Problem Statement

We want to evaluate whether the Olympic games are objectively fair, or the winning results are skewed or influenced by the wealth of competing countries and other factors.

Hypotheses

- The country's Gross Domestic Product Per Capita (GDPPC), Human Development Index (HDI), and Population can be used to predict the **total number of medals** won.
- The Gross Domestic Product Per Capita (GDPPC) and the ratio of women to men in participation for countries can be used to predict the **total number of gold medals** won.

Potential Use of Results

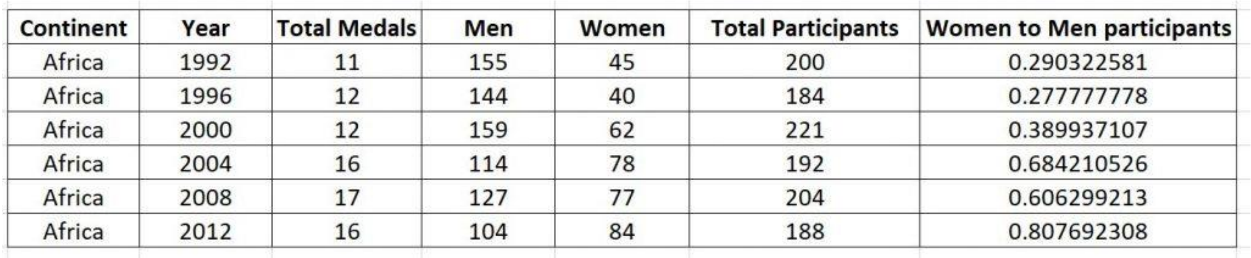
- Athletes with multiple citizenships can choose which country they want to represent in the Olympics. Governments can use the results and allocate more budget for the sports sector like infrastructure development, capacity building, etc.
- The International Olympic Committee can help countries to formulate and implement inclusive policies to organize the event in a fair manner.

Exploratory Data Analysis

Firstly, we selected 26 countries from different continents based on the availability of data on the country including the number of Olympic medals won, and the distribution of participants by gender. We also determined the ratio of men to women who participated in the Olympic games each year from 1999-2012, from each continent. The countries selected are as follows:

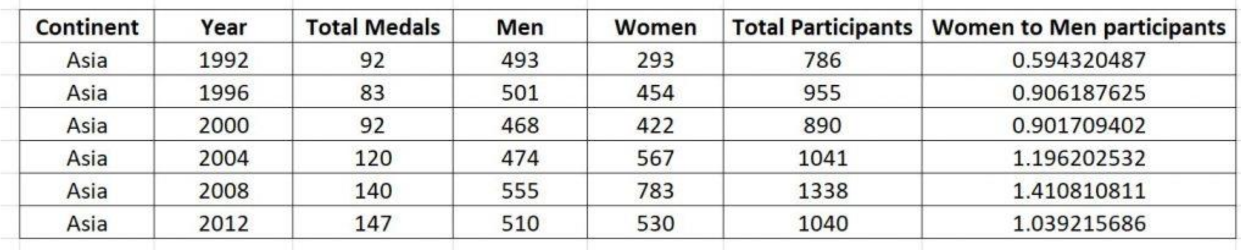
- **Africa** – South Africa, Ghana, Kenya, Zimbabwe
- **Asia** – India, Japan, China, Vietnam, South Korea
- **Australia** – Australia, and New Zealand
- **Europe** – Germany, Sweden, Great Britain, Poland, Bulgaria
- **North America** – United States, Canada, Cuba, Jamaica, Mexico
- **South America** – Brazil, Argentina, Colombia, Venezuela, Chile,

The following summary tables show the number of Olympic medals won, and the distribution of participants by gender for every continent. From our summary table, we can observe that the ratio of women to men participants in the Olympic games has generally been increasing for each continent, year to year. In some cases, the number of women competing in the Olympics came to surpass the number of men over the years. Additionally, for Asia, the number of women participating in the Olympic games is seen to exceed the number of men from 2004 onwards.



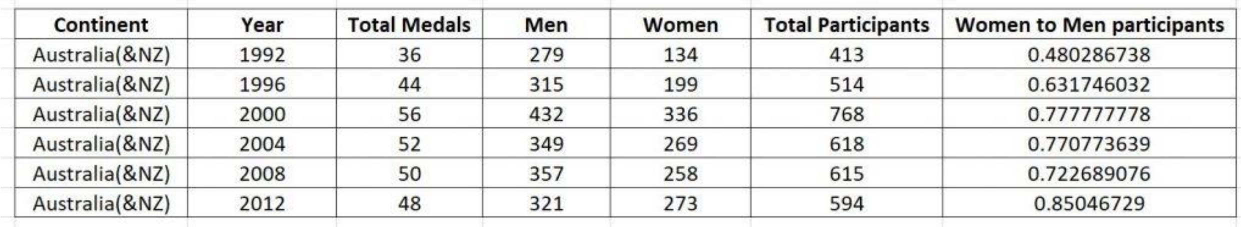
Continent	Year	Total Medals	Men	Women	Total Participants	Women to Men participants
Africa	1992	11	155	45	200	0.290322581
Africa	1996	12	144	40	184	0.277777778
Africa	2000	12	159	62	221	0.389937107
Africa	2004	16	114	78	192	0.684210526
Africa	2008	17	127	77	204	0.606299213
Africa	2012	16	104	84	188	0.807692308

