5BUIS006C

Data Visualization and Communication

Data Analysis, visualization narrative and presentation

Portfolio (2024)

NAME : Heashalla Baanu Sundaresan

UoW Number : w2083670

IIT ID : 20230983

Degree Program : BSc in Business Data Analytics

Module Name: Data Visualization and Communication

Module Code: 5BUIS006C

Contents

Research Question and Data Sourcing

Research Question

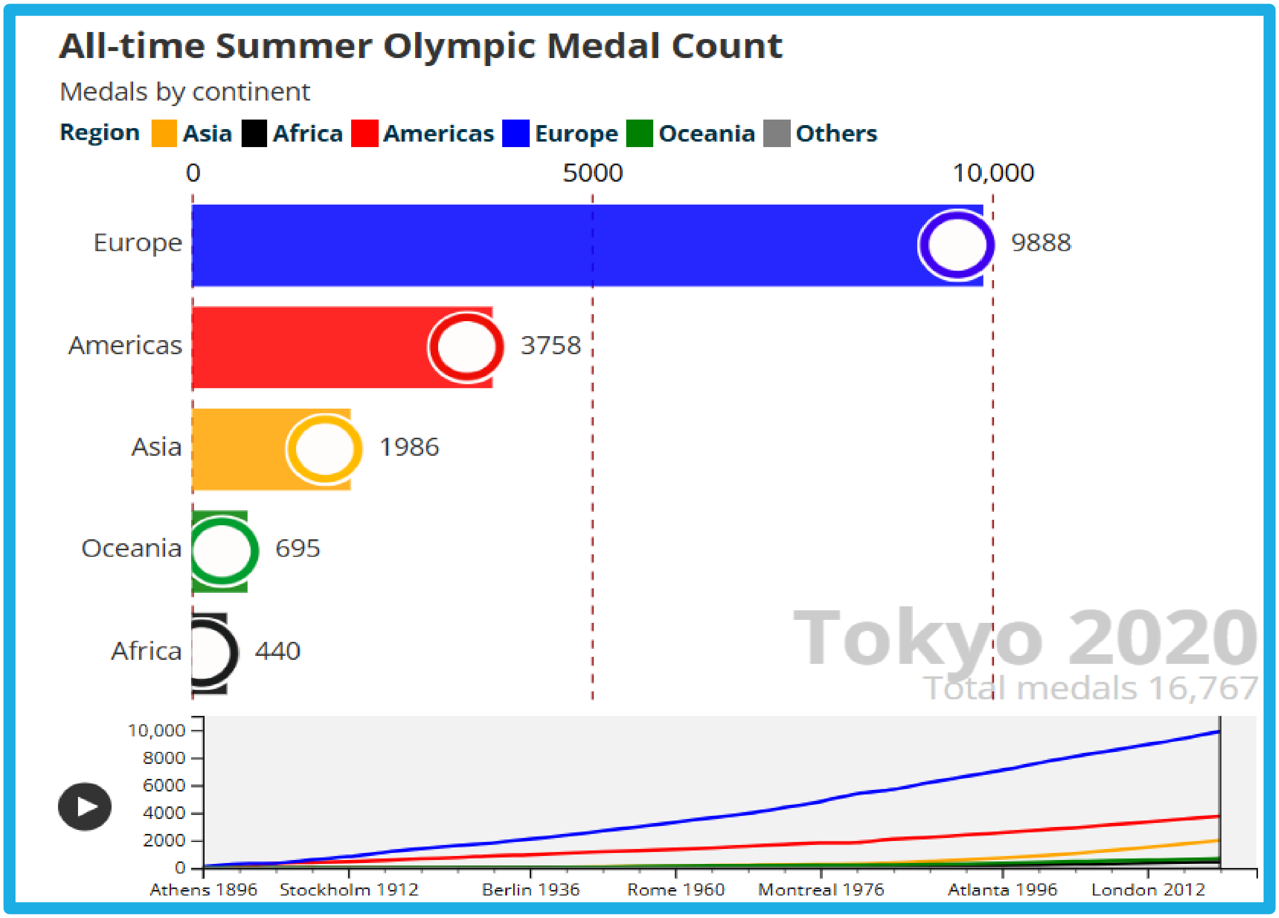
“In terms of medal success, how do different continents compare in the Tokyo 2020 Olympics?”

The Olympics is not only a unique event that happens every four years but also a space where athletes from across the world compete, representing their countries and continents (Shintaro Kano, 2020).

This creates an opportunity for EDA, focusing athlete participation, counts per Medal Types (Gold, Silver, and Bronze), Gender wise involvement and Event types (Individual, group).

Five Rings in Olympics flag represents the Union of the five continents which consists of Africa, Americas, Asia, Europe, and Oceania.

Therefore, this is an opportunity to uncover spatial analysis, disparities in sports performance and provide impactful insights on how all continents fare in the Olympics 2020.



(Euronews, 2024)

There is a lack of comprehensive research and depth analysis that focuses highly on continental performance across the Olympic games, and I had the availability of large dataset of athletes' details of Olympics 2020 for analysis.

Data Sourcing

Dataset

Amiri, A.A.(2021). Tokyo 2020 Olympics dataset: Results, events, ranks, and medals. *Kaggle*. Available at: https://www.kaggle.com/datasets/aliaamiri/2020-summer-olympics-dataset/data?select=2020_Olympics_Dataset.csv [Accessed 7 December 2024].

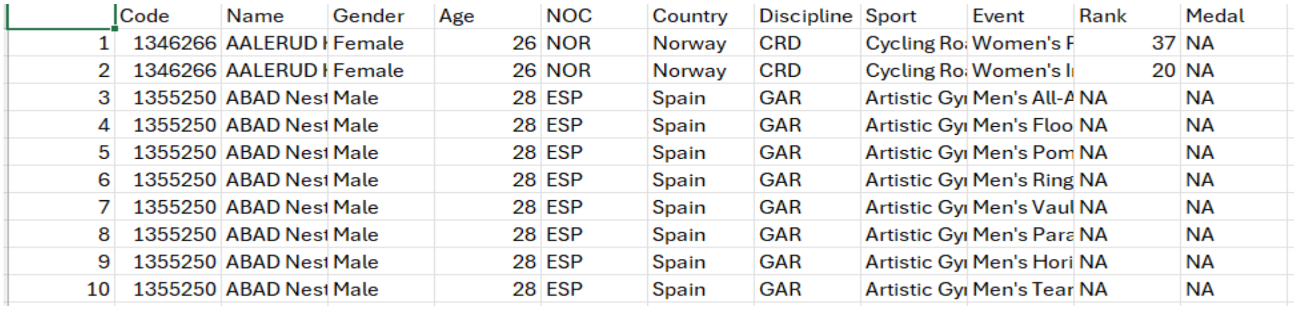
Websites

Shintaro Kano (2020). Game on: Tokyo 2020 competition schedule unveiled for Olympics in 2021.*Olympics*. Available at: https://olympics.com/en/news/tokyo-2020-olympics-2021-competition-schedule [Accessed 8 December 2024]

Euronews, (2024). 128 years of games: Which continent is the most successful in the history of the Olympics?. *Euronews*. Available at: https://www.euronews.com/2024/08/12/128-years-of-games-which-continent-is-the-most-successful-in-the-history-of-the-olympics [Accessed 8 December 2024]

Data Preparation

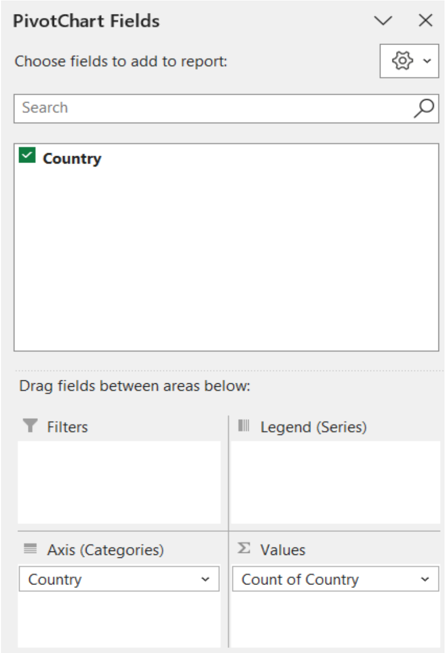
Link to Untidy dataset - Olympics_Untidy (15121 observations)



	Code	Name	Gender	Age	NOC	Country	Discipline	Sport	Event	Rank	Medal
1	1346266	AALERUD I	Female	26	NOR	Norway	CRD	Cycling Ro:	Women's F	37	NA
2	1346266	AALERUD I	Female	26	NOR	Norway	CRD	Cycling Ro:	Women's I	20	NA
3	1355250	ABAD Nest	Male	28	ESP	Spain	GAR	Artistic Gy	Men's All-A	NA	NA
4	1355250	ABAD Nest	Male	28	ESP	Spain	GAR	Artistic Gy	Men's Floo	NA	NA
5	1355250	ABAD Nest	Male	28	ESP	Spain	GAR	Artistic Gy	Men's Pom	NA	NA
6	1355250	ABAD Nest	Male	28	ESP	Spain	GAR	Artistic Gy	Men's Ring	NA	NA
7	1355250	ABAD Nest	Male	28	ESP	Spain	GAR	Artistic Gy	Men's Vaul	NA	NA
8	1355250	ABAD Nest	Male	28	ESP	Spain	GAR	Artistic Gy	Men's Para	NA	NA
9	1355250	ABAD Nest	Male	28	ESP	Spain	GAR	Artistic Gy	Men's Hori	NA	NA
10	1355250	ABAD Nest	Male	28	ESP	Spain	GAR	Artistic Gy	Men's Tear	NA	NA

Data cleaning

- Accuracy increased by comparing countries with actual NOC countries given in the Official Olympic website. For example, China is changed as People's Republic of China.
- Pivot chart to bring all unique countries (206) and counts.



PivotChart Fields

Choose fields to add to report:

Search

☑ Country

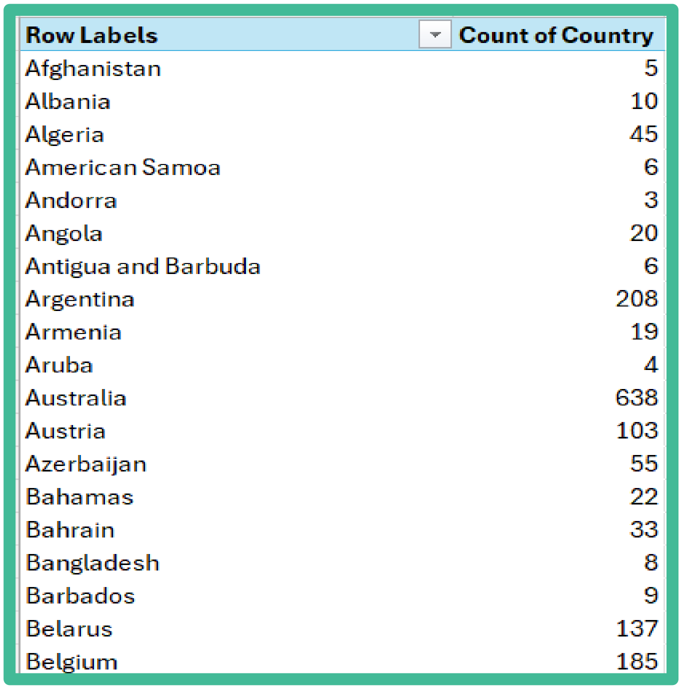
Drag fields between areas below:

Filters

Legend (Series)

Axis (Categories): Country

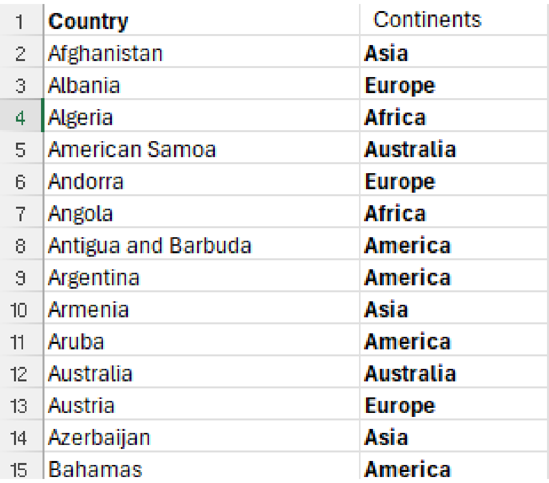
Values: Count of Country



Row Labels	Count of Country
Afghanistan	5
Albania	10
Algeria	45
American Samoa	6
Andorra	3
Angola	20
Antigua and Barbuda	6
Argentina	208
Armenia	19
Aruba	4
Australia	638
Austria	103
Azerbaijan	55
Bahamas	22
Bahrain	33
Bangladesh	8
Barbados	9
Belarus	137
Belgium	185

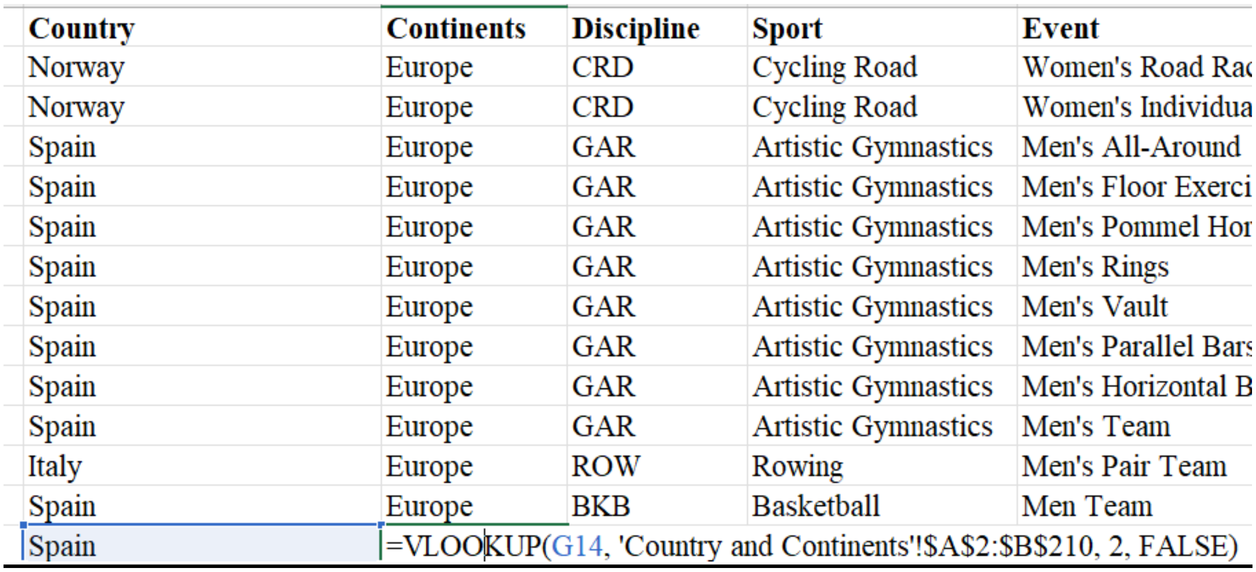
- Since my RQ is based on Continents I added column which should be derived from Countries. Therefore, individually found each country belongs to which continents from https://en.wikipedia.org/wiki/Olympic_symbols.
- Among 7 continents, the International Olympic Committee (IOC) uses a five-continent model that focuses on inhabited continents. Where North America and South America come together as America. Antarctica is a continent not a country, and there are no permanent

populations. Other than that, there is **Olympics refugees Team,** which is apart from Continents, so it considered **as N/A (Outlier)**. Türkiye and Russia wide enough to spread across Asia and Europe, but IOC decided to fit it in Europe. Link to Continents & Countries



