Journey Into Olympic Analytics: Exploring Trends and Projections

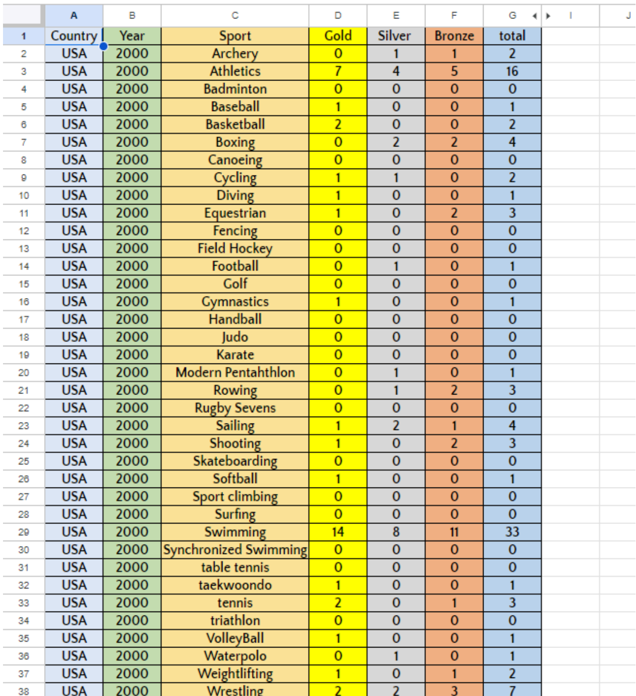
By : Ayoub Bakali | Adnan Sahli

Introduction

In this project, our aim is to delve into the intricate patterns of the Olympic Games, leveraging various analytical techniques to gain insights and forecast the performance of countries in the upcoming Olympic event. By harnessing both traditional analytical methods and cutting-edge artificial intelligence approaches, we seek to unravel hidden trends and anticipate future outcomes. This endeavor serves not only to deepen our understanding of Olympic dynamics but also to harness the potential of artificial intelligence for predictive analysis.

Data Collection:"Behind the Scenes: Gathering Insights from Olympic Data"

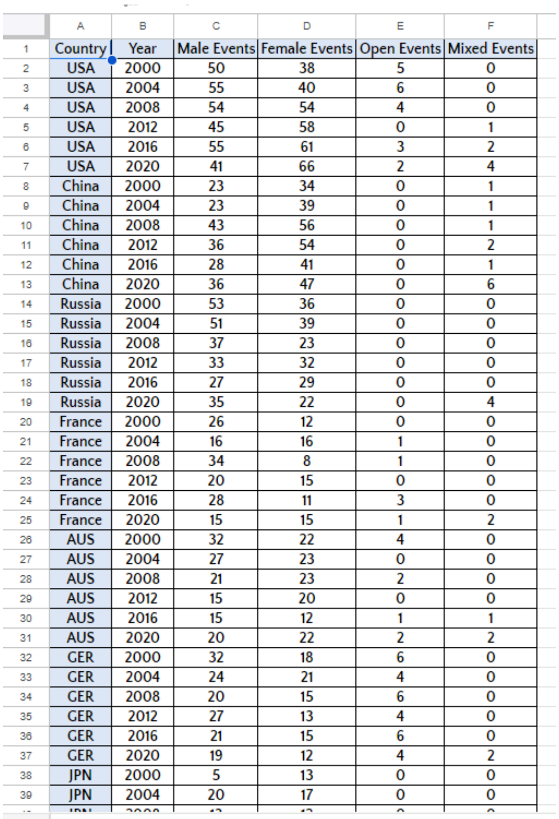
In our data collection phase, we embarked on a meticulous quest to gather comprehensive information from diverse sources. Our primary objective was to curate a well-structured dataset containing detailed medal distributions by sport, year, and country for the past six Summer Olympic editions (2000-2020),