Table for Africa



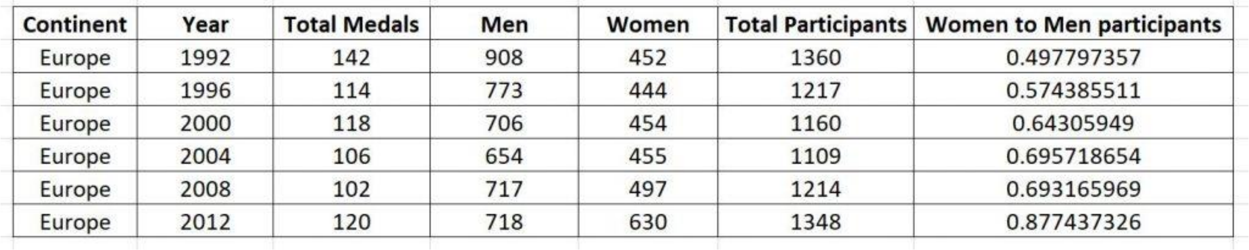
Continent	Year	Total Medals	Men	Women	Total Participants	Women to Men participants
Asia	1992	92	493	293	786	0.594320487
Asia	1996	83	501	454	955	0.906187625
Asia	2000	92	468	422	890	0.901709402
Asia	2004	120	474	567	1041	1.196202532
Asia	2008	140	555	783	1338	1.410810811
Asia	2012	147	510	530	1040	1.039215686

Table for Asia



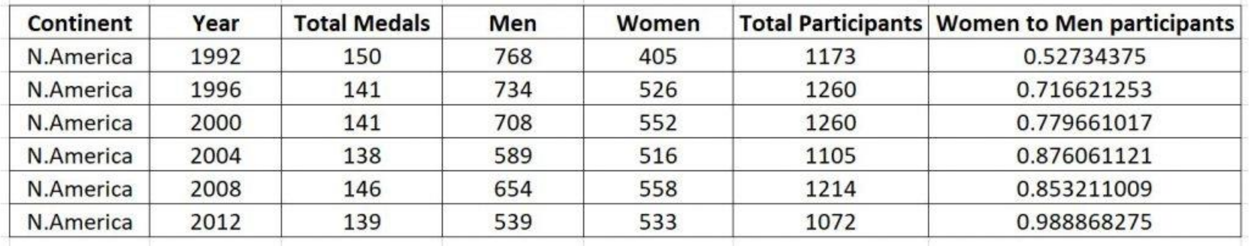
Continent	Year	Total Medals	Men	Women	Total Participants	Women to Men participants
Australia(&NZ)	1992	36	279	134	413	0.480286738
Australia(&NZ)	1996	44	315	199	514	0.631746032
Australia(&NZ)	2000	56	432	336	768	0.777777778
Australia(&NZ)	2004	52	349	269	618	0.770773639
Australia(&NZ)	2008	50	357	258	615	0.722689076
Australia(&NZ)	2012	48	321	273	594	0.85046729

Table for Australia and New Zealand



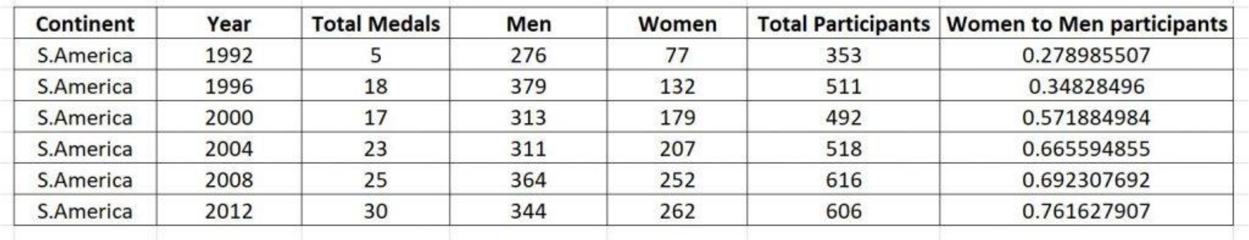
Continent	Year	Total Medals	Men	Women	Total Participants	Women to Men participants
Europe	1992	142	908	452	1360	0.497797357
Europe	1996	114	773	444	1217	0.574385511
Europe	2000	118	706	454	1160	0.64305949
Europe	2004	106	654	455	1109	0.695718654
Europe	2008	102	717	497	1214	0.693165969
Europe	2012	120	718	630	1348	0.877437326

Table for Europe



Continent	Year	Total Medals	Men	Women	Total Participants	Women to Men participants
N.America	1992	150	768	405	1173	0.52734375
N.America	1996	141	734	526	1260	0.716621253
N.America	2000	141	708	552	1260	0.779661017
N.America	2004	138	589	516	1105	0.876061121
N.America	2008	146	654	558	1214	0.853211009
N.America	2012	139	539	533	1072	0.988868275

