Predicting nations Olympic medal performance

DATA7001 GROUP 9

LINYUAN XING (47163740), JINGXIN NIE (46002002),
MARK MENDOZA (43573154), TIMOTHY LE PERS (43907021)

THE UNIVERSITY OF QUEENSLAND | St Lucia, Queensland, Australia

We give consent for this report to be used as a teaching resource.

EXECUTIVE SUMMARY

This report details the efforts Group 9 made to investigate and model nations Olympic medal performances. The main stakeholders considered throughout the investigation are governments, as they have the greatest power to influence spending on a country's sports programs. Several other stakeholders exist such as international and national sporting regulatory bodies and the athletes themselves.

In Section 1, these stakeholders and the two main research questions are outlined. Section 2 describes the datasets that were used and where they were obtained from. Section 3 details issues that were faced with the data and how they were overcome. Section 4 outlines the exploratory data analysis (EDA) that was performed as well as the model that was made. Section 5 details changes to the project scope that were made following feedback on the project pitch. Section 6 contains the conclusions from EDA and model building, in addition to recommendations that could be made to stakeholders. The Appendix includes excerpts of R language scripts that were used for Sections 3 and 4.

The project was an overall success with many insights gained regarding medal performance from the EDA, and the relationship of the number of medals a country on its gross domestic product (GDP) and population. It was recommended to stakeholders that they place more emphasis on female participation at the Games due to the almost equal likelihood that they can achieve medals as males. Additionally, countries should increase participation in high event sports such as athletics, swimming, wrestling, and gymnastics. One limitation of the project was that there were many more factors that could not be accounted for in the model to increase its predictive power – this was a stretch goal that will have to be further studied.

TABLE OF CONTENTS

Executive Summary i

1 Design Thinking 1

1.1 Introduction 1

1.2 Research Questions 1

1.3 Key Stakeholders 1

2 Getting the Data 2

2.1 Data Sources 2

3 Making the Data Fit for Use 2

3.1 Data Quality Issues 2

3.1.1 Non-Unique Data and Accuracy 2

3.1.2 Incomplete Data 2

3.1.3 Incorrect Data 3

3.1.4 Inconsistent Data 3

3.2 Data Use Considerations 3

3.2.1 Athletes with Multiple Events 3

3.2.2 Medal Count for Team Sports 3

3.2.3 Comparisons via NOC Code 3

3.3 Final Datasets 4

4 Making the Data Confess 5

4.1 Visual Exploratory Data Analysis 5

4.1.1 Relationship between total medals awarded and year 5

4.1.2 Medal split by country 6

4.1.3 Medal split by country as a percentage 6

4.1.4 Gender participation 7

4.1.5 Medals awarded per sport 8

4.1.6 Medal performance by sport 9

4.2 Insights from Visual EDA 9

4.3 Statistical Exploratory Data Analysis 10

4.4 Model Generation 11

4.4.1 Data processing and linear correlation analysis 11

4.4.2 Linear model and variable selection 12

4.4.3 Initial Model Validation 12

4.4.4 Model Improvements 13

4.4.5 Model Summary 13

4.5 Model Validation 14

5 Management of Change 15

6 Conclusion and Recommendations 15

References 16

Appendix 16

Appendix A: Code and Datasets 16

Appendix B: Data Cleaning 17

Appendix C: Modelling 21

Appendix D: Visualisation 22

1 Design Thinking

1.1 Introduction

Participating in the Olympic Games is the pinnacle of numerous athletes' careers. Watching athletes compete at the top of their game is a great source of entertainment and national pride for many that watch the Games globally. As athletes battle it out for the renown of winning a much-coveted Olympic medal, it begs the question, why do some countries seem to always secure more medals than others? The answer to this question is of course not a simple one.

This report details our project team's investigation into this topic and the results we obtained. It includes information on how we defined our research direction, where relevant data was obtained from, how we ensured the quality of our data, the analysis that was performed and finally recommendations to key stakeholders who have interests in improving their nations medal success.

1.2 Research Questions

This investigation was carried out with the aim of following the data science process (see Figure 1) to answer the two following research questions:

- What factors are most influential in understanding how many medals a country is likely to win?
- How accurately can the number of medals won by each country be predicted, using Summer Olympic data combined with measures of a country's overall wealth (gross domestic product or GDP) and population?

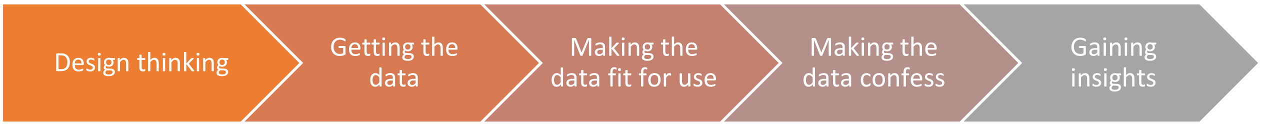


Figure 1: The data science process for this project

1.3 Key Stakeholders

The following table explores key stakeholders of the project, and a description of their potential interests in this work.

Table 1: Key stakeholders of the data science project

Stakeholder	Interest
National Governments	To inform decisions regarding best allocation of spending for sports programs.
National Olympic Committees (NOCs)	Understanding which sports a country's medals are likely to come from, and what areas need to be focused on.
Sporting regulatory bodies	The potential to flag significant irregularities in medal tallies for investigation.
Participating athletes	To add breadth to their understanding of their likelihood of winning a medal.
Citizens of participating countries	Understanding how many medals their country is likely to win at a given Olympics.

2 GETTING THE DATA

2.1 DATA SOURCES

Due to lack of a single repository for all factors which were hypothesised to impact Olympic performance, several discrete sets had to be obtained. See Appendix A: Code and Datasets for a table outlining the file names for the below datasets.

1. Data for all participating Olympic athletes from 1896 to 2016
   This source contains 15 attributes regarding participating athletes from modern Olympic games, Summer and Winter. Attributes include country represented, event competed, sex, age and medal won. There are approximately 271,000 records.

2. Data of country statistics published by World Bank from 1960 to 2018
   This dataset contains 11 economic and social indicators for 264 countries, published by the World Bank. Records are organised by country and the year it was published.

3. Data for National Olympic Committee (NOC) codes from 1896 to 2020
   This supplementary source lists all current and historical NOC codes with country name.

3 MAKING THE DATA FIT FOR USE

3.1 DATA QUALITY ISSUES

The three datasets were imported to R and were inspected using the `head` and `tail` functions. It became evident that the data was unclean. To ensure that the data was fit for use, various steps were taken to rectify key issues and ensure data was of good enough quality to allow further exploration. See Appendix B: Data Cleaning for an excerpt of code used to clean the data.

3.1.1 Non-Unique Data and Accuracy

For further analysis, it is a requirement that each record in [1] is unique. When doing a year-by-year filter of the data and attempting to do a `sum` of the medals, countries such as Germany had an irregular number of medals – higher than the actual results published for that year by the IOC. Duplicate records were removed using `unique.data.frame`. This step was also important in ensuring the accuracy of the data – a count of medals in [1] must match with medal tables published by the IOC.

Dataset [3] was utilised as a master list of NOCs by which [2] could be linked to [1]. Therefore, its accuracy had to be verified manually.

3.1.2 Incomplete Data

All attributes with missing data were inspected using the `is.na` function. For [1], age height and weight data were missing for approximately 20% of records. However, these were not key attributes for further exploration. Medal data, however, was missing for almost 85% of records. It was found through `factor` that the medal attribute contained Gold, Silver, Bronze and NA. Therefore, the string "DNW" or "did not win" was assigned to these NA fields.

A limitation of [2] was that data only encompassed the year 1960 and beyond, and therefore would not be available for Olympic games prior to this. A decision was made to consider only Olympic games from 1960 and beyond for data in [1] for model building. Furthermore, not all countries had published data for every attribute in [2], but these would be dealt with on a case-by-case basis – depending on the year of the Olympic games being explored. Some NOCs were also missing in [3] for countries such as Singapore and Hong Kong, which were manually appended.

3.1.3 Incorrect Data

Typos were found in [3], such as “Bolivia” being misspelt as “Boliva” and were corrected manually.

3.1.4 Inconsistent Data

Lastly, to ease model building, the desired attributes from [1] and [2] were initially joined into a single data frame with NOC as primary key. After initial matching showed not all statistics were joined, an attempt was done using country name. Remaining countries whose names did not match were altered manually.

3.2 Data Use Considerations

Aside from data cleaning, there were various organisational aspects of the dataset that had to be accounted for to prevent statistical misinterpretation of the data.

3.2.1 Athletes with Multiple Events

Some athletes would have participated in more than one event, especially for disciplines such as aquatics and athletics. Therefore, a distinction had to be made between an entire country’s Olympic *contingent*, which would only count each individual athlete regardless of participation in multiple events, versus the Olympic *participation*, which would consider a single athlete entering multiple events.

3.2.2 Medal Count for Team Sports

When attempting to organise the number of medals won by event for a single Olympics, it became apparent that the count for some countries were inflated due to medals for team events (e.g. basketball) counting for each team member, rather than one medal per event. Employing the `unique` function, two new parameters were aggregated: total medals *with teams* and total medals *actual*.

3.2.3 Comparisons via NOC Code

When attempting to compare historical country performance, the country name rather than NOC code should be employed due to countries changing NOC codes over time. For example, Russia is assigned the codes EUN, URS and RUS.

3.3 FINAL DATASETS

Final datasets that proceeded on to exploratory data analysis were as follows:

1. Data for all participating Olympic athletes from 1896 to 2016
   This source contains 15 attributes regarding participating athletes from modern Olympic games, Summer and Winter. Attributes include country represented, event competed, sex, age and medal won.

2. Combined data from 1960 to 2016
   This source contains 6 relevant economic and social indicators for each Olympic country, the year published by the World Bank, in addition to: total medals with teams, total medals actual, Olympic contingent and Olympic participation.

Table 2 shows an excerpt of dataset [2]. Note that "NA" values still exist in the final dataset under various attributes due to a decision made to exclude these only when analysis was done.

Table 2: Excerpt of cleaned Olympic medal data combined with key world statistics

Country	Year	Season	Total Medals Teams	Total Medals Actual	Event Participations	Contingent	Male Contingent	Female Contingent	GDP per capita (US$)	Total Population	Rural Population (%)	Land Area (km^2)	Electricity Access (%)	Precipitation (mm)
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
Afghanistan	2016	Summer	0	0	3	3	2	1	547.2	35383128	75.0	652860	97.7	NA
Albania	2016	Summer	0	0	6	6	3	3	4124.1	2876101	41.6	27400	100.0	NA
Algeria	2016	Summer	2	2	74	64	54	10	3946.4	40551404	28.5	2381740	100.0	NA
Andorra	2016	Summer	0	0	4	4	2	2	37474.7	77297	11.8	470	100.0	NA
Angola	2016	Summer	0	0	26	26	8	18	3506.1	28842484	35.9	1246700	40.7	NA
Antigua	2016	Summer	0	0	9	8	6	2	15197.6	94527	75.2	440	100.0	NA
Argentina	2016	Summer	22	5	232	215	140	75	12790.2	43590368	8.4	2736690	100.0	NA
Armenia	2016	Summer	4	4	34	31	24	7	3591.8	2936146	36.9	28470	100.0	NA
Aruba	2016	Summer	0	0	7	7	3	4	28281.4	104872	56.8	180	100.0	NA
American Samoa	2016	Summer	0	0	4	4	3	1	11697.0	55741	12.8	200	NA	NA
Australia	2016	Summer	82	30	518	420	208	212	49971.1	24190907	14.2	7692020	100.0	NA
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

4 Making the Data Confess

4.1 Visual Exploratory Data Analysis

Exploratory data analysis (EDA) was conducted on Summer Olympic medal data only from 1896 to 2016, unless otherwise stated. The actual medals awarded were considered, where team events count as 1 medal.

4.1.1 Relationship between total medals awarded and year

Firstly, the total number of medals awarded in each year was graphed to help give an overall understanding of how the number of medals being awarded has changed with time. Figure 2 shows this relationship. The y-axis shows the number of medals, and the x-axis shows the year (1896 – 2016).

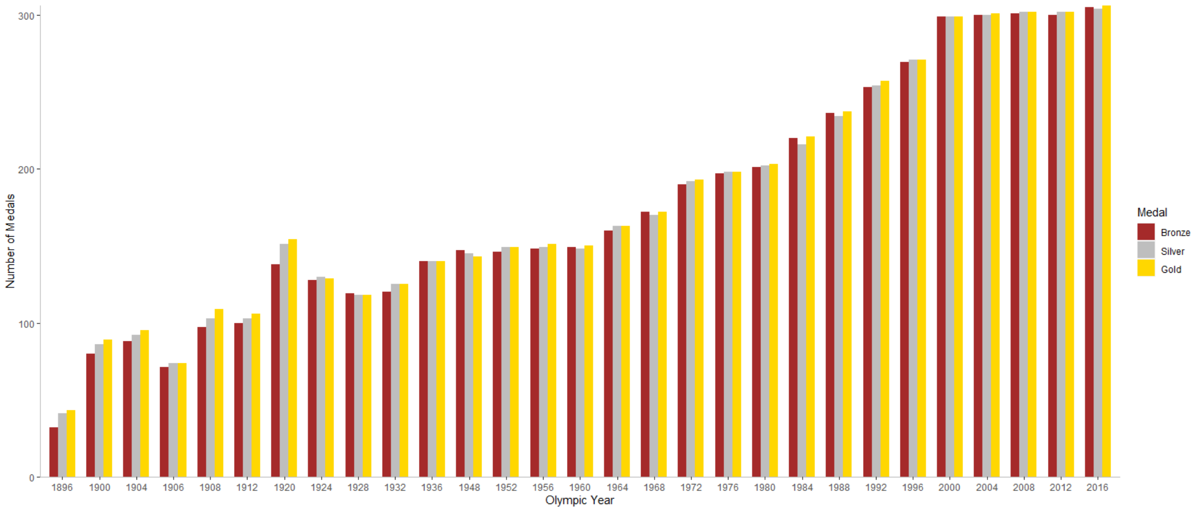


Figure 2: Number of medals awarded at each Summer Olympic Games (1896 to 2012)

A positive trend between total medals awarded and time is illustrated. In 1896, the total number of medals awarded was less than 100. In 1896, very few countries participated in the Olympics and only a few sports were included, leading to a small number of medals awarded. In 1900, the number of medals awarded increases significantly due to more countries participating stemming from observation of the previous Games' success. In 1908, there was another sharp increase in the number of medals awarded, and since that time, the size of the Olympic Games has been gradually and steadily developing. The exception to this trend is the decrease observed around 1932 mostly attributed the onset of the Great Depression. Notably, the Olympic Games were cancelled in 1916, 1940 and 1944 due to World War I and II but this did not seem to affect the overall trend significantly.

4.1.2 Medal split by country

Following the analysis of the total number of medals awarded for each year, we investigated how many medals were won by each country in all Summer Olympics combined. Figure 3 below only shows countries whose medal count is greater than 20 for visibility.

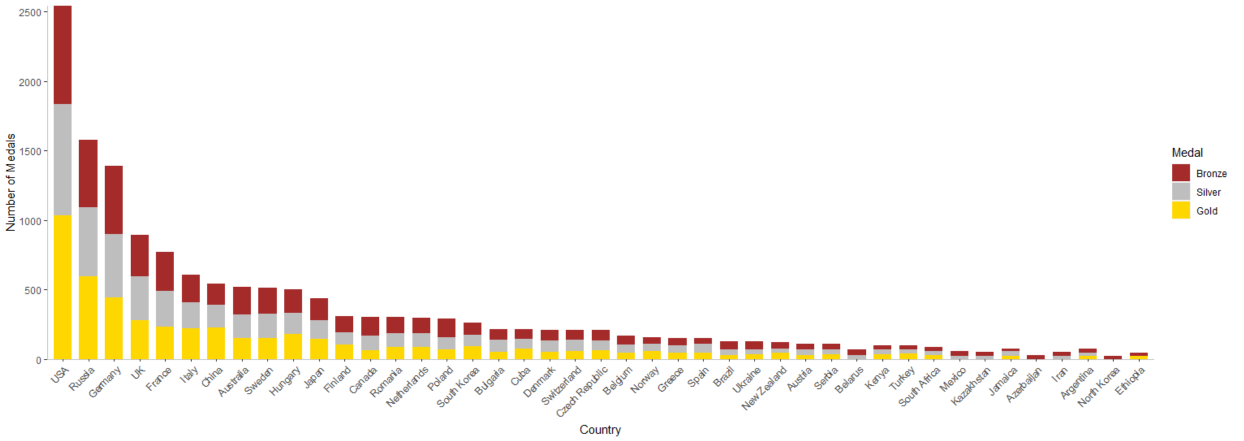


Figure 3: Total number of medals won by country (1896 to 2016)

4.1.3 Medal split by country as a percentage

Another interesting metric we decided to investigate was the split of gold, silver, and bronze medals of winning countries as a percentage of the total number of medals won shown in Figure 5. The chart aims to illustrate the "quality" of a country's medal success, rather than quantity.

