2008 Summer Olympics: A Use Case Study on Data Product

Xinqi Wang 2003316
Abduvosid Malikov 2002824
Li Jia 2000692

The Project

Questions:

- ➔ Which factor has a correlation with the number of Summer Olympic medals won by countries in year 2008?
  - GDP Per capita?
  - Population?
  - Government Expenditure on Sports?
- ➔ Any fun facts about Olympic medal winners?

Data Sources:

- ➔ Olympic_medals_1992-2008 (collected from Kaggle)
- ➔ European government expenditure (collected from Eurostat)
- ➔ Countries' GDP Per Capita and Population (collected from The World Bank API)

How we answer these questions

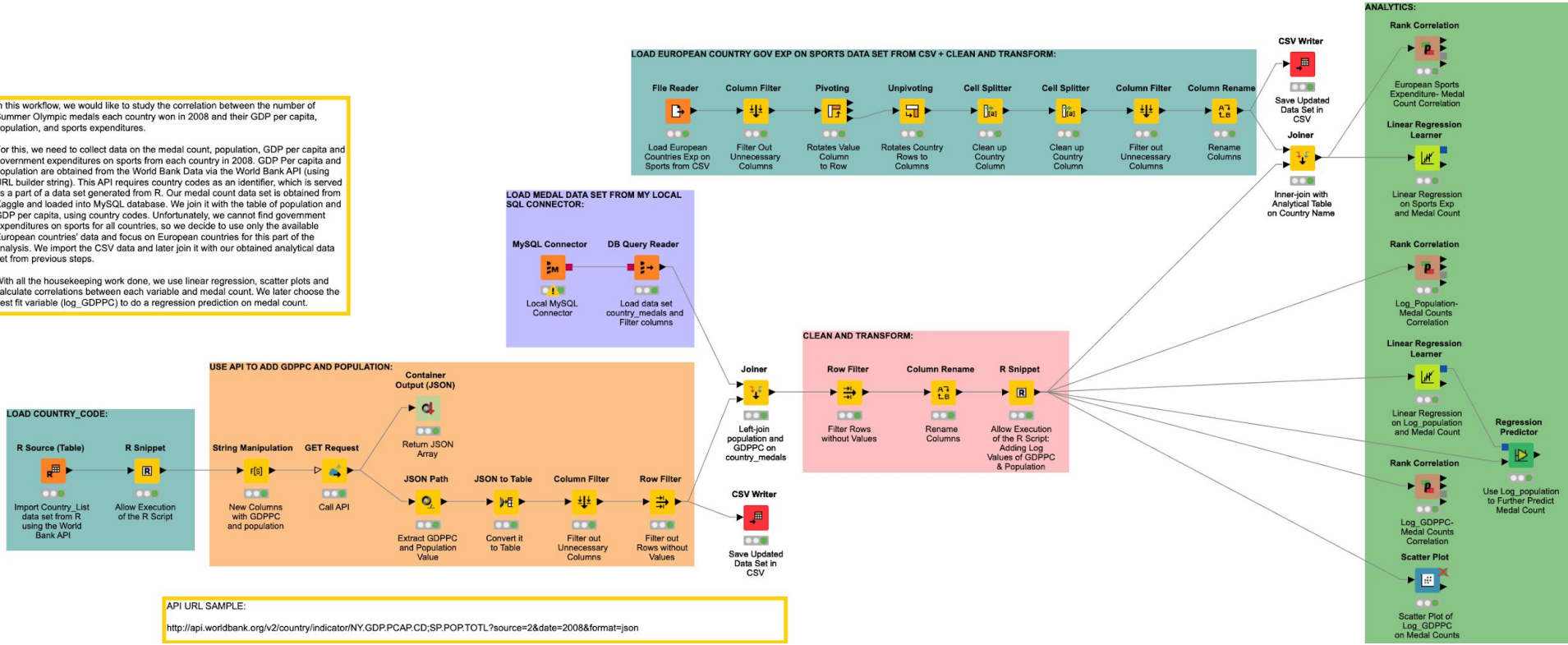
- MySQL WorkBench
  - Load Olympic_medals_1992-2008 data
- R
  - Get country_code data set using the World Bank API
- Knime
  - Build workflow and ETL data pipeline
- API
  - Used to get GDP per capita and population data from the World Bank
- Tableau
  - Create interactive summary graphs

Knime Workflow

In this workflow, we would like to study the correlation between the number of Summer Olympic medals each country won in 2008 and their GDP per capita, population, and sports expenditures.