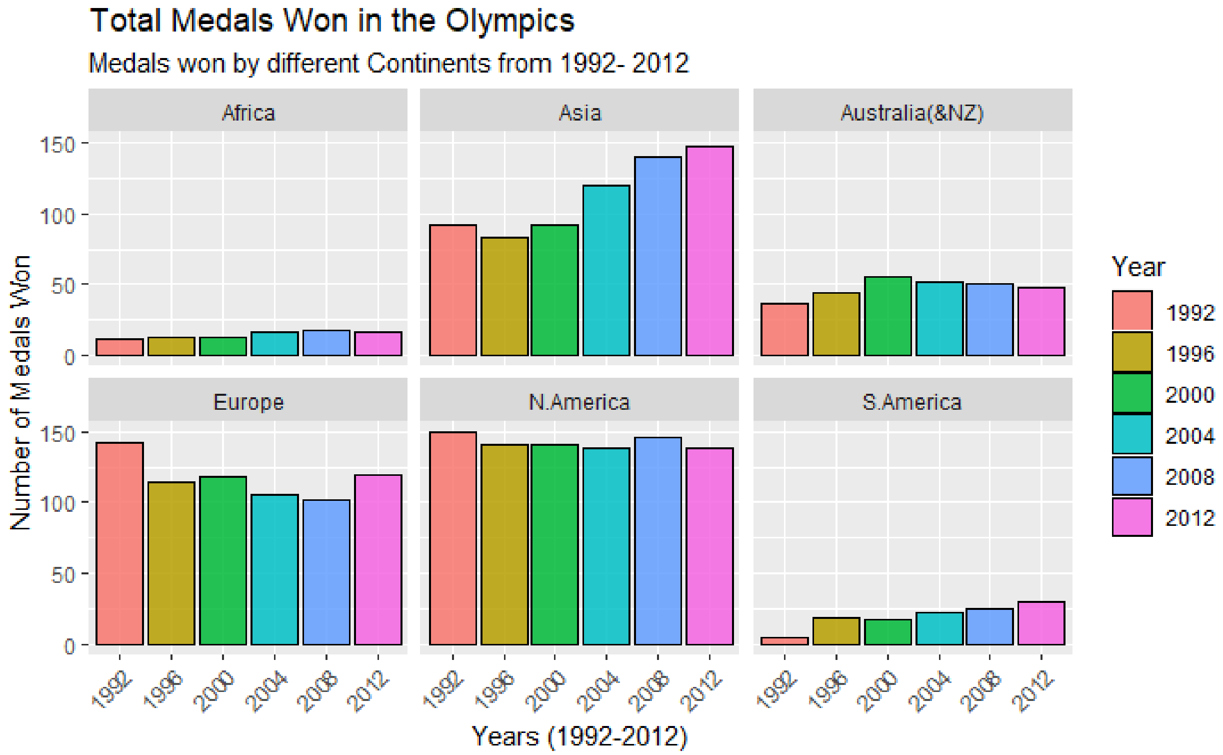
Table for North America



Continent	Year	Total Medals	Men	Women	Total Participants	Women to Men participants
S.America	1992	5	276	77	353	0.278985507
S.America	1996	18	379	132	511	0.34828496
S.America	2000	17	313	179	492	0.571884984
S.America	2004	23	311	207	518	0.665594855
S.America	2008	25	364	252	616	0.692307692
S.America	2012	30	344	262	606	0.761627907

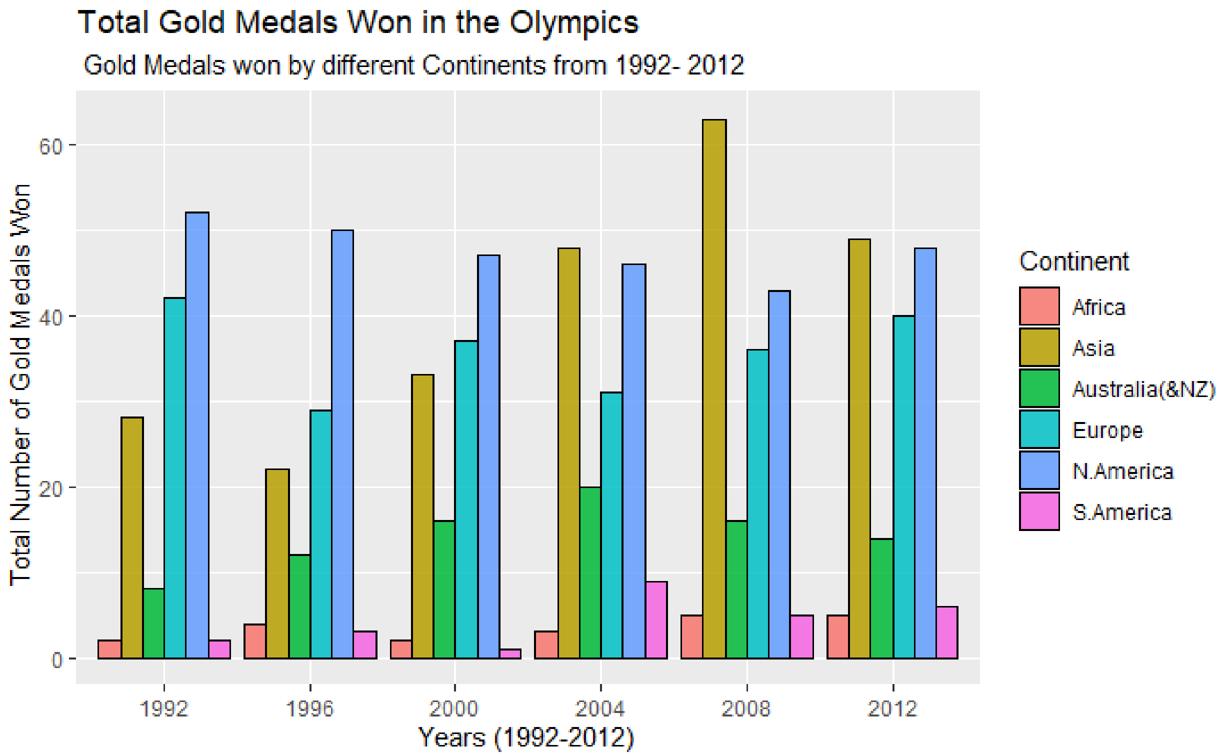
Table for South America

Table: Showing the total medals won, the ratio of women to men in participation for different continents from 1992-2012



Bar Plots Showing the Total Number of Medals won for different continents over different years

To show the total number of medals won for different continents over the years, we opted to use bar plots to display the variables in our dataset. For our first visualization, we used a bar plot to visualize the relationship between the Years (1992- 2012) and the Total Number of Medals won for each continent. From the bar plot, we can see that the total number of medals won in the Olympic games has been increasing over the years for Africa, Asia, and South America. However, the total number of medals won is seen to be decreasing for Europe up to 2008 and increasing in the year 2012. Finally, the total number of medals for Australia (including New Zealand) and North American has been fluctuating over the years. For Australia, the number of medals is seen to be increasing up to 2000 but is seen to be decreasing 2000 onwards. On the other hand, the total number of medals for North America has been decreasing until 2004 but increasing in 2008 and decreasing again in 2012.



