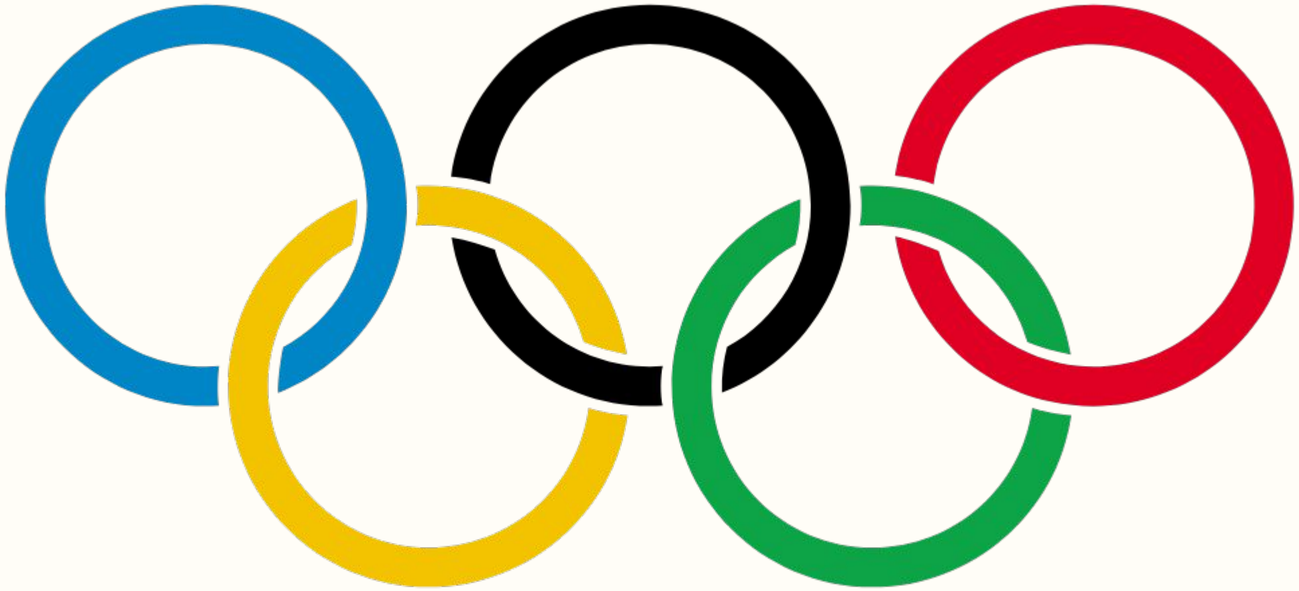
Predicting Olympic Results



By: Rohan Giri, Jennifer Gonzalez, Sameer Khan, Kristen Lowe, María Laura Peña, Scott Stempak

Table Of Contents

1

Intro / Data Cleaning and Merging

To what extent does a country's resources impact its performance at the Olympics?

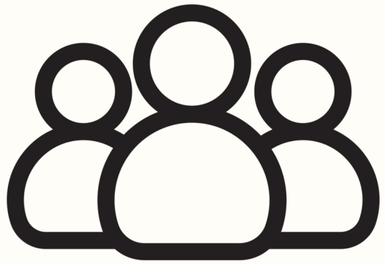
Datasets from Kaggle

Olympic Summer & Winter Games, 1896-2022

medal counts, host country, etc.