For this, we need to collect data on the medal count, population, GDP per capita and government expenditures on sports from each country in 2008. GDP Per capita and population are obtained from the World Bank Data via the World Bank API (using URL builder string). This API requires country codes as an identifier, which is served as a part of a data set generated from R. Our medal count data set is obtained from Kaggle and loaded into MySQL database. We join it with the table of population and GDP per capita, using country codes. Unfortunately, we cannot find government expenditures on sports for all countries, so we decide to use only the available European countries' data and focus on European countries for this part of the analysis. We import the CSV data and later join it with our obtained analytical data set from previous steps.

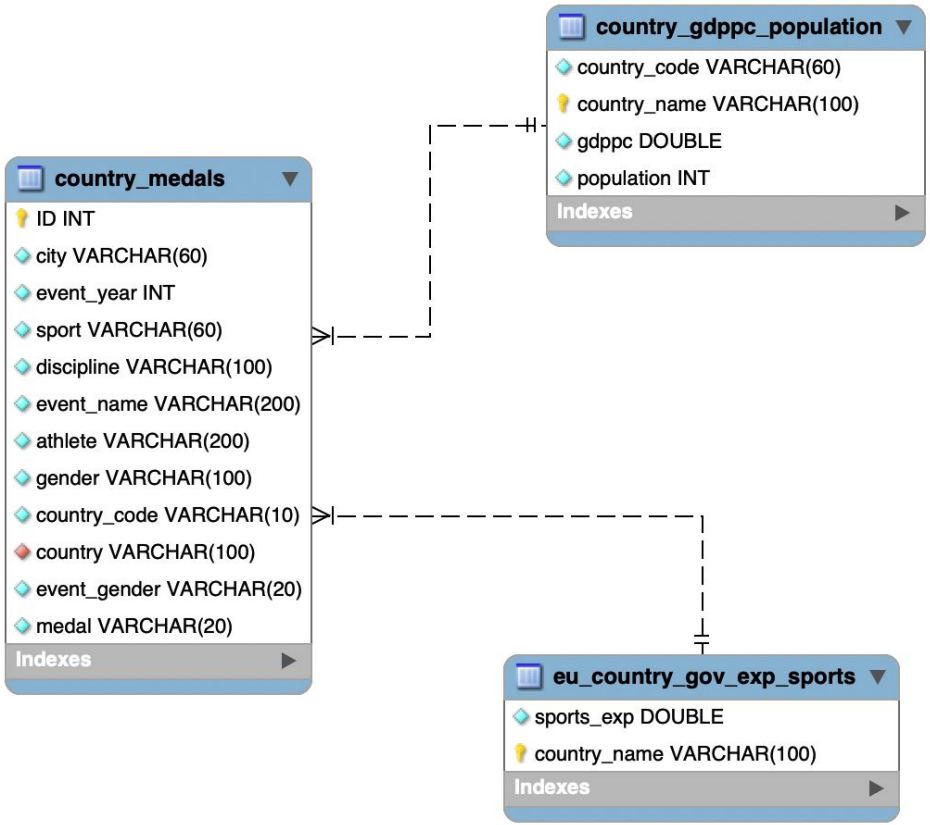
With all the housekeeping work done, we use linear regression, scatter plots and calculate correlations between each variable and medal count. We later choose the best fit variable (log_GDPPC) to do a regression prediction on medal count.



API URL SAMPLE:

http://api.worldbank.org/v2/country/indicator/NY.GDP.PCAP.CD;SP.POP.TOTL?source=2&date=2008&format=json