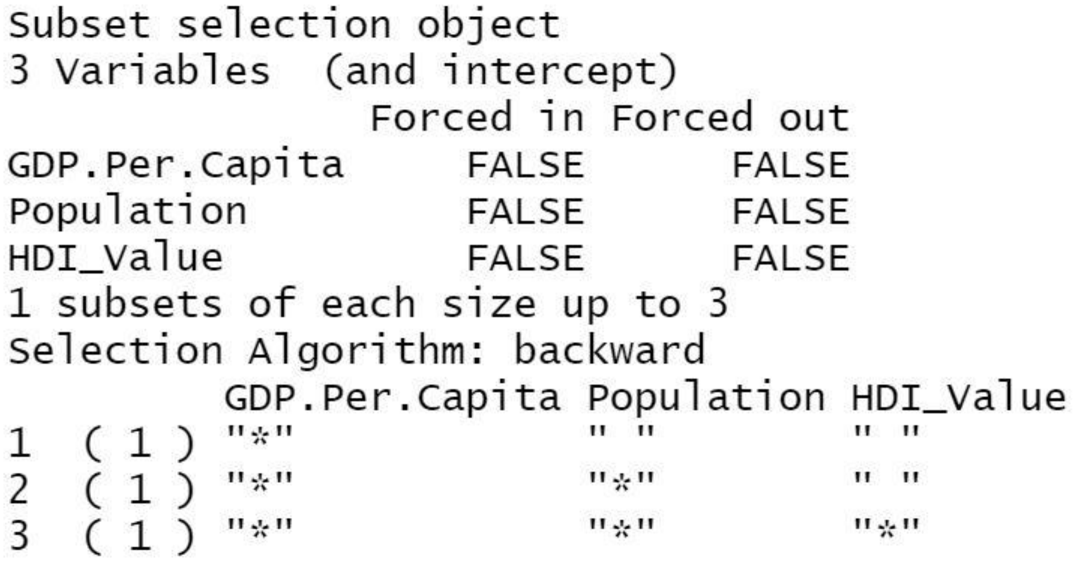
Total Gold Medals Won for different continents in the Olympic Games

For our second visualization, we plotted a bar plot visualizing the relationship between the total number of gold medals won and the Years (1992 – 2012). For each continent, except for Europe, the number of gold medals won at the Olympic games increased over time. This is true for Africa, Asia, South America, Europe, and Australia. The number of gold medals won by North America was in a downtrend since 2000, and only broke out of this downtrend in 2012. The number of gold medals won by Europe declined in 2004 but resumed their uptrend again in the following years. For South America, these results declined in the year 2000, resuming their uptrend again in 2004. For Asia, the number of gold medals declined in 1996, and then again in 2012. *These results primarily point to that for most of the continents excluding North America, the average number of gold medals won by each continent had a positive linear relationship with time. For North America, the number of gold medals had a negative linear relationship with time*.

Modeling

Since all our predictor and response variables are quantitative, we used multiple linear regression to predict the total number of medals won based on the GDP Per Capita, Population, and HDI Value. We also predicted the total number of gold medals won based on GDP Per Capita and the ratio of Women to Men participating in the Olympic events for different countries.

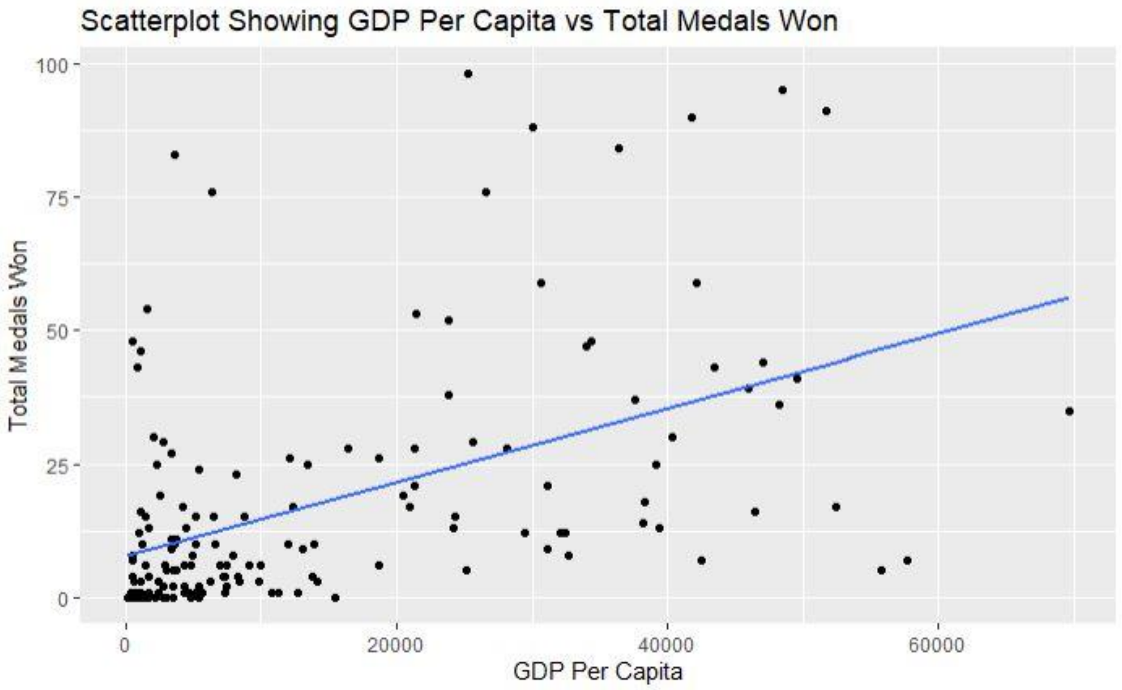
Initially, we hypothesized that GDP Per Capita, Population, and HDI Value would be good predictors for the total number of medals won. However, using backward selection, we found that GDP Per Capita and Population were the most accurate predictors. So, we only used GDP Per Capita and Population for our multiple linear regression model to see their effects on the total number of medals won.