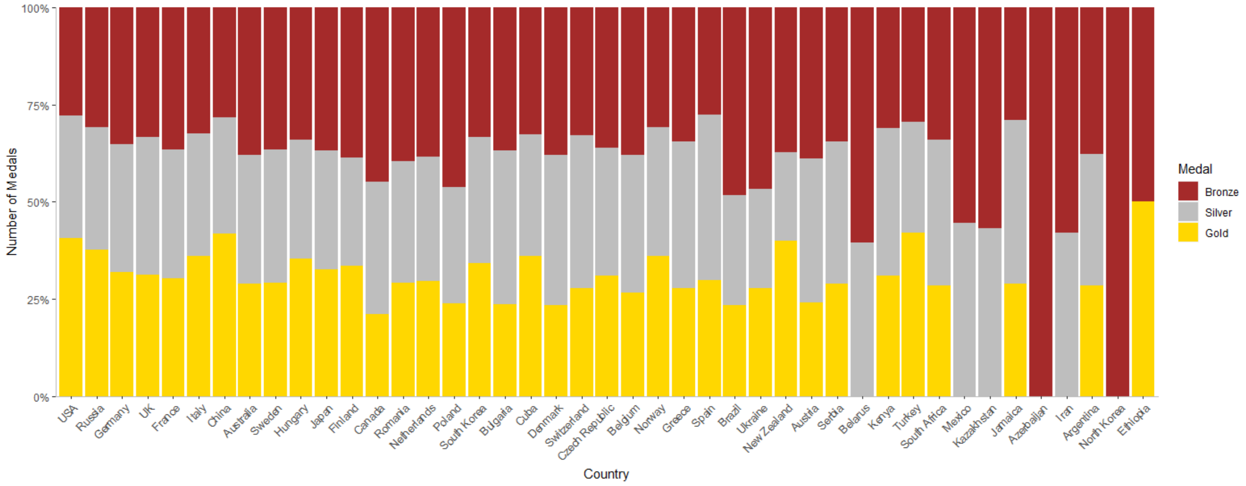


Figure 4: Type of medal won as a percentage of total number of medals won for top 20 countries (1896 to 2016)

We can from this graph see for example that almost 50% of New Zealand and Turkey's medals are gold medals, whereas the previously top performing USA is approximately 40%.

4.1.4 Gender participation

It was also possible to explore the how the gender split of total contingent in the Olympics has changed with time, shown in Figure 5 below.

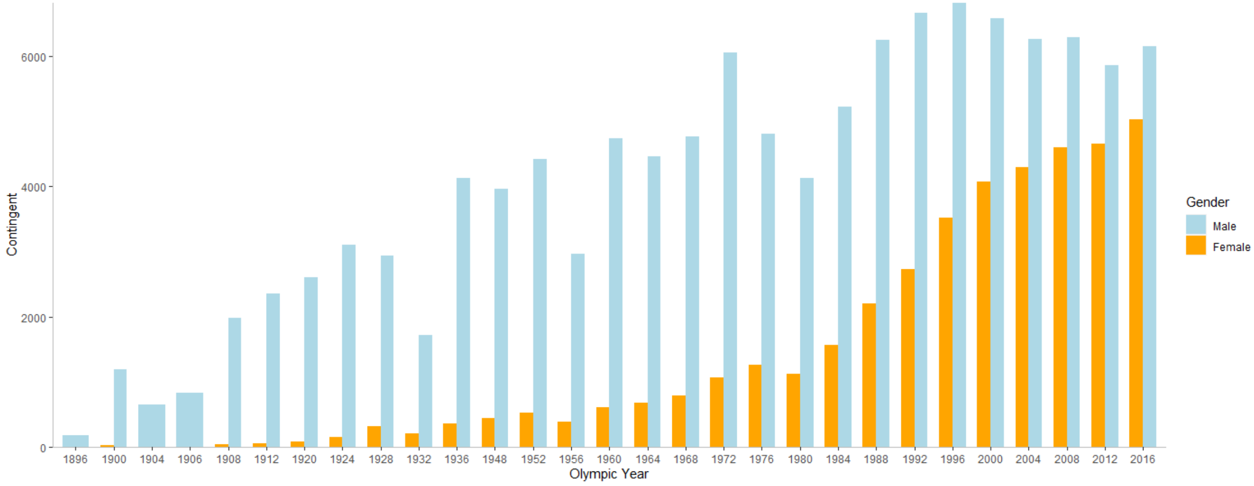


Figure 5: Number of Olympic athletes each year by gender

The light blue shows the male contingent and the orange show the female contingent. We can clearly see in the first ten Olympics, there is an extremely small number of female athletes participating in the Olympics. With time, more female athletes have been able to participate in the Olympics. In 2016, the number of male and female athletes is approaching parity.

Figure 6 below shows the proportions of medals awarded to each gender as a proportion of total medals awarded each Olympic year. The Figure is also reflective of the number of female-only and male-only events that exist (as one medal is awarded per event, regardless of whether it is team or individual sport). Although the number of medals awarded to females is reaching parity, it is still only approximately 45% of total medals awarded.

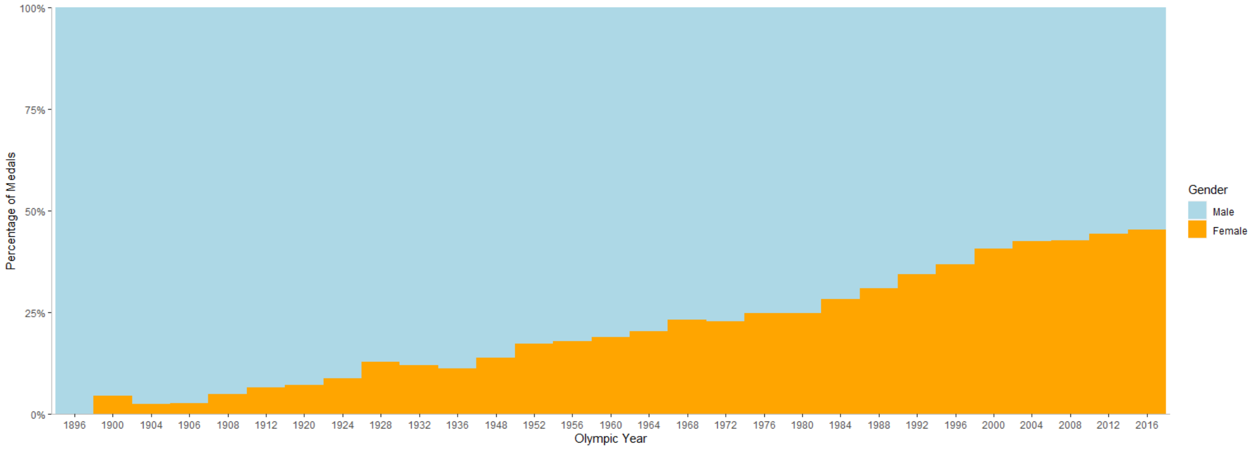


Figure 6: Percentage of medals awarded each year by gender

4.1.5 Medals awarded per sport

Figure 7 below shows the number of medals awarded for each sport over all Olympics. Note that team event awards (e.g., rugby, baseball, etc.) count as 1 medal, and not all sports hosted events at every Olympics.

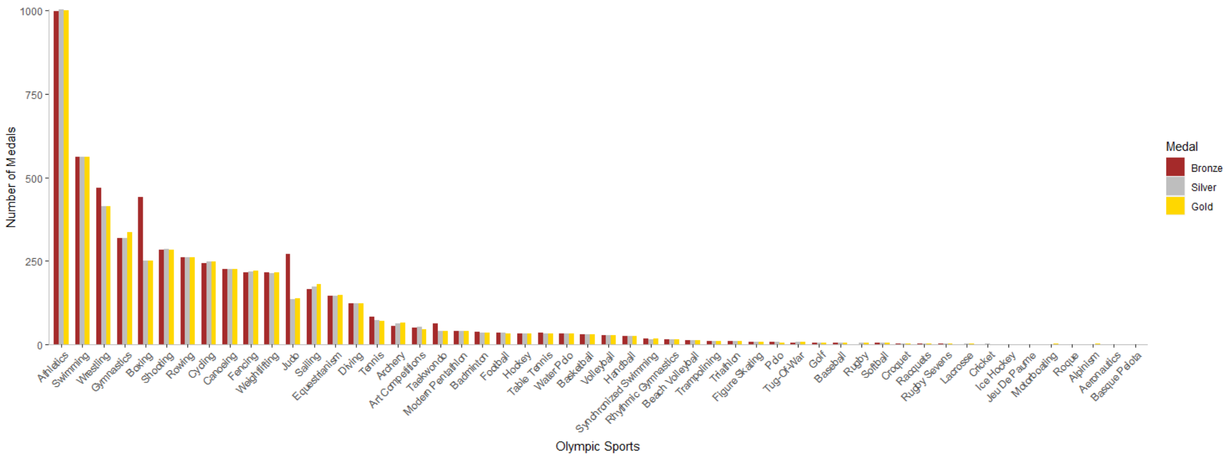


Figure 7: Number of medals awarded grouped by Olympic sport (1896 to 2016)

We can see that athletics is the sport with most awarded medals historically, with approximately 1000 each of gold, silver and bronze medals awarded. This is followed by swimming, wrestling and gymnastics. The large discrepancy can be attributed to the larger number of events in athletics compared to other sports (including discus, javelin, high jump, long jump, steeplechase, sprints, etc.) whereas swimming encompasses variations of the four strokes (freestyle, breaststroke, butterfly, backstroke).

One interesting feature of this chart is that for boxing and judo, the number of bronze medals awarded is almost 50% greater that of gold and silver medals. In 1970, the Finland Boxing Association recommended to the IOC that two bronze medals be awarded to the two losing semi-finalists instead of an Olympic diploma (Ansari, 2021).

4.1.6 Medal performance by sport

We also explored the number of medals that each country has won in different sports over all previous Summer Olympic Games to observe each country’s most successful sports, and its proportion of all-time medals won. Figure 8 only includes sports where ≥ 20 medals were obtained for aided visibility. Countries such as Kenya and Jamaica have only been successful in athletics. Other countries such as Australia, Germany, and USA, show a greater balance, with a significant proportion coming from swimming. The “NA” sport colour refers to medals won from sports that are unlisted due to having smaller medal counts.

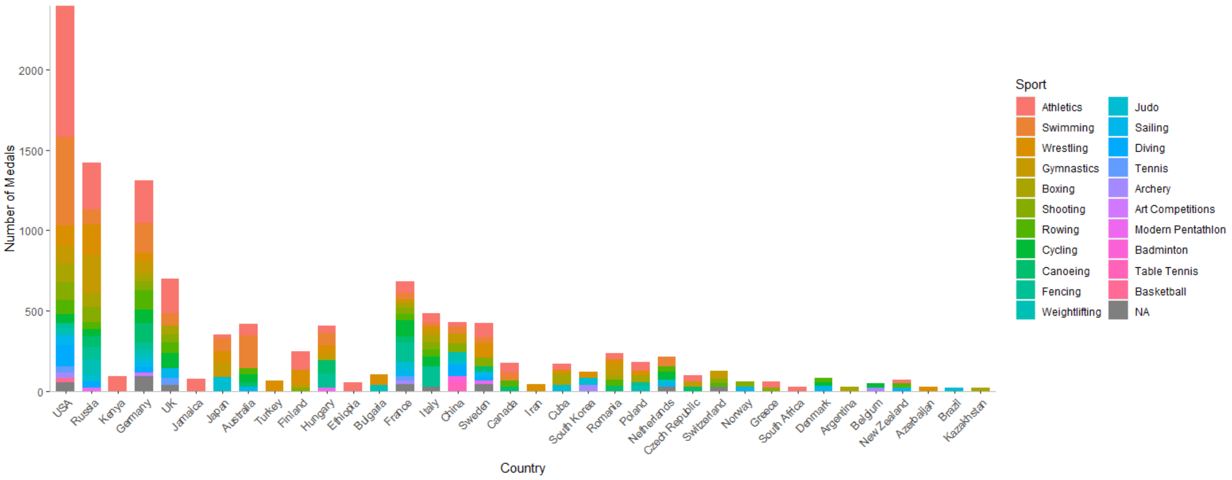


Figure 8: Total number of medals awarded grouped by country (1896 to 2016)

Figure 8 shows that successful countries gain a large percentage of their medals from disciplines with multiple events. For example, athletics, swimming, and wrestling. This indicates that if countries are interested in improving their overall medal tally, it is likely to be worthwhile investing in the development of athletes in these disciplines. Lower GDP countries such as Kenya, Jamaica and Ethiopia are also well represented in athletics. It can be argued that since athletics requires less investment to train athletes (e.g., track and field) it requires less of a barrier for such countries to participate.

4.2 Insights from Visual EDA

Based on this exploratory analysis, it is possible to make recommendations to stakeholders regarding optimising their medal tallies.

One solution for a country to improve its overall medal tally would be to focus on the sports with the largest number of medals awarded such as athletics, swimming, wrestling, and gymnastics. Focussing on developing and increasing participating athletes in these disciplines is likely to produce more medals as athletes can compete in more events.

Another solution would be to increase focus on the development of female athletes. Recent Olympic games have shown female participation to almost reach parity with male participation. Generally, female athletes have been under-represented at the Olympics, indicating that there is more room for improvement in terms of engaging female athletes than males. Therefore, an increased focus on the development and recruitment of female athletes is more likely to increase a country’s contingent than focussing on male athletes, which has plateaued and dipped throughout the Games’ history.

4.3 Statistical Exploratory Data Analysis

Combined medal data for Summer Olympic games (1960 – 2016) was matched with relevant statistics from the World Bank, leading to the following correlation plot in Figure 9. It is evident that total medals (teams or actual), is positively correlated by Olympic participation and contingent. Therefore, these would be the most accurate predictors for medal performance. To a lesser extent are GDP per capita and total population, which are inaccurate predictors for the largest NOCs China, USA, Russia and India, but may provide useful predictive power for remaining nations. Surprisingly, land area was correlated with Olympic medal performance.