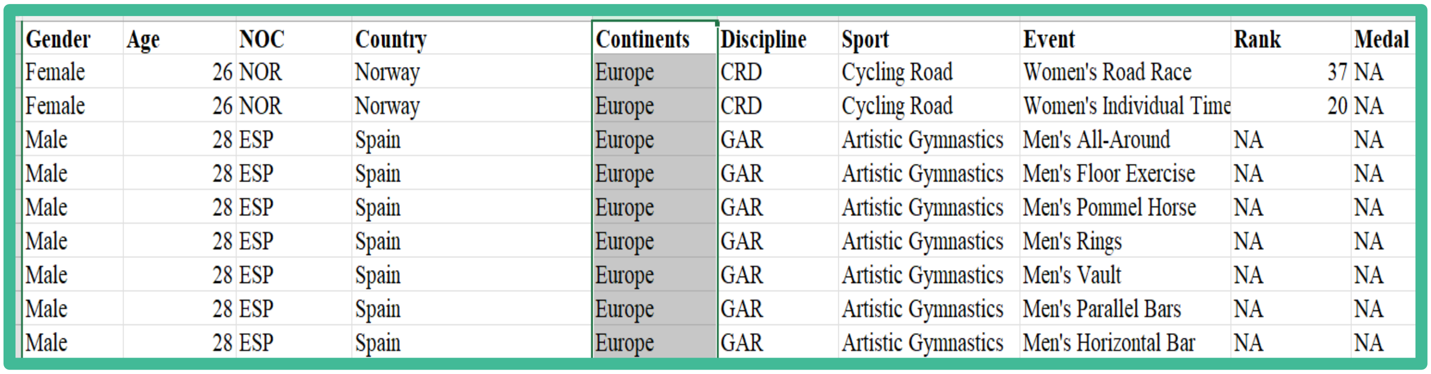
	Country	Continents
2	Afghanistan	Asia
3	Albania	Europe
4	Algeria	Africa
5	American Samoa	Australia
6	Andorra	Europe
7	Angola	Africa
8	Antigua and Barbuda	America
9	Argentina	America
10	Armenia	Asia
11	Aruba	America
12	Australia	Australia
13	Austria	Europe
14	Azerbaijan	Asia
15	Bahamas	America

Using the created sheet of Continents and Countries applying that to the main dataset.



Country	Continents	Discipline	Sport	Event
Norway	Europe	CRD	Cycling Road	Women's Road Rac
Norway	Europe	CRD	Cycling Road	Women's Individua
Spain	Europe	GAR	Artistic Gymnastics	Men's All-Around
Spain	Europe	GAR	Artistic Gymnastics	Men's Floor Exerci
Spain	Europe	GAR	Artistic Gymnastics	Men's Pommel Hor
Spain	Europe	GAR	Artistic Gymnastics	Men's Rings
Spain	Europe	GAR	Artistic Gymnastics	Men's Vault
Spain	Europe	GAR	Artistic Gymnastics	Men's Parallel Bars
Spain	Europe	GAR	Artistic Gymnastics	Men's Horizontal B
Spain	Europe	GAR	Artistic Gymnastics	Men's Team
Italy	Europe	ROW	Rowing	Men's Pair Team
Spain	Europe	BKB	Basketball	Men Team
Spain	=VLOOKUP(G14, 'Country and Continents'!$A$2:$B$210, 2, FALSE)			

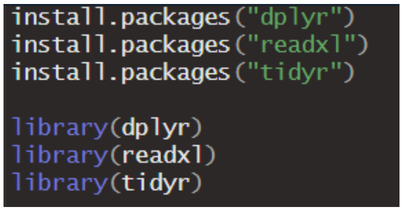
Output



Gender	Age	NOC	Country	Continents	Discipline	Sport	Event	Rank	Medal
Female	26	NOR	Norway	Europe	CRD	Cycling Road	Women's Road Race	37	NA
Female	26	NOR	Norway	Europe	CRD	Cycling Road	Women's Individual Time	20	NA
Male	28	ESP	Spain	Europe	GAR	Artistic Gymnastics	Men's All-Around	NA	NA
Male	28	ESP	Spain	Europe	GAR	Artistic Gymnastics	Men's Floor Exercise	NA	NA
Male	28	ESP	Spain	Europe	GAR	Artistic Gymnastics	Men's Pommel Horse	NA	NA
Male	28	ESP	Spain	Europe	GAR	Artistic Gymnastics	Men's Rings	NA	NA
Male	28	ESP	Spain	Europe	GAR	Artistic Gymnastics	Men's Vault	NA	NA
Male	28	ESP	Spain	Europe	GAR	Artistic Gymnastics	Men's Parallel Bars	NA	NA
Male	28	ESP	Spain	Europe	GAR	Artistic Gymnastics	Men's Horizontal Bar	NA	NA

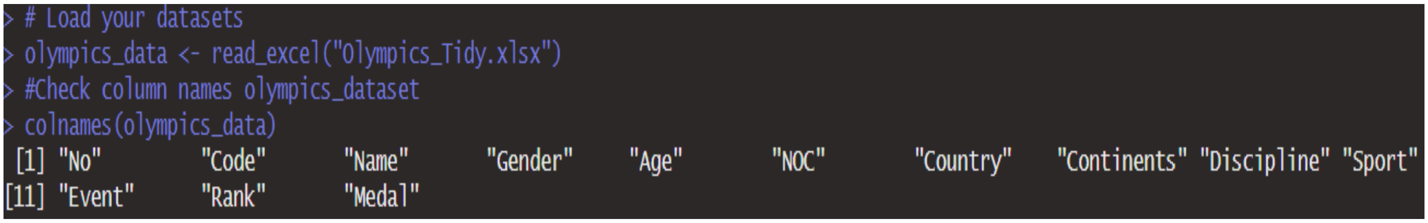
Transforming it into tidy format

Running necessary packages



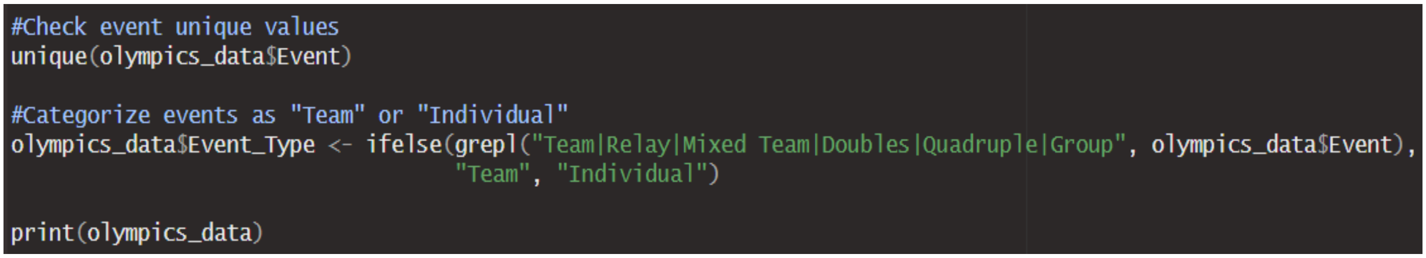
```
install.packages("dplyr")
install.packages("readxl")
install.packages("tidyr")

library(dplyr)
library(readxl)
library(tidyr)
```

All column names



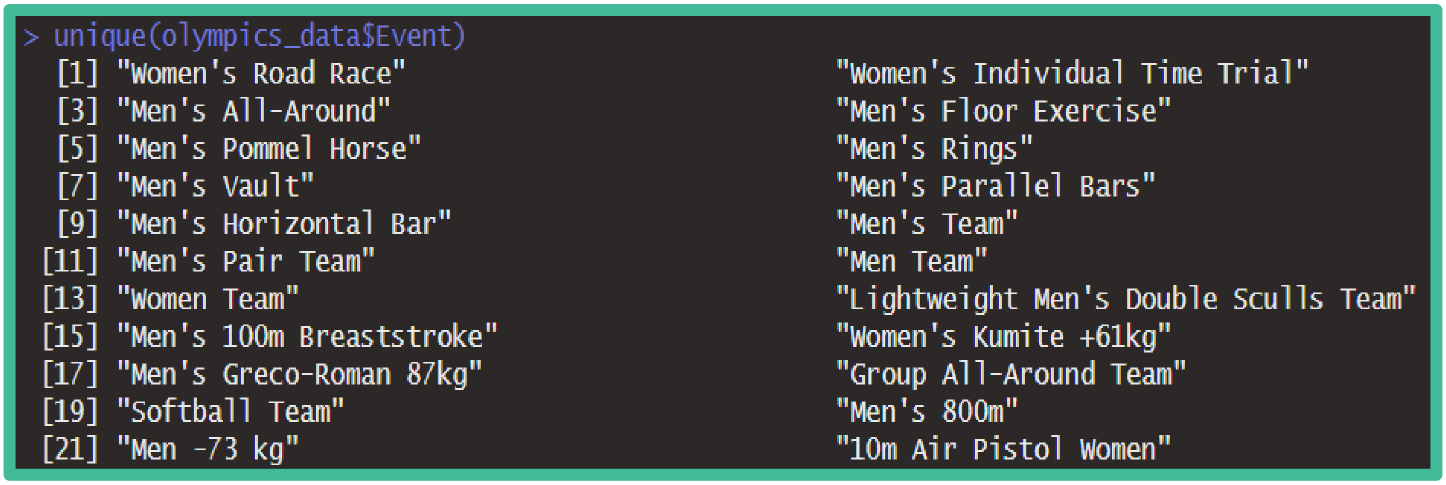
```
> # Load your datasets
> olympics_data <- read_excel("Olympics_Tidy.xlsx")
> #Check column names olympics_dataset
> colnames(olympics_data)
 [1] "No"         "Code"       "Name"       "Gender"     "Age"        "NOC"        "Country"    "Continents" "Discipline" "Sport"
[11] "Event"      "Rank"       "Medal"
```

Finding unique events identified, certain terms are used for team games. for example, Team, Relay, Mixed Team, Doubles, Quadruple, Group. Therefore, added column that derives from Event as team and individual. It can be a broad analysis



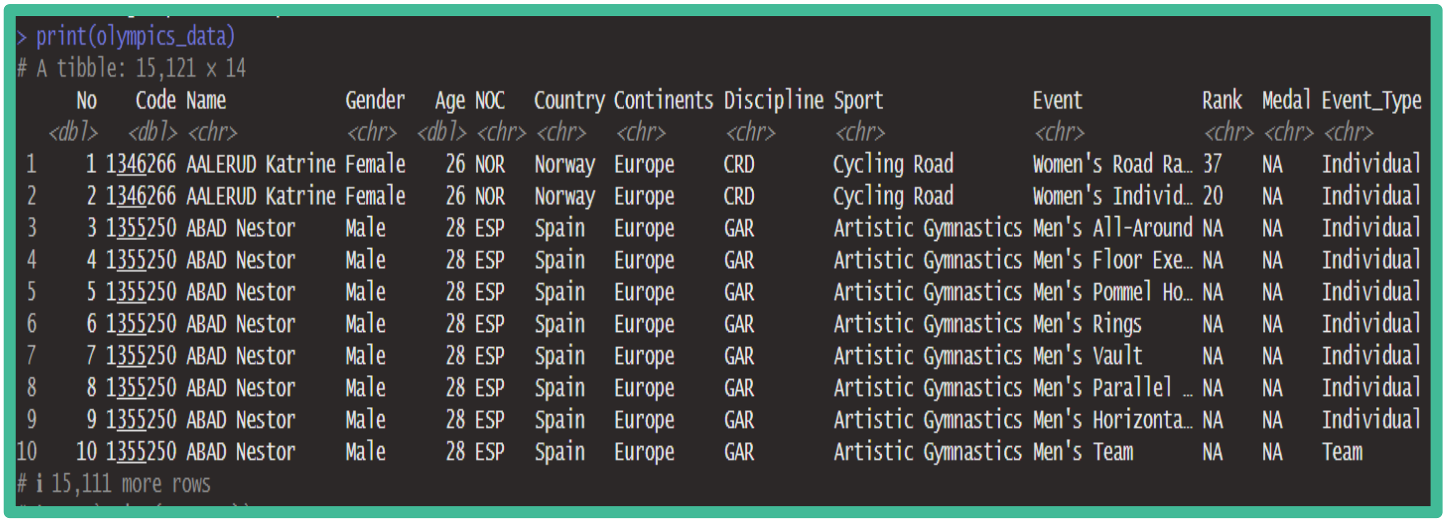
```
#Check event unique values
unique(olympics_data$Event)

#Categorize events as "Team" or "Individual"
olympics_data$Event_Type <- ifelse(grepl("Team|Relay|Mixed Team|Doubles|Quadruple|Group", olympics_data$Event),
                                   "Team", "Individual")

print(olympics_data)
```

Sample of unique values in events,



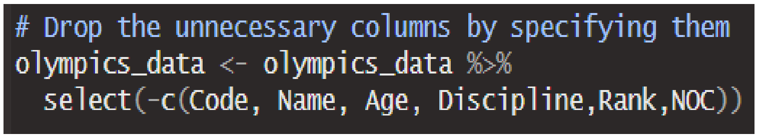
```
> unique(olympics_data$Event)
  [1] "Women's Road Race"                         "Women's Individual Time Trial"
  [3] "Men's All-Around"                          "Men's Floor Exercise"
  [5] "Men's Pommel Horse"                        "Men's Rings"
  [7] "Men's Vault"                               "Men's Parallel Bars"
  [9] "Men's Horizontal Bar"                      "Men's Team"
 [11] "Men's Pair Team"                           "Men Team"
 [13] "Women Team"                                "Lightweight Men's Double Sculls Team"
 [15] "Men's 100m Breaststroke"                   "Women's Kumite +61kg"
 [17] "Men's Greco-Roman 87kg"                    "Group All-Around Team"
 [19] "Softball Team"                             "Men's 800m"
 [21] "Men -73 kg"                                "10m Air Pistol Women"
```