Our EER Diagram



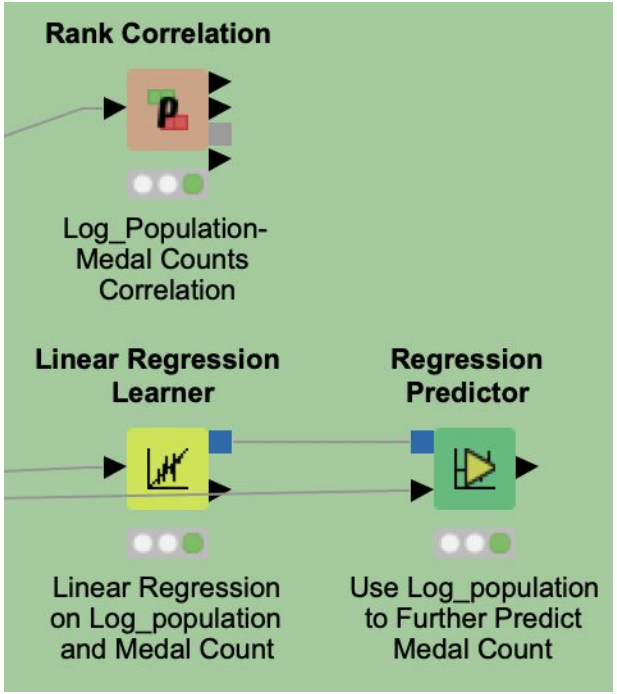
R Snippet

```
library(WDI)
library(tidyverse)

data <- WDI_data$country
codes <- data[,2]
country_names <- data[,3]
df <- data.frame("codes" = codes, "countries" = country_names)

df <- df %>% filter(!grepl("[[:digit:]]", df$codes))  # Filter numbers and grouping observations out
drop_id <- c("EU","HK","OE")
df <- df %>% filter( !grepl( paste( drop_id , collapse="|"), df$codes ) )

fl_iso2c <- substr(df$codes, 1, 1)  # drop values with certain starting char

retain_id <- c("XK","ZA","ZM","ZW")  # Check again if any grouping observations
d1 <- df %>% filter( grepl( "X", fl_iso2c ) | grepl( "Z", fl_iso2c ) &
                     !grepl( paste( retain_id , collapse="|"), df$codes ) )

# Save observations which are the opposite of our check results
df <- df %>% filter( !( grepl( "X", fl_iso2c ) | grepl( "Z", fl_iso2c ) &
                       !grepl( paste( retain_id , collapse="|"), df$codes ) ) )

rm( d1 , drop_id, fl_iso2c , retain_id )  # Clear no longer needed variables
df$countries[df$codes == "TW"] <- "Taiwan" # rename Taiwan Province
```

Correlation Between Medal Count and Population

Correlation Value	P value	Degrees of Freedom
0.290730	0.00805	80

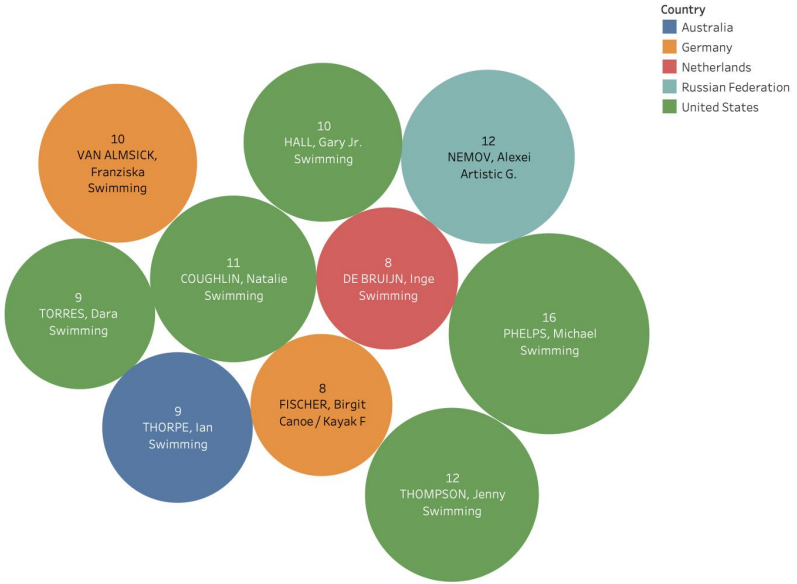
Variable	Coefficient	Std.Err	t-value	P > |t|
log_population	14.386	3.077	4.675	0
intercept	-215.37	51.492	-4.183	0