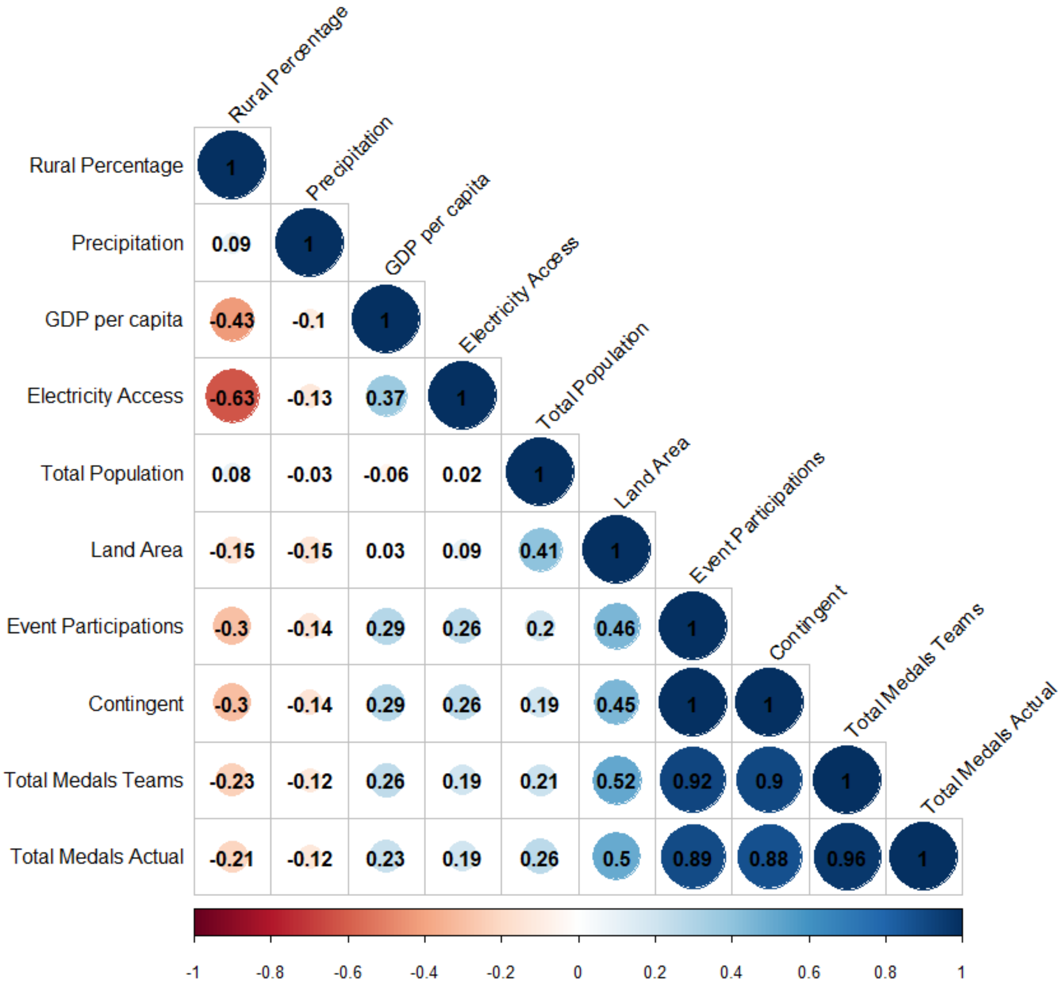


Figure 9: Correlation matrix showing multiple data attributes

4.4 Model Generation

The code used for the modelling outlined in this section can be found in Appendix C.

4.4.1 Data processing and linear correlation analysis

To simplify the model, the following decisions were made:

- The model will only consider the GDP and total population, to explore non-athlete related attributes.
- The model will be generated for data pertaining to the London 2012 Summer Olympics.
- Medals (gold, silver, and bronze) were aggregated and so that the model considers total medals won without distinguishing placing.
- Each member of a winning team will be awarded a medal (team events counted for more than 1 medal).
- The model will omit "NA" values for countries that do not have statistics, as well as countries that do not have a medal count.

Since the units of each variable are different, the data was first standardized. This was done using the scale function in R: each number in a group of numbers is subtracted from the average of the group of numbers and then divided by the root mean square of the group of numbers to eliminate the difference caused by different unit variables.

We then checked the degree of linear correlation between the variables and obtained the linear correlation coefficient. The output from R is shown in Figure 10.

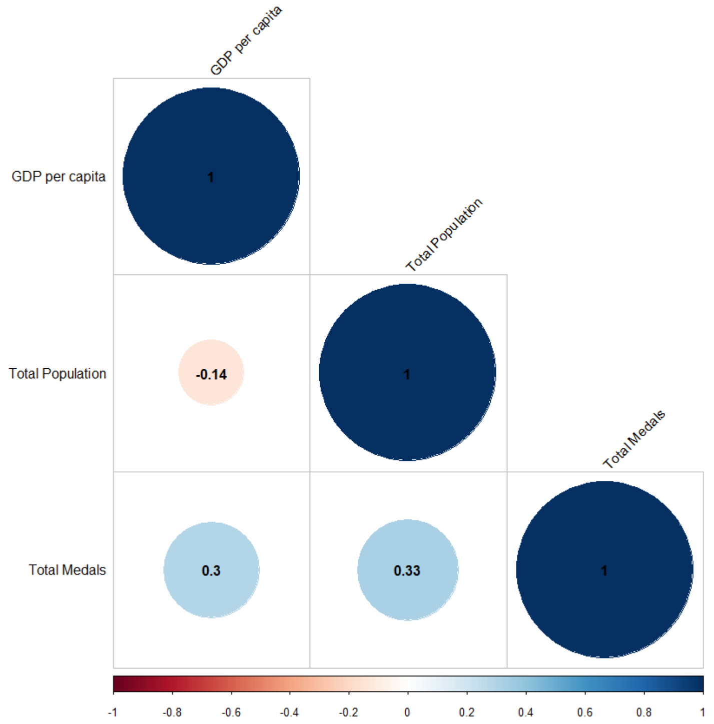


Figure 10: Correlation matrix showing initial regression coefficients

As seen prior, there is a slight positive linear correlation between GDP, population, and medals. This shows that we can expect a country with higher GDP and population to have a higher medal count.

4.4.2 Linear model and variable selection

We did not try to create a single linear regression model of medals against each variable, but rather attempted a multivariate linear model first, and then checked the results for selection. This aided in deciding whether to discard certain variables. We then carried out a multiple linear regression analysis of medals won on GDP and population. The output from this analysis is shown in Table 3.

Table 3: Summary statistics from multiple linear regression

p-value	R^2	Intercept	GDP	Population
4.58×10^{-5}	0.2015	Not significant	Significant	Significant

Table 3 shows that firstly, the p-value is less than 0.05, indicating that the model is statistically significant. Secondly, looking at the coefficients, we can see that the p-value of the intercept is not significant, but GDP and population are both significant. Inaccuracies are expected as the model tends towards 0, because there does not exist a country where GDP and population are 0. Finally, we can see that the R^2 value of this model is approximately 0.20, which shows that the residual error between the model and data remains very high.

From the above results, both the GDP and population variables are correlated with total medals won. To better demonstrate this point, we used stepwise regression to select variables. The results also show that when GDP and population exist at the same time, the Akaike Information Criterion (AIC) value of the model is the smallest, that is, the current model is the most optimised. The AIC is an estimator of prediction error and thereby relative quality of statistical models for a given set of data.

4.4.3 Initial Model Validation

Figure 11 below shows graphs that were used for model validation. The plot in the upper left does not show any regularity, which means that there is a linear relationship. The scatter points in the upper right graph are roughly concentrated on the straight line in the QQ graph, indicating that the residuals are normal. The plot in the lower left graph may have rules, and the variance increases with the mean. In the bottom right plot, the larger the upper and lower spacing, and the more obvious the difference in variance. Check the abnormal points in the lower right picture, there may be abnormal points.

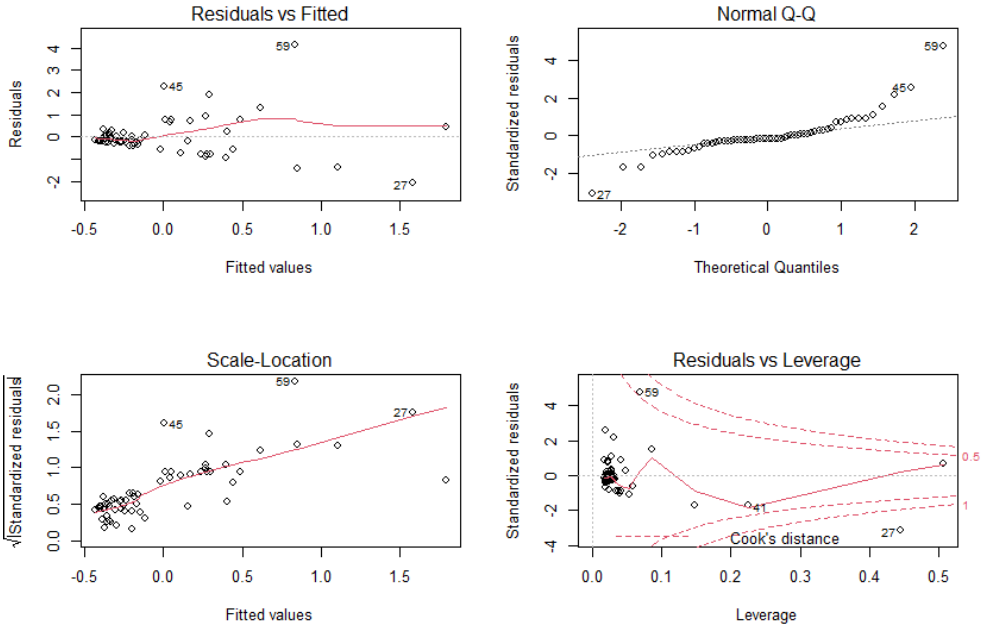


Figure 11: Graphs used for model checking

To make a more accurate diagnosis of the model fit, we performed various tests with R such as multicollinearity test, heteroscedasticity test and independence test. In these tests, we found that the model did not satisfy the assumption of homoscedasticity. To eliminate heteroscedasticity, we needed to add additional transformations.

4.4.4 Model Improvements

To optimise the model, we performed a scene transformation on the variables to eliminate heteroscedasticity and used the `CAR` package to determine the power of GDP and Population. The MLE of lambda is the best power of each variable and is shown in Table 4.

Table 4: Variable transformation

Initial variables	Variables after power transformation
GDP	GDP$^{0.10468}$
Population	Population$^{0.16127}$

Summary statistics of the new model obtained is shown in Table 5.

Table 5: Summary statistics from the new model

p-value	R^2	Intercept	GDP$^{0.16127}$	Population$^{0.10468}$
4.54×10^{-11}	0.4349	Significant	Significant	Significant

We can see that the intercept term has become significant, and the R^2 value has also been improved from 0.20 to 0.43.

4.4.5 Model Summary

Figure 12 shows a three-dimensional plot of the two regression models created. It is evident that the nonlinear model is more representative of the data.

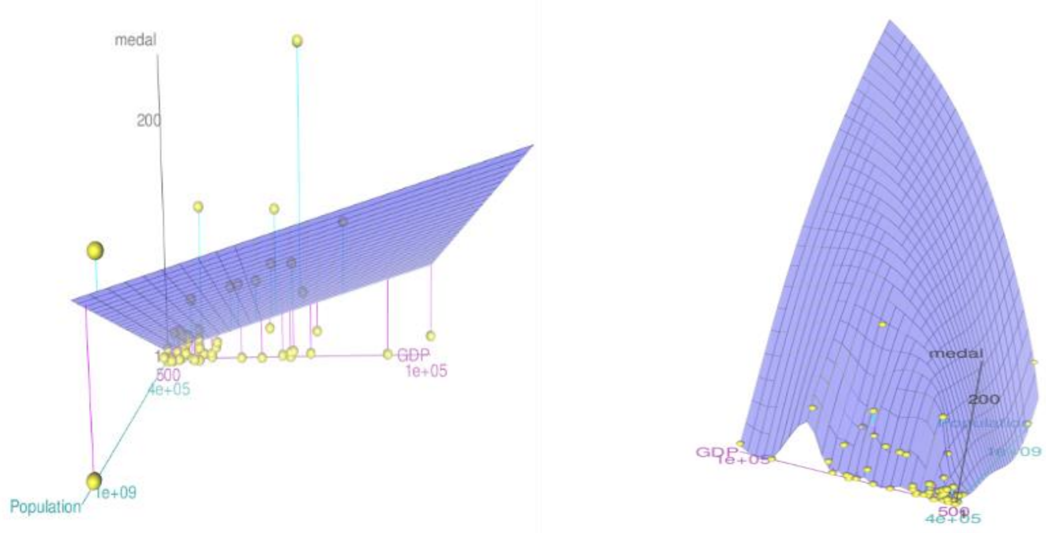


Figure 12: Three-dimensional plot of the linear and non-linear regression models

GDP represents the economic strength of a country. It has a non-linear positive correlation with the number of medals won. This is because every country needs to invest a lot of money to produce high level athletes. Good economic strength can enable the athletes of the country to obtain advanced training, and also facilitate the enthusiasm of the athletes in a variety of ways. The population size is also positively correlated with medals. The larger the population, the greater the chance that there are potential high-level athletes in the country. It follows that the country's chances of winning a medal would also be higher. Both variables however exhibit non-linear relationships with the number of medals won. This is likely because these variables do not directly influence an athlete's performance. GDP is a measure of a country's wealth, not the spending that went into the various relevant sports programs. The population attribute indicates the probability of the existence of potential high-level athletes at an Olympics.

To summarise, GDP and population both affect the number of medals won. However, it can be seen from the coefficients and indices of the model that when 1 point GDP or population changes, it has little effect on the number of medals. This is representative of the complexity of the task at hand, in that GDP and population are not sufficient for creating a truly representative model of nations Olympic medal performances. It can be said however, that they are strong candidates for variables that should be included in any such model.

4.5 Model Validation

The fit of this non-linear model was checked against data from the 2000 Summer Olympics to see if the model could be extended to other years. The results are summarized in Table 6.

Table 6: Summary statistics from model validation with 2000 Summer Olympics data

p-value	R^2	Intercept	$GDP^{0.16127}$	$Population^{0.10468}$
5.35×10^{-16}	0.3116	Significant	Significant	Significant

This result shows that the nonlinear regression model is still significant, and the constant term and variables are also significant. However, that the model has a lower predictive power for total medals.

This offers further confirmation that the relationship between GDP, population and medals won is non-linear. It also shows that the relationship is likely to be non-linear every year. The changed R^2 value however indicates that the degree of non-linearity may differ each year. However, this may also be the result of the fact that GDP and population alone are insufficient for generating a full description of a country's Olympic medal performance. There is little doubt though, that the nonlinear relationship between GDP and population on the number of medals has been affirmed by our model.

5 MANAGEMENT OF CHANGE

Throughout this project, the Group sought the feedback of peers and teaching staff for continuous improvement towards the final project submission. From the initial pitch, our feedback from staff was to focus more on how our project could be driven by stakeholders' need for insights. After closer considerations of stakeholders' needs and their ability to action change for the Olympics, we decided to focus on investigating GDP and population as well as a smaller subset of factors. Initially, we had planned to perform clustering on the data as well, however this was relegated to future exploration due to time constraints, as well as not seeming to generate significant insight related to our aims.

In the final stages of the project, the Group submitted a trial presentation for peer review. The feedback obtained from peers and staff centred around creating better "story telling", via restructuring of our presentation, interspersing visualisations all throughout, and reformatting the "making the data confess" section as to not display raw code outputs. An additional visualisation was also added, in the form of a 3D plot, aimed at improving communication of our final model. In terms of this report, we incorporated advice not to include screenshots of code within the report body but relegate these to the Appendix to improve communication.

6 CONCLUSION AND RECOMMENDATIONS

This report aimed to answer the research questions:

- What factors are most influential in understanding how many medals a country is likely to win?
- How accurately can the number of medals won by each country be predicted using Summer Olympic data combined with measures of a country's overall wealth and population?

Two datasets and one supplementary dataset were cleaned to ensure EDA could be performed accurately. Then, a predictive model of medal count in 2012 using two factors, country GDP and total population, was made. EDA revealed some possible ways for a country to improve their medal tally. Governments and NOCs should focus on increasing female contingent, as well as increasing event participations in sports that award more medals (e.g., athletics, swimming, wrestling and gymnastics). Sporting regulatory bodies can advertise the various classes of competitive events as opportunities for inclusion at future Olympic games, as well as inspiring younger citizens to take up their respective sports while at primary school.

The model that was developed found a non-linear relationship between medals won and both GDP and population with high levels of significance but low R squared values. This indicates that while the number of medals won is influenced by a country's GDP and population, it does not tell the whole story. While the model offers sound insight into two factors that predict how successful a country is at a single Olympics, future studies could account for other attributes such as contingent and event participations, to increase its predictive power.