After grouping team and individual



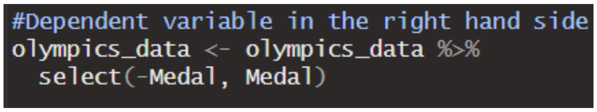
```
> print(olympics_data)
# A tibble: 15,121 x 14
      No    Code Name           Gender   Age NOC   Country Continents Discipline Sport             Event             Rank  Medal Event_Type
   <dbl>   <dbl> <chr>          <chr>  <dbl> <chr> <chr>   <chr>      <chr>      <chr>             <chr>             <chr> <chr> <chr>
 1     1 1346266 AALERUD Katrine Female   26 NOR   Norway  Europe     CRD        Cycling Road      Women's Road Ra… 37    NA    Individual
 2     2 1346266 AALERUD Katrine Female   26 NOR   Norway  Europe     CRD        Cycling Road      Women's Individ… 20    NA    Individual
 3     3 1355250 ABAD Nestor    Male      28 ESP   Spain   Europe     GAR        Artistic Gymnastics Men's All-Around NA    NA    Individual
 4     4 1355250 ABAD Nestor    Male      28 ESP   Spain   Europe     GAR        Artistic Gymnastics Men's Floor Exe… NA    NA    Individual
 5     5 1355250 ABAD Nestor    Male      28 ESP   Spain   Europe     GAR        Artistic Gymnastics Men's Pommel Ho… NA    NA    Individual
 6     6 1355250 ABAD Nestor    Male      28 ESP   Spain   Europe     GAR        Artistic Gymnastics Men's Rings      NA    NA    Individual
 7     7 1355250 ABAD Nestor    Male      28 ESP   Spain   Europe     GAR        Artistic Gymnastics Men's Vault      NA    NA    Individual
 8     8 1355250 ABAD Nestor    Male      28 ESP   Spain   Europe     GAR        Artistic Gymnastics Men's Parallel … NA    NA    Individual
 9     9 1355250 ABAD Nestor    Male      28 ESP   Spain   Europe     GAR        Artistic Gymnastics Men's Horizonta… NA    NA    Individual
10    10 1355250 ABAD Nestor    Male      28 ESP   Spain   Europe     GAR        Artistic Gymnastics Men's Team       NA    NA    Team
# ℹ 15,111 more rows
```

Selecting wise demographic and geographic variables and removed rest of them

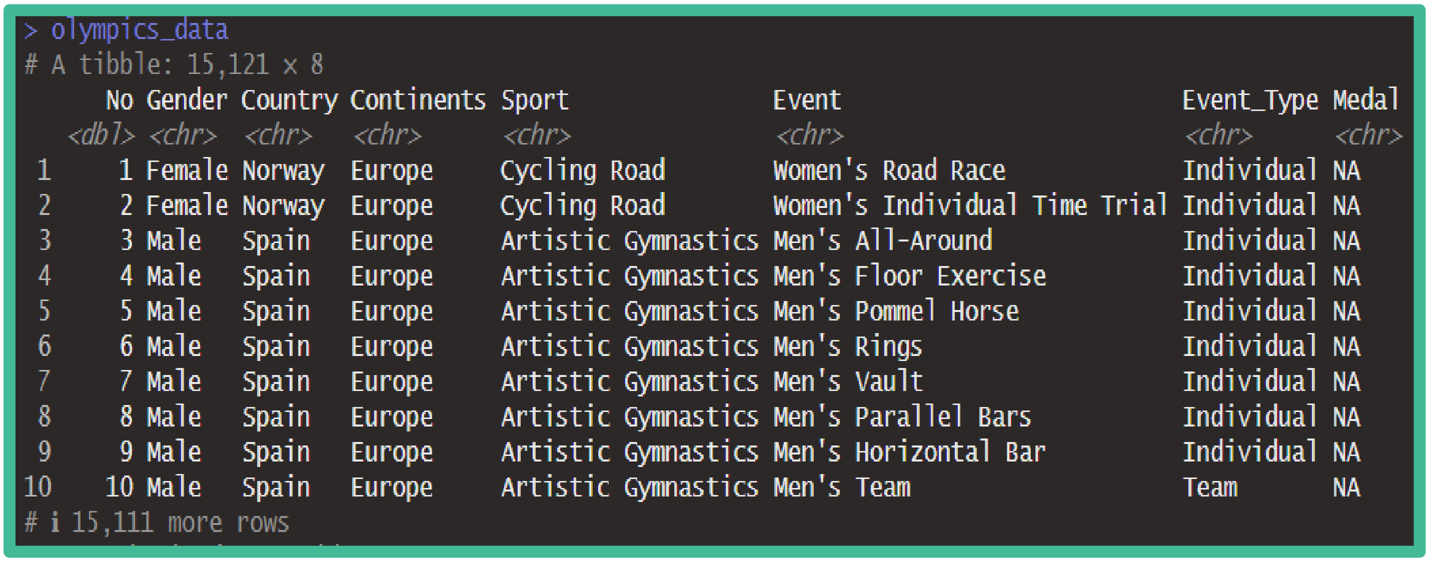


```
# Drop the unnecessary columns by specifying them
olympics_data <- olympics_data %>%
  select(-c(Code, Name, Age, Discipline,Rank,NOC))
```

Arranging the columns as per IVs and DVs where DV on the right-hand side

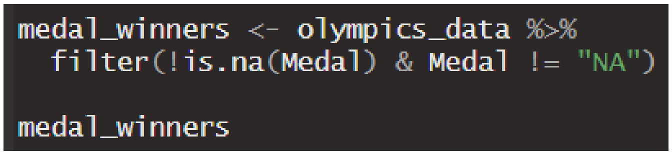


```
#Dependent variable in the right hand side
olympics_data <- olympics_data %>%
  select(-Medal, Medal)
```

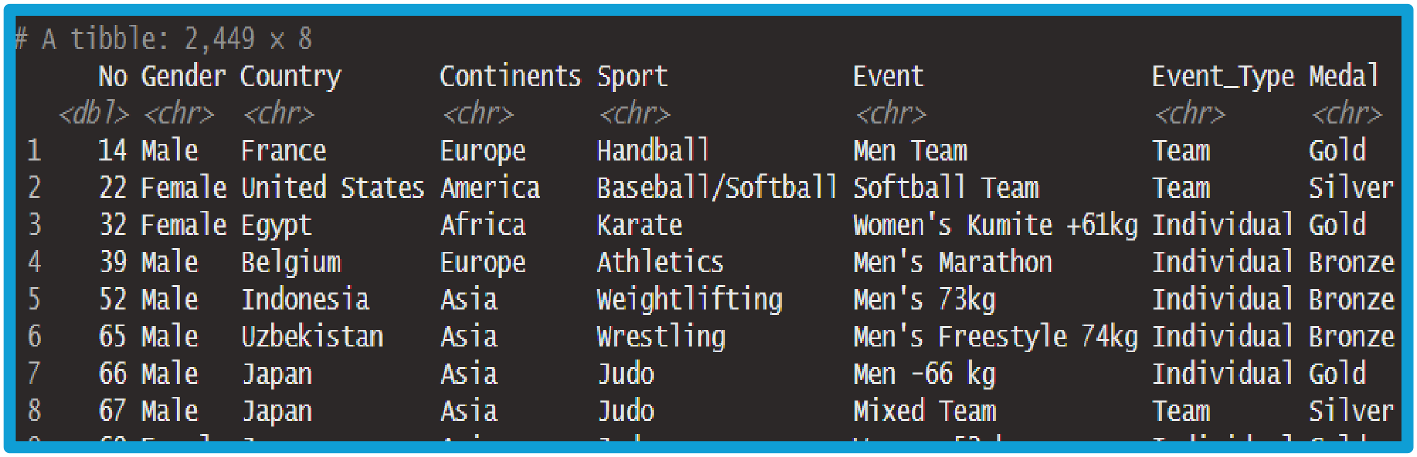


```
> olympics_data
# A tibble: 15,121 x 8
      No Gender Country Continents Sport               Event                        Event_Type Medal
   <dbl> <chr>  <chr>   <chr>      <chr>               <chr>                        <chr>      <chr>
 1     1 Female Norway  Europe     Cycling Road        Women's Road Race            Individual NA
 2     2 Female Norway  Europe     Cycling Road        Women's Individual Time Trial Individual NA
 3     3 Male   Spain   Europe     Artistic Gymnastics Men's All-Around             Individual NA
 4     4 Male   Spain   Europe     Artistic Gymnastics Men's Floor Exercise         Individual NA
 5     5 Male   Spain   Europe     Artistic Gymnastics Men's Pommel Horse           Individual NA
 6     6 Male   Spain   Europe     Artistic Gymnastics Men's Rings                  Individual NA
 7     7 Male   Spain   Europe     Artistic Gymnastics Men's Vault                  Individual NA
 8     8 Male   Spain   Europe     Artistic Gymnastics Men's Parallel Bars          Individual NA
 9     9 Male   Spain   Europe     Artistic Gymnastics Men's Horizontal Bar         Individual NA
10    10 Male   Spain   Europe     Artistic Gymnastics Men's Team                   Team       NA
# ℹ 15,111 more rows
```

Filter NA –Not applicable, participants did not receive any medals at all, been removed.

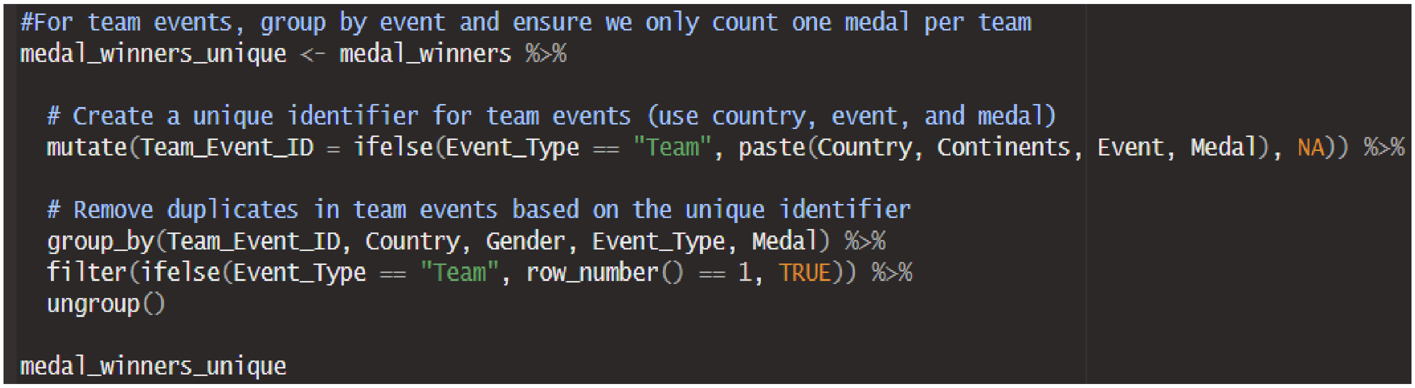


```
medal_winners <- olympics_data %>%
  filter(!is.na(Medal) & Medal != "NA")

medal_winners
```

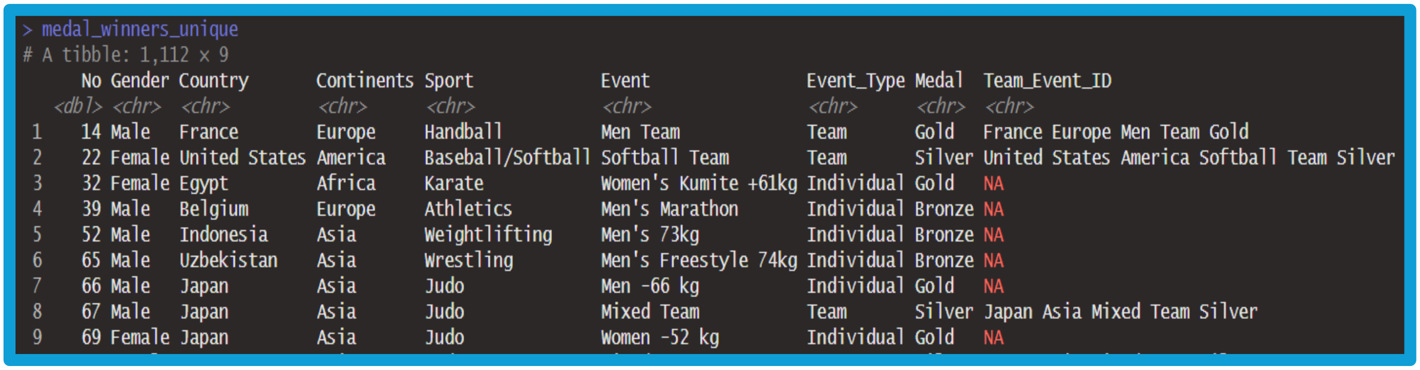


```
# A tibble: 2,449 x 8
     No Gender Country       Continents Sport             Event              Event_Type Medal
  <dbl> <chr>  <chr>         <chr>      <chr>             <chr>              <chr>      <chr>
1    14 Male   France        Europe     Handball          Men Team           Team       Gold
2    22 Female United States America    Baseball/Softball Softball Team      Team       Silver
3    32 Female Egypt         Africa     Karate            Women's Kumite +61kg Individual Gold
4    39 Male   Belgium       Europe     Athletics         Men's Marathon     Individual Bronze
5    52 Male   Indonesia     Asia       Weightlifting     Men's 73kg         Individual Bronze
6    65 Male   Uzbekistan    Asia       Wrestling         Men's Freestyle 74kg Individual Bronze
7    66 Male   Japan         Asia       Judo              Men -66 kg         Individual Gold
8    67 Male   Japan         Asia       Judo              Mixed Team         Team       Silver
```

group athletes per Gender, Country, Continents, Event_Type and Medal type victory counts to create unique observations.