	A	B	C	D	E	F	G
1	Country	Year	Sport	Gold	Silver	Bronze	total
2	USA	2000	Archery	0	1	1	2
3	USA	2000	Athletics	7	4	5	16
4	USA	2000	Badminton	0	0	0	0
5	USA	2000	Baseball	1	0	0	1
6	USA	2000	Basketball	2	0	0	2
7	USA	2000	Boxing	0	2	2	4
8	USA	2000	Canoeing	0	0	0	0
9	USA	2000	Cycling	1	1	0	2
10	USA	2000	Diving	1	0	0	1
11	USA	2000	Equestrian	1	0	2	3
12	USA	2000	Fencing	0	0	0	0
13	USA	2000	Field Hockey	0	0	0	0
14	USA	2000	Football	0	1	0	1
15	USA	2000	Golf	0	0	0	0
16	USA	2000	Gymnastics	1	0	0	1
17	USA	2000	Handball	0	0	0	0
18	USA	2000	Judo	0	0	0	0
19	USA	2000	Karate	0	0	0	0
20	USA	2000	Modern Pentahthlon	0	1	0	1
21	USA	2000	Rowing	0	1	2	3
22	USA	2000	Rugby Sevens	0	0	0	0
23	USA	2000	Sailing	1	2	1	4
24	USA	2000	Shooting	1	0	2	3
25	USA	2000	Skateboarding	0	0	0	0
26	USA	2000	Softball	1	0	0	1
27	USA	2000	Sport climbing	0	0	0	0
28	USA	2000	Surfing	0	0	0	0
29	USA	2000	Swimming	14	8	11	33
30	USA	2000	Synchronized Swimming	0	0	0	0
31	USA	2000	table tennis	0	0	0	0
32	USA	2000	taekwoondo	1	0	0	1
33	USA	2000	tennis	2	0	1	3
34	USA	2000	triathlon	0	0	0	0
35	USA	2000	VolleyBall	1	0	0	1
36	USA	2000	Waterpolo	0	1	0	1
37	USA	2000	Weightlifting	1	0	1	2
38	USA	2000	Wrestling	2	2	3	7

encompassing a total of 11 sports. Despite exhaustive searches, we encountered challenges in finding a pre-existing dataset meeting our criteria, prompting us to compile the data manually. This meticulous approach yielded a dataset comprising approximately 2500 entries, meticulously organized to cover every conceivable year-country combination while maintaining a consistent roster of sports.

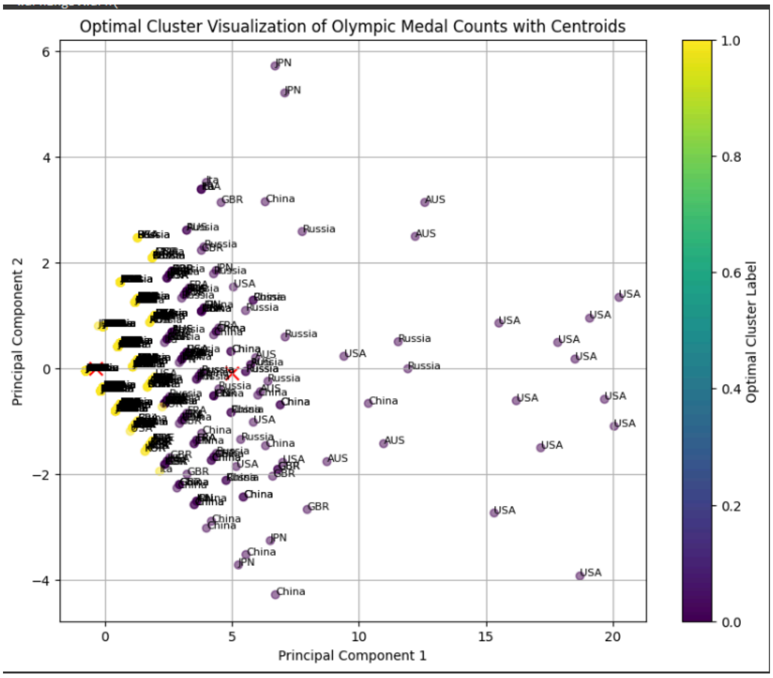
In addition to our primary dataset, we assembled three supplementary datasets, each serving a specific analytical purpose. The first dataset cataloged medal distributions based on event gender categories (Male, Female, Mixed, and Open), providing nuanced insights into gender-specific Olympic performances. The second dataset quantified the cumulative medal tallies for each country across all sports from 2000 to 2020, offering a macroscopic view of national Olympic success. Finally, the third dataset presented an array of factors potentially influencing country-level performance, including GDP per Capita, Population Size, Competitors, Sports Participation, Past Rank, Hosting status, and Social Development Index. While endeavoring to incorporate additional factors such as Sports Budget and Number of Stadiums, their unavailability posed constraints.



Country	Year	Male Events	Female Events	Open Events	Mixed Events
USA	2000	50	38	5	0
USA	2004	55	40	6	0
USA	2008	54	54	4	0
USA	2012	45	58	0	1
USA	2016	55	61	3	2
USA	2020	41	66	2	4
China	2000	23	34	0	1
China	2004	23	39	0	1
China	2008	43	56	0	1
China	2012	36	54	0	2
China	2016	28	41	0	1
China	2020	36	47	0	6
Russia	2000	53	36	0	0
Russia	2004	51	39	0	0
Russia	2008	37	23	0	0
Russia	2012	33	32	0	0
Russia	2016	27	29	0	0
Russia	2020	35	22	0	4
France	2000	26	12	0	0
France	2004	16	16	1	0
France	2008	34	8	1	0
France	2012	20	15	0	0
France	2016	28	11	3	0
France	2020	15	15	1	2
AUS	2000	32	22	4	0
AUS	2004	27	23	0	0
AUS	2008	21	23	2	0
AUS	2012	15	20	0	0
AUS	2016	15	12	1	1
AUS	2020	20	22	2	2
GER	2000	32	18	6	0
GER	2004	24	21	4	0
GER	2008	20	15	6	0
GER	2012	27	13	4	0
GER	2016	21	15	6	0
GER	2020	19	12	4	2
JPN	2000	5	13	0	0
JPN	2004	20	17	0	0