REFERENCES

Ansari, A. (2021, August 1). *Explained: Two bronze medals are awarded in the Olympics boxing competition.* Retrieved from Olympics: https://olympics.com/en/featured-news/why-two-bronze-medals-boxing

History.com. (2021, July 21). *The Olympic Games*. Retrieved from History.com: https://www.history.com/topics/sports/olympic-games

IOC. (2018, May). *120 years of Olympic history: athletes and results.* Retrieved from Kaggle: https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results

World Bank. (2020, October). *World Statistics dataset from World Bank.* Retrieved from Kaggle: https://www.kaggle.com/mutindafestus/world-statistics-dataset-from-world-bank

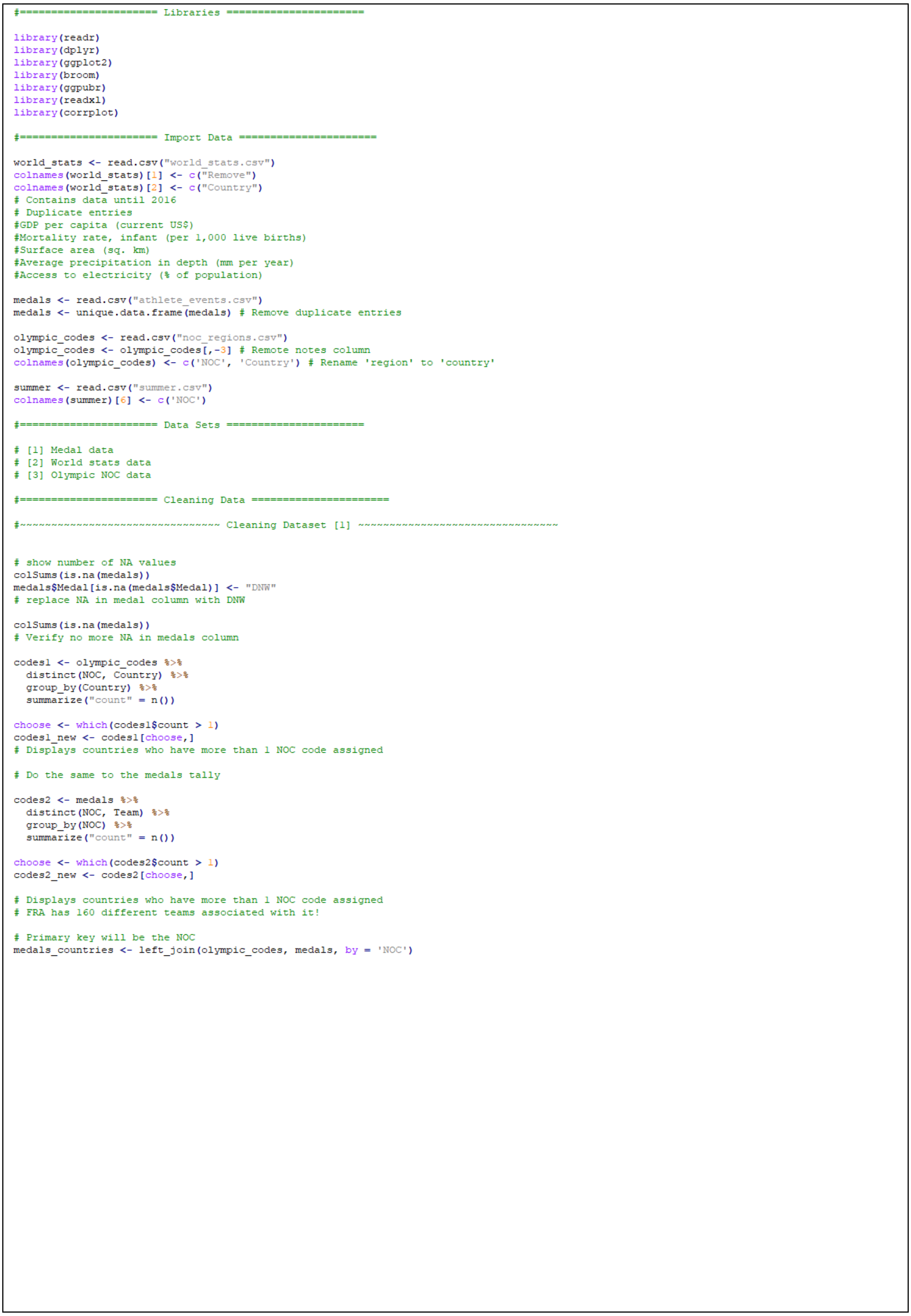
APPENDIX

APPENDIX A: CODE AND DATASETS

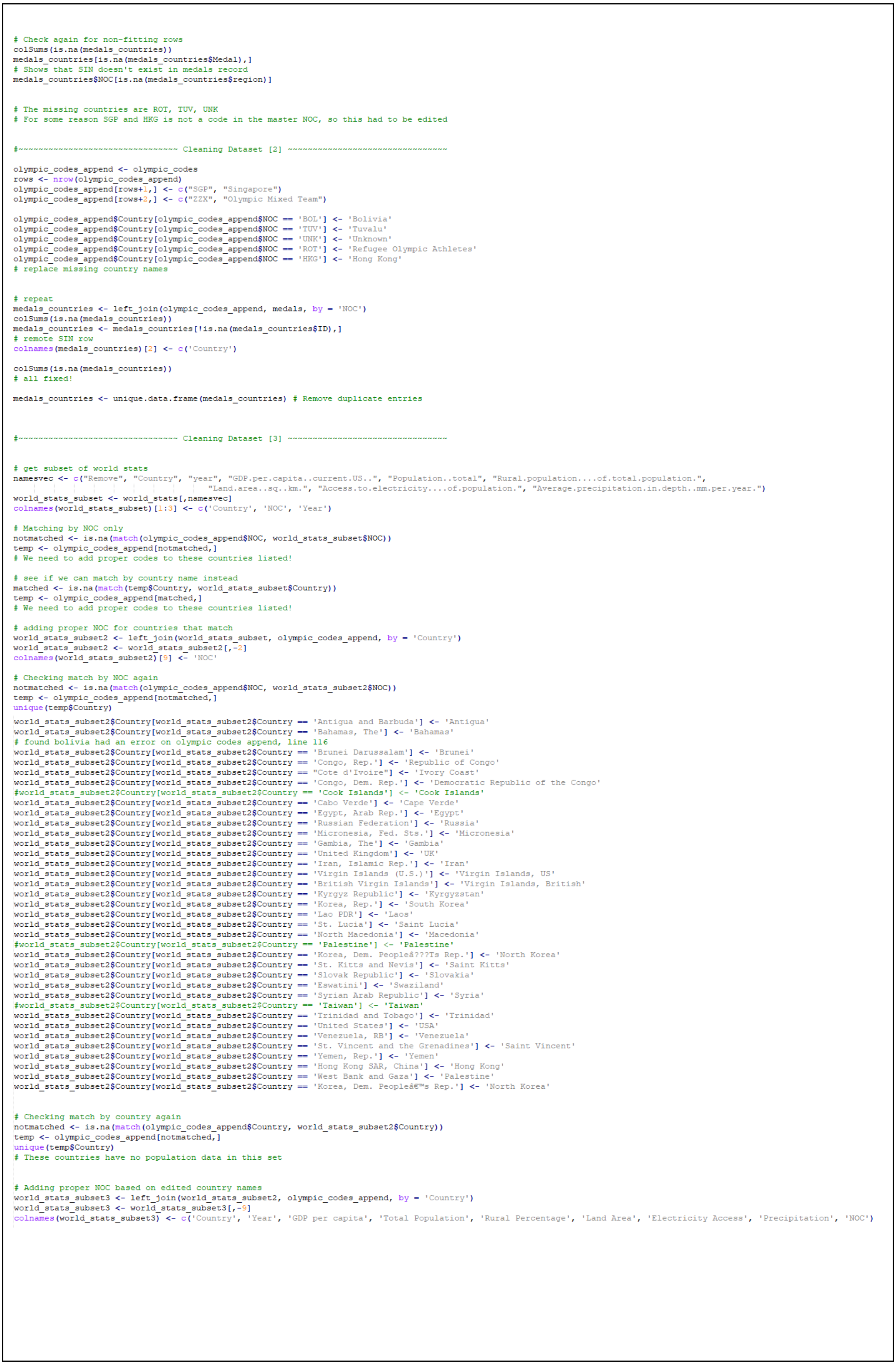
The following table lists the code scripts and datasets utilised in this project.

File Name	Description	Source
medals_athletes.csv	CSV file of athlete and medal data	(IOC, 2018)
world_stats.csv	CSV file of world statistics data	(World Bank, 2020)
project.r	R language file containing code for data cleaning. Excerpts shown in Appendix B.	
model.r	R language file containing code for modelling. Excerpts shown in Appendix C.	
visualisation.r	R language file containing code for visualisations. Excerpts shown in Appendix D.	

