter['Event'].value_counts())

Event
Ice Hockey Men's Ice Hockey          4702
Bobsleigh Men's Four                 1715
Alpine Skiing Men's Slalom           1593
Alpine Skiing Men's Giant Slalom     1535
Alpine Skiing Men's Downhill         1164
                                     ...
Speed Skating Men's Allround           23
Freestyle Skiing Women's Halfpipe      23
Snowboarding Women's Slopestyle        23
Freestyle Skiing Women's Slopestyle    22
Alpinism Mixed Alpinism                21
Name: count, Length: 119, dtype: int64
```

Variables

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Year	Season	City	Sport	Event	Medal	Received_Medal
0	107607	Fritz Richard Gustav Schuft	M	19	175.0	70.0	Germany	GER	1896	Summer	Athina	Gymnastics	Gymnastics Men's Pommelled Horse	NaN	0
1	122526	Pierre Alexandre Tuffri	M	19	175.0	70.0	France	FRA	1896	Summer	Athina	Athletics	Athletics Men's Triple Jump	Silver	1
2	122526	Pierre Alexandre Tuffri	M	19	175.0	70.0	France	FRA	1896	Summer	Athina	Athletics	Athletics Men's Long Jump	NaN	0
3	12563	Conrad Helmut Fritz Bcker	M	25	175.0	70.0	Germany	GER	1896	Summer	Athina	Gymnastics	Gymnastics Men's Horse Vault	NaN	0
4	12563	Conrad Helmut Fritz Bcker	M	25	175.0	70.0	Germany	GER	1896	Summer	Athina	Gymnastics	Gymnastics Men's Parallel Bars	NaN	0

Exploratory Analysis: Fun Facts

Number of Participants in Summer and Winter Olympics

Total Medal Count by Year for FRG, GDR, URS, and YUG
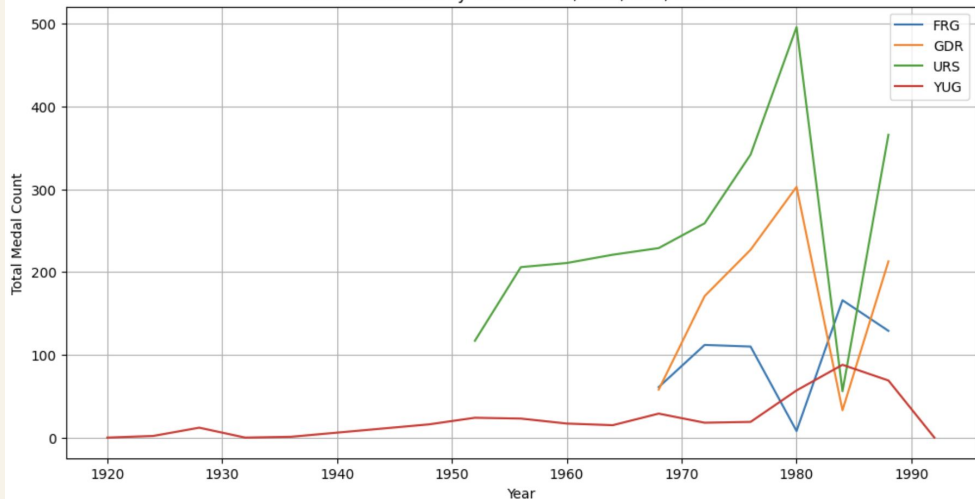
	Year	Athletes	Male	Female	PropFemale	Summer	Winter	AvgHeight	AvgWeight	AvgAge	NOCMostMedals	Countries	SummerHostCity	WinterHostCity
0	1896	176	176	NaN	NaN	176	0	174.73	70.18	23.76	GRE	12	Athina	NaN
1	1900	1220	1197	23.0	1.89	1220	0	175.10	70.19	26.98	FRA	31	Paris	NaN
2	1904	650	644	6.0	0.92	650	0	175.13	70.25	26.13	USA	15	St. Louis	NaN
3	1906	841	835	6.0	0.71	841	0	175.48	70.70	25.79	GRE	21	Athina	NaN
4	1908	2024	1980	44.0	2.17	2024	0	175.39	70.84	26.35	GBR	22	London	NaN
5	1912	2409	2356	53.0	2.20	2409	0	175.44	70.46	27.40	SWE	29	Stockholm	NaN
6	1920	2675	2597	78.0	2.92	2675	0	175.13	70.34	28.25	USA	29	Antwerpen	NaN

FRG - West Germany
GDR- East Germany
URS - Soviet Union
YUG - Yugoslavia

Relevant Historical Trends

Trend in Proportion of Female Athletes Over the Years
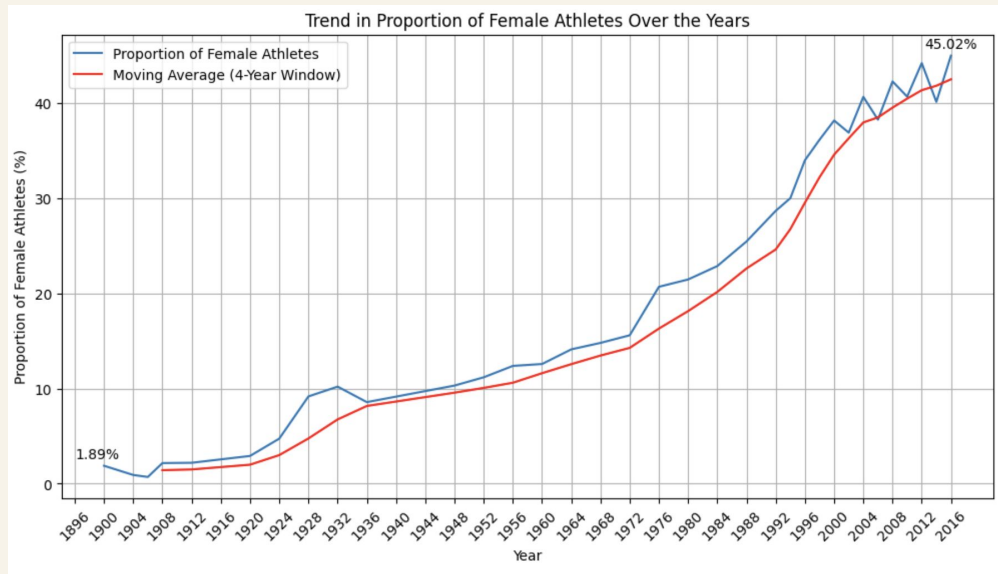
- Increased 1.89%-45.02% from 1900-2016
- Big jump in 1976 (~5%, slope = 1.28)