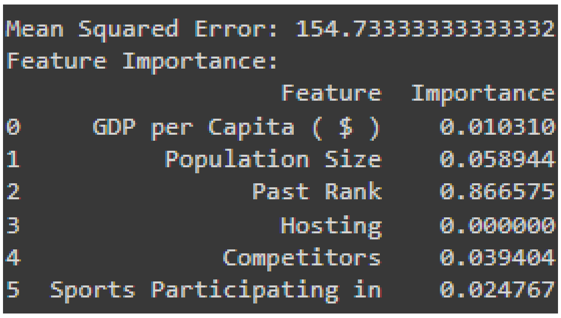
To ensure data accuracy and reliability, we cross-referenced our compiled data with reputable sources such as Wikipedia and Olympedia, meticulously validating each entry. The culmination of these efforts resulted in the consolidation of all datasets into a single Excel file named 'Olympics,' comprising four distinct sheets. In total, our dataset comprises over 3000 entries, collectively offering invaluable insights across a spectrum of Olympic variables, amounting to a staggering 15,000 individual data points.

Data Analysis:Deciphering Olympic Trends: Insights from Data Analysis

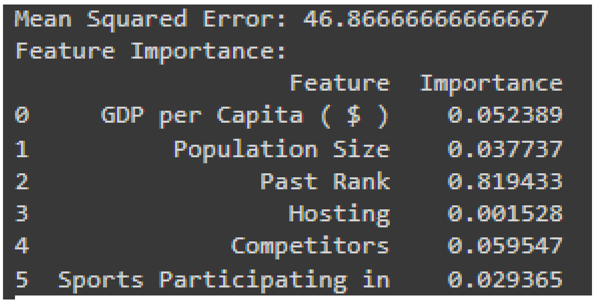


In our data analysis, we employed three key algorithms: K-means clustering, decision tree analysis, and K-nearest neighbors (KNN). Our exploration began with K-means clustering applied to the comprehensive dataset, utilizing the elbow method to identify the optimal number of clusters. Following this approach, we determined that two clusters provided the most suitable grouping, as illustrated in the accompanying graph generated through Principal Component Analysis (PCA).

As depicted in the graph, our dataset exhibits a distinct pattern of clustering. A densely populated area on the left signifies a diverse mix of countries competing across various sports, whereas the right side is dominated by instances labeled as 'USA,' indicating the country's unequivocal dominance in specific disciplines such as swimming and athletics. The middle section portrays a competitive landscape, featuring countries such as Russia, Great Britain, China, Australia, and Japan, where a semblance of parity exists. This clustering strategy was further validated through silhouette scoring, with the two-cluster configuration yielding a commendable silhouette score of 0.815.

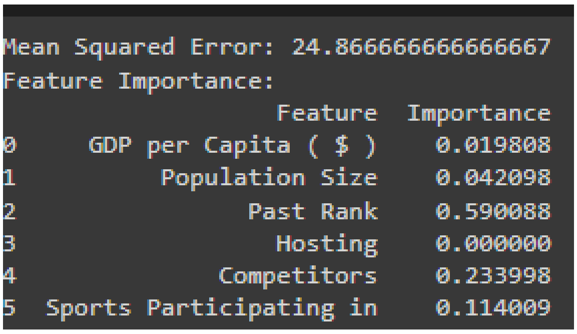


```
Mean Squared Error: 154.73333333333332
Feature Importance:
                  Feature  Importance
0     GDP per Capita ( $ )    0.010310
1          Population Size    0.058944
2                Past Rank    0.866575
3                  Hosting    0.000000
4              Competitors    0.039404
5  Sports Participating in    0.024767
```

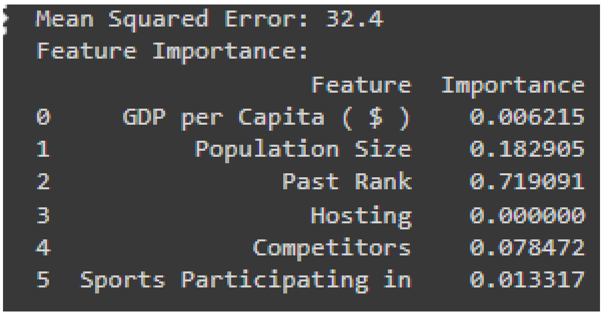


```
Mean Squared Error: 46.86666666666667
Feature Importance:
                  Feature  Importance
0     GDP per Capita ( $ )    0.052389
1          Population Size    0.037737
2                Past Rank    0.819433
3                  Hosting    0.001528
4              Competitors    0.059547
5  Sports Participating in    0.029365
```