World GDP by Country: 1960-2022

GDP by year



Country Population from 1960 to 2022

population by year

Countries of the World

area in sq. km., region, coastline ratio

Data Cleaning and Merging

Clean

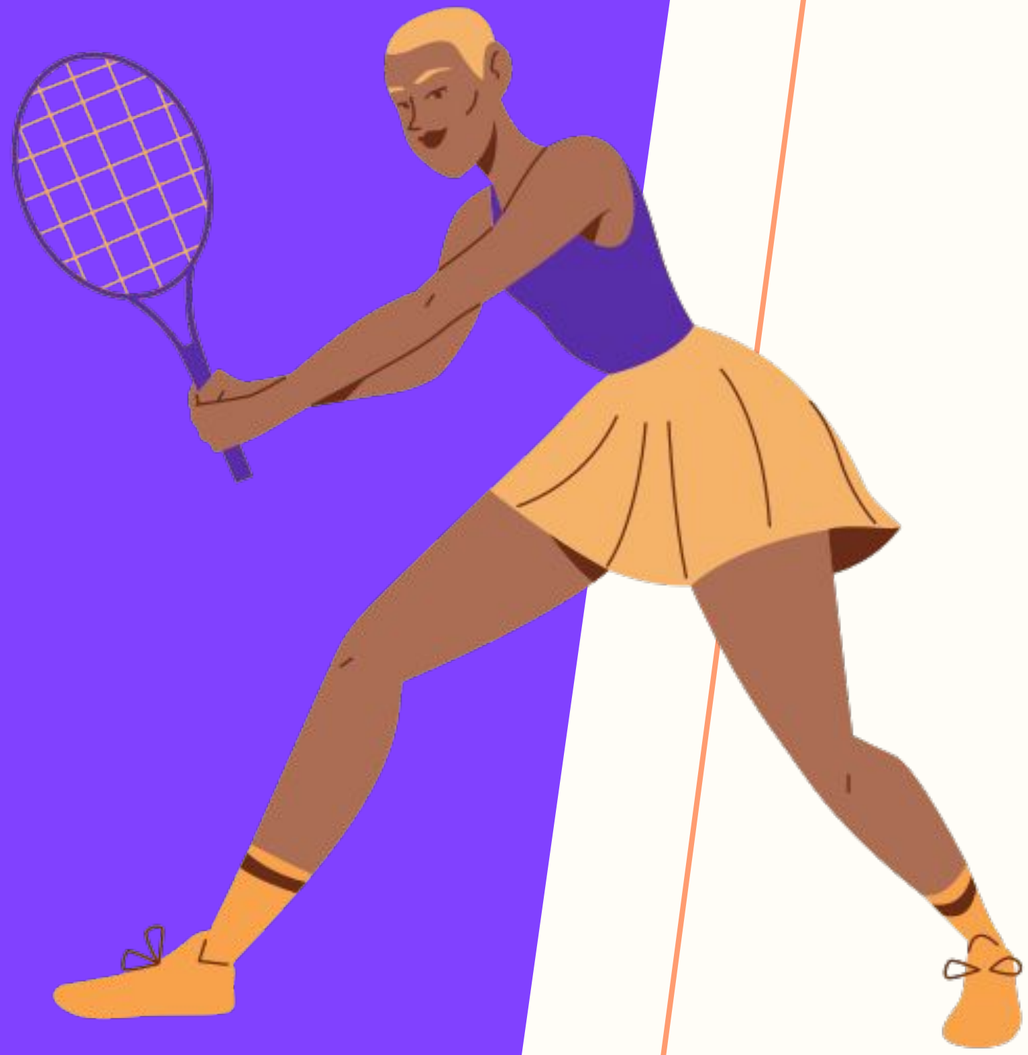
- mapped country names
- pivoted data frames
- removed missing values

Merge

- 2 olympics + GDP
- + population
- + other country resources
- feature engineering

Resulting Variables

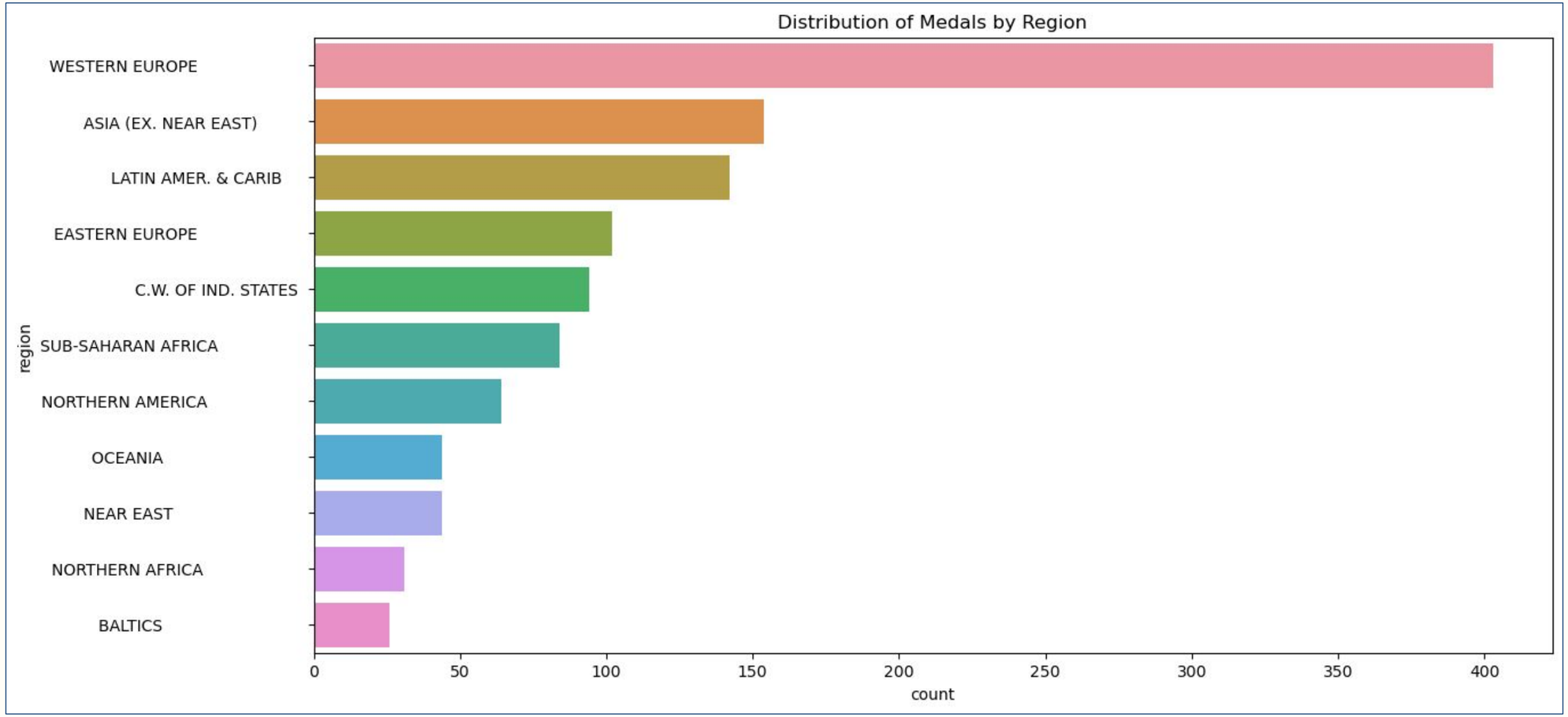
- country
- year
- medal count (gold, silver, bronze)
- host country
- game season (winter/summer)
- GDP
- population
- population density
- region
- area (sq. km.)
- coastline (coast-to-area ratio)
- GDP per capita
- host country status (binary)