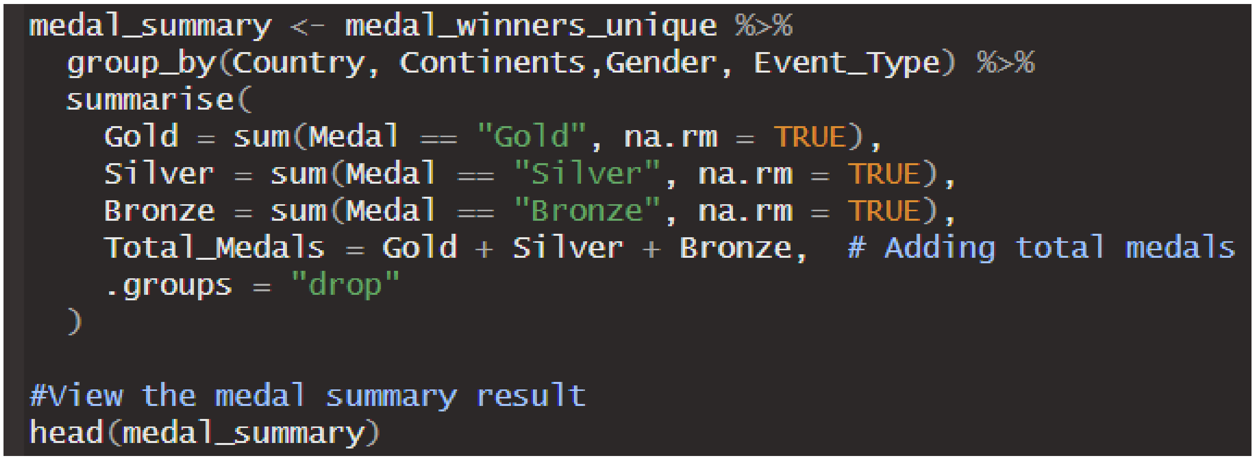


```
#For team events, group by event and ensure we only count one medal per team
medal_winners_unique <- medal_winners %>%

  # Create a unique identifier for team events (use country, event, and medal)
  mutate(Team_Event_ID = ifelse(Event_Type == "Team", paste(Country, Continents, Event, Medal), NA)) %>%

  # Remove duplicates in team events based on the unique identifier
  group_by(Team_Event_ID, Country, Gender, Event_Type, Medal) %>%
  filter(ifelse(Event_Type == "Team", row_number() == 1, TRUE)) %>%
  ungroup()

medal_winners_unique
```

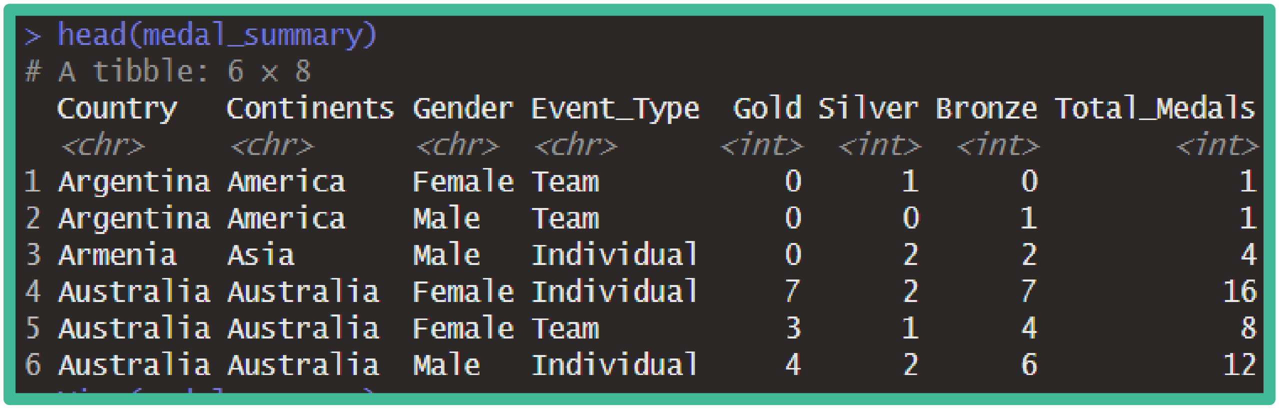


```
> medal_winners_unique
# A tibble: 1,112 x 9
     No Gender Country       Continents Sport             Event                Event_Type Medal  Team_Event_ID
  <dbl> <chr>  <chr>         <chr>      <chr>             <chr>                <chr>      <chr>  <chr>
1    14 Male   France        Europe     Handball          Men Team             Team       Gold   France Europe Men Team Gold
2    22 Female United States America    Baseball/Softball Softball Team        Team       Silver United States America Softball Team Silver
3    32 Female Egypt         Africa     Karate            Women's Kumite +61kg Individual Gold   NA
4    39 Male   Belgium       Europe     Athletics         Men's Marathon       Individual Bronze NA
5    52 Male   Indonesia     Asia       Weightlifting     Men's 73kg           Individual Bronze NA
6    65 Male   Uzbekistan    Asia       Wrestling         Men's Freestyle 74kg Individual Bronze NA
7    66 Male   Japan         Asia       Judo              Men -66 kg           Individual Gold   NA
8    67 Male   Japan         Asia       Judo              Mixed Team           Team       Silver Japan Asia Mixed Team Silver
9    69 Female Japan         Asia       Judo              Women -52 kg         Individual Gold   NA
```

Widening dataset by bringing Gold, Silver and Bronze as separate variables.

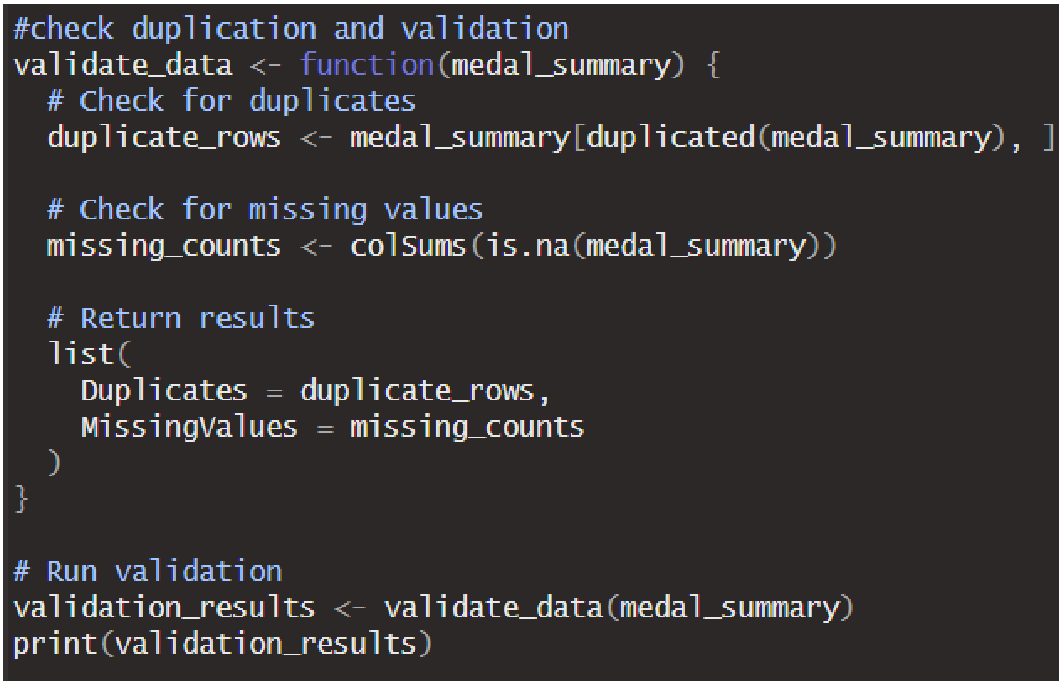


```
medal_summary <- medal_winners_unique %>%
  group_by(Country, Continents,Gender, Event_Type) %>%
  summarise(
    Gold = sum(Medal == "Gold", na.rm = TRUE),
    Silver = sum(Medal == "Silver", na.rm = TRUE),
    Bronze = sum(Medal == "Bronze", na.rm = TRUE),
    Total_Medals = Gold + Silver + Bronze,  # Adding total medals
    .groups = "drop"
  )

#View the medal summary result
head(medal_summary)
```



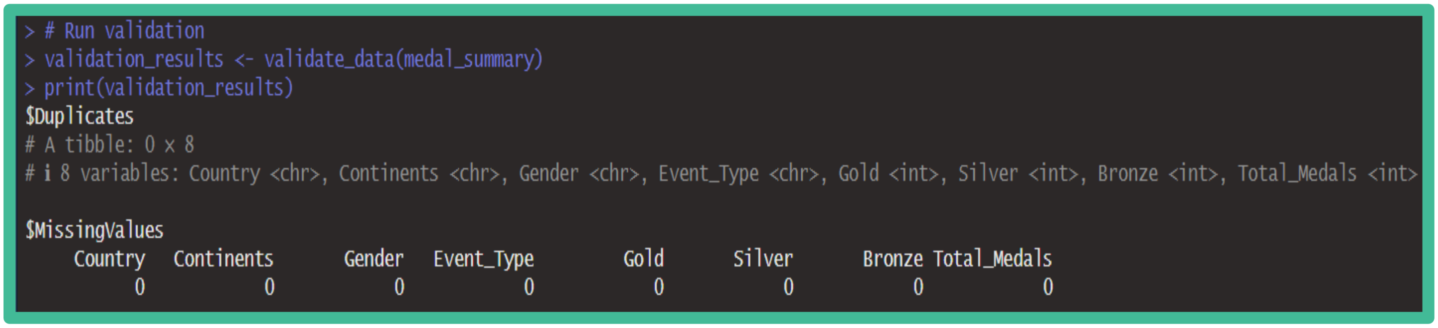
```
> head(medal_summary)
# A tibble: 6 × 8
  Country   Continents Gender Event_Type  Gold Silver Bronze Total_Medals
  <chr>     <chr>      <chr>  <chr>      <int>  <int>  <int>        <int>
1 Argentina America    Female Team           0      1      0            1
2 Argentina America    Male   Team           0      0      1            1
3 Armenia   Asia       Male   Individual     0      2      2            4
4 Australia Australia  Female Individual     7      2      7           16
5 Australia Australia  Female Team           3      1      4            8
6 Australia Australia  Male   Individual     4      2      6           12
```

Check duplication and validation



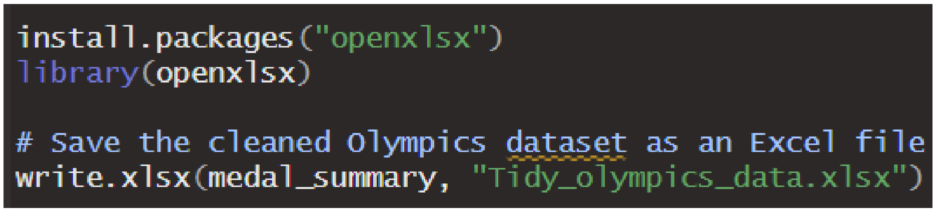
```
#check duplication and validation
validate_data <- function(medal_summary) {
  # Check for duplicates
  duplicate_rows <- medal_summary[duplicated(medal_summary), ]

  # Check for missing values
  missing_counts <- colSums(is.na(medal_summary))

  # Return results
  list(
    Duplicates = duplicate_rows,
    MissingValues = missing_counts
  )
}

# Run validation
validation_results <- validate_data(medal_summary)
print(validation_results)
```

No duplications found



```
> # Run validation
> validation_results <- validate_data(medal_summary)
> print(validation_results)
$Duplicates
# A tibble: 0 x 8
# i 8 variables: Country <chr>, Continents <chr>, Gender <chr>, Event_Type <chr>, Gold <int>, Silver <int>, Bronze <int>, Total_Medals <int>

$MissingValues
     Country   Continents       Gender   Event_Type         Gold       Silver       Bronze Total_Medals
           0            0            0            0            0            0            0            0
```

Converted to excel,



```
install.packages("openxlsx")
library(openxlsx)

# Save the cleaned Olympics dataset as an Excel file
write.xlsx(medal_summary, "Tidy_olympics_data.xlsx")
```

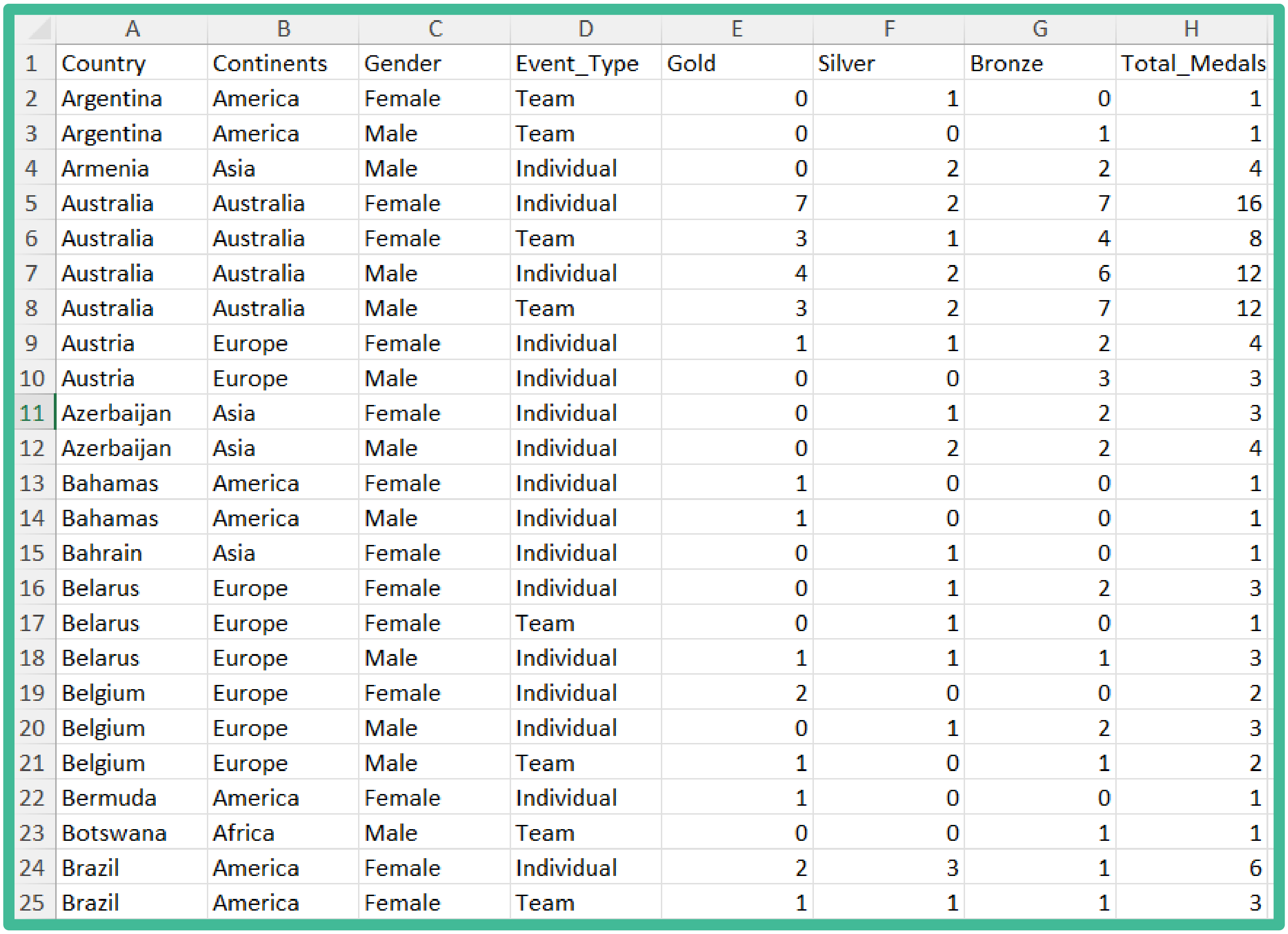
- ○ Link to R script - Transform_Data_Tidy.R
- ○ Link to the Final cleaned dataset - Final_olympics_data.xlsx (226 observations)

Variables	Olympics
Independent variable	Country, Continents, Gender, Event_Type , Gold, Silver, Bronze
Dependent variable	Total_Medals

Variables And Justification

I. Country	Identifies, nations the athletes represent.
II. Continents	Grouping by Five ring continents enables a broader spatial analysis.
III. Gender	Measure the number of winners based on gender-based analysis.
IV. Event_Type	Differentiated between individuals and group events.
V. Gold	Number of gold winners per country, Continents, Gender and Event_Type
VI. Silver	Number of silver winners per country, Continents, Gender and Event_Type
VII. Bronze	Number of bronze winners per country, Continents, Gender and Event_Type
VIII. Total_Medals	Gold, silver and bronze medals received participants counts.