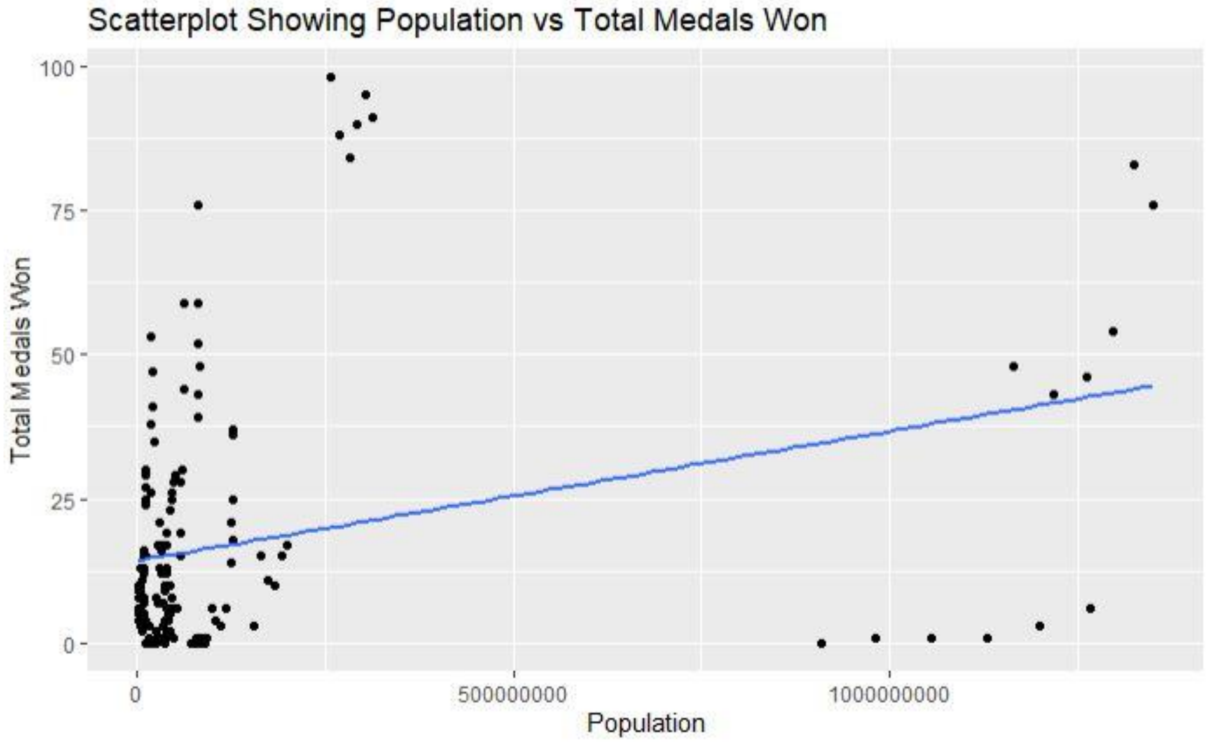


```
Subset selection object
3 Variables  (and intercept)
               Forced in Forced out
GDP.Per.Capita     FALSE      FALSE
Population         FALSE      FALSE
HDI_Value          FALSE      FALSE
1 subsets of each size up to 3
Selection Algorithm: backward
         GDP.Per.Capita Population HDI_Value
1  ( 1 ) "*"            " "        " "
2  ( 1 ) "*"            "*"        " "
3  ( 1 ) "*"            "*"        "*"
```

Results of the backward selection to show the most accurate predictor variables.

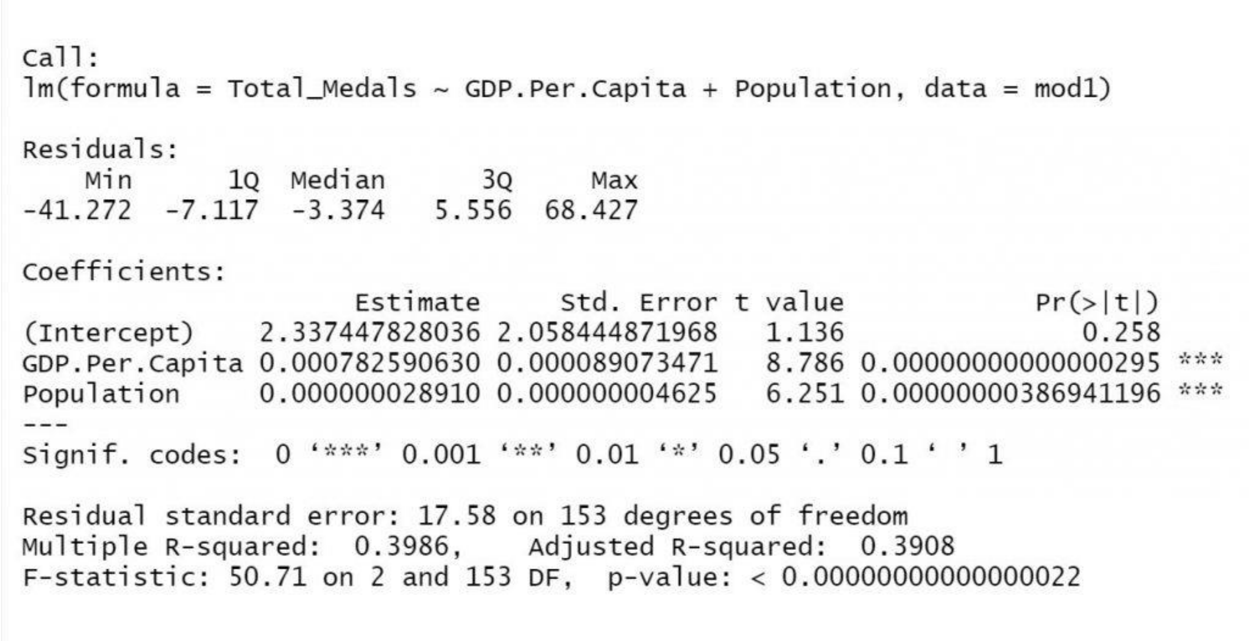
Using the results of backward selection, we plotted a scatterplot to determine if there was a relationship between the predictors and response variables.





There does not seem to be an obvious relationship between the plots for GDP Per Capita and the total number of medals. However, there is no relationship between Population and the total number of medals. Even though the relationship is ambiguous, we still want to use both GDP Per Capita and Population in the multiple linear regression model to see if they have any effect on the total number of medals.