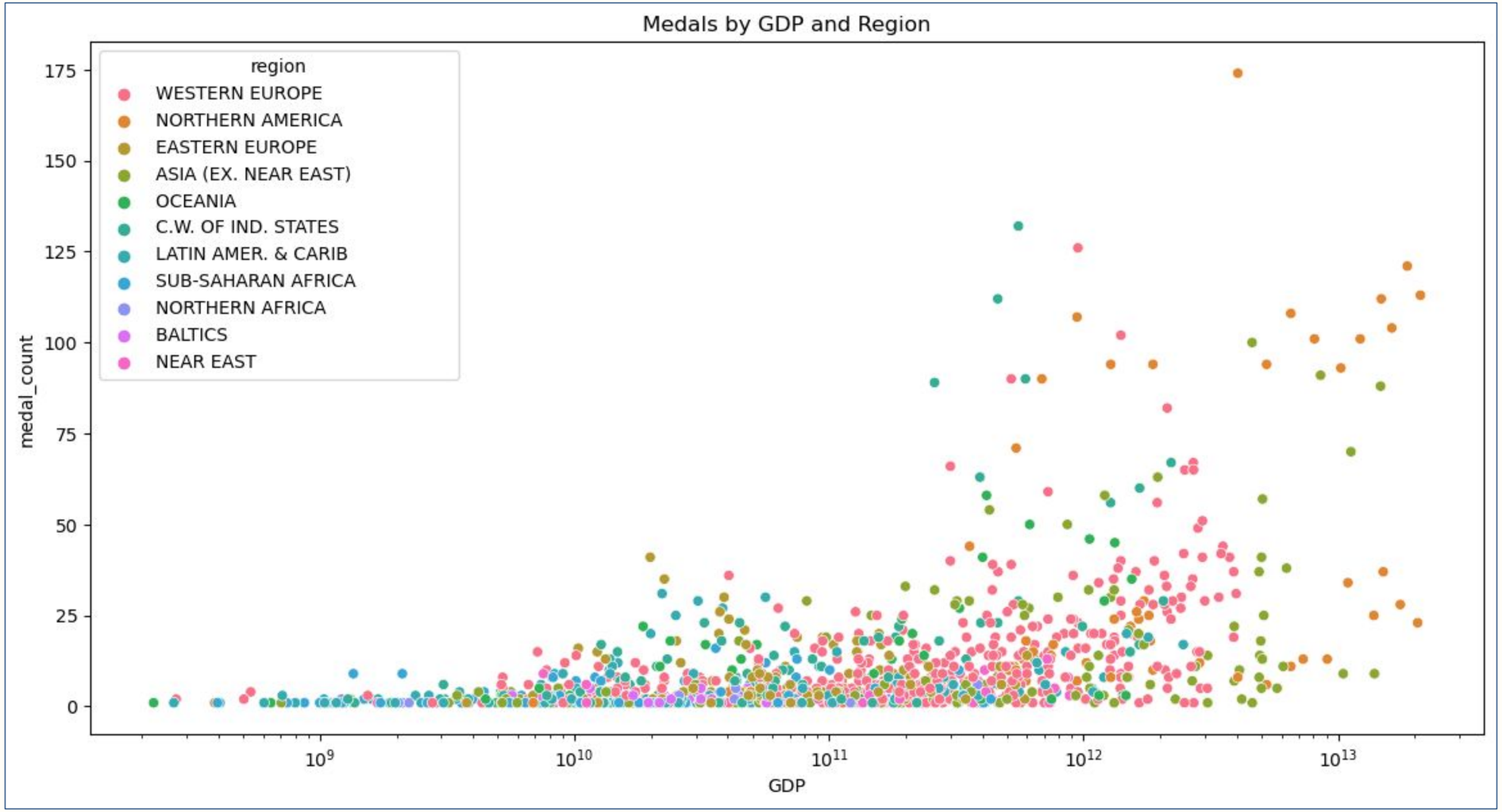
2

Exploratory Data Analysis

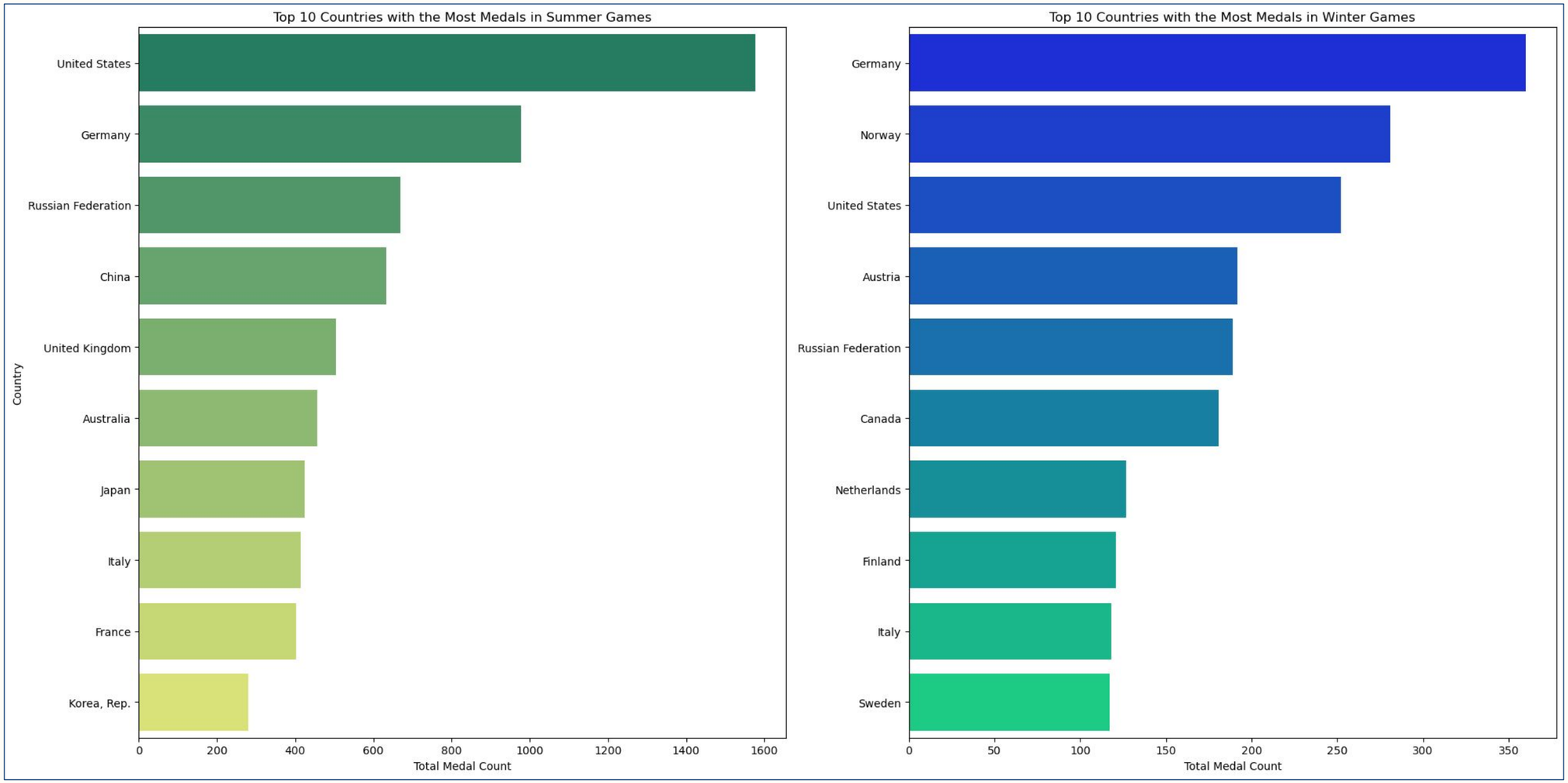
EDA


Distribution of Medals by Region
WESTERN EUROPE
ASIA (EX. NEAR EAST)
LATIN AMER. & CARIB
EASTERN EUROPE
C.W. OF IND. STATES
SUB-SAHARAN AFRICA
NORTHERN AMERICA
OCEANIA
NEAR EAST
NORTHERN AFRICA
BALTICS
region
0
50
100
150
200
250
300
350
400
count