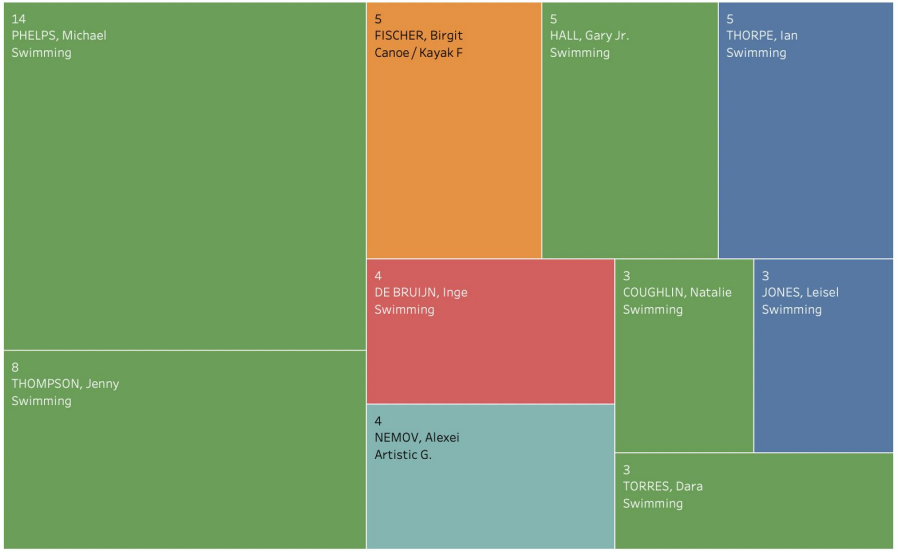
Who Won the Most Medals/Golden Medals for His/Her Country?