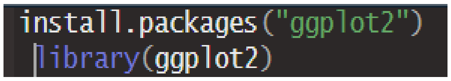
Output



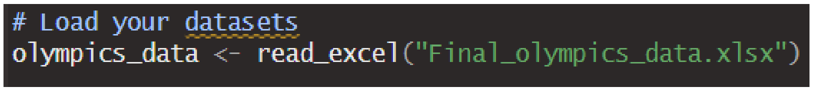
	A	B	C	D	E	F	G	H
1	Country	Continents	Gender	Event_Type	Gold	Silver	Bronze	Total_Medals
2	Argentina	America	Female	Team	0	1	0	1
3	Argentina	America	Male	Team	0	0	1	1
4	Armenia	Asia	Male	Individual	0	2	2	4
5	Australia	Australia	Female	Individual	7	2	7	16
6	Australia	Australia	Female	Team	3	1	4	8
7	Australia	Australia	Male	Individual	4	2	6	12
8	Australia	Australia	Male	Team	3	2	7	12
9	Austria	Europe	Female	Individual	1	1	2	4
10	Austria	Europe	Male	Individual	0	0	3	3
11	Azerbaijan	Asia	Female	Individual	0	1	2	3
12	Azerbaijan	Asia	Male	Individual	0	2	2	4
13	Bahamas	America	Female	Individual	1	0	0	1
14	Bahamas	America	Male	Individual	1	0	0	1
15	Bahrain	Asia	Female	Individual	0	1	0	1
16	Belarus	Europe	Female	Individual	0	1	2	3
17	Belarus	Europe	Female	Team	0	1	0	1
18	Belarus	Europe	Male	Individual	1	1	1	3
19	Belgium	Europe	Female	Individual	2	0	0	2
20	Belgium	Europe	Male	Individual	0	1	2	3
21	Belgium	Europe	Male	Team	1	0	1	2
22	Bermuda	America	Female	Individual	1	0	0	1
23	Botswana	Africa	Male	Team	0	0	1	1
24	Brazil	America	Female	Individual	2	3	1	6
25	Brazil	America	Female	Team	1	1	1	3

Exploratory Data Analysis

R-script for EDA = EDA.R

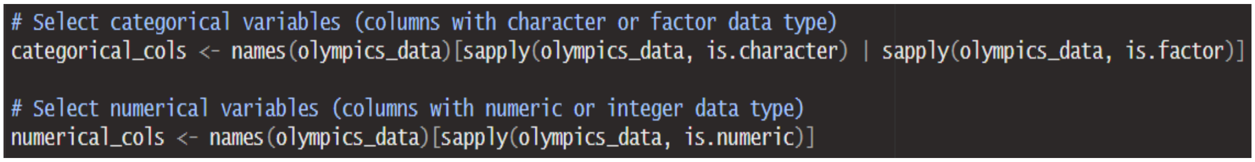


```
install.packages("ggplot2")
 library(ggplot2)
```

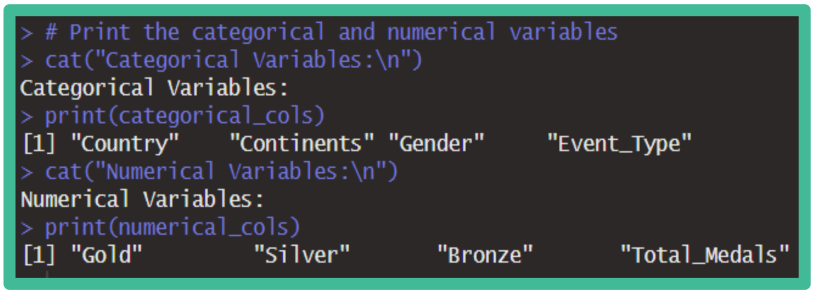


```
# Load your datasets
olympics_data <- read_excel("Final_olympics_data.xlsx")
```

Categorical and Numerical variables

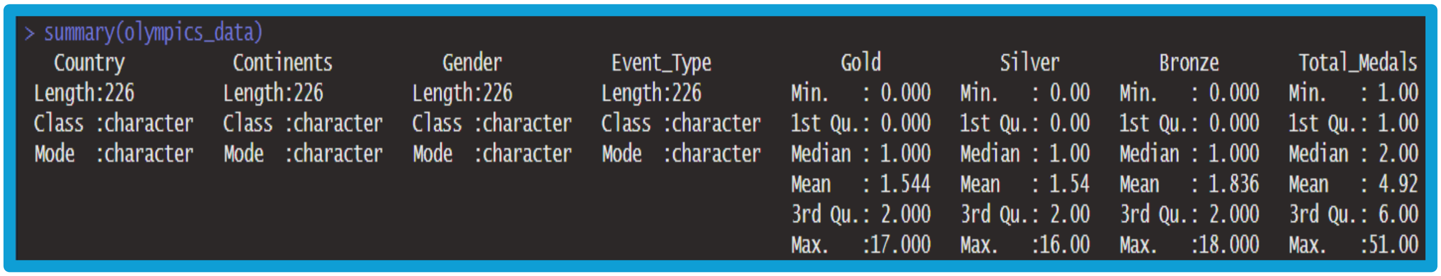


```
# Select categorical variables (columns with character or factor data type)
categorical_cols <- names(olympics_data)[sapply(olympics_data, is.character) | sapply(olympics_data, is.factor)]

# Select numerical variables (columns with numeric or integer data type)
numerical_cols <- names(olympics_data)[sapply(olympics_data, is.numeric)]
```

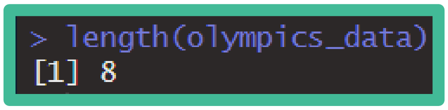


```
> # Print the categorical and numerical variables
> cat("Categorical Variables:\n")
Categorical Variables:
> print(categorical_cols)
[1] "Country"    "Continents" "Gender"     "Event_Type"
> cat("Numerical Variables:\n")
Numerical Variables:
> print(numerical_cols)
[1] "Gold"          "Silver"       "Bronze"       "Total_Medals"
```

Descriptive Statistics



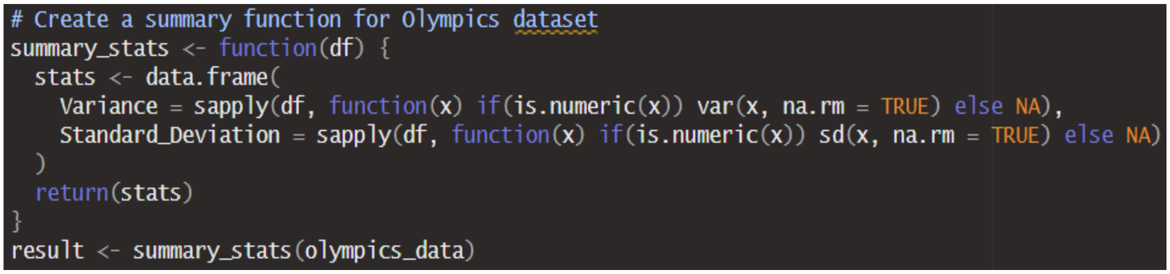
```
> summary(olympics_data)
   Country           Continents          Gender           Event_Type             Gold            Silver          Bronze          Total_Medals
 Length:226         Length:226         Length:226         Length:226         Min.   : 0.000   Min.   : 0.00   Min.   : 0.000   Min.   : 1.00
 Class :character   Class :character   Class :character   Class :character   1st Qu.: 0.000   1st Qu.: 0.00   1st Qu.: 0.000   1st Qu.: 1.00
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median : 1.000   Median : 1.00   Median : 1.000   Median : 2.00
                                                                             Mean   : 1.544   Mean   : 1.54   Mean   : 1.836   Mean   : 4.92
                                                                             3rd Qu.: 2.000   3rd Qu.: 2.00   3rd Qu.: 2.000   3rd Qu.: 6.00
                                                                             Max.   :17.000   Max.   :16.00   Max.   :18.000   Max.   :51.00
```



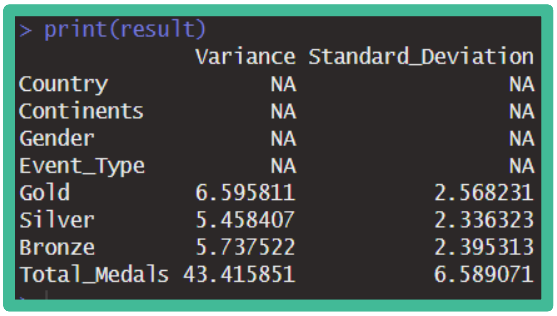
```
> length(olympics_data)
[1] 8
```

- There are Categorical data such as Country, Continents, Gender and Event_Type. Whereas Gold, Silver, Bronze and Total_Medals are Numerical Data that contains Minimum, Maximum and Quartile values. Summary shows 226 observations and 8 variables with errorless values.

Variance and Standard deviation

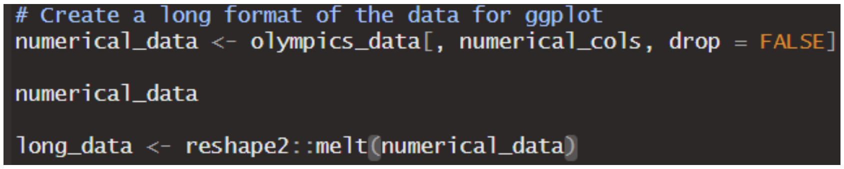


```r
# Create a summary function for Olympics dataset
summary_stats <- function(df) {
  stats <- data.frame(
    Variance = sapply(df, function(x) if(is.numeric(x)) var(x, na.rm = TRUE) else NA),
    Standard_Deviation = sapply(df, function(x) if(is.numeric(x)) sd(x, na.rm = TRUE) else NA)
  )
  return(stats)
}
result <- summary_stats(olympics_data)
```



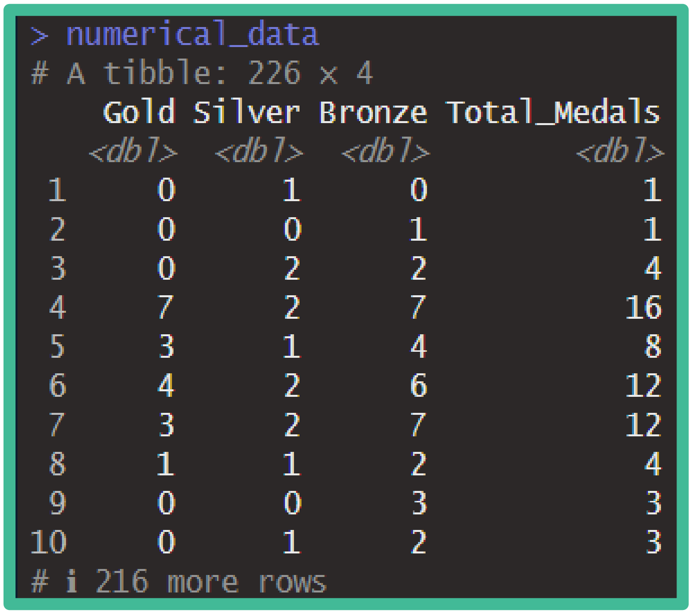
```
> print(result)
               Variance Standard_Deviation
Country              NA                 NA
Continents           NA                 NA
Gender               NA                 NA
Event_Type           NA                 NA
Gold           6.595811           2.568231
Silver         5.458407           2.336323
Bronze         5.737522           2.395313
Total_Medals  43.415851           6.589071
```

- NAs' represents Categorical data. Except for Total_Medals, other variables have low variance and Standard Deviation which means datapoints are close to mean.