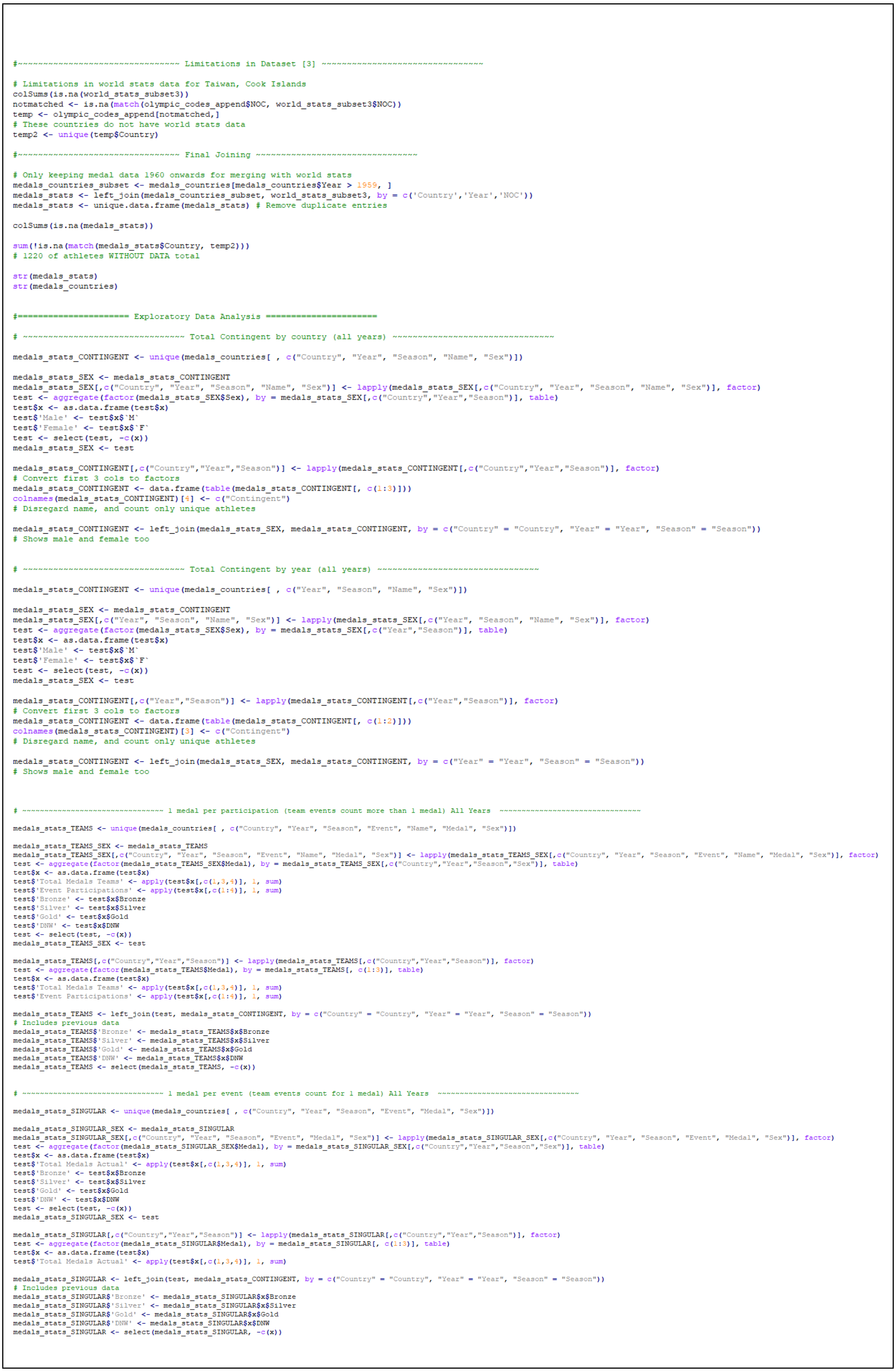
APPENDIX B: DATA CLEANING



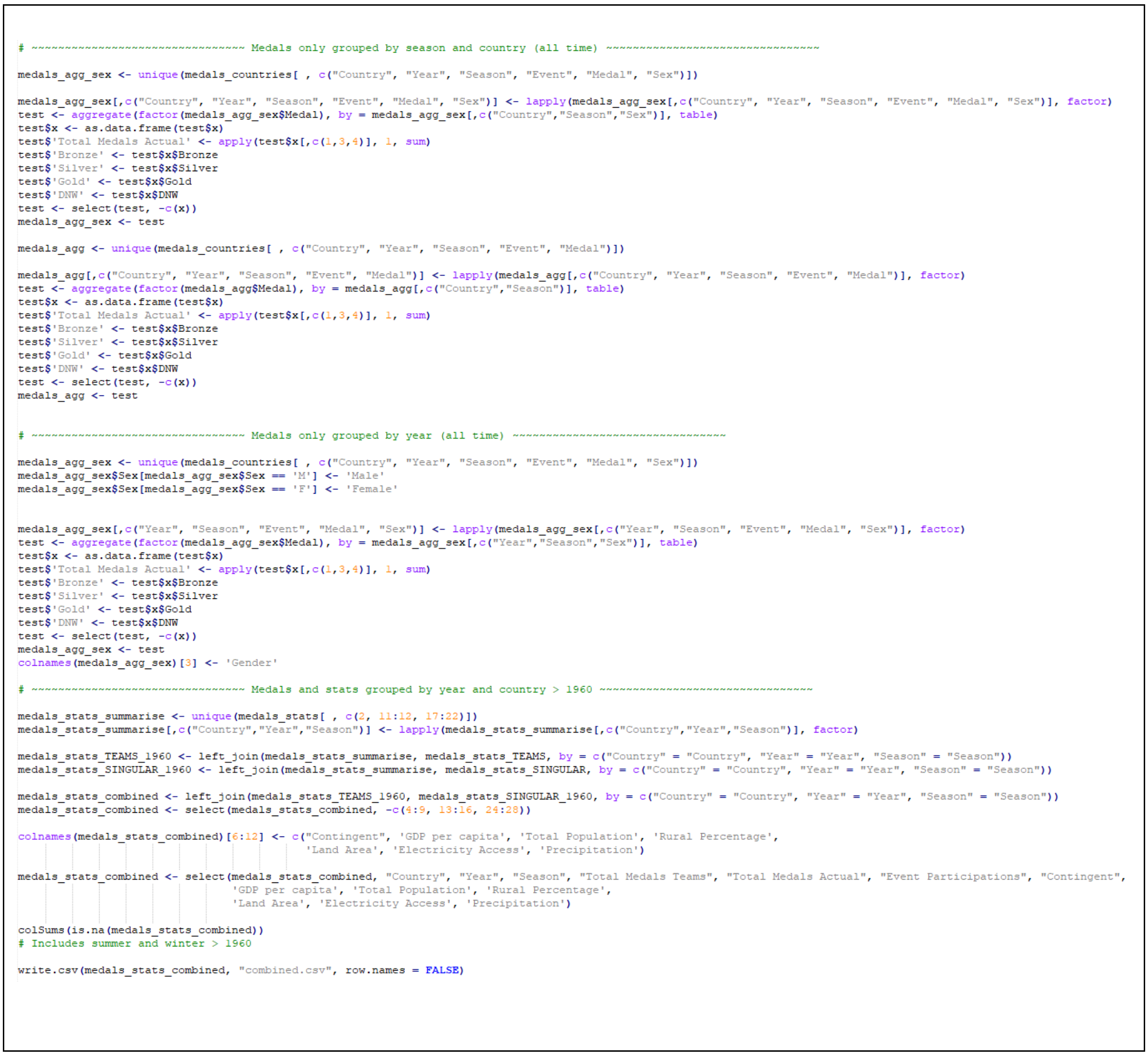
```
#====================== Libraries ======================

library(readr)
library(dplyr)
library(ggplot2)
library(broom)
library(ggpubr)
library(readxl)
library(corrplot)

#====================== Import Data ======================

world_stats <- read.csv("world_stats.csv")
colnames(world_stats)[1] <- c("Remove")
colnames(world_stats)[2] <- c("Country")
# Contains data until 2016
# Duplicate entries
#GDP per capita (current US$)
#Mortality rate, infant (per 1,000 live births)
#Surface area (sq. km)
#Average precipitation in depth (mm per year)
#Access to electricity (% of population)

medals <- read.csv("athlete_events.csv")
medals <- unique.data.frame(medals) # Remove duplicate entries

olympic_codes <- read.csv("noc_regions.csv")
olympic_codes <- olympic_codes[,-3] # Remote notes column
colnames(olympic_codes) <- c('NOC', 'Country') # Rename 'region' to 'country'

summer <- read.csv("summer.csv")
colnames(summer)[6] <- c('NOC')

#====================== Data Sets ======================

# [1] Medal data
# [2] World stats data
# [3] Olympic NOC data

#====================== Cleaning Data ======================

#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ Cleaning Dataset [1] ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~


# show number of NA values
colSums(is.na(medals))
medals$Medal[is.na(medals$Medal)] <- "DNW"
# replace NA in medal column with DNW

colSums(is.na(medals))
# Verify no more NA in medals column

codes1 <- olympic_codes %>%
  distinct(NOC, Country) %>%
  group_by(Country) %>%
  summarize("count" = n())

choose <- which(codes1$count > 1)
codes1_new <- codes1[choose,]
# Displays countries who have more than 1 NOC code assigned

# Do the same to the medals tally

codes2 <- medals %>%
  distinct(NOC, Team) %>%
  group_by(NOC) %>%
  summarize("count" = n())

choose <- which(codes2$count > 1)
codes2_new <- codes2[choose,]

# Displays countries who have more than 1 NOC code assigned
# FRA has 160 different teams associated with it!

# Primary key will be the NOC
medals_countries <- left_join(olympic_codes, medals, by = 'NOC')
```



```r
# Check again for non-fitting rows
colSums(is.na(medals_countries))
medals_countries[is.na(medals_countries$Medal),]
# Shows that SIN doesn't exist in medals record
medals_countries$NOC[is.na(medals_countries$region)]


# The missing countries are ROT, TUV, UNK
# For some reason SGP and HKG is not a code in the master NOC, so this had to be edited


#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ Cleaning Dataset [2] ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

olympic_codes_append <- olympic_codes
rows <- nrow(olympic_codes_append)
olympic_codes_append[rows+1,] <- c("SGP", "Singapore")
olympic_codes_append[rows+2,] <- c("ZZX", "Olympic Mixed Team")

olympic_codes_append$Country[olympic_codes_append$NOC == 'BOL'] <- 'Bolivia'
olympic_codes_append$Country[olympic_codes_append$NOC == 'TUV'] <- 'Tuvalu'
olympic_codes_append$Country[olympic_codes_append$NOC == 'UNK'] <- 'Unknown'
olympic_codes_append$Country[olympic_codes_append$NOC == 'ROT'] <- 'Refugee Olympic Athletes'
olympic_codes_append$Country[olympic_codes_append$NOC == 'HKG'] <- 'Hong Kong'
# replace missing country names


# repeat
medals_countries <- left_join(olympic_codes_append, medals, by = 'NOC')
colSums(is.na(medals_countries))
medals_countries <- medals_countries[!is.na(medals_countries$ID),]
# remote SIN row
colnames(medals_countries)[2] <- c('Country')

colSums(is.na(medals_countries))
# all fixed!

medals_countries <- unique.data.frame(medals_countries) # Remove duplicate entries



#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ Cleaning Dataset [3] ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~


# get subset of world stats
namesvec <- c("Remove", "Country", "year", "GDP.per.capita..current.US..", "Population..total", "Rural.population....of.total.population.",
                                       "Land.area..sq..km.", "Access.to.electricity....of.population.", "Average.precipitation.in.depth..mm.per.year.")
world_stats_subset <- world_stats[,namesvec]
colnames(world_stats_subset)[1:3] <- c('Country', 'NOC', 'Year')

# Matching by NOC only
notmatched <- is.na(match(olympic_codes_append$NOC, world_stats_subset$NOC))
temp <- olympic_codes_append[notmatched,]
# We need to add proper codes to these countries listed!

# see if we can match by country name instead
matched <- is.na(match(temp$Country, world_stats_subset$Country))
temp <- olympic_codes_append[matched,]
# We need to add proper codes to these countries listed!

# adding proper NOC for countries that match
world_stats_subset2 <- left_join(world_stats_subset, olympic_codes_append, by = 'Country')
world_stats_subset2 <- world_stats_subset2[,-2]
colnames(world_stats_subset2)[9] <- 'NOC'

# Checking match by NOC again
notmatched <- is.na(match(olympic_codes_append$NOC, world_stats_subset2$NOC))
temp <- olympic_codes_append[notmatched,]
unique(temp$Country)
world_stats_subset2$Country[world_stats_subset2$Country == 'Antigua and Barbuda'] <- 'Antigua'
world_stats_subset2$Country[world_stats_subset2$Country == 'Bahamas, The'] <- 'Bahamas'
# found bolivia had an error on olympic codes append, line 116
world_stats_subset2$Country[world_stats_subset2$Country == 'Brunei Darussalam'] <- 'Brunei'
world_stats_subset2$Country[world_stats_subset2$Country == 'Congo, Rep.'] <- 'Republic of Congo'
world_stats_subset2$Country[world_stats_subset2$Country == "Cote d'Ivoire"] <- 'Ivory Coast'
world_stats_subset2$Country[world_stats_subset2$Country == 'Congo, Dem. Rep.'] <- 'Democratic Republic of the Congo'
#world_stats_subset2$Country[world_stats_subset2$Country == 'Cook Islands'] <- 'Cook Islands'
world_stats_subset2$Country[world_stats_subset2$Country == 'Cabo Verde'] <- 'Cape Verde'
world_stats_subset2$Country[world_stats_subset2$Country == 'Egypt, Arab Rep.'] <- 'Egypt'
world_stats_subset2$Country[world_stats_subset2$Country == 'Russian Federation'] <- 'Russia'
world_stats_subset2$Country[world_stats_subset2$Country == 'Micronesia, Fed. Sts.'] <- 'Micronesia'
world_stats_subset2$Country[world_stats_subset2$Country == 'Gambia, The'] <- 'Gambia'
world_stats_subset2$Country[world_stats_subset2$Country == 'United Kingdom'] <- 'UK'
world_stats_subset2$Country[world_stats_subset2$Country == 'Iran, Islamic Rep.'] <- 'Iran'
world_stats_subset2$Country[world_stats_subset2$Country == 'Virgin Islands (U.S.)'] <- 'Virgin Islands, US'
world_stats_subset2$Country[world_stats_subset2$Country == 'British Virgin Islands'] <- 'Virgin Islands, British'
world_stats_subset2$Country[world_stats_subset2$Country == 'Kyrgyz Republic'] <- 'Kyrgyzstan'
world_stats_subset2$Country[world_stats_subset2$Country == 'Korea, Rep.'] <- 'South Korea'
world_stats_subset2$Country[world_stats_subset2$Country == 'Lao PDR'] <- 'Laos'
world_stats_subset2$Country[world_stats_subset2$Country == 'St. Lucia'] <- 'Saint Lucia'
world_stats_subset2$Country[world_stats_subset2$Country == 'North Macedonia'] <- 'Macedonia'
#world_stats_subset2$Country[world_stats_subset2$Country == 'Palestine'] <- 'Palestine'
world_stats_subset2$Country[world_stats_subset2$Country == 'Korea, Dem. Peopleâ???Ts Rep.'] <- 'North Korea'
world_stats_subset2$Country[world_stats_subset2$Country == 'St. Kitts and Nevis'] <- 'Saint Kitts'
world_stats_subset2$Country[world_stats_subset2$Country == 'Slovak Republic'] <- 'Slovakia'
world_stats_subset2$Country[world_stats_subset2$Country == 'Eswatini'] <- 'Swaziland'
world_stats_subset2$Country[world_stats_subset2$Country == 'Syrian Arab Republic'] <- 'Syria'
#world_stats_subset2$Country[world_stats_subset2$Country == 'Taiwan'] <- 'Taiwan'
world_stats_subset2$Country[world_stats_subset2$Country == 'Trinidad and Tobago'] <- 'Trinidad'
world_stats_subset2$Country[world_stats_subset2$Country == 'United States'] <- 'USA'
world_stats_subset2$Country[world_stats_subset2$Country == 'Venezuela, RB'] <- 'Venezuela'
world_stats_subset2$Country[world_stats_subset2$Country == 'St. Vincent and the Grenadines'] <- 'Saint Vincent'
world_stats_subset2$Country[world_stats_subset2$Country == 'Yemen, Rep.'] <- 'Yemen'
world_stats_subset2$Country[world_stats_subset2$Country == 'Hong Kong SAR, China'] <- 'Hong Kong'
world_stats_subset2$Country[world_stats_subset2$Country == 'West Bank and Gaza'] <- 'Palestine'
world_stats_subset2$Country[world_stats_subset2$Country == 'Korea, Dem. Peopleâ€™s Rep.'] <- 'North Korea'


# Checking match by country again
notmatched <- is.na(match(olympic_codes_append$Country, world_stats_subset2$Country))
temp <- olympic_codes_append[notmatched,]
unique(temp$Country)
# These countries have no population data in this set


# Adding proper NOC based on edited country names
world_stats_subset3 <- left_join(world_stats_subset2, olympic_codes_append, by = 'Country')
world_stats_subset3 <- world_stats_subset3[,-9]
colnames(world_stats_subset3) <- c('Country', 'Year', 'GDP per capita', 'Total Population', 'Rural Percentage', 'Land Area', 'Electricity Access', 'Precipitation', 'NOC')
```

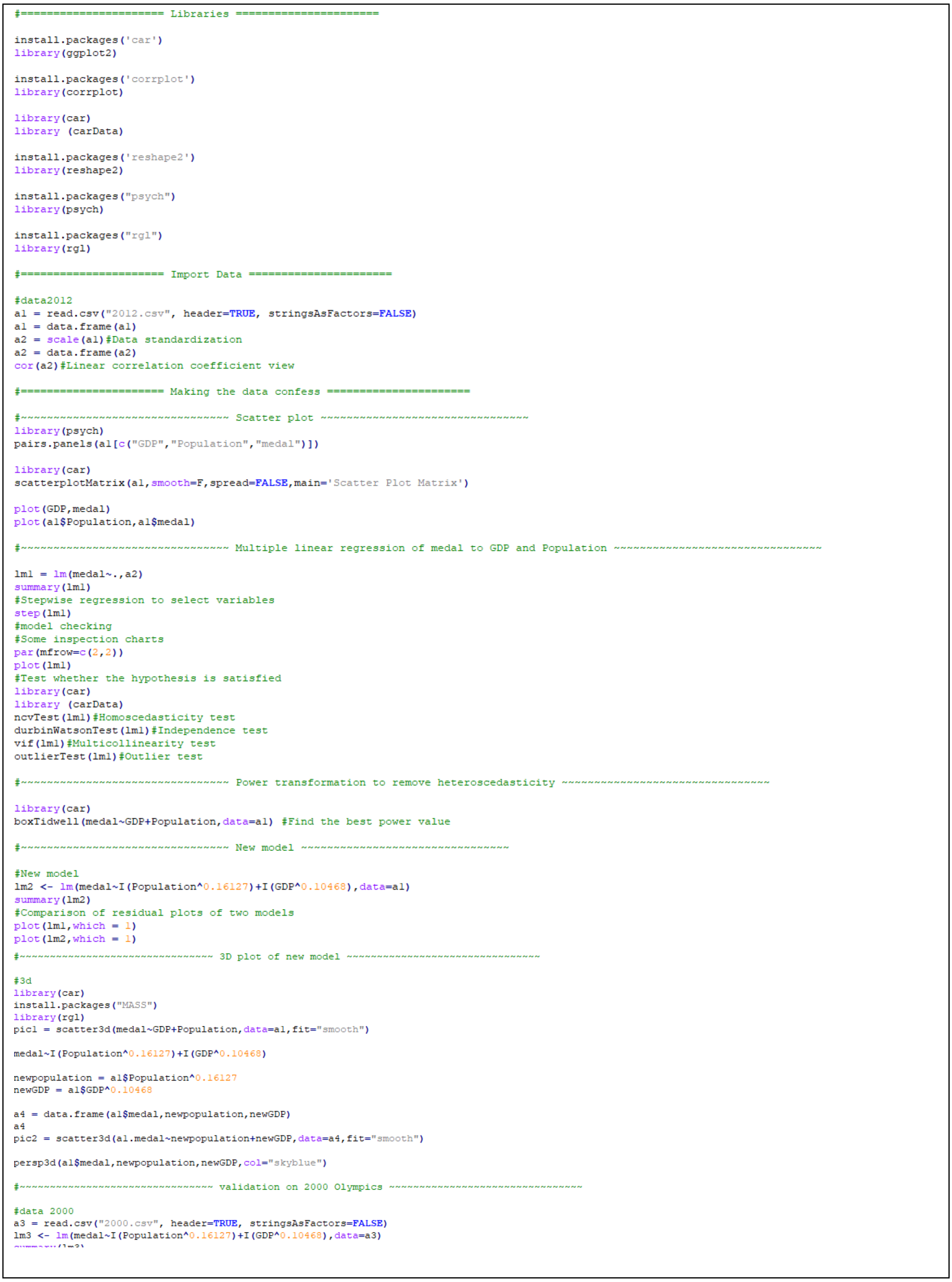


```r
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ Limitations in Dataset [3] ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

# Limitations in world stats data for Taiwan, Cook Islands
colSums(is.na(world_stats_subset3))
notmatched <- is.na(match(olympic_codes_append$NOC, world_stats_subset3$NOC))
temp <- olympic_codes_append[notmatched,]
# These countries do not have world stats data
temp2 <- unique(temp$Country)

#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ Final Joining ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

# Only keeping medal data 1960 onwards for merging with world stats
medals_countries_subset <- medals_countries[medals_countries$Year > 1959, ]
medals_stats <- left_join(medals_countries_subset, world_stats_subset3, by = c('Country','Year','NOC'))
medals_stats <- unique.data.frame(medals_stats) # Remove duplicate entries

colSums(is.na(medals_stats))

sum(!is.na(match(medals_stats$Country, temp2)))
# 1220 of athletes WITHOUT DATA total

str(medals_stats)
str(medals_countries)


#===================== Exploratory Data Analysis =====================

# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ Total Contingent by country (all years) ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

medals_stats_CONTINGENT <- unique(medals_countries[ , c("Country", "Year", "Season", "Name", "Sex")])

medals_stats_SEX <- medals_stats_CONTINGENT
medals_stats_SEX[,c("Country", "Year", "Season", "Name", "Sex")] <- lapply(medals_stats_SEX[,c("Country", "Year", "Season", "Name", "Sex")], factor)
test <- aggregate(factor(medals_stats_SEX$Sex), by = medals_stats_SEX[,c("Country","Year","Season")], table)
test$x <- as.data.frame(test$x)
test$'Male' <- test$x$`M`
test$'Female' <- test$x$`F`
test <- select(test, -c(x))
medals_stats_SEX <- test

medals_stats_CONTINGENT[,c("Country","Year","Season")] <- lapply(medals_stats_CONTINGENT[,c("Country","Year","Season")], factor)
# Convert first 3 cols to factors
medals_stats_CONTINGENT <- data.frame(table(medals_stats_CONTINGENT[, c(1:3)]))
colnames(medals_stats_CONTINGENT)[4] <- c("Contingent")
# Disregard name, and count only unique athletes

medals_stats_CONTINGENT <- left_join(medals_stats_SEX, medals_stats_CONTINGENT, by = c("Country" = "Country", "Year" = "Year", "Season" = "Season"))
# Shows male and female too


# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ Total Contingent by year (all years) ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

medals_stats_CONTINGENT <- unique(medals_countries[ , c("Year", "Season", "Name", "Sex")])

medals_stats_SEX <- medals_stats_CONTINGENT
medals_stats_SEX[,c("Year", "Season", "Name", "Sex")] <- lapply(medals_stats_SEX[,c("Year", "Season", "Name", "Sex")], factor)
test <- aggregate(factor(medals_stats_SEX$Sex), by = medals_stats_SEX[,c("Year","Season")], table)
test$x <- as.data.frame(test$x)
test$'Male' <- test$x$`M`
test$'Female' <- test$x$`F`
test <- select(test, -c(x))
medals_stats_SEX <- test

medals_stats_CONTINGENT[,c("Year","Season")] <- lapply(medals_stats_CONTINGENT[,c("Year","Season")], factor)
# Convert first 3 cols to factors
medals_stats_CONTINGENT <- data.frame(table(medals_stats_CONTINGENT[, c(1:2)]))
colnames(medals_stats_CONTINGENT)[3] <- c("Contingent")
# Disregard name, and count only unique athletes

medals_stats_CONTINGENT <- left_join(medals_stats_SEX, medals_stats_CONTINGENT, by = c("Year" = "Year", "Season" = "Season"))
# Shows male and female too



# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ 1 medal per participation (team events count more than 1 medal) All Years  ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

medals_stats_TEAMS <- unique(medals_countries[ , c("Country", "Year", "Season", "Event", "Name", "Medal", "Sex")])

medals_stats_TEAMS_SEX <- medals_stats_TEAMS
medals_stats_TEAMS_SEX[,c("Country", "Year", "Season", "Event", "Name", "Medal", "Sex")] <- lapply(medals_stats_TEAMS_SEX[,c("Country", "Year", "Season", "Event", "Name", "Medal", "Sex")], factor)
test <- aggregate(factor(medals_stats_TEAMS_SEX$Medal), by = medals_stats_TEAMS_SEX[,c("Country","Year","Season","Sex")], table)
test$x <- as.data.frame(test$x)
test$'Total Medals Teams' <- apply(test$x[,c(1,3,4)], 1, sum)
test$'Event Participations' <- apply(test$x[,c(1:4)], 1, sum)
test$'Bronze' <- test$x$Bronze
test$'Silver' <- test$x$Silver
test$'Gold' <- test$x$Gold
test$'DNW' <- test$x$DNW
test <- select(test, -c(x))
medals_stats_TEAMS_SEX <- test

medals_stats_TEAMS[,c("Country","Year","Season")] <- lapply(medals_stats_TEAMS[,c("Country","Year","Season")], factor)
test <- aggregate(factor(medals_stats_TEAMS$Medal), by = medals_stats_TEAMS[, c(1:3)], table)
test$x <- as.data.frame(test$x)
test$'Total Medals Teams' <- apply(test$x[,c(1,3,4)], 1, sum)
test$'Event Participations' <- apply(test$x[,c(1:4)], 1, sum)

medals_stats_TEAMS <- left_join(test, medals_stats_CONTINGENT, by = c("Country" = "Country", "Year" = "Year", "Season" = "Season"))
# Includes previous data
medals_stats_TEAMS$'Bronze' <- medals_stats_TEAMS$x$Bronze
medals_stats_TEAMS$'Silver' <- medals_stats_TEAMS$x$Silver
medals_stats_TEAMS$'Gold' <- medals_stats_TEAMS$x$Gold
medals_stats_TEAMS$'DNW' <- medals_stats_TEAMS$x$DNW
medals_stats_TEAMS <- select(medals_stats_TEAMS, -c(x))


# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ 1 medal per event (team events count for 1 medal) All Years  ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

medals_stats_SINGULAR <- unique(medals_countries[ , c("Country", "Year", "Season", "Event", "Medal", "Sex")])

medals_stats_SINGULAR_SEX <- medals_stats_SINGULAR
medals_stats_SINGULAR_SEX[,c("Country", "Year", "Season", "Event", "Medal", "Sex")] <- lapply(medals_stats_SINGULAR_SEX[,c("Country", "Year", "Season", "Event", "Medal", "Sex")], factor)
test <- aggregate(factor(medals_stats_SINGULAR_SEX$Medal), by = medals_stats_SINGULAR_SEX[,c("Country","Year","Season","Sex")], table)
test$x <- as.data.frame(test$x)
test$'Total Medals Actual' <- apply(test$x[,c(1,3,4)], 1, sum)
test$'Bronze' <- test$x$Bronze
test$'Silver' <- test$x$Silver
test$'Gold' <- test$x$Gold
test$'DNW' <- test$x$DNW
test <- select(test, -c(x))
medals_stats_SINGULAR_SEX <- test

medals_stats_SINGULAR[,c("Country","Year","Season")] <- lapply(medals_stats_SINGULAR[,c("Country","Year","Season")], factor)
test <- aggregate(factor(medals_stats_SINGULAR$Medal), by = medals_stats_SINGULAR[, c(1:3)], table)
test$x <- as.data.frame(test$x)
test$'Total Medals Actual' <- apply(test$x[,c(1,3,4)], 1, sum)

medals_stats_SINGULAR <- left_join(test, medals_stats_CONTINGENT, by = c("Country" = "Country", "Year" = "Year", "Season" = "Season"))
# Includes previous data
medals_stats_SINGULAR$'Bronze' <- medals_stats_SINGULAR$x$Bronze
medals_stats_SINGULAR$'Silver' <- medals_stats_SINGULAR$x$Silver
medals_stats_SINGULAR$'Gold' <- medals_stats_SINGULAR$x$Gold
medals_stats_SINGULAR$'DNW' <- medals_stats_SINGULAR$x$DNW
medals_stats_SINGULAR <- select(medals_stats_SINGULAR, -c(x))
```