```python
[497]: # were any new sports added in 1976 that influenced this growth?

data_1972 = athletes[athletes['Year'] == 1972]
data_1976 = athletes[athletes['Year'] == 1976]

sports_1972 = set(data_1972['Event'])
sports_1976 = set(data_1976['Event'])

new_sports_1976 = sports_1976 - sports_1972

new_sports_1976
#women's bball, handball, cross country skiing, Rowing and some other mixed gender sports
```

```
[497]: {"Basketball Women's Basketball",
       "Canoeing Men's Canadian Doubles, 500 metres",
       "Canoeing Men's Canadian Singles, 500 metres",
       "Canoeing Men's Kayak Doubles, 500 metres",
       "Canoeing Men's Kayak Singles, 500 metres",
       "Cross Country Skiing Women's 4 x 5 kilometres Relay",
       'Figure Skating Mixed Ice Dancing',
       "Handball Women's Handball",
       "Rowing Men's Quadruple Sculls",
       "Rowing Women's Coxed Eights",
       "Rowing Women's Coxed Fours",
       "Rowing Women's Coxed Quadruple Sculls",
       "Rowing Women's Coxless Pairs",
       "Rowing Women's Double Sculls",
       "Rowing Women's Single Sculls",
       'Sailing Mixed Multihull',
       'Sailing Mixed Two Person Dinghy',
       "Speed Skating Men's 1,000 metres"}
```

New women's sport additions in 1976: basketball, handball, cross country skiing, rowing, other mixed gender sports

Approach #1: New Medalist

Filter data after year 2000 and determine a list of countries with 0 prior medals, and 1 or more after 2000