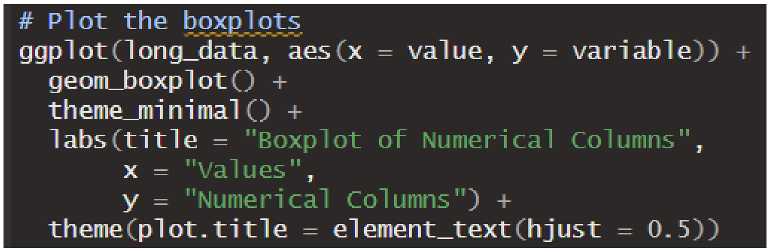
```r
# Create a long format of the data for ggplot
numerical_data <- olympics_data[, numerical_cols, drop = FALSE]

numerical_data

long_data <- reshape2::melt(numerical_data)
```

- Extacts numerical columns and bring wide format for visualisations

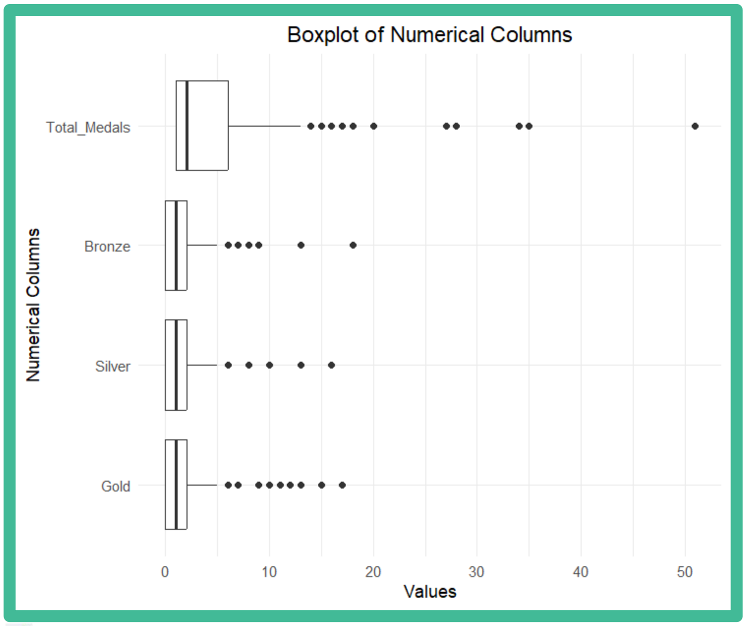


```
> numerical_data
# A tibble: 226 × 4
    Gold Silver Bronze Total_Medals
   <dbl>  <dbl>  <dbl>        <dbl>
 1     0      1      0            1
 2     0      0      1            1
 3     0      2      2            4
 4     7      2      7           16
 5     3      1      4            8
 6     4      2      6           12
 7     3      2      7           12
 8     1      1      2            4
 9     0      0      3            3
10     0      1      2            3
# ℹ 216 more rows
```

Boxplot of Numerical variables

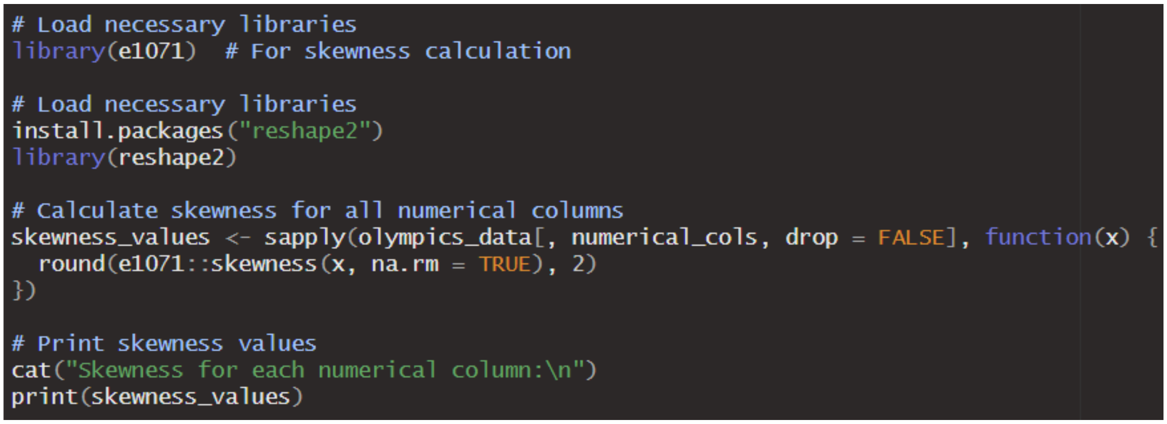


```
# Plot the boxplots
ggplot(long_data, aes(x = value, y = variable)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Boxplot of Numerical Columns",
       x = "Values",
       y = "Numerical Columns") +
  theme(plot.title = element_text(hjust = 0.5))
```

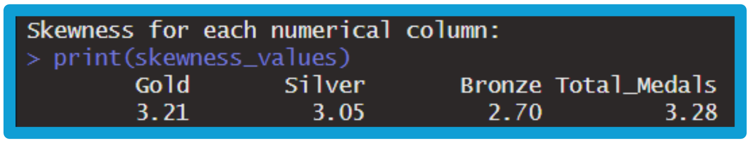


- Descriptive statistics of numerical columns are presented in box plot.

Skewness



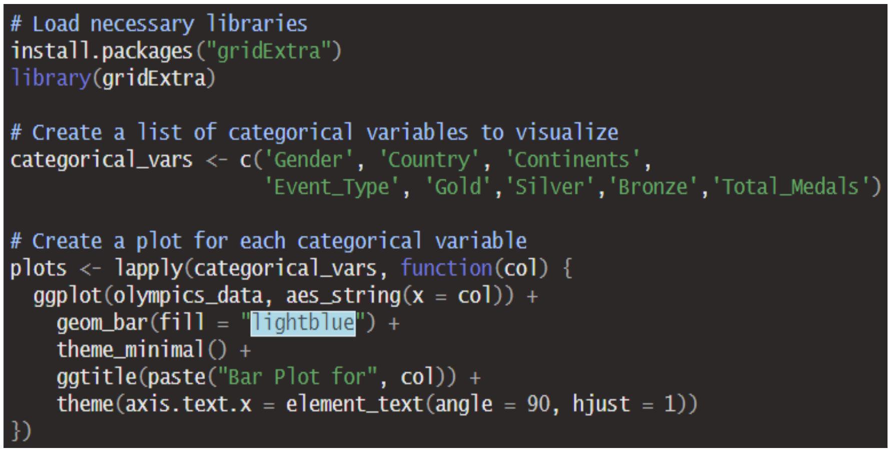
```
# Load necessary libraries
library(e1071)  # For skewness calculation

# Load necessary libraries
install.packages("reshape2")
library(reshape2)

# Calculate skewness for all numerical columns
skewness_values <- sapply(olympics_data[, numerical_cols, drop = FALSE], function(x) {
  round(e1071::skewness(x, na.rm = TRUE), 2)
})

# Print skewness values
cat("Skewness for each numerical column:\n")
print(skewness_values)
```



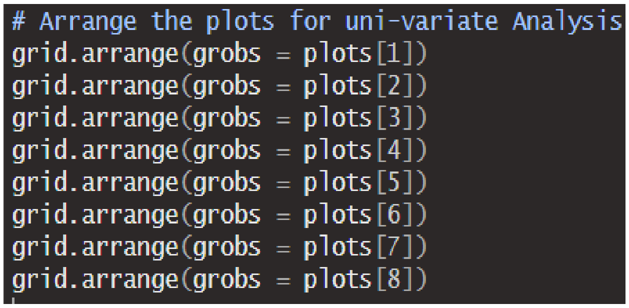
```
Skewness for each numerical column:
> print(skewness_values)
        Gold       Silver       Bronze Total_Medals
        3.21         3.05         2.70         3.28
```

- Since all numerical columns are greater than 0, Positive skewed.

Univariate Analysis

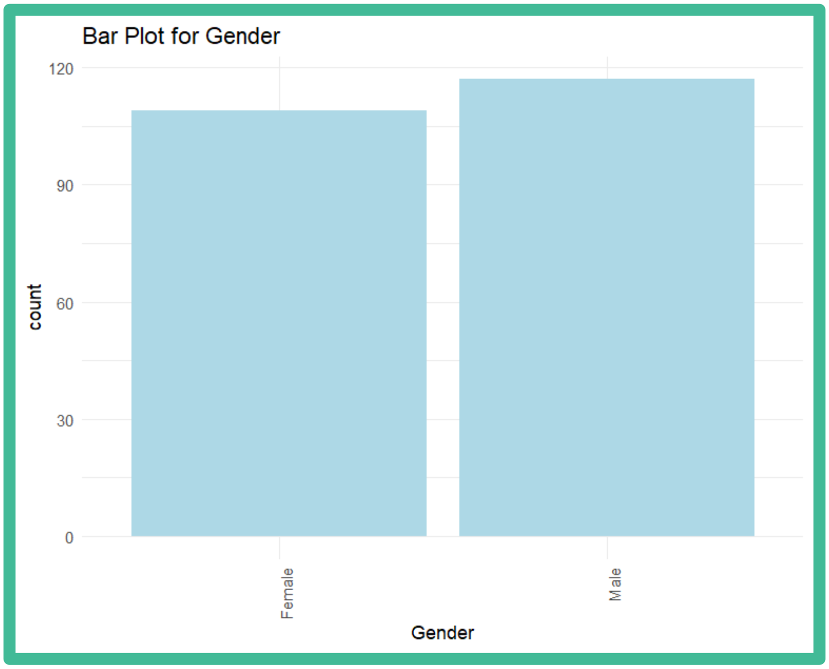


```
# Load necessary libraries
install.packages("gridExtra")
library(gridExtra)

# Create a list of categorical variables to visualize
categorical_vars <- c('Gender', 'Country', 'Continents',
                      'Event_Type', 'Gold','Silver','Bronze','Total_Medals')

# Create a plot for each categorical variable
plots <- lapply(categorical_vars, function(col) {
  ggplot(olympics_data, aes_string(x = col)) +
    geom_bar(fill = "lightblue") +
    theme_minimal() +
    ggtitle(paste("Bar Plot for", col)) +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
})
```



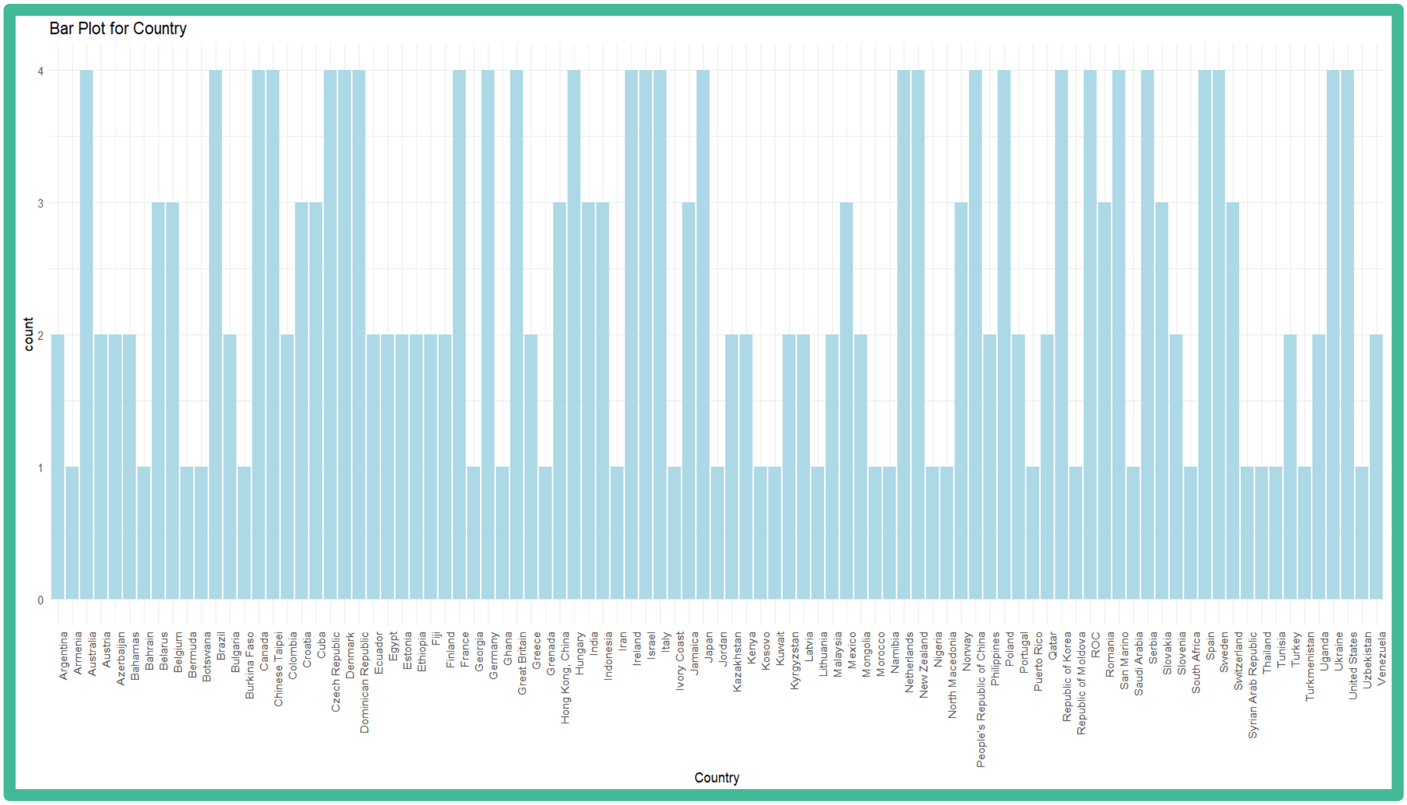
```
# Arrange the plots for uni-variate Analysis
grid.arrange(grobs = plots[1])
grid.arrange(grobs = plots[2])
grid.arrange(grobs = plots[3])
grid.arrange(grobs = plots[4])
grid.arrange(grobs = plots[5])
grid.arrange(grobs = plots[6])
grid.arrange(grobs = plots[7])
grid.arrange(grobs = plots[8])
```

I.



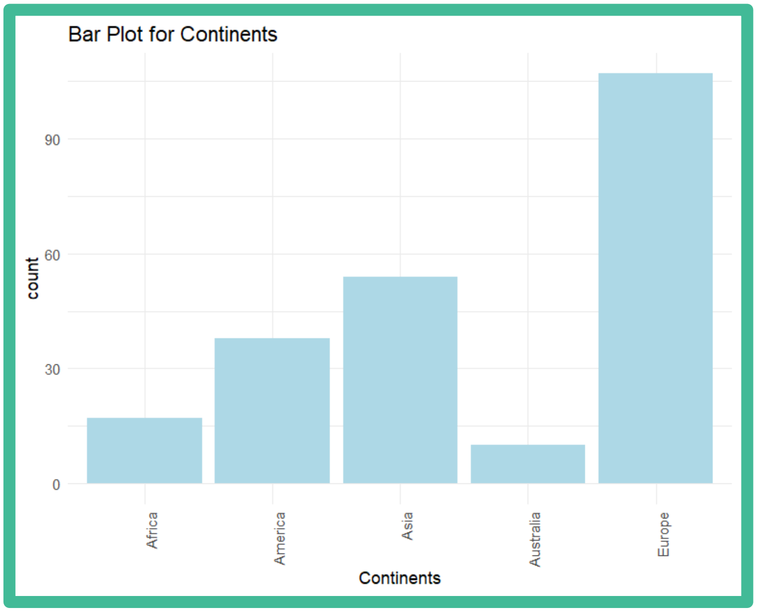
Revealing males slightly higher winning than females.

II.