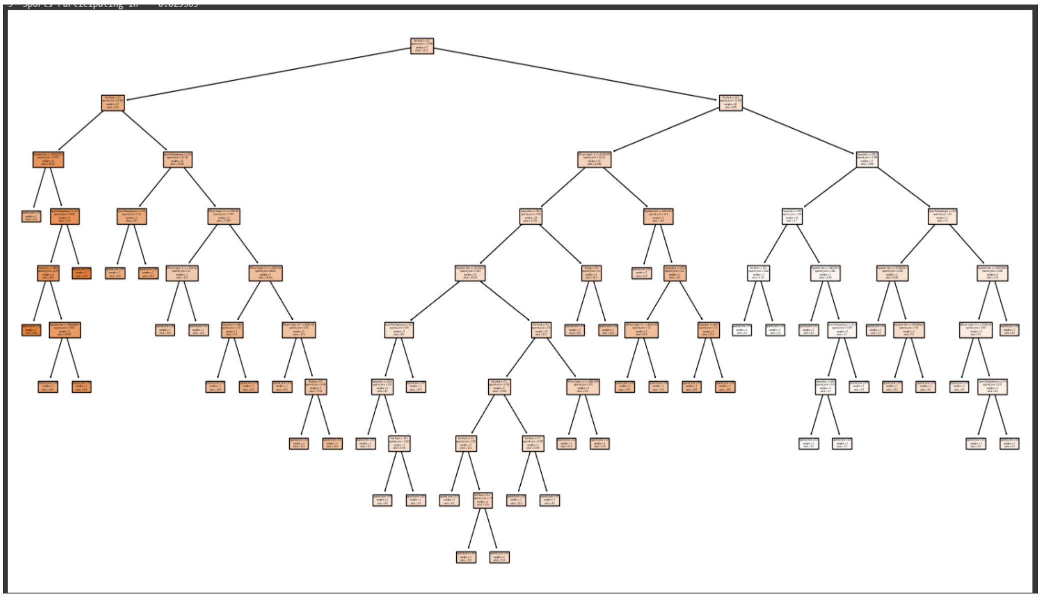
Subsequently, we employed decision tree analysis on a dataset comprising diverse parameters, aiming to discern their influence on medal tallies. Despite varying degrees of success, with Mean Squared Error (MSE) ranging from 25 to 155 across different medal categories, our analysis unearthed noteworthy insights. Notably, factors such as GDP per Capita exhibited limited influence on medal counts, suggesting that financial resources alone do not guarantee success in Olympic competition. Conversely, historical performance, as indicated by Past Rank, emerged as a consistently influential factor, underscoring the significance of a country's sporting pedigree. Additionally, the number of competitors emerged as a significant predictor, affirming the intuitive notion that larger delegations tend to accrue more medals. Although the hosting parameter appeared less influential on medal counts directly, it indirectly impacted performance by influencing the number of competitors.



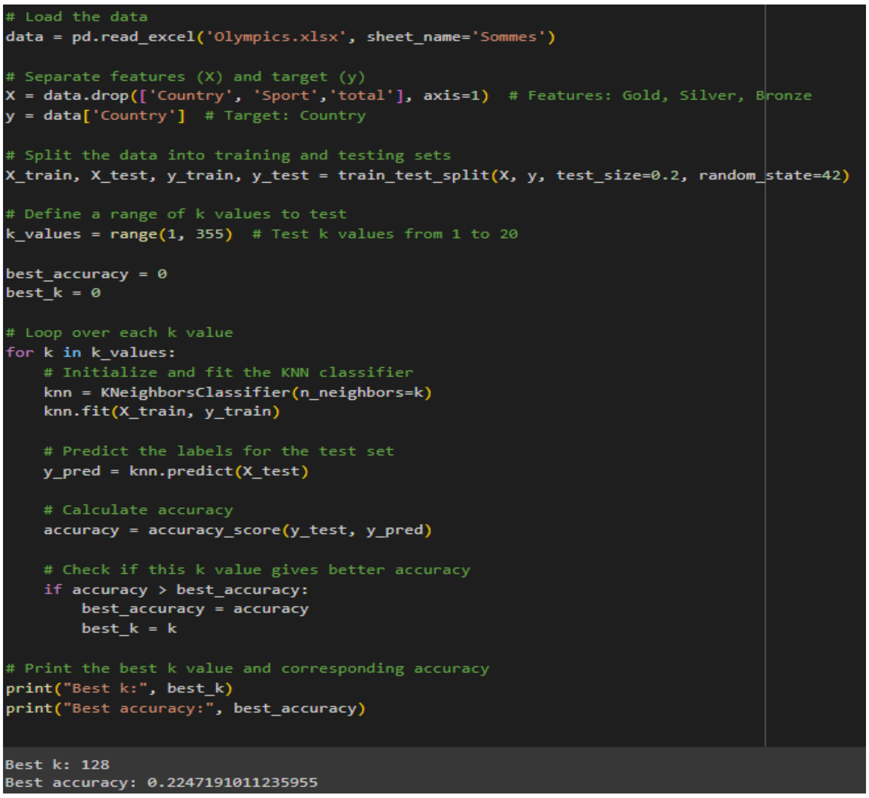
```
Mean Squared Error: 24.866666666666667
Feature Importance:
                  Feature  Importance
0     GDP per Capita ( $ )    0.019808
1          Population Size    0.042098
2                Past Rank    0.590088
3                  Hosting    0.000000
4              Competitors    0.233998
5  Sports Participating in    0.114009
```



```
Mean Squared Error: 32.4
Feature Importance:
                  Feature  Importance
0     GDP per Capita ( $ )    0.006215
1          Population Size    0.182905
2                Past Rank    0.719091
3                  Hosting    0.000000
4              Competitors    0.078472
5  Sports Participating in    0.013317
```