```python
# Finding countries that earned medals after 2000
recent_medals_after_2000 = athletes[athletes['Year'] > 2000].groupby(['NOC', 'Sport'])['Received_Medal'].sum().reset_index()
recent_medals_countries_after_2000 = recent_medals_after_2000[recent_medals_after_2000['Received_Medal'] > 0]['NOC']

# Countries with zero medals before the year 2000
zero_medals_before_2000 = athletes[athletes['Year'] <= 2000].groupby('NOC')['Received_Medal'].sum().reset_index()
zero_medals_countries_before_2000 = zero_medals_before_2000[zero_medals_before_2000['Received_Medal'] == 0]['NOC']

# Countries that started winning medals after 2000
newly_winning_countries_after_2000 = pd.DataFrame(zero_medals_countries_before_2000[zero_medals_countries_before_2000.isin(recent_medals_countries_after_2000)])

# Filter newly winning countries by female athletes
newly_winning_countries_after_2000_female = athletes[(athletes['NOC'].isin(newly_winning_countries_after_2000['NOC'])) & (athletes['Sex'] == 'F')]

# Count the number of female medals for each country
female_medal_counts = newly_winning_countries_after_2000_female.groupby('NOC')['Received_Medal'].sum()

# Sort and get the top 10 countries
top_10_countries_after_2000_female = female_medal_counts.sort_values(ascending=False).head(10)