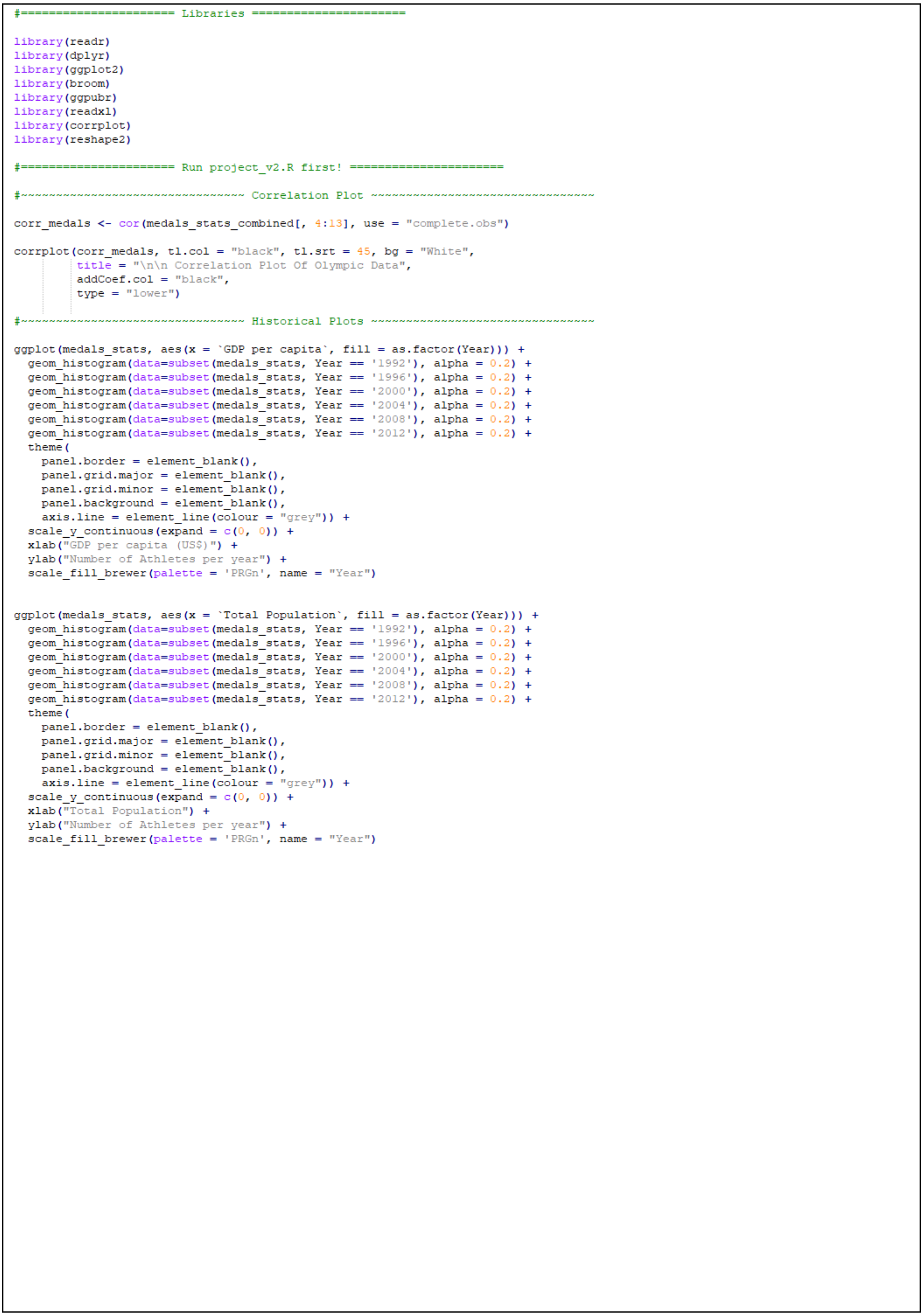
```r
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ Medals only grouped by season and country (all time) ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

medals_agg_sex <- unique(medals_countries[ , c("Country", "Year", "Season", "Event", "Medal", "Sex")])

medals_agg_sex[,c("Country", "Year", "Season", "Event", "Medal", "Sex")] <- lapply(medals_agg_sex[,c("Country", "Year", "Season", "Event", "Medal", "Sex")], factor)
test <- aggregate(factor(medals_agg_sex$Medal), by = medals_agg_sex[,c("Country","Season","Sex")], table)
test$x <- as.data.frame(test$x)
test$'Total Medals Actual' <- apply(test$x[,c(1,3,4)], 1, sum)
test$'Bronze' <- test$x$Bronze
test$'Silver' <- test$x$Silver
test$'Gold' <- test$x$Gold
test$'DNW' <- test$x$DNW
test <- select(test, -c(x))
medals_agg_sex <- test

medals_agg <- unique(medals_countries[ , c("Country", "Year", "Season", "Event", "Medal")])

medals_agg[,c("Country", "Year", "Season", "Event", "Medal")] <- lapply(medals_agg[,c("Country", "Year", "Season", "Event", "Medal")], factor)
test <- aggregate(factor(medals_agg$Medal), by = medals_agg[,c("Country","Season")], table)
test$x <- as.data.frame(test$x)
test$'Total Medals Actual' <- apply(test$x[,c(1,3,4)], 1, sum)
test$'Bronze' <- test$x$Bronze
test$'Silver' <- test$x$Silver
test$'Gold' <- test$x$Gold
test$'DNW' <- test$x$DNW
test <- select(test, -c(x))
medals_agg <- test


# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ Medals only grouped by year (all time) ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

medals_agg_sex <- unique(medals_countries[ , c("Country", "Year", "Season", "Event", "Medal", "Sex")])
medals_agg_sex$Sex[medals_agg_sex$Sex == 'M'] <- 'Male'
medals_agg_sex$Sex[medals_agg_sex$Sex == 'F'] <- 'Female'


medals_agg_sex[,c("Year", "Season", "Event", "Medal", "Sex")] <- lapply(medals_agg_sex[,c("Year", "Season", "Event", "Medal", "Sex")], factor)
test <- aggregate(factor(medals_agg_sex$Medal), by = medals_agg_sex[,c("Year","Season","Sex")], table)
test$x <- as.data.frame(test$x)
test$'Total Medals Actual' <- apply(test$x[,c(1,3,4)], 1, sum)
test$'Bronze' <- test$x$Bronze
test$'Silver' <- test$x$Silver
test$'Gold' <- test$x$Gold
test$'DNW' <- test$x$DNW
test <- select(test, -c(x))
medals_agg_sex <- test
colnames(medals_agg_sex)[3] <- 'Gender'

# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ Medals and stats grouped by year and country > 1960 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

medals_stats_summarise <- unique(medals_stats[ , c(2, 11:12, 17:22)])
medals_stats_summarise[,c("Country","Year","Season")] <- lapply(medals_stats_summarise[,c("Country","Year","Season")], factor)

medals_stats_TEAMS_1960 <- left_join(medals_stats_summarise, medals_stats_TEAMS, by = c("Country" = "Country", "Year" = "Year", "Season" = "Season"))
medals_stats_SINGULAR_1960 <- left_join(medals_stats_summarise, medals_stats_SINGULAR, by = c("Country" = "Country", "Year" = "Year", "Season" = "Season"))

medals_stats_combined <- left_join(medals_stats_TEAMS_1960, medals_stats_SINGULAR_1960, by = c("Country" = "Country", "Year" = "Year", "Season" = "Season"))
medals_stats_combined <- select(medals_stats_combined, -c(4:9, 13:16, 24:28))

colnames(medals_stats_combined)[6:12] <- c("Contingent", 'GDP per capita', 'Total Population', 'Rural Percentage',
                                           'Land Area', 'Electricity Access', 'Precipitation')

medals_stats_combined <- select(medals_stats_combined, "Country", "Year", "Season", "Total Medals Teams", "Total Medals Actual", "Event Participations", "Contingent",
                                'GDP per capita', 'Total Population', 'Rural Percentage',
                                'Land Area', 'Electricity Access', 'Precipitation')

colSums(is.na(medals_stats_combined))
# Includes summer and winter > 1960

write.csv(medals_stats_combined, "combined.csv", row.names = FALSE)
```

APPENDIX C: MODELLING



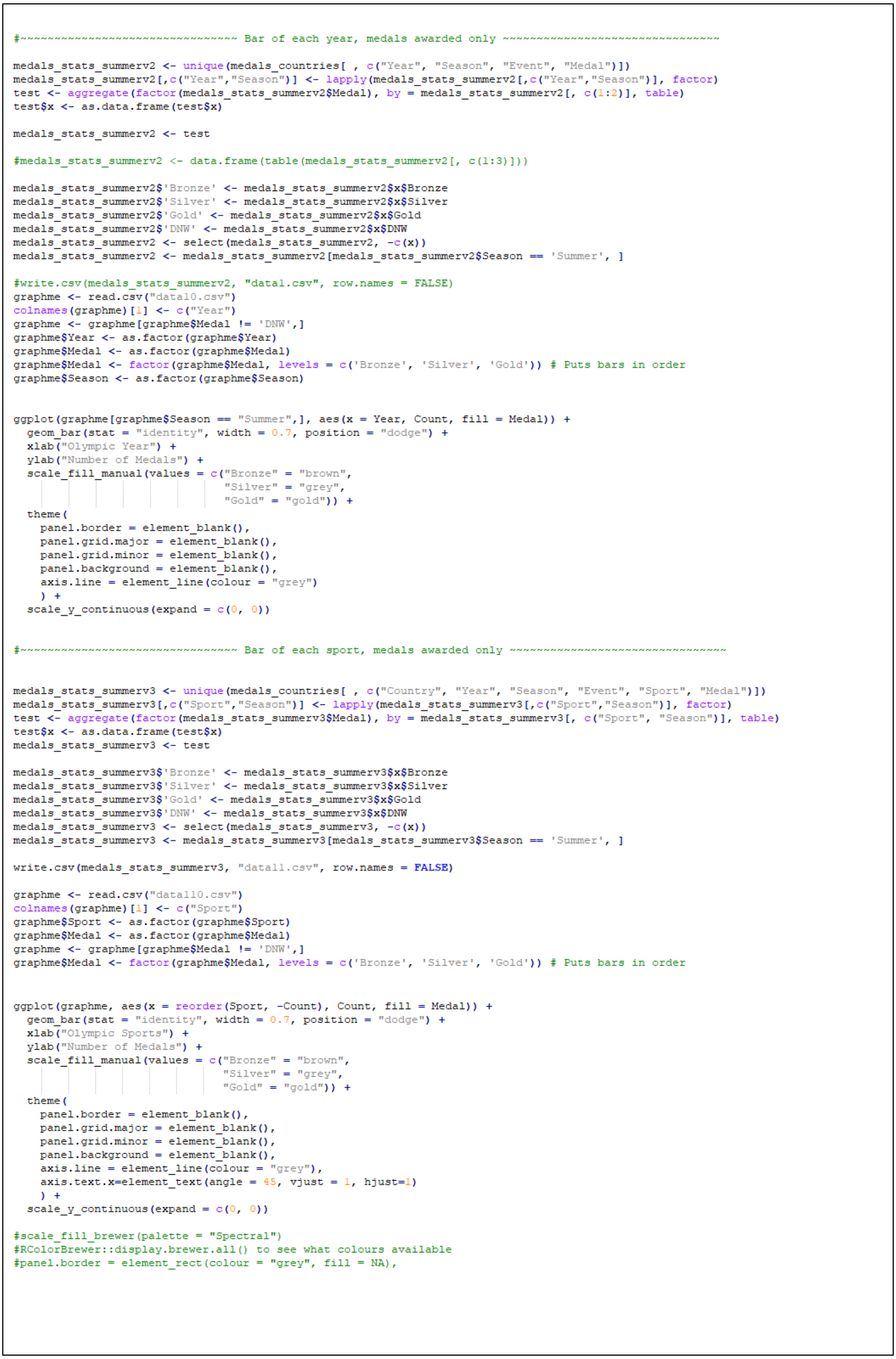
```
#====================== Libraries ======================

install.packages('car')
library(ggplot2)

install.packages('corrplot')
library(corrplot)

library(car)
library (carData)

install.packages('reshape2')
library(reshape2)

install.packages("psych")
library(psych)

install.packages("rgl")
library(rgl)

#====================== Import Data ======================

#data2012
a1 = read.csv("2012.csv", header=TRUE, stringsAsFactors=FALSE)
a1 = data.frame(a1)
a2 = scale(a1)#Data standardization
a2 = data.frame(a2)
cor(a2)#Linear correlation coefficient view

#====================== Making the data confess ======================

#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ Scatter plot ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
library(psych)
pairs.panels(a1[c("GDP","Population","medal")])

library(car)
scatterplotMatrix(a1,smooth=F,spread=FALSE,main='Scatter Plot Matrix')

plot(GDP,medal)
plot(a1$Population,a1$medal)

#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ Multiple linear regression of medal to GDP and Population ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

lm1 = lm(medal~.,a2)
summary(lm1)
#Stepwise regression to select variables
step(lm1)
#model checking
#Some inspection charts
par(mfrow=c(2,2))
plot(lm1)
#Test whether the hypothesis is satisfied
library(car)
library (carData)
ncvTest(lm1)#Homoscedasticity test
durbinWatsonTest(lm1)#Independence test
vif(lm1)#Multicollinearity test
outlierTest(lm1)#Outlier test

#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ Power transformation to remove heteroscedasticity ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

library(car)
boxTidwell(medal~GDP+Population,data=a1) #Find the best power value

#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ New model ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

#New model
lm2 <- lm(medal~I(Population^0.16127)+I(GDP^0.10468),data=a1)
summary(lm2)
#Comparison of residual plots of two models
plot(lm1,which = 1)
plot(lm2,which = 1)
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ 3D plot of new model ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

#3d
library(car)
install.packages("MASS")
library(rgl)
pic1 = scatter3d(medal~GDP+Population,data=a1,fit="smooth")

medal~I(Population^0.16127)+I(GDP^0.10468)

newpopulation = a1$Population^0.16127
newGDP = a1$GDP^0.10468

a4 = data.frame(a1$medal,newpopulation,newGDP)
a4
pic2 = scatter3d(a1.medal~newpopulation+newGDP,data=a4,fit="smooth")

persp3d(a1$medal,newpopulation,newGDP,col="skyblue")

#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ validation on 2000 Olympics ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

#data 2000
a3 = read.csv("2000.csv", header=TRUE, stringsAsFactors=FALSE)
lm3 <- lm(medal~I(Population^0.16127)+I(GDP^0.10468),data=a3)
summary(lm3)
```