The results of the multiple linear regression are shown below

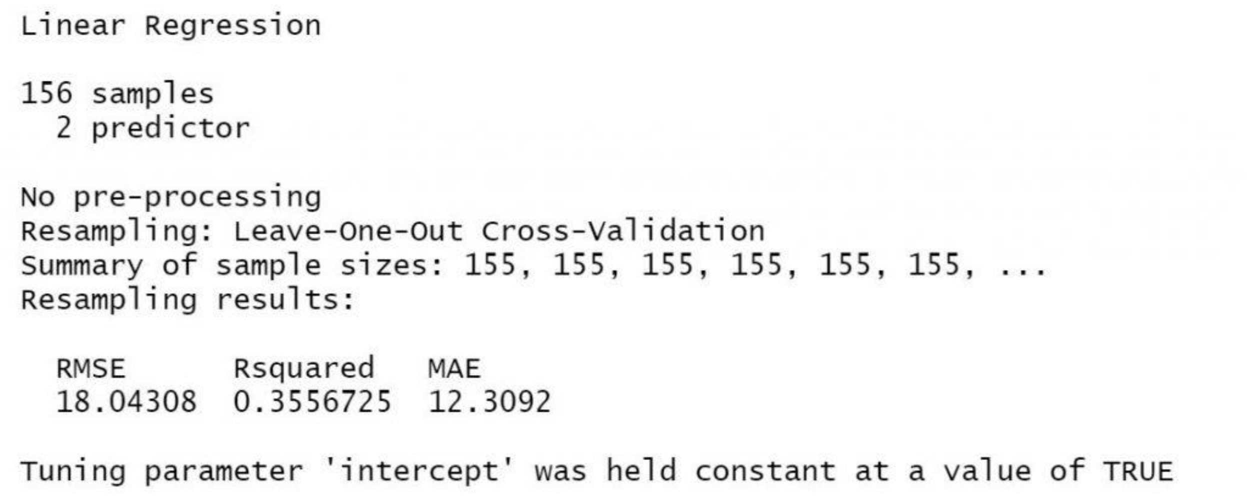


```
Call:
lm(formula = Total_Medals ~ GDP.Per.Capita + Population, data = mod1)

Residuals:
    Min      1Q  Median      3Q     Max 
-41.272  -7.117  -3.374   5.556  68.427 

Coefficients:
                   Estimate     Std. Error t value           Pr(>|t|)    
(Intercept)    2.337447828036 2.058444871968   1.136              0.258    
GDP.Per.Capita 0.000782590630 0.000089073471   8.786 0.0000000000000295 ***
Population     0.000000028910 0.000000004625   6.251 0.00000000386941196 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.58 on 153 degrees of freedom
Multiple R-squared:  0.3986,    Adjusted R-squared:  0.3908 
F-statistic: 50.71 on 2 and 153 DF,  p-value: < 0.00000000000000022
```

Multiple Linear Regression Using GDP Per Capita and Population

Observing the adjusted R-squared value, we can see that there is a relationship between the predictors (GDP Per Capita, and Population) and the response variable (total number of medals). However, the relationship between these variables is not very strong since the adjusted R-squared value is 0.3908, which is closer to 0 than 1. Furthermore, looking at the intercepts, we can say that for every $10,000 increase in the GDP Per Capita of a country, the total number of medals won increases by 7. Additionally, for every 1 million increases in population, the number of medals won increases by 0.028.

To determine the accuracy of the model we did leave-one-out cross-validation and the results of the cross-validation are shown below.



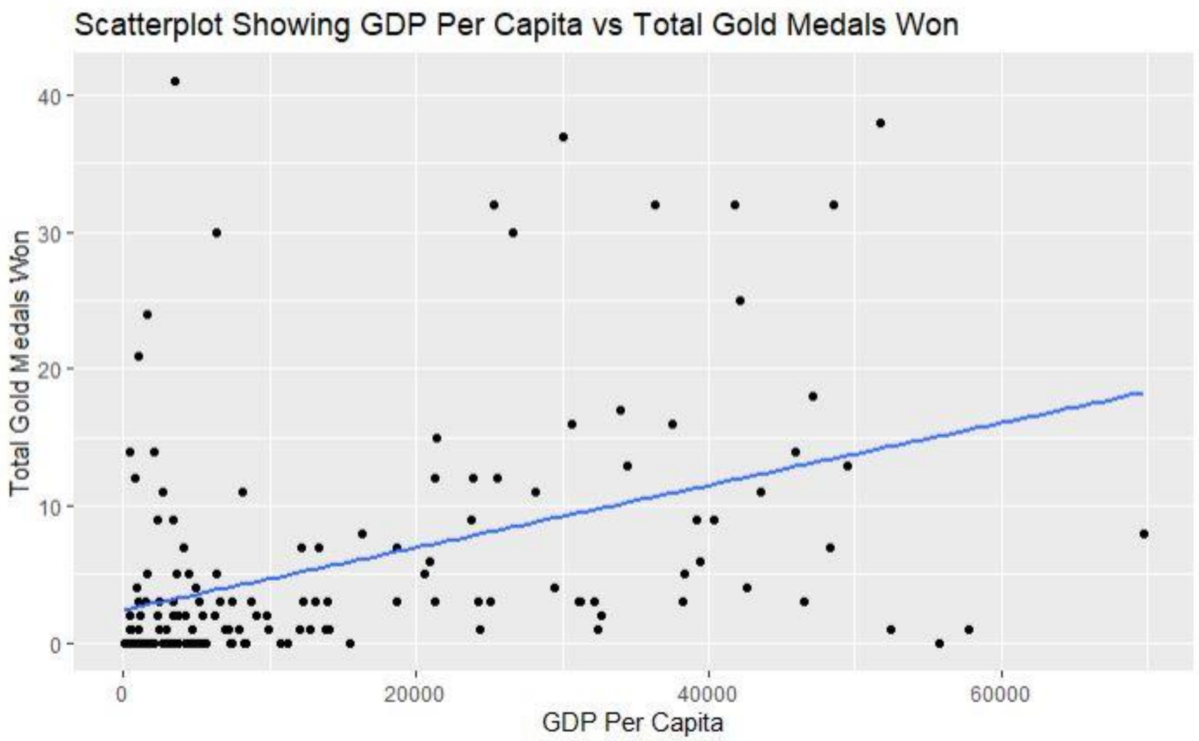
```
Linear Regression

156 samples
  2 predictor

No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 155, 155, 155, 155, 155, 155, ...
Resampling results:

  RMSE      Rsquared   MAE
  18.04308  0.3556725  12.3092

Tuning parameter 'intercept' was held constant at a value of TRUE
```