top_10_countries_after_2000_female
```
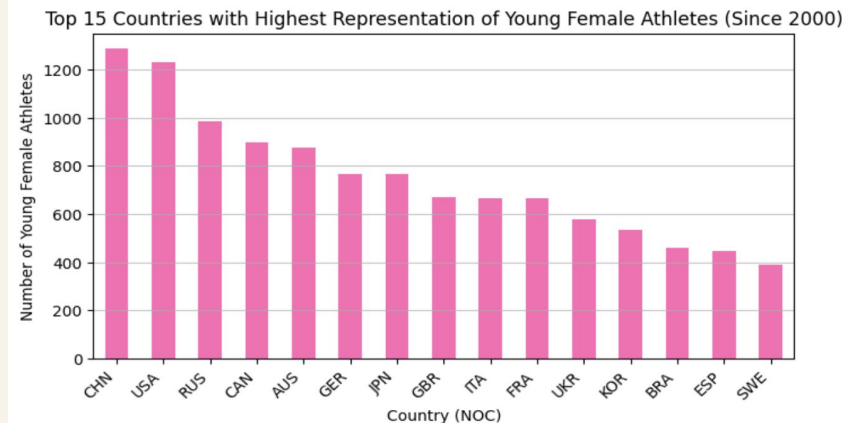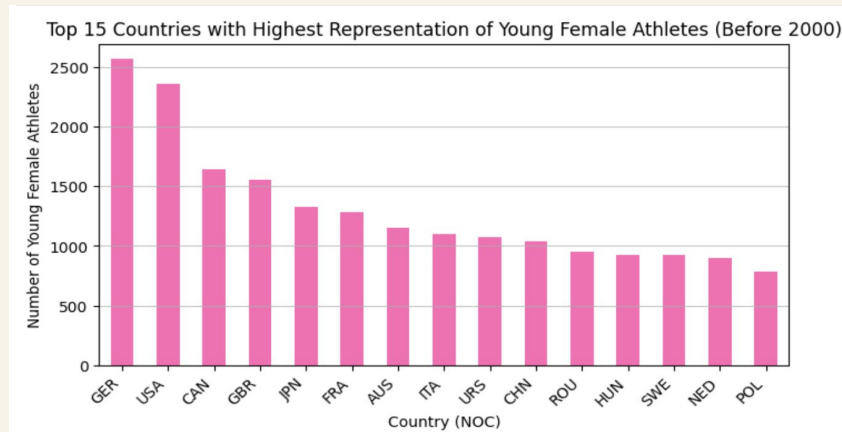
```
NOC
SRB    28
BRN     3
TJK     1
AFG     0
JOR     0
TOG     0
SUD     0
PAR     0
MRI     0
GUA     0
Name: Received_Medal, dtype: int64
```

	Year	NOC	Medal	Sport	Sex	TotalMedalCount
10	2016	SRB	Gold	Wrestling	M	1
11	2016	SRB	Silver	Taekwondo	F	1
12	2016	SRB	Silver	Basketball	M	12
13	2016	SRB	Silver	Canoeing	M	2
14	2016	SRB	Silver	Volleyball	F	12

Approach #2: Young Female Participant

Focus on emerging countries in representation of young female athletes (< 25) in recent years

Top 15 Countries with Highest Representation of Young Female Athletes (Before 2000)

Top 15 Countries with Highest Representation of Young Female Athletes (Since 2000)

* South Korea (KOR), Brazil (BRA), and Spain (ESP), but KOR has the highest female participation

Results

Sports with Most Female Medals (Under 25) in Korea Since 2000
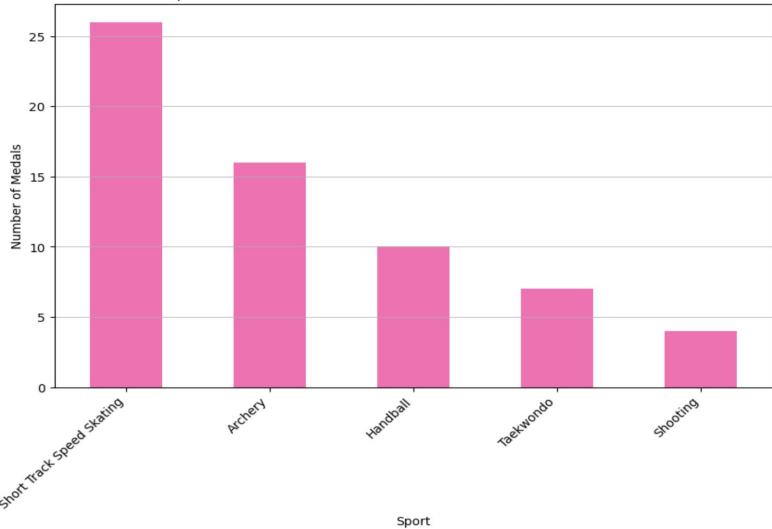
```
[536]: # Filter data for female athletes with medals in short track speed skating from South Korea (NOC=='KOR') since 2000
       female_athletes_kor_short_track_since_2000 = athletes[(athletes['Sex'] == 'F') &
                                                             (athletes['Sport'] == 'Short Track Speed Skating') &
                                                             (athletes['Medal'].notnull()) &
                                                             (athletes['Year'] >= 2000) &
                                                             (athletes['NOC'] == 'KOR')]

       result_df = female_athletes_kor_short_track_since_2000.groupby(['Name', 'Year']).agg({
           'Medal': 'count',
           'Age': 'max'
       }).reset_index().sort_values(by='Year', ascending=False)