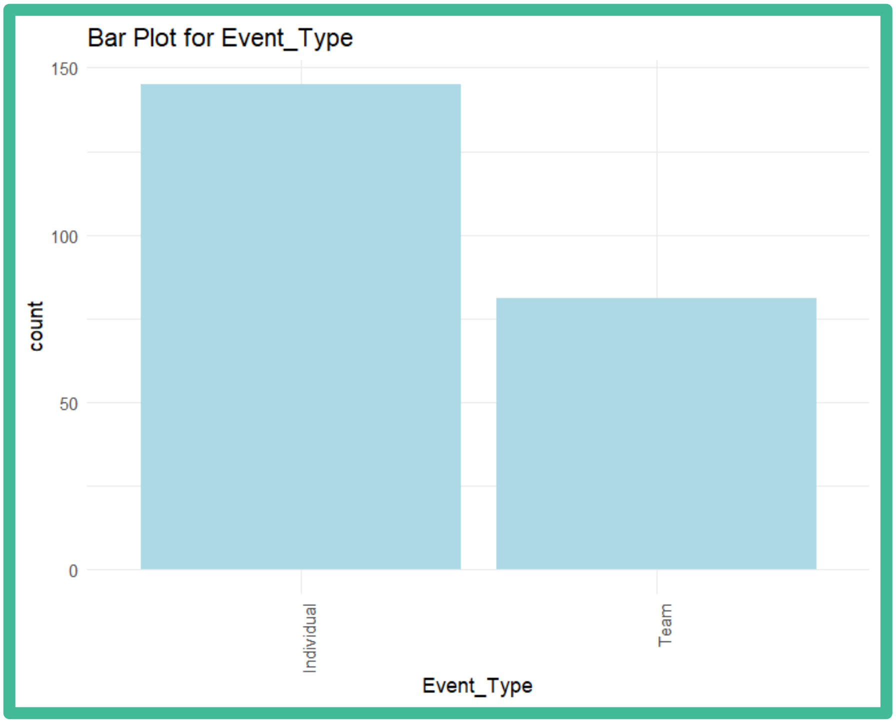
- Some countries not presented since those removed while data cleaning. All the countries that won medals are uneven distribution and counts represent the repetition occurred because of gender and Event_Type were broken-down further.

III.



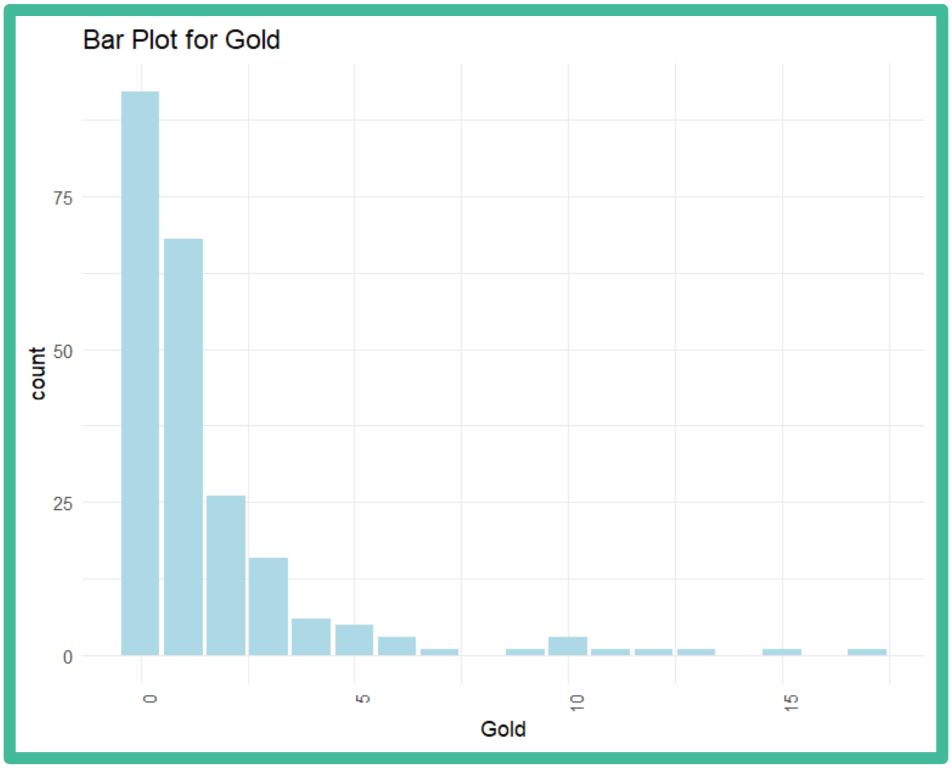
- Five ring continents are brought here, Olympics Refugees Team is completely removed and not taken into any continent.

IV.



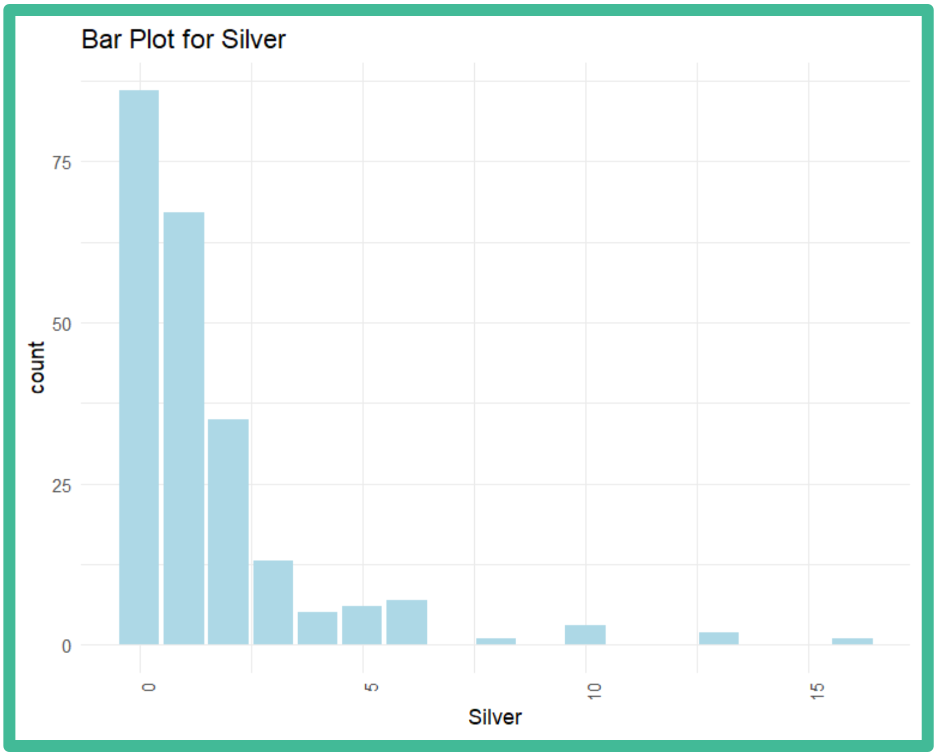
- Individual events are more frequent than team events, indicating a greater emphasis on individual performance.

V.



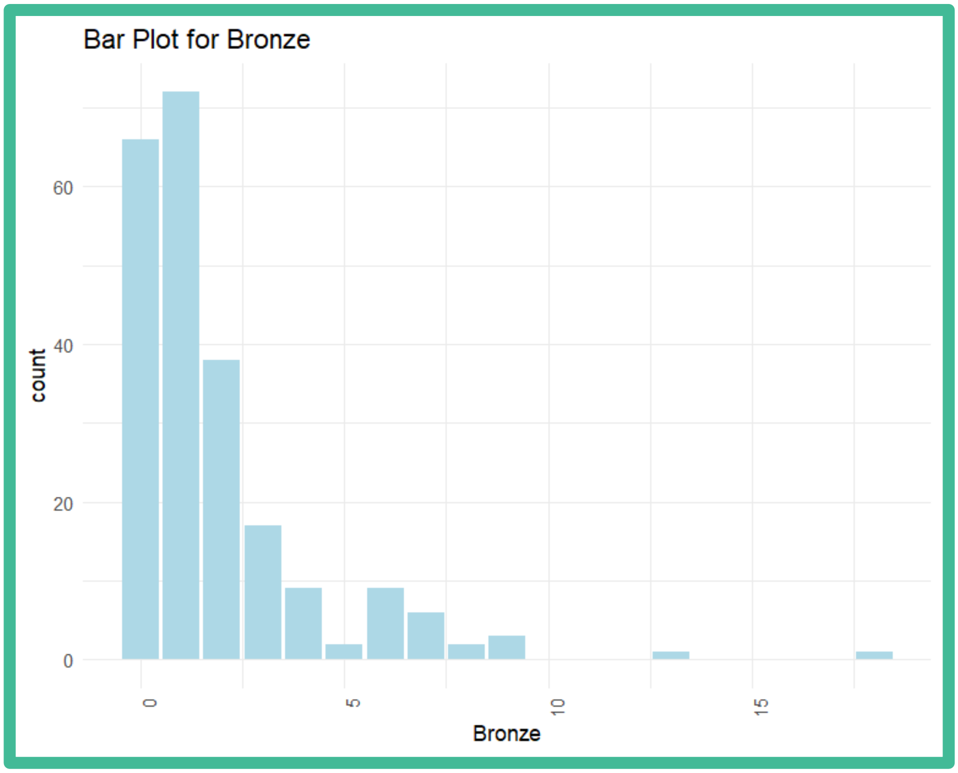
- After grouping dataset by gender, country and Event_Type, most of the observations contain 0s and for example, a very few rows taken greater than 10 gold medals.

VI.



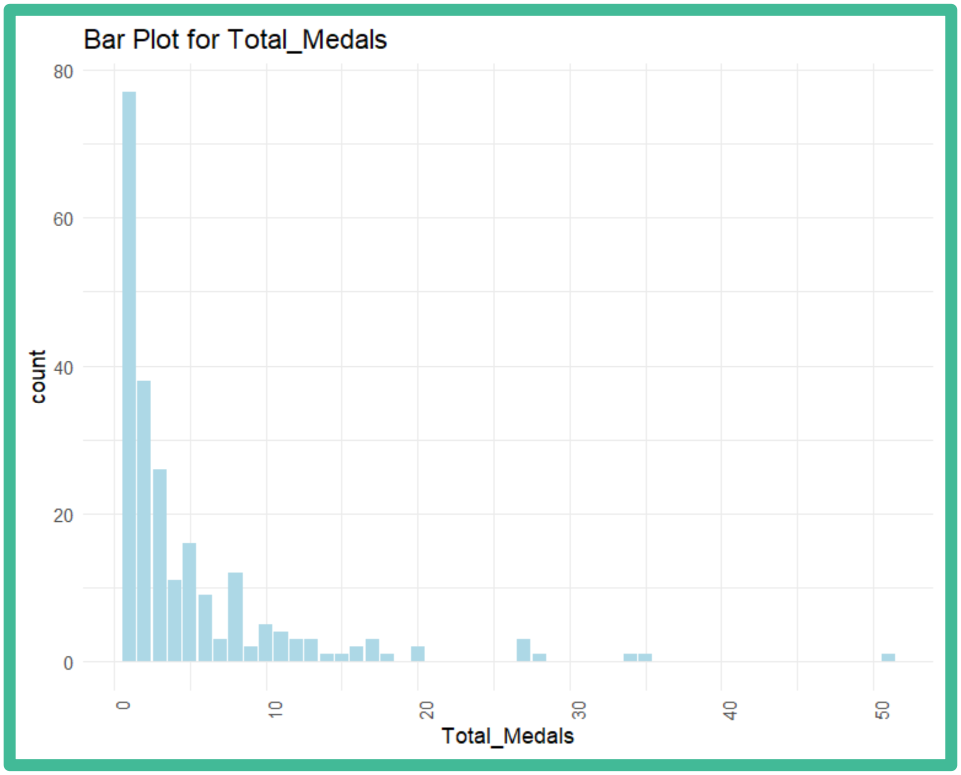
- After grouping dataset by gender, country and Event_Type, more than 75 observations contain 0s. for example, the rows that won 1 Silver medal is greater than 60 counts.

VII.



- 0s are lesser than 1s, the number of counts range is lower than other medal types.

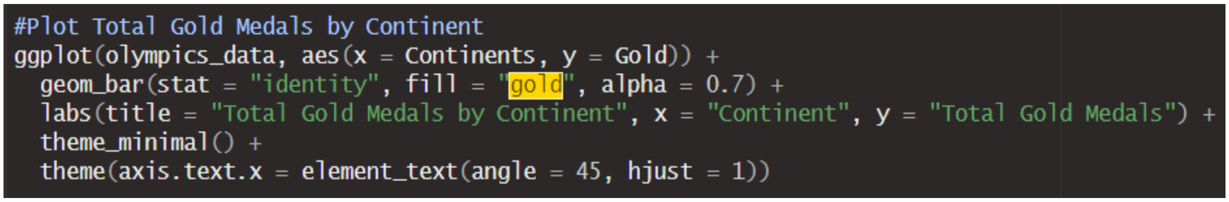
VIII.



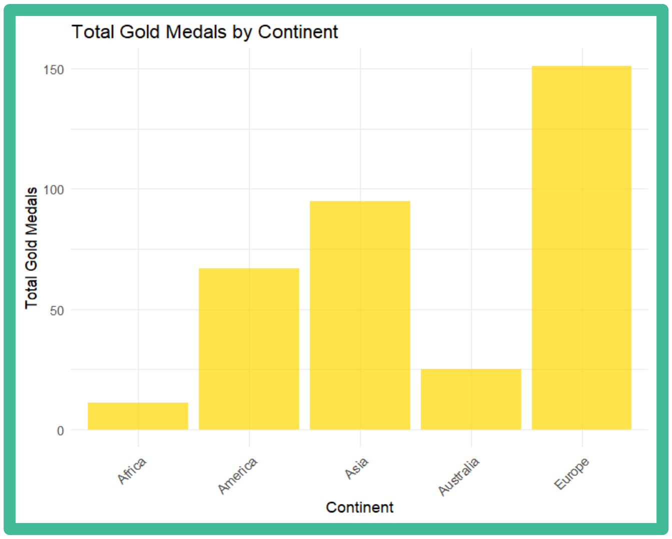
- Adding up all medal types there are no 0s and it represents many observations have 1 medal, because we widely separated columns by male/ female, Individual/ Team.

Bivariate Analysis

I.