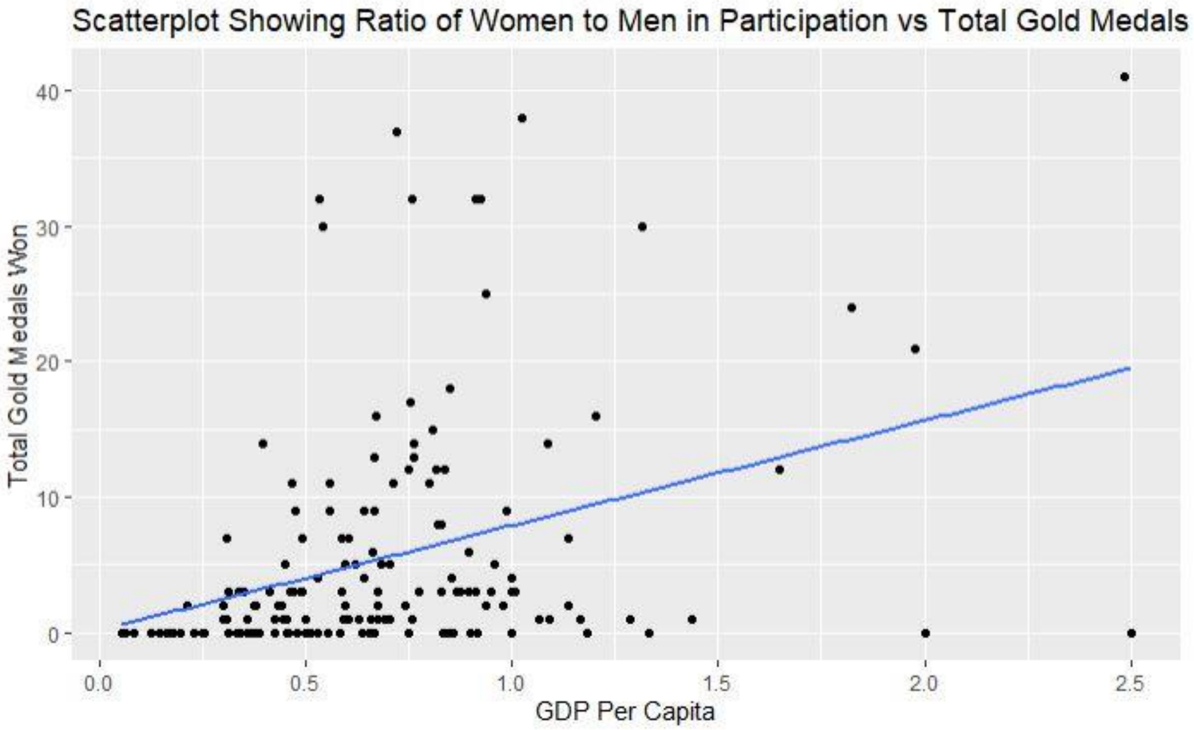
Leave-One-Out Cross-Validation for GDPPC and Population

The RMSE value is 18.04308 so we can conclude that GDP Per Capita and Population are good predictors of the total medals won.

For our second model, we hypothesized that GDP Per Capita, the ratio of Women to Men participation would be good predictors for the total number of gold medals won.

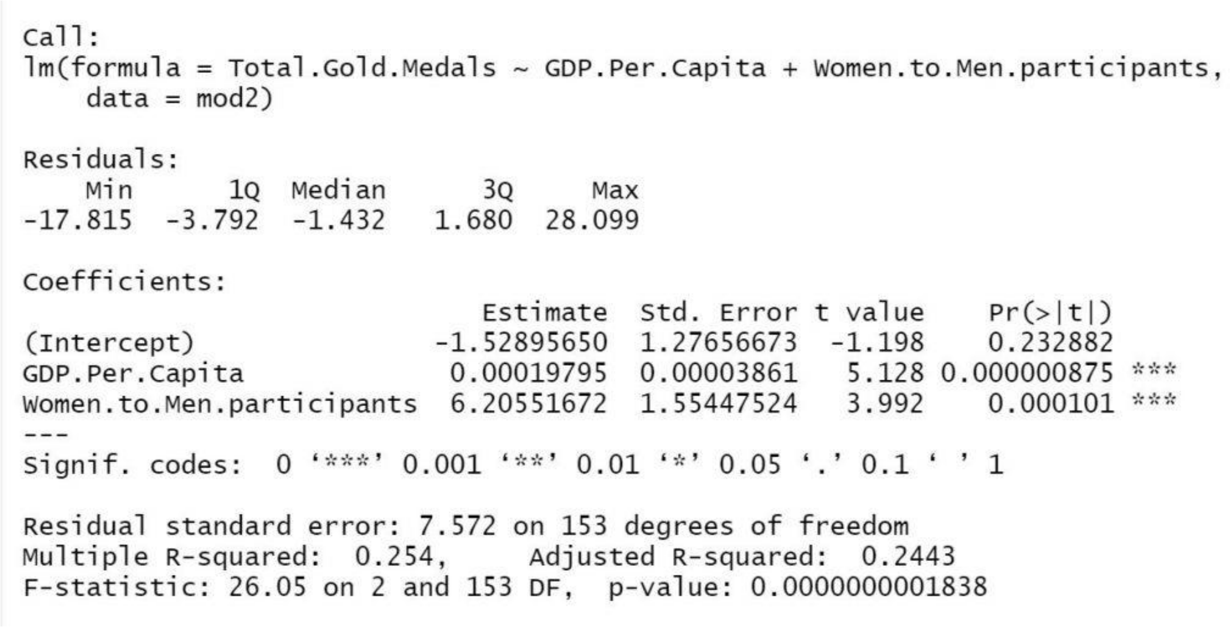
We plotted 2 different scatterplots to determine if there was a relationship between the predictors and response variables.