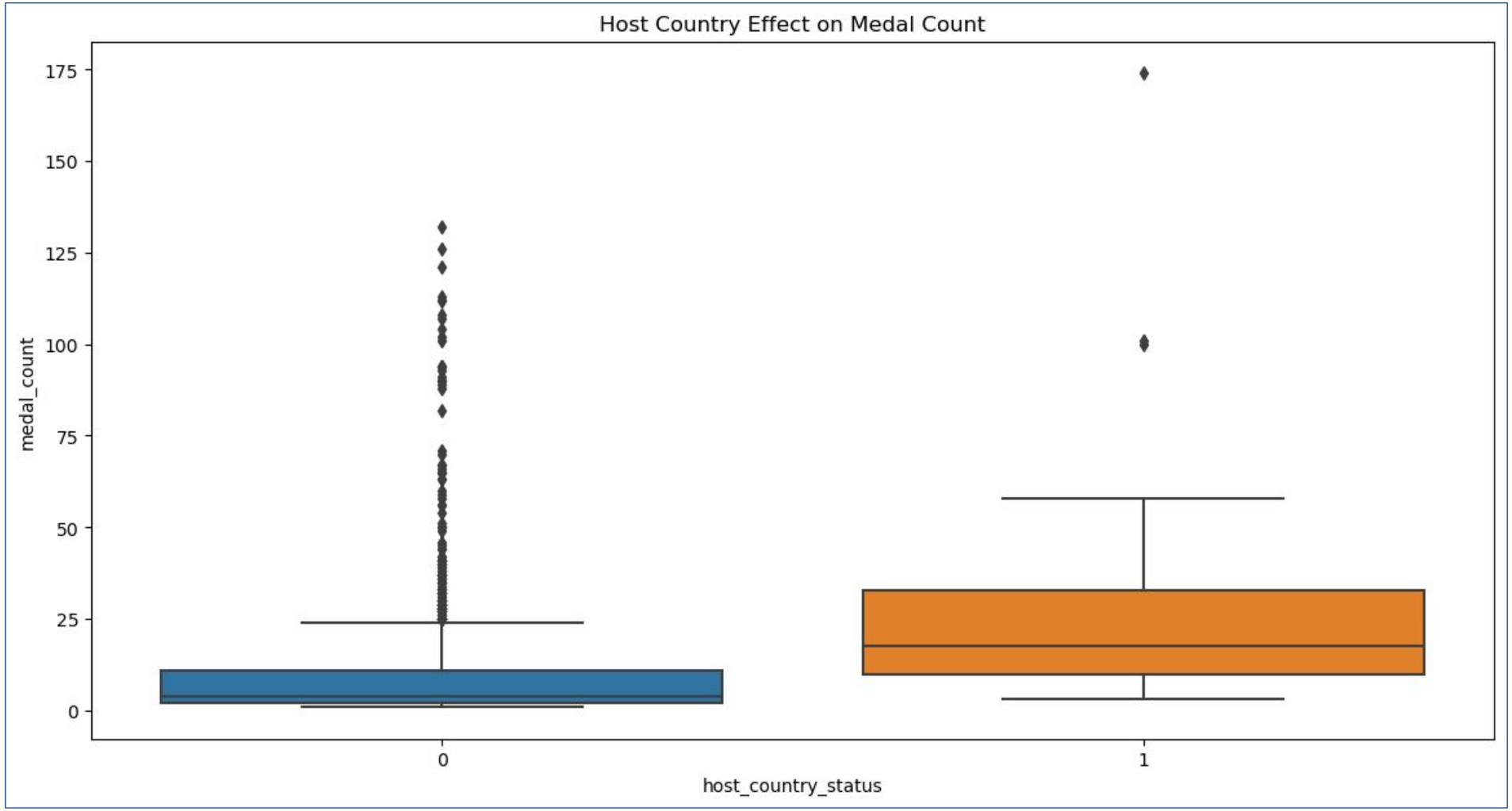
EDA


Medals by GDP and Region
region
WESTERN EUROPE
NORTHERN AMERICA
EASTERN EUROPE
ASIA (EX. NEAR EAST)
OCEANIA
C.W. OF IND. STATES
LATIN AMER. & CARIB
SUB-SAHARAN AFRICA
NORTHERN AFRICA
BALTICS
NEAR EAST
medal_count
0
25
50
75
100
125
150
175
10^9
10^{10}
10^{11}
10^{12}
10^{13}
GDP