CEU CENTRAL EUROPEAN UNIVERSITY

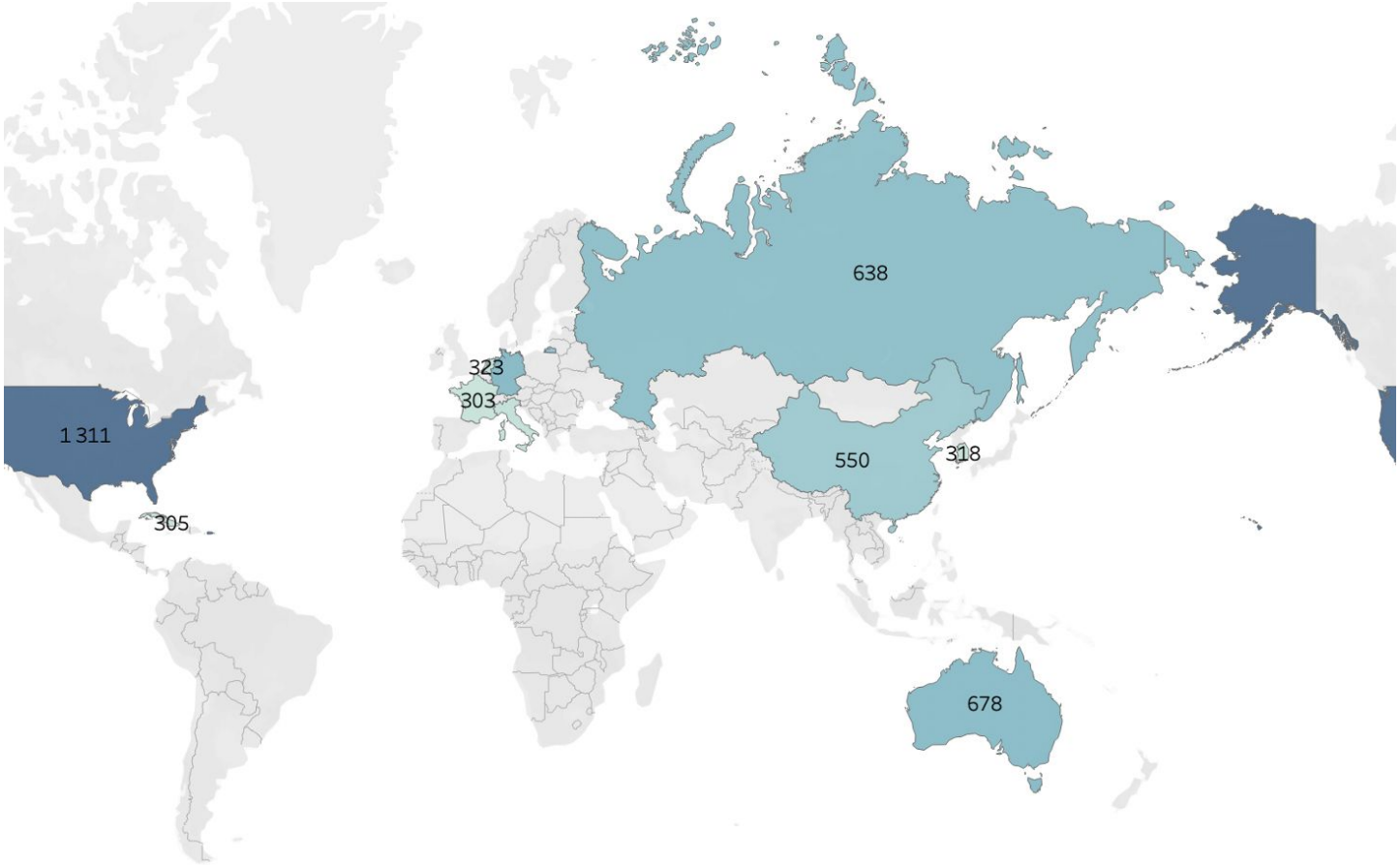
Olympic Medal Winners, Top 10



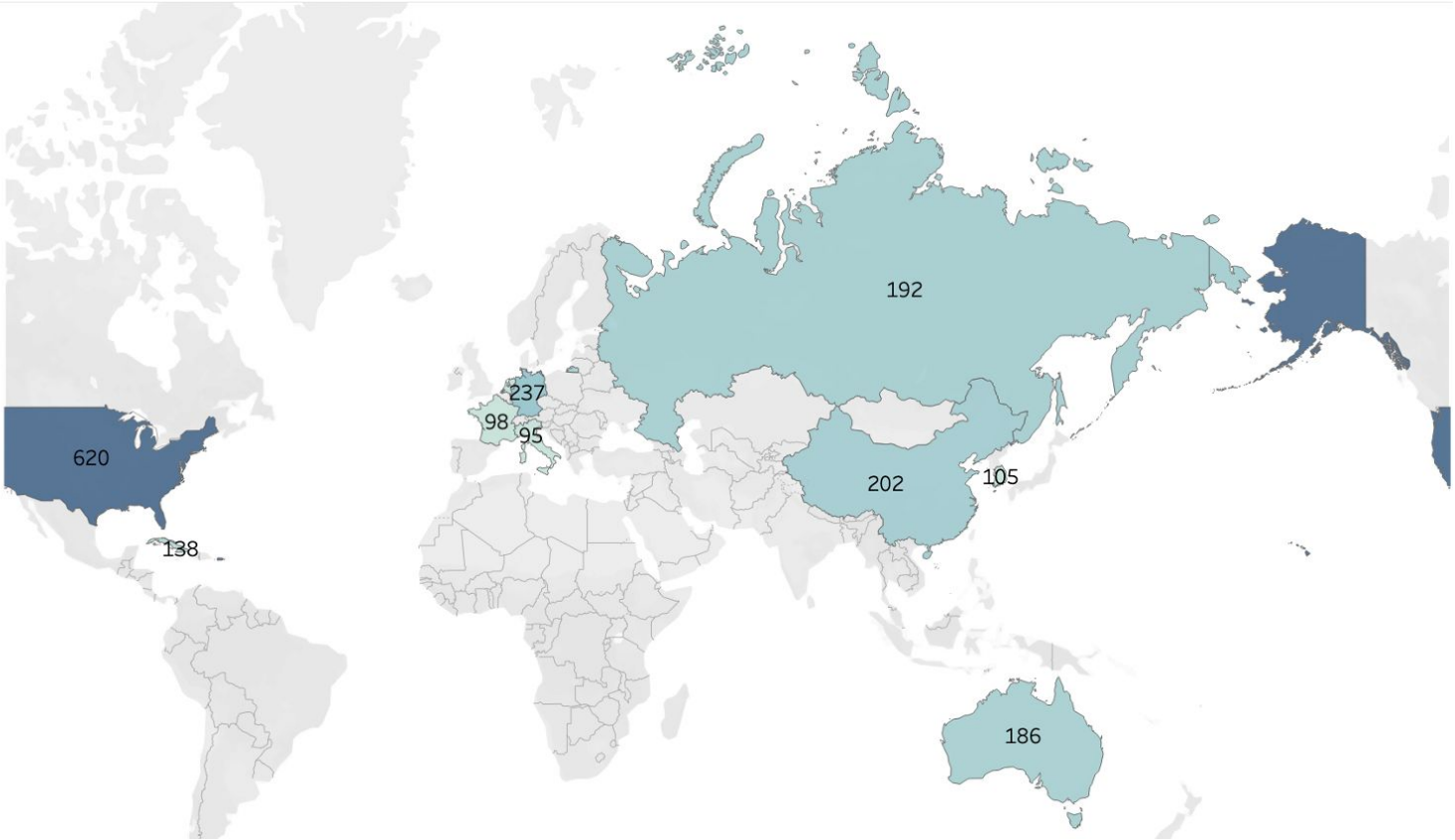