       result_df.head()
```

	Name	Year	Medal	Age
16	Shim Seok-Hui	2014	3	17
9	Jo Hae-Ri	2014	1	27
15	Park Seung-Hui	2014	3	21
11	Kim A-Lang	2014	1	18
6	Gong Sang-Jeong	2014	1	17

Shim Seok-Hui won 3 medals in 2014 at the age of 17 in Short Track Speed Skating

EmpowHER Campaign

Uplifting female athletes and empowering younger generations of girls around the world

The Face(s)

Shim Seok-Hui (KOR)

Maja Ognjenovic (SRB), Captain/Setter

Sponsees

Shim Seok-Hui

Serbian Women's Volleyball Team

The campaign entails becoming the official sponsor of the Serbian (SRB) Women's Volleyball Team, and of Shim Seok-Hui, a South Korean (KOR) skater and 3-time medalist in recent history. Maja Ognjenovic, captain (setter) of SRB volleyball team, and KOR's Seok-Hui can both serve as the faces of the campaign.

Future Directions

1. Implement ML models for anomaly detection or clustering to identify unusual patterns in the data
2. Train models to predict future trends based on historical data (e.g. time series forecasting)
3. Obtain more up-to-date data to improve accuracy of trends, analysis, and predictions; former athletes from 2016 have (probably) already retired
4. Consider merging external population or GDP data for exhaustive analysis
5. Track real time data; 5 new sports introduced in 2024
6. Follow upcoming trends (e.g. emphasis on creativity and youth)

One More Statistic

Top 10 Performances by Overall Medal Count for FRA
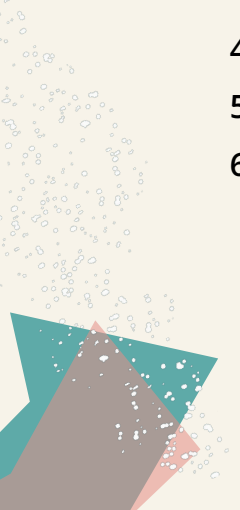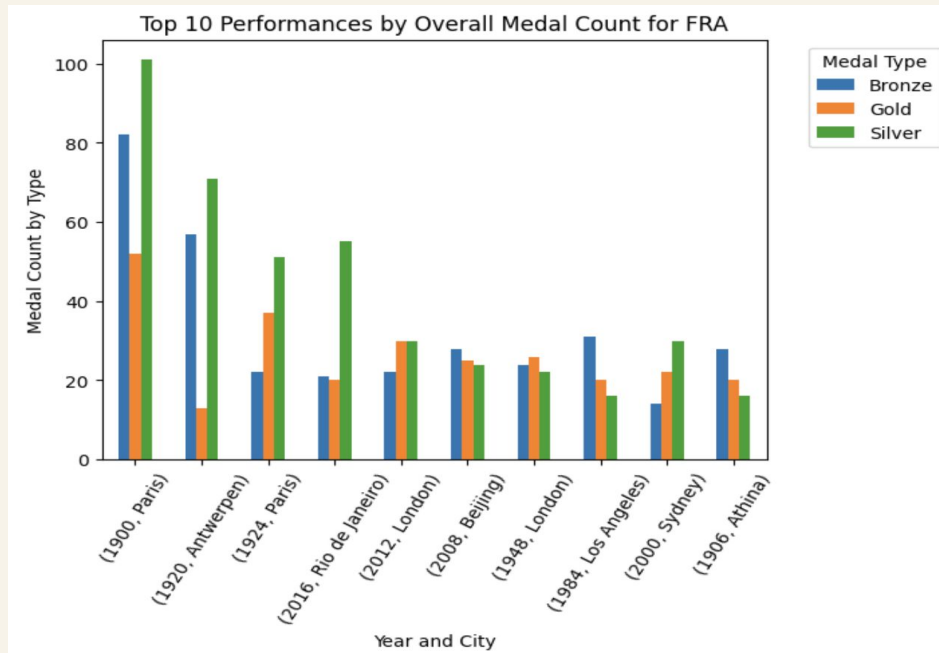
2024 Paris Olympics: France takes shocking lead in Medals

THANK YOU!

Do you have any questions?

Email: tmg6jda@virginia.edu

SOURCES

1. https://en.wikipedia.org/wiki/List_of_Olympic_Games_boycotts
2. https://www.sportspromedia.com/insights/analysis/paris-2024-olympics-sponsors-lvmh-carrefour-revenue/#:~:text=Below%20that%2C%20there%20are%2014,official%20supporters%20of%20Paris%202024
3. https://thewalrus.ca/loveofthegame-why-olympians-are-struggling-to-get-sponsorships/