APPENDIX D: VISUALISATION

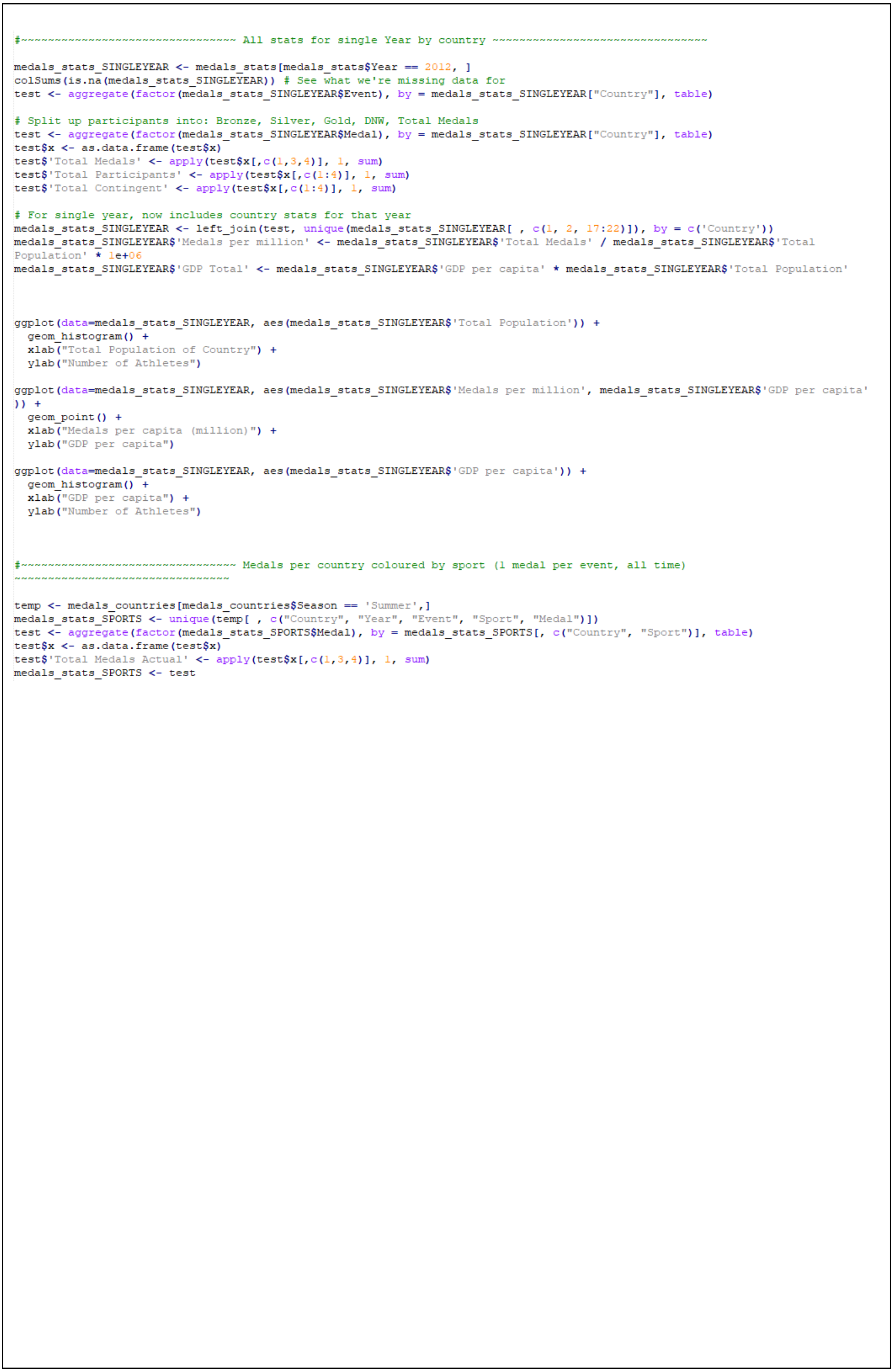


```r
#====================== Libraries ======================

library(readr)
library(dplyr)
library(ggplot2)
library(broom)
library(ggpubr)
library(readxl)
library(corrplot)
library(reshape2)

#====================== Run project_v2.R first! ======================

#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ Correlation Plot ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

corr_medals <- cor(medals_stats_combined[, 4:13], use = "complete.obs")

corrplot(corr_medals, tl.col = "black", tl.srt = 45, bg = "White",
         title = "\n\n Correlation Plot Of Olympic Data",
         addCoef.col = "black",
         type = "lower")

#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ Historical Plots ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

ggplot(medals_stats, aes(x = `GDP per capita`, fill = as.factor(Year))) +
  geom_histogram(data=subset(medals_stats, Year == '1992'), alpha = 0.2) +
  geom_histogram(data=subset(medals_stats, Year == '1996'), alpha = 0.2) +
  geom_histogram(data=subset(medals_stats, Year == '2000'), alpha = 0.2) +
  geom_histogram(data=subset(medals_stats, Year == '2004'), alpha = 0.2) +
  geom_histogram(data=subset(medals_stats, Year == '2008'), alpha = 0.2) +
  geom_histogram(data=subset(medals_stats, Year == '2012'), alpha = 0.2) +
  theme(
    panel.border = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    axis.line = element_line(colour = "grey")) +
  scale_y_continuous(expand = c(0, 0)) +
  xlab("GDP per capita (US$)") +
  ylab("Number of Athletes per year") +
  scale_fill_brewer(palette = 'PRGn', name = "Year")


ggplot(medals_stats, aes(x = `Total Population`, fill = as.factor(Year))) +
  geom_histogram(data=subset(medals_stats, Year == '1992'), alpha = 0.2) +
  geom_histogram(data=subset(medals_stats, Year == '1996'), alpha = 0.2) +
  geom_histogram(data=subset(medals_stats, Year == '2000'), alpha = 0.2) +
  geom_histogram(data=subset(medals_stats, Year == '2004'), alpha = 0.2) +
  geom_histogram(data=subset(medals_stats, Year == '2008'), alpha = 0.2) +
  geom_histogram(data=subset(medals_stats, Year == '2012'), alpha = 0.2) +
  theme(
    panel.border = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    axis.line = element_line(colour = "grey")) +
  scale_y_continuous(expand = c(0, 0)) +
  xlab("Total Population") +
  ylab("Number of Athletes per year") +
  scale_fill_brewer(palette = 'PRGn', name = "Year")
```

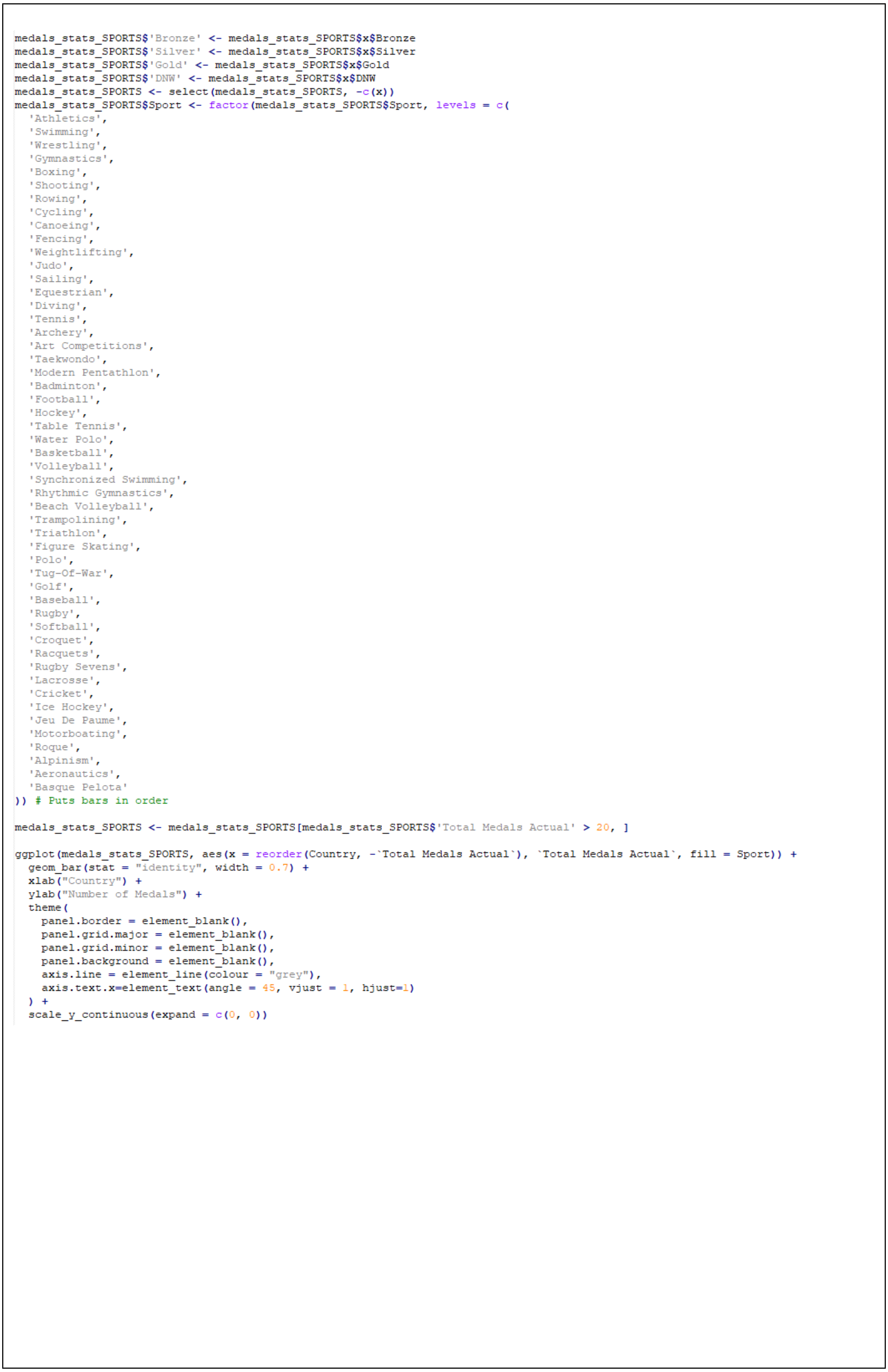
```r
ggplot(medals_stats_combined, aes(x = `Contingent`, fill = as.factor(Year))) +
  geom_histogram(data=subset(medals_stats_combined, Year == '1992'), alpha = 0.2) +
  geom_histogram(data=subset(medals_stats_combined, Year == '1996'), alpha = 0.2) +
  geom_histogram(data=subset(medals_stats_combined, Year == '2000'), alpha = 0.2) +
  geom_histogram(data=subset(medals_stats_combined, Year == '2004'), alpha = 0.2) +
  geom_histogram(data=subset(medals_stats_combined, Year == '2008'), alpha = 0.2) +
  geom_histogram(data=subset(medals_stats_combined, Year == '2012'), alpha = 0.2) +
  theme(
    panel.border = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    axis.line = element_line(colour = "grey")) +
  scale_y_continuous(expand = c(0, 0)) +
  xlab("Size of Contingent") +
  ylab("Number of Athletes per year") +
  scale_fill_brewer(palette = 'RdBu', name = "Year")

mean(medals_stats_combined$Contingent[medals_stats_combined$Year == '1992'])

ggplot(medals_stats_combined, aes(x = `Event Participations`, fill = as.factor(Year))) +
  geom_histogram(data=subset(medals_stats_combined, Year == '1992'), alpha = 0.2) +
  geom_histogram(data=subset(medals_stats_combined, Year == '1996'), alpha = 0.2) +
  geom_histogram(data=subset(medals_stats_combined, Year == '2000'), alpha = 0.2) +
  geom_histogram(data=subset(medals_stats_combined, Year == '2004'), alpha = 0.2) +
  geom_histogram(data=subset(medals_stats_combined, Year == '2008'), alpha = 0.2) +
  geom_histogram(data=subset(medals_stats_combined, Year == '2012'), alpha = 0.2) +
  theme(
    panel.border = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    axis.line = element_line(colour = "grey")) +
  scale_y_continuous(expand = c(0, 0)) +
  xlab("Total Event Participations") +
  ylab("Number of Athletes per year") +
  scale_fill_brewer(palette = 'RdBu', name = "Year")



ggplot(medals_stats_combined[medals_stats_combined$Country == c('Australia', 'USA', 'Afghanistan'),], aes(x=Year, y=`GDP per
capita`, colour=`Country`, na.rm = TRUE)) +
  geom_point() +
  geom_smooth(fullrange=TRUE)

ggplot(data=medals_stats_combined, aes(medals_stats_combined$'Total Medals Actual', medals_stats_combined$'GDP per capita')) +
  geom_point() +
  xlab("Total Medals") +
  ylab("GDP per capita") +
  geom_smooth(method='lm')

ggplot(data=medals_stats_combined, aes(medals_stats_combined$'Total Medals Actual', medals_stats_combined$'Contingent')) +
  geom_point() +
  xlab("Total Medals") +
  ylab("Number of Athletes in Contingent") +
  geom_smooth(method='lm')

ggplot(data=medals_stats_combined, aes(medals_stats_combined$'Total Medals Actual', medals_stats_combined$'Event
Participations')) +
  geom_point() +
  xlab("Total Medals") +
  ylab("Number of Event Participations") +
  geom_smooth(method='lm')
```