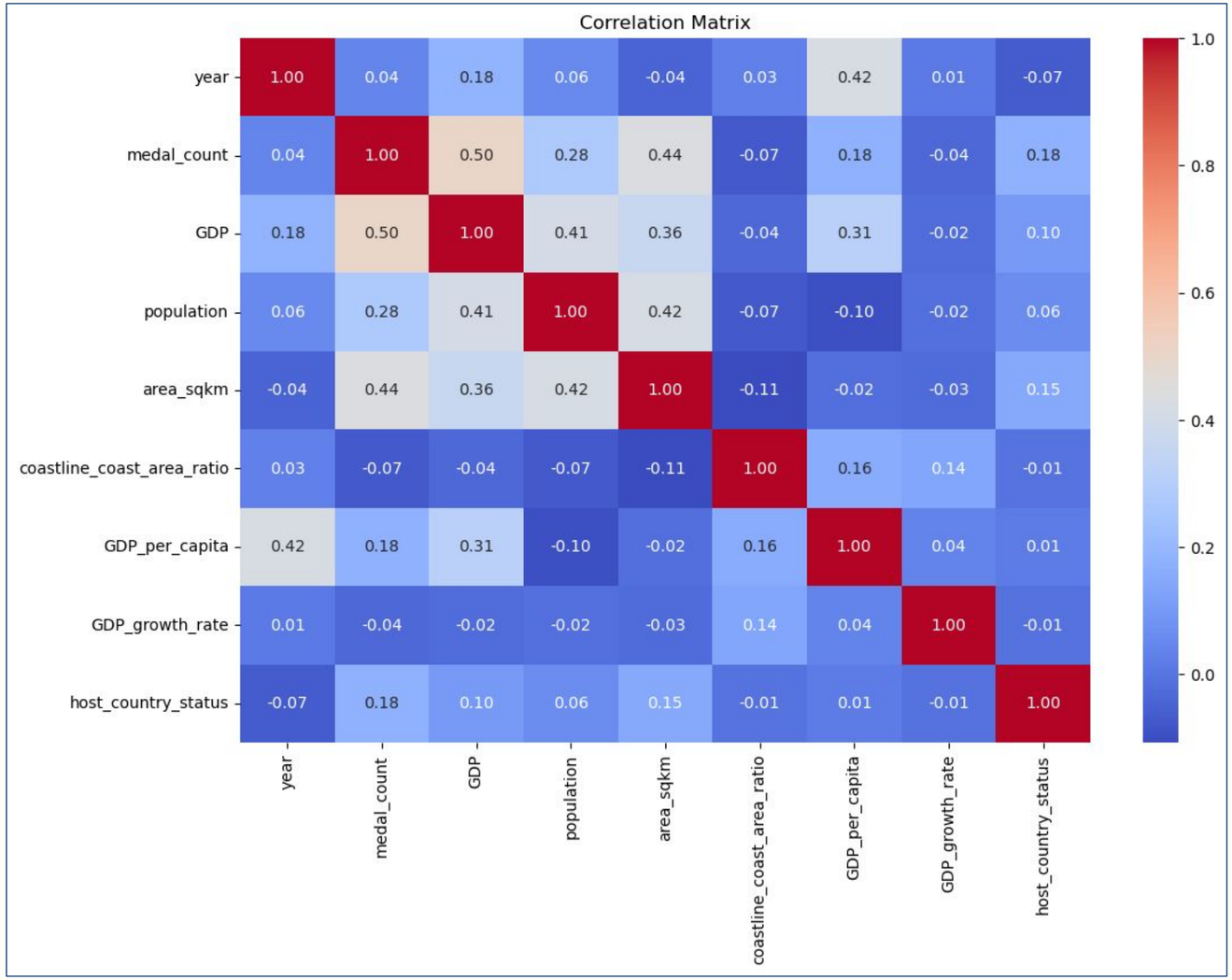
EDA


Top 10 Countries with the Most Medals in Summer Games
United States
Germany
Russian Federation
China
United Kingdom
Australia
Japan
Italy
France
Korea, Rep.
Country
0
200
400
600
800
1000
1200
1400
1600
Total Medal Count
Top 10 Countries with the Most Medals in Winter Games
Germany
Norway
United States
Austria
Russian Federation
Canada
Netherlands
Finland
Italy
Sweden
0
50
100
150
200
250
300
350
Total Medal Count

EDA


Host Country Effect on Medal Count
medal_count
175
150
125
100
75
50
25
0
0
1
host_country_status

EDA


Correlation Matrix
year 1.00 0.04 0.18 0.06 -0.04 0.03 0.42 0.01 -0.07
medal_count 0.04 1.00 0.50 0.28 0.44 -0.07 0.18 -0.04 0.18
GDP 0.18 0.50 1.00 0.41 0.36 -0.04 0.31 -0.02 0.10
population 0.06 0.28 0.41 1.00 0.42 -0.07 -0.10 -0.02 0.06
area_sqkm -0.04 0.44 0.36 0.42 1.00 -0.11 -0.02 -0.03 0.15
coastline_coast_area_ratio 0.03 -0.07 -0.04 -0.07 -0.11 1.00 0.16 0.14 -0.01
GDP_per_capita 0.42 0.18 0.31 -0.10 -0.02 0.16 1.00 0.04 0.01
GDP_growth_rate 0.01 -0.04 -0.02 -0.02 -0.03 0.14 0.04 1.00 -0.01
host_country_status -0.07 0.18 0.10 0.06 0.15 -0.01 0.01 -0.01 1.00
year
medal_count
GDP
population
area_sqkm
coastline_coast_area_ratio
GDP_per_capita
GDP_growth_rate
host_country_status
1.0
0.8
0.6
0.4
0.2
0.0

3

Machine Learning Models

Models Overview

Decision Tree

Bagging

Random Forest

Gradient Boosting

Summer Decision Tree Model

Evaluation Metrics
MAE: 7.785714
MSE: 224.551948
RMSE: 14.985057
R^2: 0.322729

Summer Bagging Model

Evaluation Metrics	Optimize Hyperparameters using Grid Search	Optimized Evaluation Metrics
MAE: 5.201948		MAE: 4.794478
MSE: 85.332792	• bootstrap: False • bootstrap_features: False • max_features: 0.7 • max_samples: 0.7 • n_estimators: 200	MSE: 57.540310
RMSE: 9.237575		RMSE: 7.585533
R^2: 0.742628		R^2: 0.826453

Summer Random Forest Model

Evaluation Metrics	Optimize Hyperparameters using Grid Search	Optimized Evaluation Metrics
MAE: 4.674238		MAE: 4.507012
MSE: 52.975255	• bootstrap: True • max_depth: 20 • max_features: sqrt • min_samples_leaf: 1 • min_samples_split: 2 • n_estimators: 100	MSE: 49.951581
RMSE: 7.278410		RMSE: 7.067643
R^2: 0.840221		R^2: 0.849341