Olympic Gold Medal Winners, Top 10



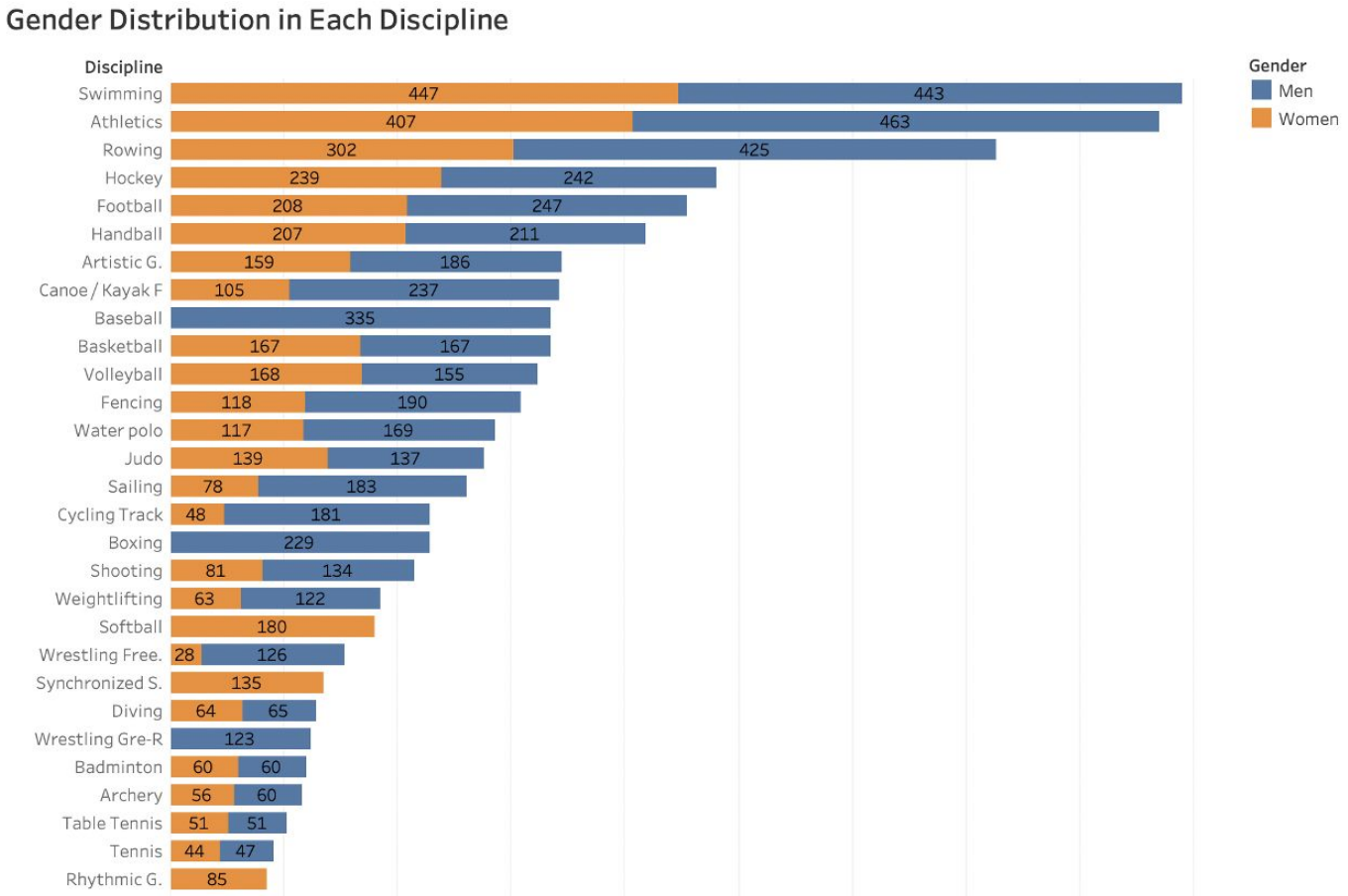
Which Country Has the Most Medals?