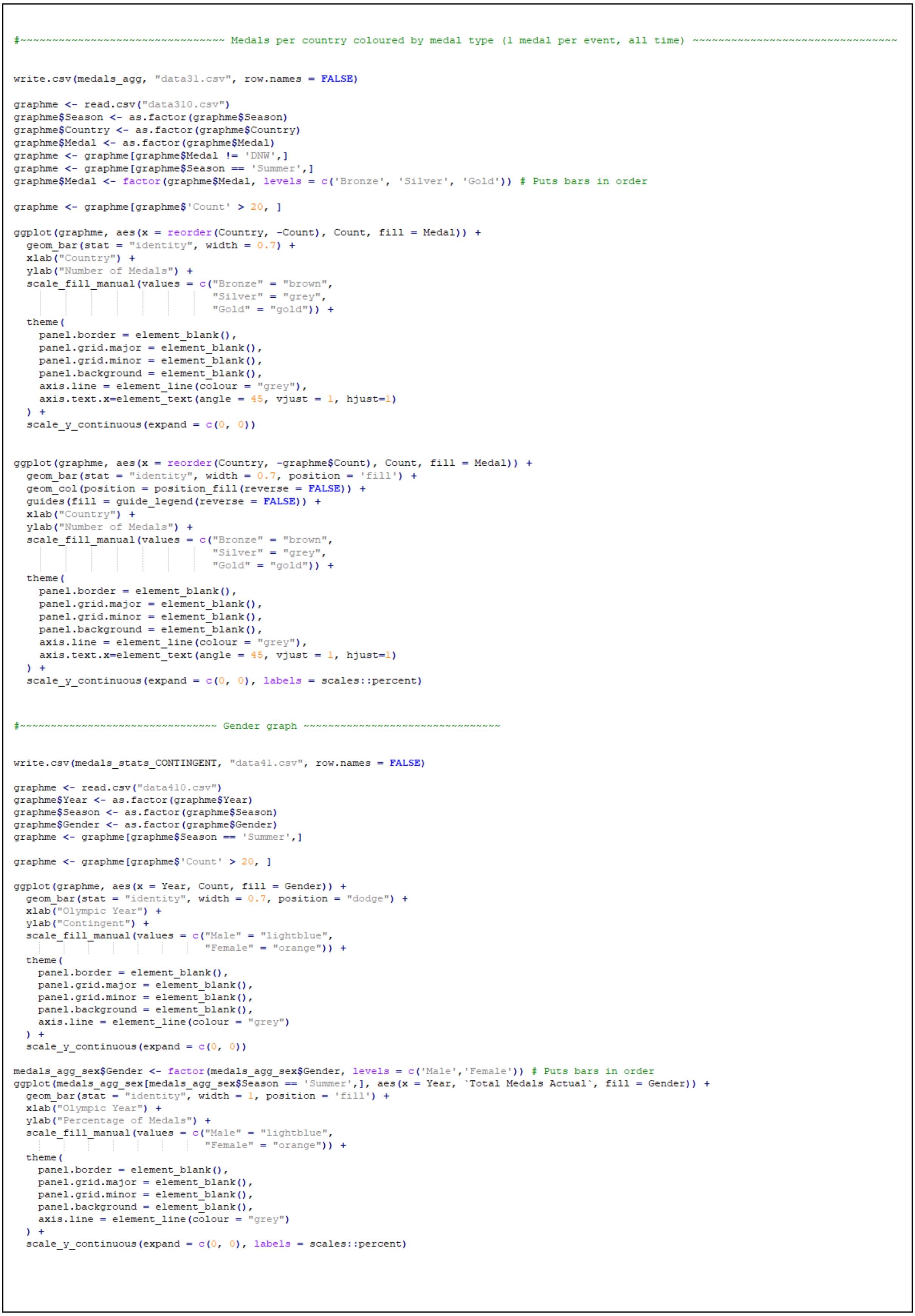
```r
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ Bar of each year, medals awarded only ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

medals_stats_summerv2 <- unique(medals_countries[ , c("Year", "Season", "Event", "Medal")])
medals_stats_summerv2[,c("Year","Season")] <- lapply(medals_stats_summerv2[,c("Year","Season")], factor)
test <- aggregate(factor(medals_stats_summerv2$Medal), by = medals_stats_summerv2[, c(1:2)], table)
test$x <- as.data.frame(test$x)

medals_stats_summerv2 <- test

#medals_stats_summerv2 <- data.frame(table(medals_stats_summerv2[, c(1:3)]))

medals_stats_summerv2$'Bronze' <- medals_stats_summerv2$x$Bronze
medals_stats_summerv2$'Silver' <- medals_stats_summerv2$x$Silver
medals_stats_summerv2$'Gold' <- medals_stats_summerv2$x$Gold
medals_stats_summerv2$'DNW' <- medals_stats_summerv2$x$DNW
medals_stats_summerv2 <- select(medals_stats_summerv2, -c(x))
medals_stats_summerv2 <- medals_stats_summerv2[medals_stats_summerv2$Season == 'Summer', ]

#write.csv(medals_stats_summerv2, "data1.csv", row.names = FALSE)
graphme <- read.csv("data10.csv")
colnames(graphme)[1] <- c("Year")
graphme <- graphme[graphme$Medal != 'DNW',]
graphme$Year <- as.factor(graphme$Year)
graphme$Medal <- as.factor(graphme$Medal)
graphme$Medal <- factor(graphme$Medal, levels = c('Bronze', 'Silver', 'Gold')) # Puts bars in order
graphme$Season <- as.factor(graphme$Season)


ggplot(graphme[graphme$Season == "Summer",], aes(x = Year, Count, fill = Medal)) +
  geom_bar(stat = "identity", width = 0.7, position = "dodge") +
  xlab("Olympic Year") +
  ylab("Number of Medals") +
  scale_fill_manual(values = c("Bronze" = "brown",
                               "Silver" = "grey",
                               "Gold" = "gold")) +
  theme(
    panel.border = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    axis.line = element_line(colour = "grey")
    ) +
  scale_y_continuous(expand = c(0, 0))


#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ Bar of each sport, medals awarded only ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~


medals_stats_summerv3 <- unique(medals_countries[ , c("Country", "Year", "Season", "Event", "Sport", "Medal")])
medals_stats_summerv3[,c("Sport","Season")] <- lapply(medals_stats_summerv3[,c("Sport","Season")], factor)
test <- aggregate(factor(medals_stats_summerv3$Medal), by = medals_stats_summerv3[, c("Sport", "Season")], table)
test$x <- as.data.frame(test$x)
medals_stats_summerv3 <- test

medals_stats_summerv3$'Bronze' <- medals_stats_summerv3$x$Bronze
medals_stats_summerv3$'Silver' <- medals_stats_summerv3$x$Silver
medals_stats_summerv3$'Gold' <- medals_stats_summerv3$x$Gold
medals_stats_summerv3$'DNW' <- medals_stats_summerv3$x$DNW
medals_stats_summerv3 <- select(medals_stats_summerv3, -c(x))
medals_stats_summerv3 <- medals_stats_summerv3[medals_stats_summerv3$Season == 'Summer', ]

write.csv(medals_stats_summerv3, "data11.csv", row.names = FALSE)

graphme <- read.csv("data110.csv")
colnames(graphme)[1] <- c("Sport")
graphme$Sport <- as.factor(graphme$Sport)
graphme$Medal <- as.factor(graphme$Medal)
graphme <- graphme[graphme$Medal != 'DNW',]
graphme$Medal <- factor(graphme$Medal, levels = c('Bronze', 'Silver', 'Gold')) # Puts bars in order


ggplot(graphme, aes(x = reorder(Sport, -Count), Count, fill = Medal)) +
  geom_bar(stat = "identity", width = 0.7, position = "dodge") +
  xlab("Olympic Sports") +
  ylab("Number of Medals") +
  scale_fill_manual(values = c("Bronze" = "brown",
                               "Silver" = "grey",
                               "Gold" = "gold")) +
  theme(
    panel.border = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    axis.line = element_line(colour = "grey"),
    axis.text.x=element_text(angle = 45, vjust = 1, hjust=1)
    ) +
  scale_y_continuous(expand = c(0, 0))

#scale_fill_brewer(palette = "Spectral")
#RColorBrewer::display.brewer.all() to see what colours available
#panel.border = element_rect(colour = "grey", fill = NA),
```



```
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ All stats for single Year by country ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

medals_stats_SINGLEYEAR <- medals_stats[medals_stats$Year == 2012, ]
colSums(is.na(medals_stats_SINGLEYEAR)) # See what we're missing data for
test <- aggregate(factor(medals_stats_SINGLEYEAR$Event), by = medals_stats_SINGLEYEAR["Country"], table)

# Split up participants into: Bronze, Silver, Gold, DNW, Total Medals
test <- aggregate(factor(medals_stats_SINGLEYEAR$Medal), by = medals_stats_SINGLEYEAR["Country"], table)
test$x <- as.data.frame(test$x)
test$'Total Medals' <- apply(test$x[,c(1,3,4)], 1, sum)
test$'Total Participants' <- apply(test$x[,c(1:4)], 1, sum)
test$'Total Contingent' <- apply(test$x[,c(1:4)], 1, sum)

# For single year, now includes country stats for that year
medals_stats_SINGLEYEAR <- left_join(test, unique(medals_stats_SINGLEYEAR[ , c(1, 2, 17:22)]), by = c('Country'))
medals_stats_SINGLEYEAR$'Medals per million' <- medals_stats_SINGLEYEAR$'Total Medals' / medals_stats_SINGLEYEAR$'Total
Population' * 1e+06
medals_stats_SINGLEYEAR$'GDP Total' <- medals_stats_SINGLEYEAR$'GDP per capita' * medals_stats_SINGLEYEAR$'Total Population'



ggplot(data=medals_stats_SINGLEYEAR, aes(medals_stats_SINGLEYEAR$'Total Population')) +
  geom_histogram() +
  xlab("Total Population of Country") +
  ylab("Number of Athletes")

ggplot(data=medals_stats_SINGLEYEAR, aes(medals_stats_SINGLEYEAR$'Medals per million', medals_stats_SINGLEYEAR$'GDP per capita'
)) +
  geom_point() +
  xlab("Medals per capita (million)") +
  ylab("GDP per capita")

ggplot(data=medals_stats_SINGLEYEAR, aes(medals_stats_SINGLEYEAR$'GDP per capita')) +
  geom_histogram() +
  xlab("GDP per capita") +
  ylab("Number of Athletes")



#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ Medals per country coloured by sport (1 medal per event, all time)
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

temp <- medals_countries[medals_countries$Season == 'Summer',]
medals_stats_SPORTS <- unique(temp[ , c("Country", "Year", "Event", "Sport", "Medal")])
test <- aggregate(factor(medals_stats_SPORTS$Medal), by = medals_stats_SPORTS[, c("Country", "Sport")], table)
test$x <- as.data.frame(test$x)
test$'Total Medals Actual' <- apply(test$x[,c(1,3,4)], 1, sum)
medals_stats_SPORTS <- test
```



```
medals_stats_SPORTS$'Bronze' <- medals_stats_SPORTS$x$Bronze
medals_stats_SPORTS$'Silver' <- medals_stats_SPORTS$x$Silver
medals_stats_SPORTS$'Gold' <- medals_stats_SPORTS$x$Gold
medals_stats_SPORTS$'DNW' <- medals_stats_SPORTS$x$DNW
medals_stats_SPORTS <- select(medals_stats_SPORTS, -c(x))
medals_stats_SPORTS$Sport <- factor(medals_stats_SPORTS$Sport, levels = c(
  'Athletics',
  'Swimming',
  'Wrestling',
  'Gymnastics',
  'Boxing',
  'Shooting',
  'Rowing',
  'Cycling',
  'Canoeing',
  'Fencing',
  'Weightlifting',
  'Judo',
  'Sailing',
  'Equestrian',
  'Diving',
  'Tennis',
  'Archery',
  'Art Competitions',
  'Taekwondo',
  'Modern Pentathlon',
  'Badminton',
  'Football',
  'Hockey',
  'Table Tennis',
  'Water Polo',
  'Basketball',
  'Volleyball',
  'Synchronized Swimming',
  'Rhythmic Gymnastics',
  'Beach Volleyball',
  'Trampolining',
  'Triathlon',
  'Figure Skating',
  'Polo',
  'Tug-Of-War',
  'Golf',
  'Baseball',
  'Rugby',
  'Softball',
  'Croquet',
  'Racquets',
  'Rugby Sevens',
  'Lacrosse',
  'Cricket',
  'Ice Hockey',
  'Jeu De Paume',
  'Motorboating',
  'Roque',
  'Alpinism',
  'Aeronautics',
  'Basque Pelota'
)) # Puts bars in order

medals_stats_SPORTS <- medals_stats_SPORTS[medals_stats_SPORTS$'Total Medals Actual' > 20, ]

ggplot(medals_stats_SPORTS, aes(x = reorder(Country, -`Total Medals Actual`), `Total Medals Actual`, fill = Sport)) +
  geom_bar(stat = "identity", width = 0.7) +
  xlab("Country") +
  ylab("Number of Medals") +
  theme(
    panel.border = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    axis.line = element_line(colour = "grey"),
    axis.text.x=element_text(angle = 45, vjust = 1, hjust=1)
  ) +
  scale_y_continuous(expand = c(0, 0))
```



```r
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ Medals per country coloured by medal type (1 medal per event, all time) ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~


write.csv(medals_agg, "data31.csv", row.names = FALSE)

graphme <- read.csv("data310.csv")
graphme$Season <- as.factor(graphme$Season)
graphme$Country <- as.factor(graphme$Country)
graphme$Medal <- as.factor(graphme$Medal)
graphme <- graphme[graphme$Medal != 'DNW',]
graphme <- graphme[graphme$Season == 'Summer',]
graphme$Medal <- factor(graphme$Medal, levels = c('Bronze', 'Silver', 'Gold')) # Puts bars in order

graphme <- graphme[graphme$'Count' > 20, ]

ggplot(graphme, aes(x = reorder(Country, -Count), Count, fill = Medal)) +
  geom_bar(stat = "identity", width = 0.7) +
  xlab("Country") +
  ylab("Number of Medals") +
  scale_fill_manual(values = c("Bronze" = "brown",
                               "Silver" = "grey",
                               "Gold" = "gold")) +
  theme(
    panel.border = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    axis.line = element_line(colour = "grey"),
    axis.text.x=element_text(angle = 45, vjust = 1, hjust=1)
  ) +
  scale_y_continuous(expand = c(0, 0))


ggplot(graphme, aes(x = reorder(Country, -graphme$Count), Count, fill = Medal)) +
  geom_bar(stat = "identity", width = 0.7, position = 'fill') +
  geom_col(position = position_fill(reverse = FALSE)) +
  guides(fill = guide_legend(reverse = FALSE)) +
  xlab("Country") +
  ylab("Number of Medals") +
  scale_fill_manual(values = c("Bronze" = "brown",
                               "Silver" = "grey",
                               "Gold" = "gold")) +
  theme(
    panel.border = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    axis.line = element_line(colour = "grey"),
    axis.text.x=element_text(angle = 45, vjust = 1, hjust=1)
  ) +
  scale_y_continuous(expand = c(0, 0), labels = scales::percent)



#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ Gender graph ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~


write.csv(medals_stats_CONTINGENT, "data41.csv", row.names = FALSE)

graphme <- read.csv("data410.csv")
graphme$Year <- as.factor(graphme$Year)
graphme$Season <- as.factor(graphme$Season)
graphme$Gender <- as.factor(graphme$Gender)
graphme <- graphme[graphme$Season == 'Summer',]

graphme <- graphme[graphme$'Count' > 20, ]

ggplot(graphme, aes(x = Year, Count, fill = Gender)) +
  geom_bar(stat = "identity", width = 0.7, position = "dodge") +
  xlab("Olympic Year") +
  ylab("Contingent") +
  scale_fill_manual(values = c("Male" = "lightblue",
                               "Female" = "orange")) +
  theme(
    panel.border = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    axis.line = element_line(colour = "grey")
  ) +
  scale_y_continuous(expand = c(0, 0))

medals_agg_sex$Gender <- factor(medals_agg_sex$Gender, levels = c('Male','Female')) # Puts bars in order
ggplot(medals_agg_sex[medals_agg_sex$Season == 'Summer',], aes(x = Year, `Total Medals Actual`, fill = Gender)) +
  geom_bar(stat = "identity", width = 1, position = 'fill') +
  xlab("Olympic Year") +
  ylab("Percentage of Medals") +
  scale_fill_manual(values = c("Male" = "lightblue",
                               "Female" = "orange")) +
  theme(
    panel.border = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    axis.line = element_line(colour = "grey")
  ) +
  scale_y_continuous(expand = c(0, 0), labels = scales::percent)
```