Lastly, our exploration culminated with KNN analysis applied to a dataset documenting the aggregate medal tallies for each country across six editions of the Olympics. Despite initial optimism, our findings revealed a limited accuracy of 0.22 for the optimal K value of 128. This modest performance can be attributed to the inherent variability in each country's performance across different sports. Notably, disparities in medal opportunities across sports contribute to this variability, as disciplines such as swimming offer significantly more medal chances compared to others like football.



```
# Load the data
data = pd.read_excel('Olympics.xlsx', sheet_name='Sommes')

# Separate features (X) and target (y)
X = data.drop(['Country', 'Sport','total'], axis=1)  # Features: Gold, Silver, Bronze
y = data['Country']  # Target: Country

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Define a range of k values to test
k_values = range(1, 355)  # Test k values from 1 to 20

best_accuracy = 0
best_k = 0

# Loop over each k value
for k in k_values:
    # Initialize and fit the KNN classifier
    knn = KNeighborsClassifier(n_neighbors=k)
    knn.fit(X_train, y_train)

    # Predict the labels for the test set
    y_pred = knn.predict(X_test)

    # Calculate accuracy
    accuracy = accuracy_score(y_test, y_pred)

    # Check if this k value gives better accuracy
    if accuracy > best_accuracy:
        best_accuracy = accuracy
        best_k = k

# Print the best k value and corresponding accuracy
print("Best k:", best_k)
print("Best accuracy:", best_accuracy)
```

```
Best k: 128
Best accuracy: 0.2247191011235955
```

Overall, our data analysis endeavors have shed light on various facets of Olympic performance, uncovering the complex interplay of factors influencing medal outcomes and offering valuable insights for future strategic planning.

Predictions:Forecasting the Future: Predictive Analytics in Olympic Performance

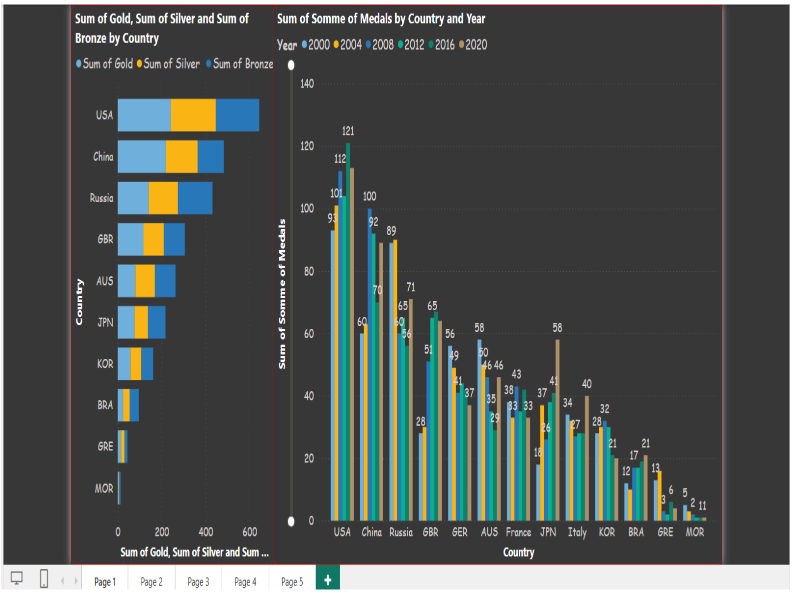
"After extensive experimentation, we sought to predict future Olympic performances for select countries, with a particular focus on the upcoming Olympics in Paris. Initially, we explored the feasibility of employing ARIMA models for time-series forecasting, given the structured nature of our data. However, despite our efforts, ARIMA did not yield satisfactory results, prompting us to explore alternative approaches.