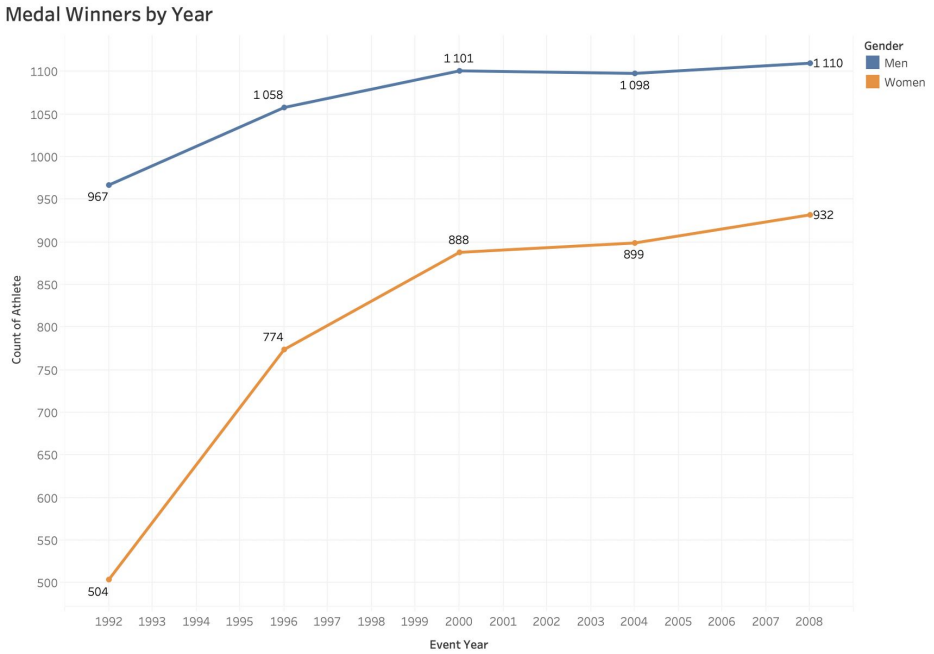
Which Country Has the Most Golden Medals?

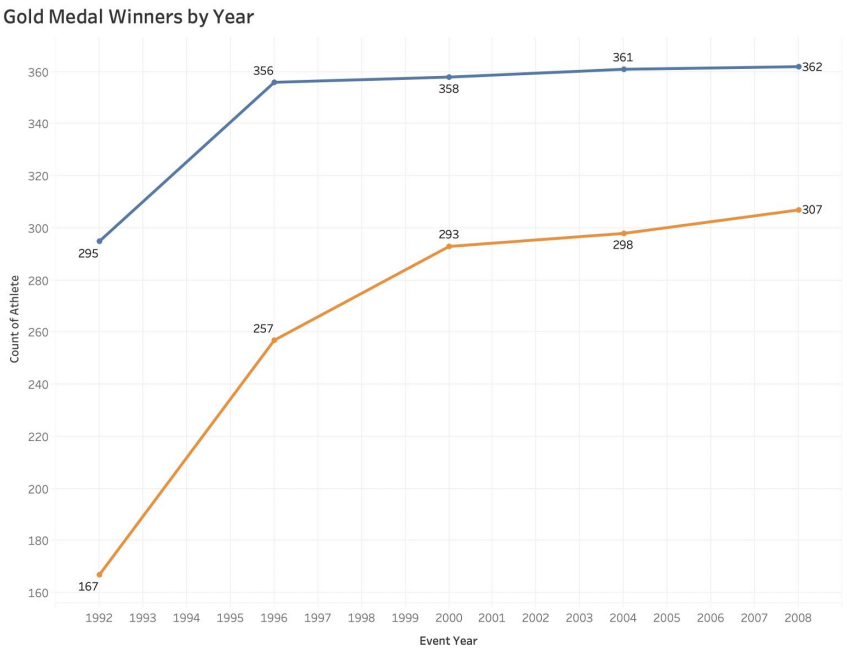


Which Discipline Has the Most Male/Female Medal Winners?



Which Year Has the Most Female Medal/Golden Medal Winner?





CEU CENTRAL EUROPEAN UNIVERSITY

Thank you