Summer Gradient Boosting Model

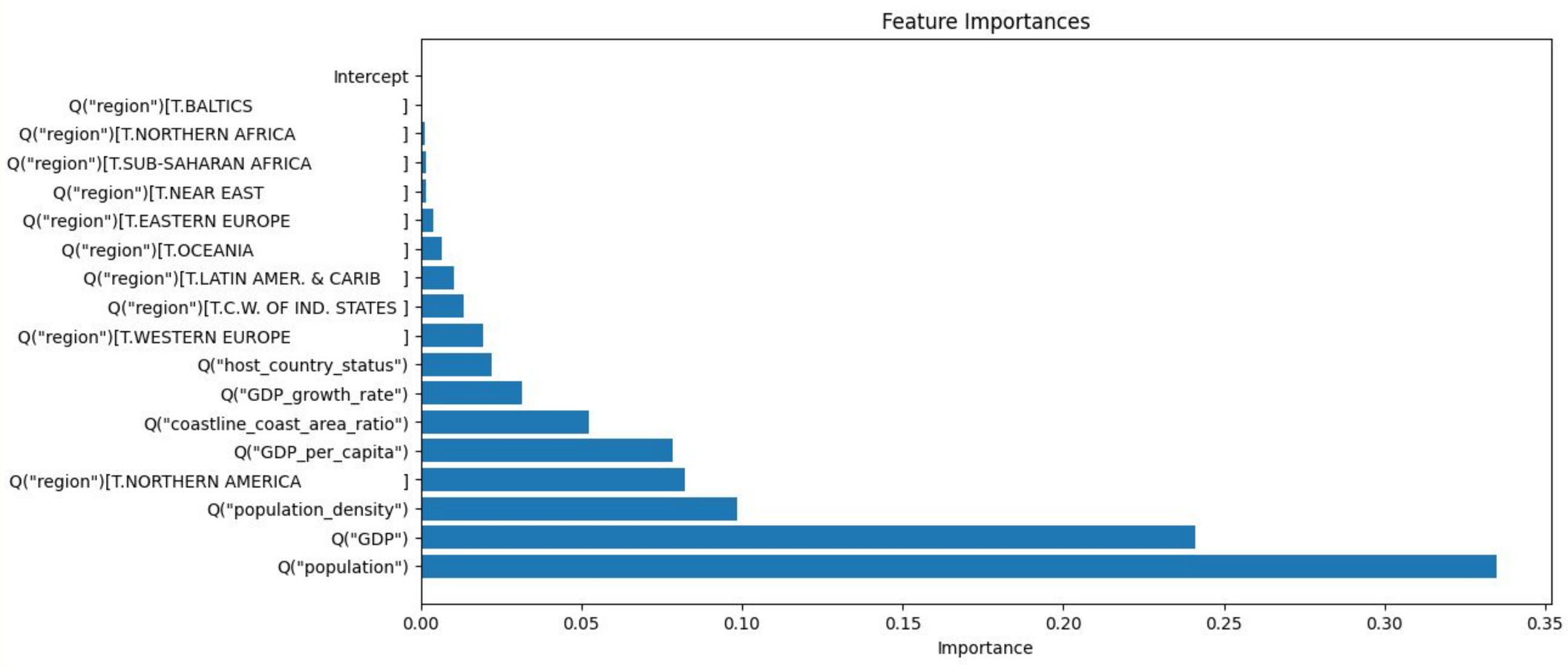
Evaluation Metrics	Optimize Hyperparameters using Grid Search	Optimized Evaluation Metrics
MAE: 5.013585		MAE: 4.538984
MSE: 66.179258		MSE: 47.197012
RMSE: 8.135063		RMSE: 6.870008
R^2: 0.800397		R^2: 0.857649

- learning_rate: 0.1
- max_depth: 5
- max_features: sqrt
- min_samples_leaf: 2
- min_samples_split: 2
- n_estimators: 250
- subsample: 0.8

Summer Olympics Models Summary

	MAE	MSE	RMSE	R^2
Decision Tree	7.786	224.552	14.985	0.323
Bagging	5.202	85.333	9.238	0.743
Random Forest	4.674	52.975	7.278	0.840
Gradient Boosting	5.014	66.179	8.135	0.800
Optimized Bagging	4.794	57.540	7.586	0.826
Optimized Random Forest	4.507	49.951	7.068	0.849
Optimized Gradient Boosting	4.539	47.197	6.870	0.858

Summer Olympics Models Feature Importance



Winter Decision Tree Model

Evaluation Metrics
MAE: 5.224138
MSE: 49.396552
RMSE: 7.028268
R^2: 0.484955

Winter Bagging Model

Evaluation Metrics	Optimize Hyperparameters using Grid Search	Optimized Evaluation Metrics
MAE: 3.950000		MAE: 3.884483
MSE: 30.017759		MSE: 30.226231
RMSE: 5.478846		RMSE: 5.497839
R^2: 0.687013		R^2: 0.684839

- bootstrap: False
- bootstrap_features: False
- max_features: 1.0
- max_samples: 0.7
- n_estimators: 100

Winter Random Forest Model

Evaluation Metrics	Optimize Hyperparameters using Grid Search	Optimized Evaluation Metrics
MAE: 4.059540		MAE: 3.615690
MSE: 31.426216	• bootstrap: False • max_depth: 20 • max_features: sqrt • min_samples_leaf: 1 • min_samples_split: 2 • n_estimators: 200	MSE: 27.948496
RMSE: 5.605909		RMSE: 5.286634
R^2: 0.672327		R^2: 0.708588