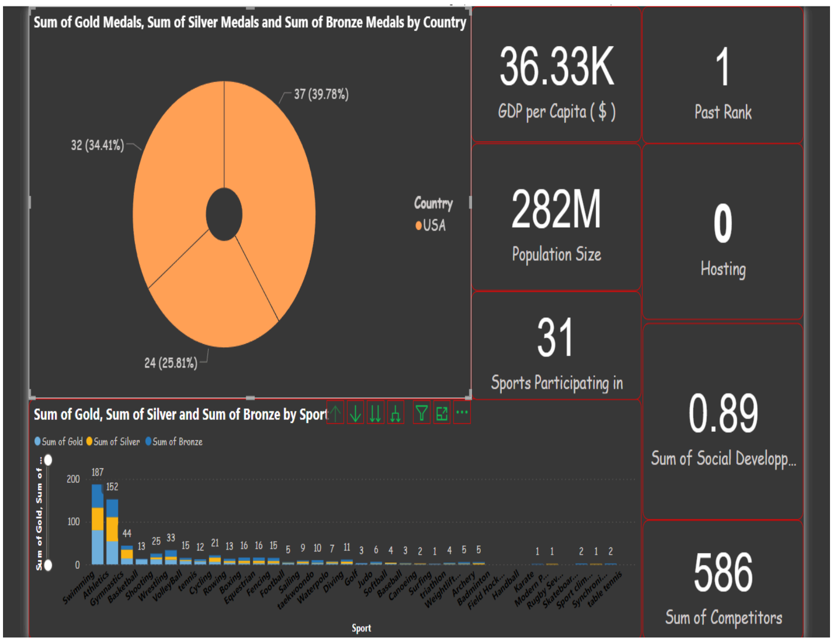
Subsequently, we turned to regression analysis, considering its flexibility in handling multivariate datasets. We formulated two distinct regression models to predict medal outcomes: one leveraging various parameters such as GDP per Capita, Population Size, Past Rank, and others, and another utilizing historical medal data spanning multiple Olympic editions.

Our regression models produced insightful predictions, offering valuable glimpses into potential Olympic outcomes. Despite the inherent complexity of predicting sporting events, our models provided useful estimates that could inform strategic planning and decision-making for participating countries.

In conclusion, while ARIMA proved unsuitable for our predictive tasks, regression analysis emerged as a viable alternative, enabling us to generate informed projections for Olympic medal tallies. Moving forward, these predictions serve as valuable tools for stakeholders in the sports community, offering valuable insights into potential outcomes and informing strategic planning efforts."

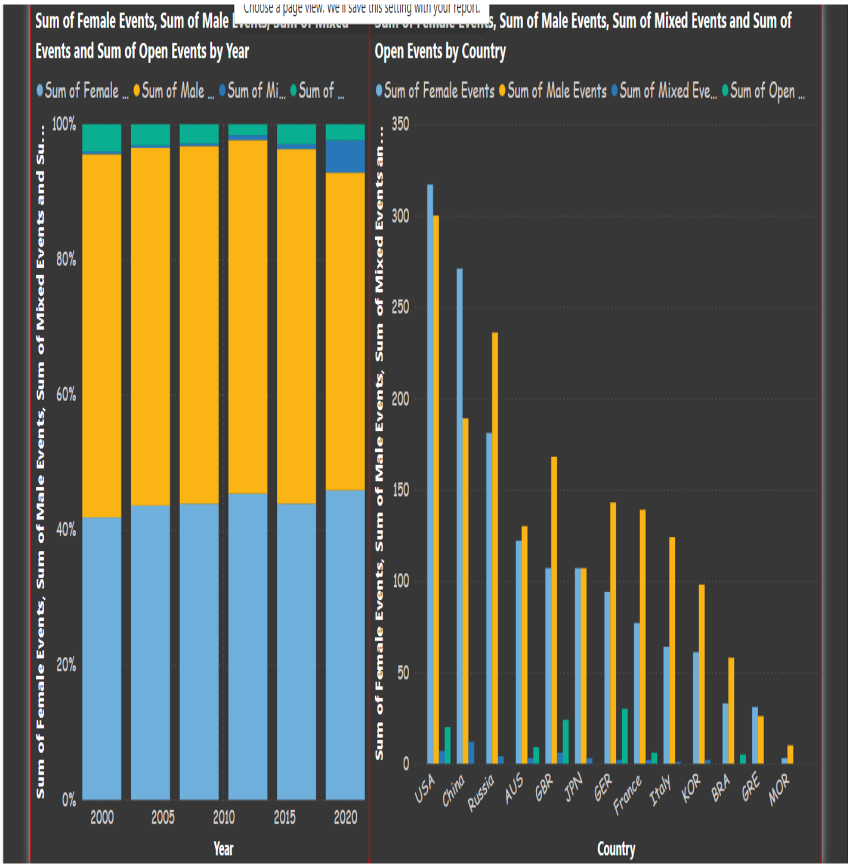
Visualizing Olympic Data: Insights through Power BI





"In our visualization efforts within Power BI, we aimed to create intuitive and informative dashboards that offer insights into Olympic data trends and predictive analytics. Our approach involved structuring the dashboards across multiple pages to facilitate easy navigation and understanding.