There does not seem to be an obvious relationship between the plots for GDP Per Capita and the total number of gold medals. However, there is no relationship between the ratio of Women to Men participants and the total number of gold medals. Even though the relationship is unclear, we still want to use all the predictors in the multiple linear regression model to see if they have any effect on the response variable.

The results of the multiple linear regression are shown below.

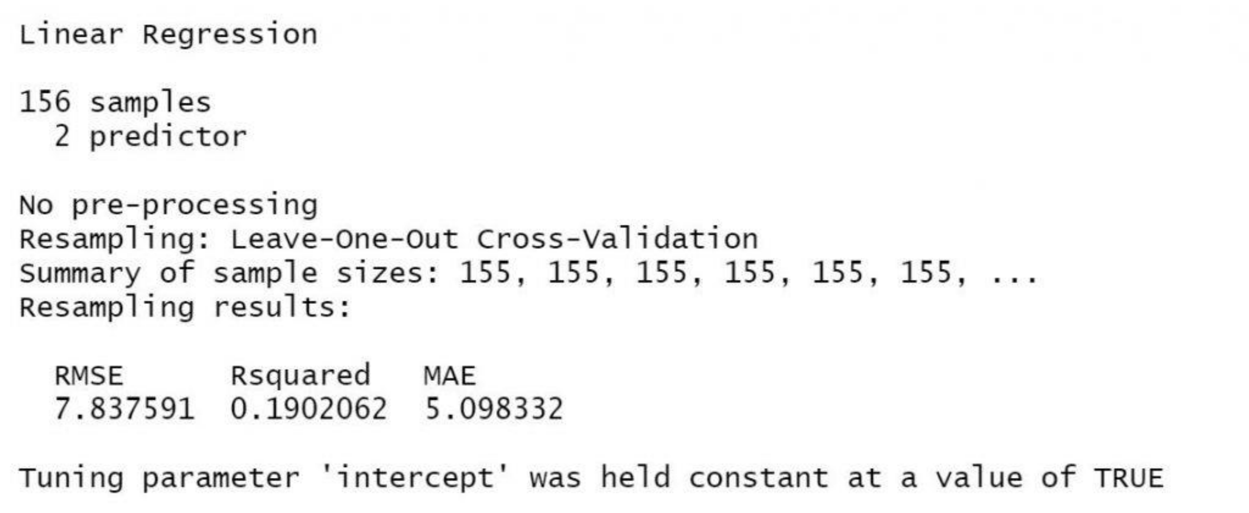


```
Call:
lm(formula = Total.Gold.Medals ~ GDP.Per.Capita + Women.to.Men.participants,
    data = mod2)

Residuals:
    Min      1Q  Median      3Q     Max 
-17.815  -3.792  -1.432   1.680  28.099 

Coefficients:
                            Estimate  Std. Error t value    Pr(>|t|)    
(Intercept)               -1.52895650  1.27656673  -1.198    0.232882    
GDP.Per.Capita             0.00019795  0.00003861   5.128 0.000000875 ***
Women.to.Men.participants  6.20551672  1.55447524   3.992    0.000101 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.572 on 153 degrees of freedom
Multiple R-squared:  0.254,	Adjusted R-squared:  0.2443 
F-statistic: 26.05 on 2 and 153 DF,  p-value: 0.0000000001838
```

Multiple linear regression using GDP Per Capita and Ratio of Women to Men in participation

Looking at the adjusted R-squared value, we can see that there is a relationship between the predictors (GDP Per Capita, and the ratio of women to men) and the response variable (total number of gold medals). However, the relationship between these variables is not very strong since the adjusted R-squared value is 0.2443, which is closer to 0 than 1. Next, looking at the intercepts, we can say that for every $10,000 increase in the GDP Per Capita of a country, the total number of gold medals won increases by 1.9 (approximately 2). Furthermore, for every 1 unit increase in the ratio of women to men in participation, the number of gold medals won increases by 6.2.

Next, we determined the accuracy of the model using leave-one-out cross-validation. The results of the cross-validation are shown below.



```
Linear Regression 

156 samples
  2 predictor

No pre-processing
Resampling: Leave-One-Out Cross-Validation 
Summary of sample sizes: 155, 155, 155, 155, 155, 155, ... 
Resampling results:

  RMSE      Rsquared   MAE     
  7.837591  0.1902062  5.098332

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Leave-One-Out Cross-Validation for GDPPC and Ratio of women to men in participation

Here, the RMSE value is 7.837 so we can conclude that GDP Per Capita and the ratio of women to men in participation are good predictors for the total number of gold medals won in the Olympic games.

Conclusion

We can conclude that GDP Per Capita and the Population of different countries are good predictors for the total number of medals won in the Olympic games. Furthermore, we also concluded that GDP Per Capita and the ratio of women to men in participation for different countries are good variables that can predict the total number of gold medals won. These results are important as athletes with multiple citizenships can choose which country they want to represent in the Olympics. Additionally, the International Olympic Committee can help countries to formulate and implement inclusive policies to organize the event in a fair manner. Next, we initially hypothesized that GDP Per Capita, Population, and HDI value has a significant impact on the total medals won. However, we found that the relationship between these variables was not very strong. This might be as we only selected a few countries from every continent. These relationships might have been stronger had we selected all the countries from different continents. In the future, we could select all the countries in the world and check if the relationship between the predictors (GDPPC, Population, HDI value) and the response variable (total medals won) are strong and if our conclusion remains unchanged.