Winter Gradient Boosting Model

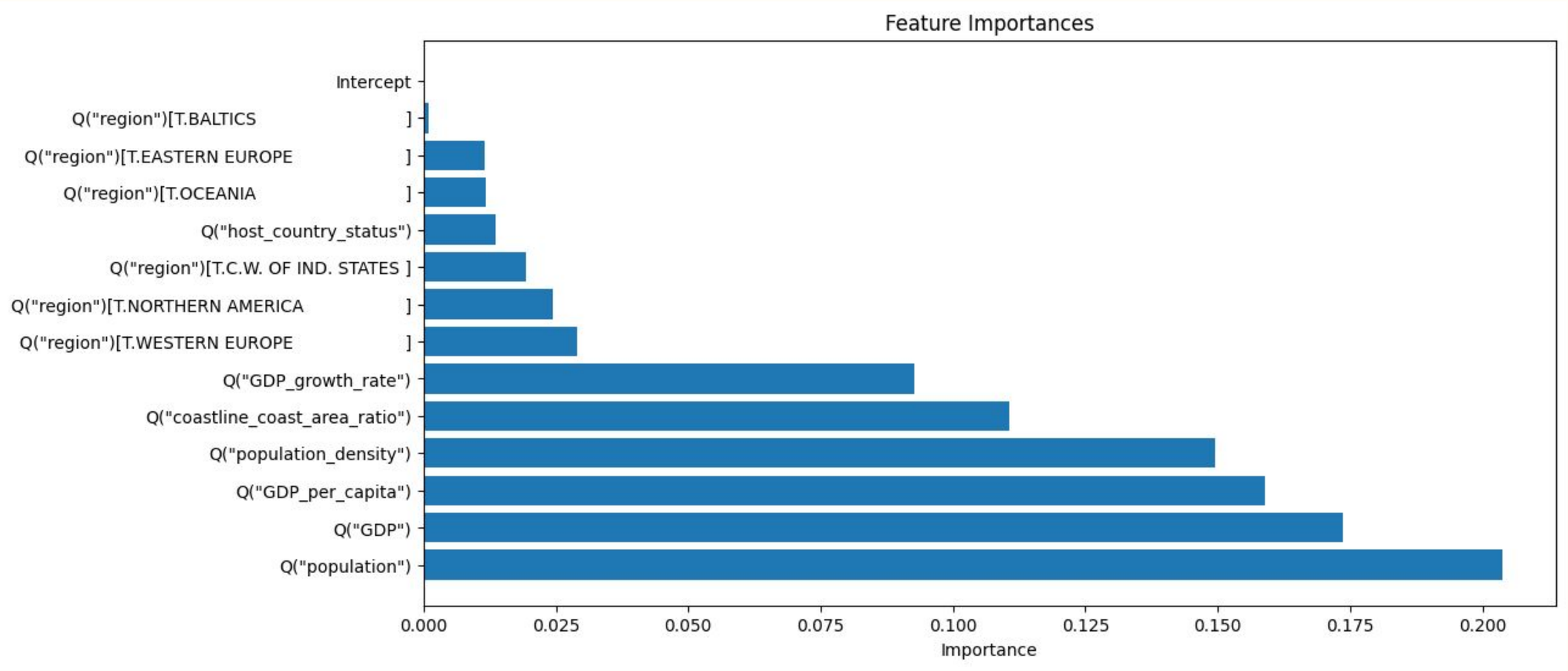
Evaluation Metrics	Optimize Hyperparameters using Grid Search	Optimized Evaluation Metrics
MAE: 4.152153		MAE: 3.675121
MSE: 32.578276		MSE: 28.368045
RMSE: 5.707738		RMSE: 5.326166
R^2: 0.660315		R^2: 0.704214

- learning_rate: 0.1
- max_depth: 3
- max_features: sqrt
- min_samples_leaf: 4
- min_samples_split: 2
- n_estimators: 250
- subsample: 0.8

Winter Olympics Models Summary

	MAE	MSE	RMSE	R^2
Decision Tree	5.224	49.397	7.028	0.485
Bagging	3.950	30.018	5.479	0.687
Random Forest	4.060	31.426	5.606	0.672
Gradient Boosting	4.152	32.578	5.708	0.660
Optimized Bagging	3.884	30.226	5.498	0.684
Optimized Random Forest	3.616	27.948	5.287	0.709
Optimized Gradient Boosting	3.675	28.368	5.326	0.704

Winter Olympics Models Feature Importance



Can our model predict how many medals the US will bring home in 2024?

Our prediction:

(from the Summer optimized gradient boosting model)

```
[127]: usa2024 = pd.read_csv('USA2024.csv').head(1).drop(columns=['Unnamed: 0'])
       grid_summer_gb.predict(usa2024)

[127]: array([113.81950708])
```

114 medals!

🥇🥈🥉🦅🦅

🇺🇸🇺🇸

Conclusion

- GDP and Population are the most predictors

- Limitations

  we weren't able to take into account a country's cultural emphasis on sports or other economic indicators besides GDP

Thanks for listening!

Any questions?