On the first page, we focused on presenting a comprehensive overview of medal tallies by country and medal type across the six Olympic editions covered in our dataset. Utilizing interactive visuals such as bar charts and slicers, viewers can explore the distribution of medals over time and across different countries. Additionally, we implemented drill-through functionality to enable users to delve deeper into specific year-country combinations, revealing detailed breakdowns of medal distributions by sport and medal type. This interactive feature enhances user engagement and facilitates deeper analysis of Olympic performance metrics.



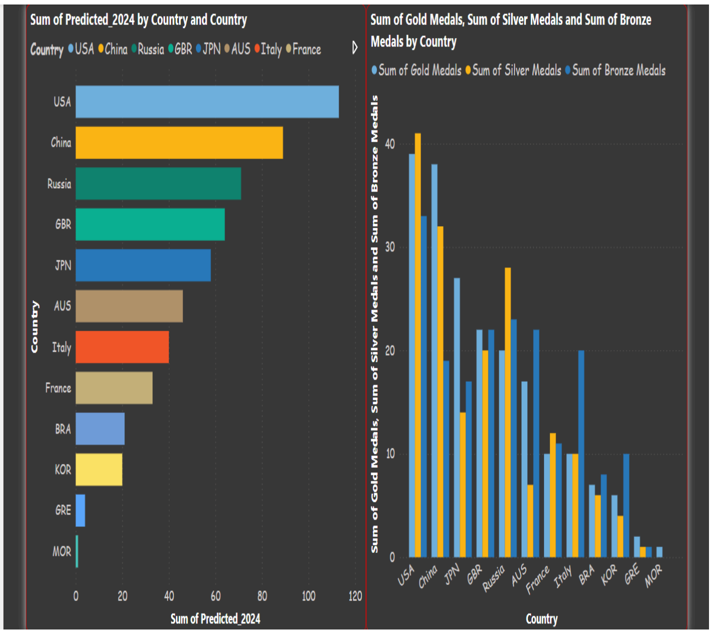
Choose a page view. We'll save this setting with your report.

Sum of Female Events, Sum of Male Events, Sum of Mixed Events and Sum of Open Events by Year

Sum of Female Events, Sum of Male Events, Sum of Mixed Events and Sum of Open Events by Country

● Sum of Female ... ● Sum of Male ... ● Sum of Mi... ● Sum of ...
Sum of Female Events, Sum of Male Events, Sum of Mixed Events and Su...
100%
80%
60%
40%
20%
0%
2000
2005
2010
2015
2020
Year
● Sum of Female Events ● Sum of Male Events ● Sum of Mixed Eve... ● Sum of Open ...
Sum of Female Events, Sum of Male Events, Sum of Mixed Events an...
350
300
250
200
150
100
50
0
USA
China
Russia
AUS
GBR
JPN
GER
France
Italy
KOR
BRA
GRE
MOR
Country

Moving to the second page, we showcased a breakdown of medals by event gender, offering insights into the gender distribution of medal winners across various countries and years. By visualizing this data through stacked bar charts and trend lines, viewers gain valuable insights into the evolving landscape of gender representation in Olympic sports.



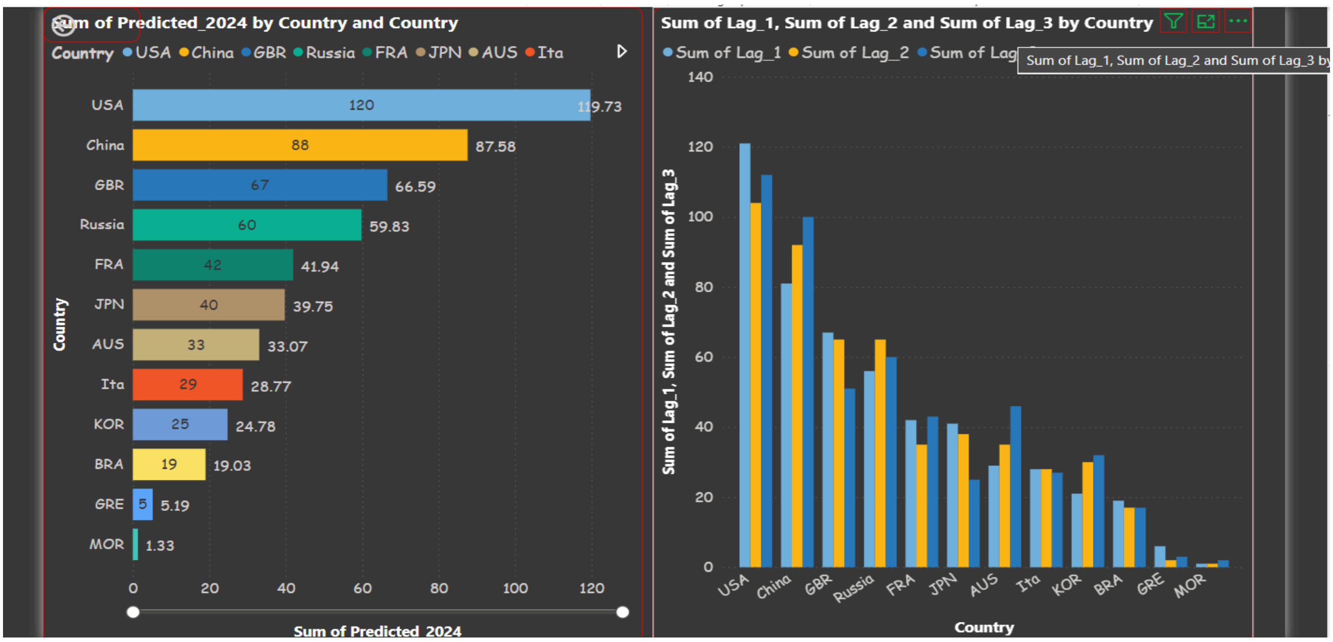
Sum of Predicted_2024 by Country and Country

Sum of Gold Medals, Sum of Silver Medals and Sum of Bronze Medals by Country

Country ● USA ● China ● Russia ● GBR ● JPN ● AUS ● Italy ● France
Country
USA
China
Russia
GBR
JPN
AUS
Italy
France
BRA
KOR
GRE
MOR
0
20
40
60
80
100
120
Sum of Predicted_2024
● Sum of Gold Medals ● Sum of Silver Medals ● Sum of Bronze Medals
Sum of Gold Medals, Sum of Silver Medals and Sum of Bronze Medals
40
30
20
10
0
USA
China
JPN
GBR
Russia
AUS
France
Italy
BRA
KOR
GRE
MOR
Country

On pages four and five, we transitioned to presenting the results of our predictive analytics efforts. Leveraging regression models, we generated forecasts for future Olympic medal tallies, providing stakeholders with valuable insights into potential outcomes for the upcoming Olympic Games in Paris. Through visually appealing line graphs and comparison tables, viewers can assess the accuracy of our predictions and gain actionable insights for strategic planning and decision-making.

By incorporating interactive elements, intuitive visualizations, and drill-through functionality, our Power BI dashboards offer a user-friendly platform for exploring Olympic data, uncovering insights, and informing strategic decision-making in the realm of sports analytics.


Sum of Predicted_2024 by Country and Country
Country ● USA ● China ● GBR ● Russia ● FRA ● JPN ● AUS ● Ita
USA 120 119.73
China 88 87.58
GBR 67 66.59
Russia 60 59.83
FRA 42 41.94
JPN 40 39.75
AUS 33 33.07
Ita 29 28.77
KOR 25 24.78
BRA 19 19.03
GRE 5 5.19
MOR 1.33
Country
0 20 40 60 80 100 120
Sum of Predicted_2024
Sum of Lag_1, Sum of Lag_2 and Sum of Lag_3 by Country
● Sum of Lag_1 ● Sum of Lag_2 ● Sum of Lag_3
Sum of Lag_1, Sum of Lag_2 and Sum of Lag_3 by
140 120 100 80 60 40 20 0
Sum of Lag_1, Sum of Lag_2 and Sum of Lag_3
USA China GBR Russia FRA JPN AUS Ita KOR BRA GRE MOR
Country

Conclusion:Journey's End: Concluding Thoughts on Olympic Insights and Predictions

In conclusion, our study delving into the intricacies of Olympic data analysis has been an enriching journey, both technically and culturally. Despite encountering challenges and hurdles along the way, navigating through various errors and complexities has provided valuable learning opportunities on the technical front. Moreover, from a cultural perspective, delving into the nuances of Olympic performance metrics has offered profound insights into the dynamics of this global sporting event.

Through meticulous data collection and analysis, we gained a deeper understanding of the patterns and trends prevalent in Olympic medal distribution. Our exploration into the influence of different factors such as GDP per capita, population size, past rank, and hosting status shed light on their varying degrees of impact on countries' performance in the Olympics. Surprisingly, certain factors that were initially assumed to have significant influence turned out to be less impactful than anticipated, while others proved to be pivotal in determining Olympic success.

Despite the imperfections inherent in any data analysis endeavor, our study has significantly enhanced our understanding of this subject matter. Leveraging data that was not readily available initially, we embarked on a real-world study that provided invaluable insights into the intricacies of Olympic performance.

As we await the unfolding of the upcoming Olympic Games, we eagerly anticipate the validation of our predictive models. While our predictions may not be perfect, the journey of exploration and discovery undertaken in this study has undoubtedly broadened our perspectives and deepened our understanding of the fascinating world of Olympic sports analytics.