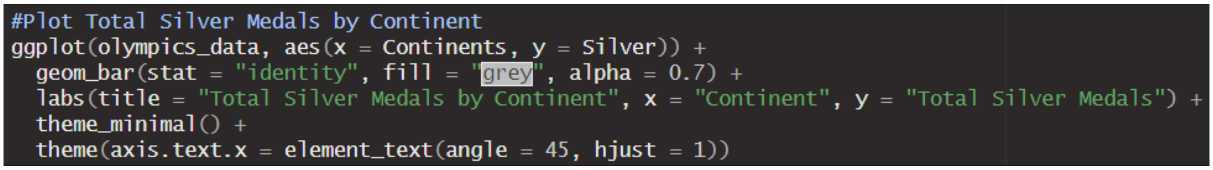


```
#Plot Total Gold Medals by Continent
ggplot(olympics_data, aes(x = Continents, y = Gold)) +
  geom_bar(stat = "identity", fill = "gold", alpha = 0.7) +
  labs(title = "Total Gold Medals by Continent", x = "Continent", y = "Total Gold Medals") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

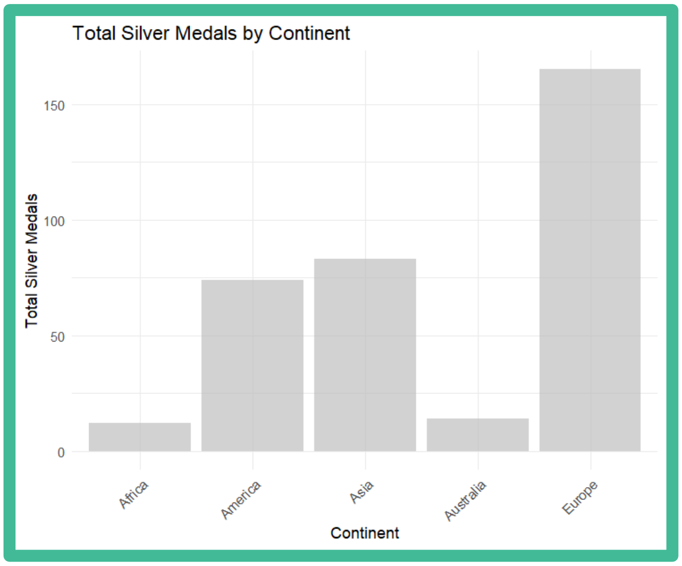


- Continents wise Europe has more, and Africa has less Gold medals

II.

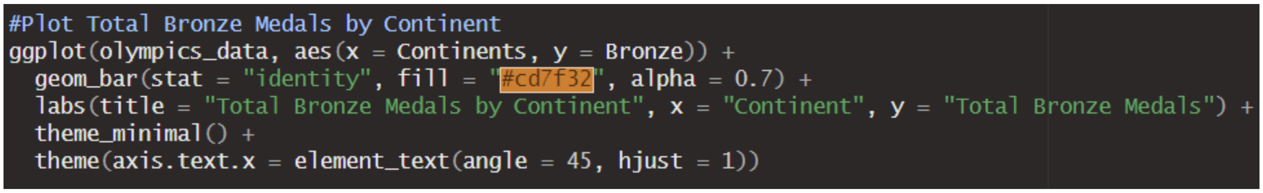


```
#Plot Total Silver Medals by Continent
ggplot(olympics_data, aes(x = Continents, y = Silver)) +
  geom_bar(stat = "identity", fill = "grey", alpha = 0.7) +
  labs(title = "Total Silver Medals by Continent", x = "Continent", y = "Total Silver Medals") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

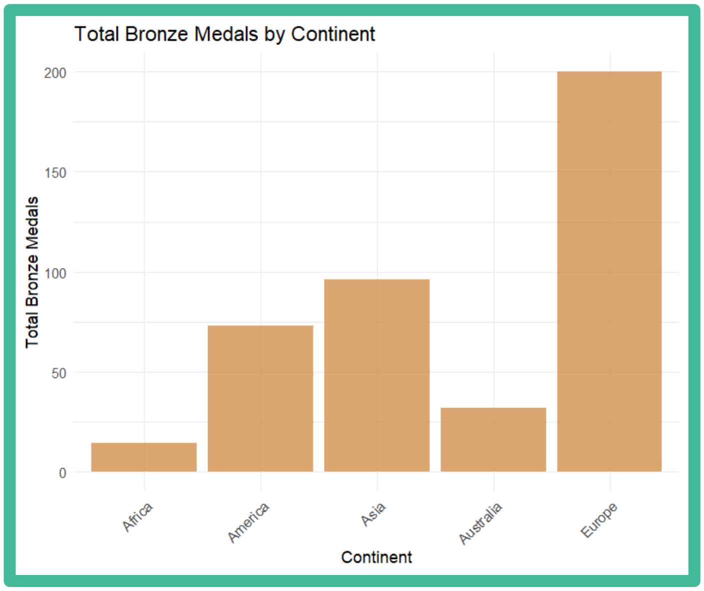


- Australia and Africa have nearly same value of Silver medals

III.

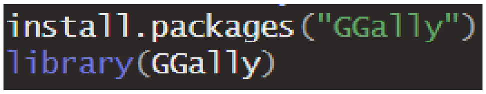


```
#Plot Total Bronze Medals by Continent
ggplot(olympics_data, aes(x = Continents, y = Bronze)) +
  geom_bar(stat = "identity", fill = "#cd7f32", alpha = 0.7) +
  labs(title = "Total Bronze Medals by Continent", x = "Continent", y = "Total Bronze Medals") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



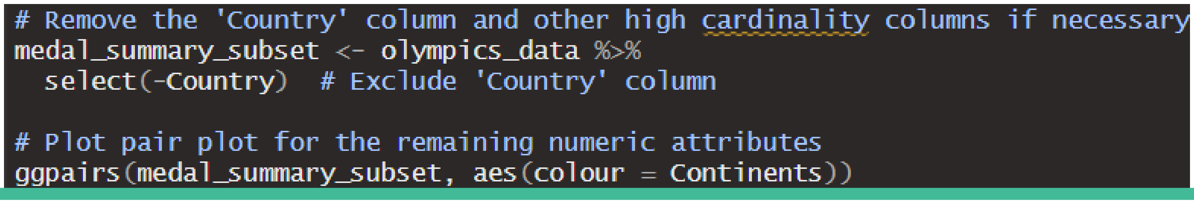
- Bronze comparatively has higher counts than other medal types.

Multivariate Analysis

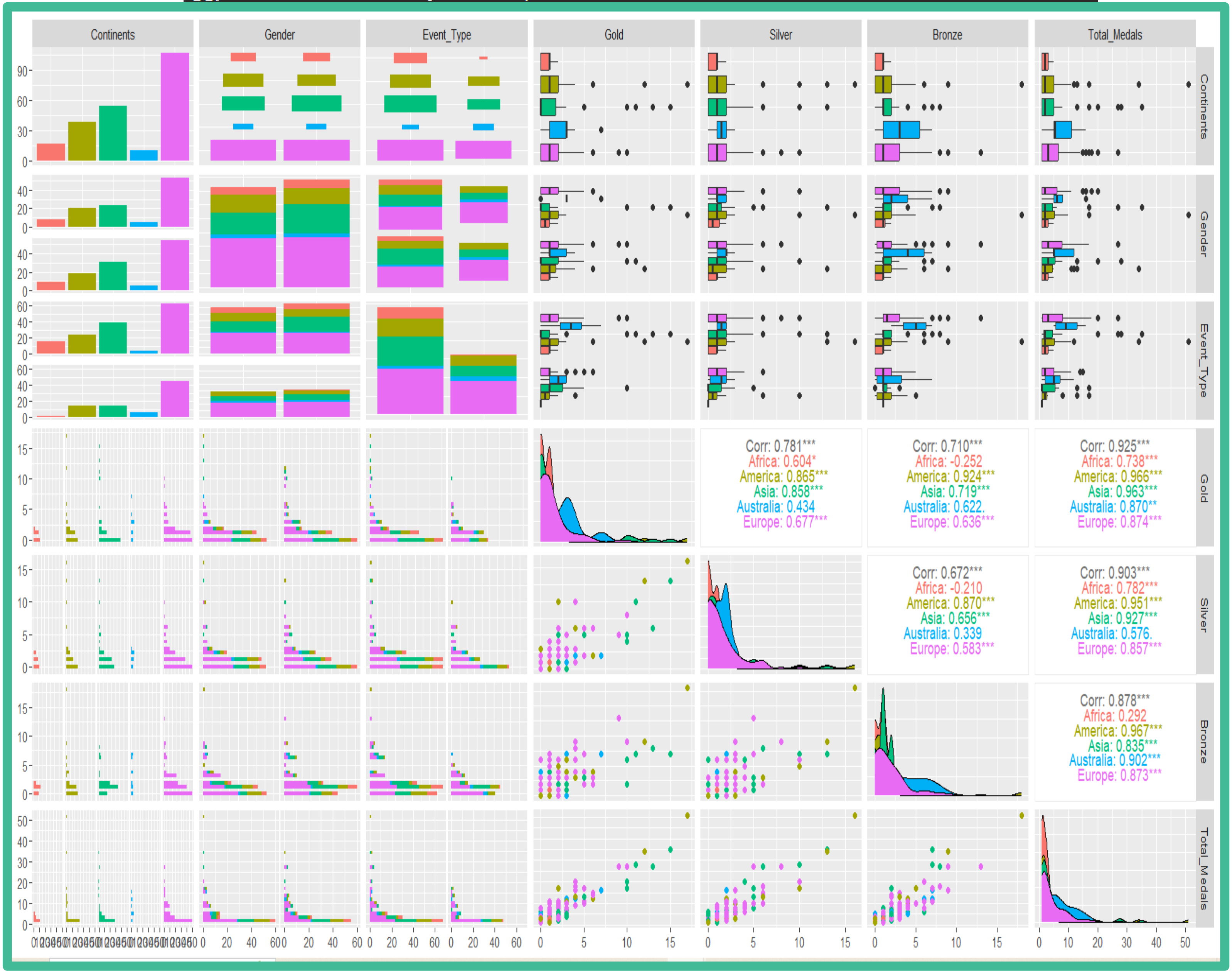


```
install.packages("GGally")
library(GGally)
```

I.

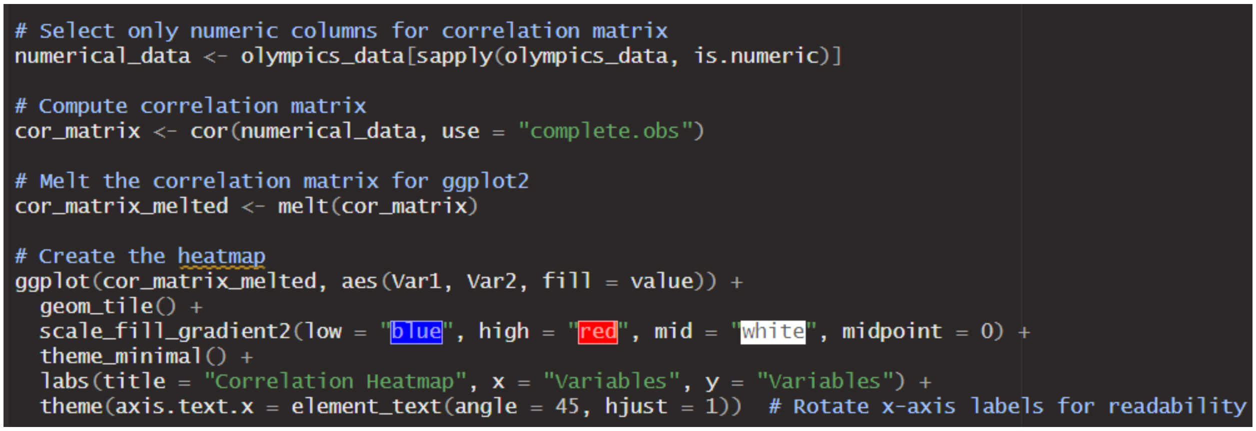


```
# Remove the 'Country' column and other high cardinality columns if necessary
medal_summary_subset <- olympics_data %>%
  select(-Country)  # Exclude 'Country' column

# Plot pair plot for the remaining numeric attributes
ggpairs(medal_summary_subset, aes(colour = Continents))
```

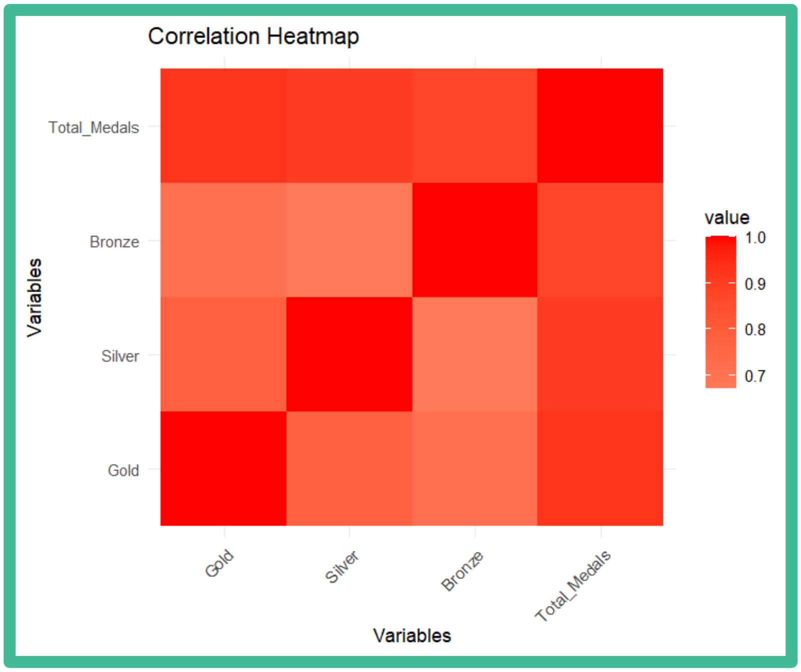


- Continents correlation with all variables except for countries since it has high cardinality. All of those are positive correlations except for Africa and Australia have moderate correlation.

II.

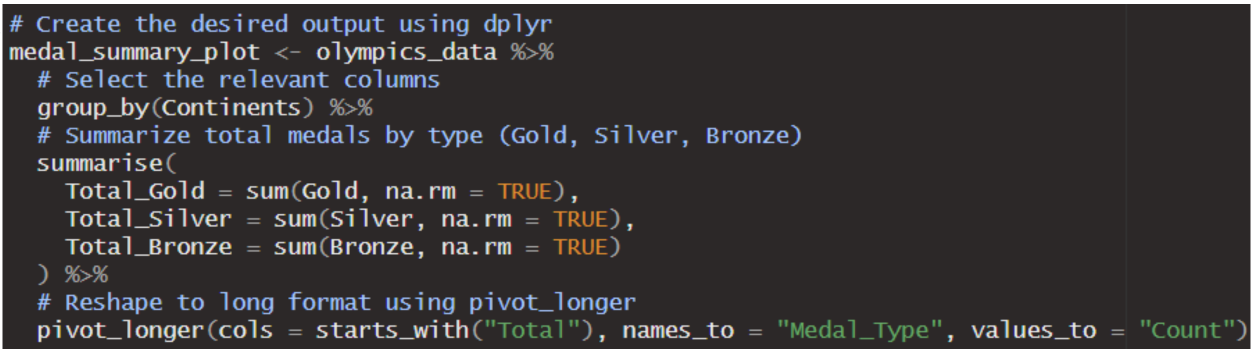


```
# Select only numeric columns for correlation matrix
numerical_data <- olympics_data[sapply(olympics_data, is.numeric)]

# Compute correlation matrix
cor_matrix <- cor(numerical_data, use = "complete.obs")

# Melt the correlation matrix for ggplot2
cor_matrix_melted <- melt(cor_matrix)

# Create the heatmap
ggplot(cor_matrix_melted, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0) +
  theme_minimal() +
  labs(title = "Correlation Heatmap", x = "Variables", y = "Variables") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels for readability
```

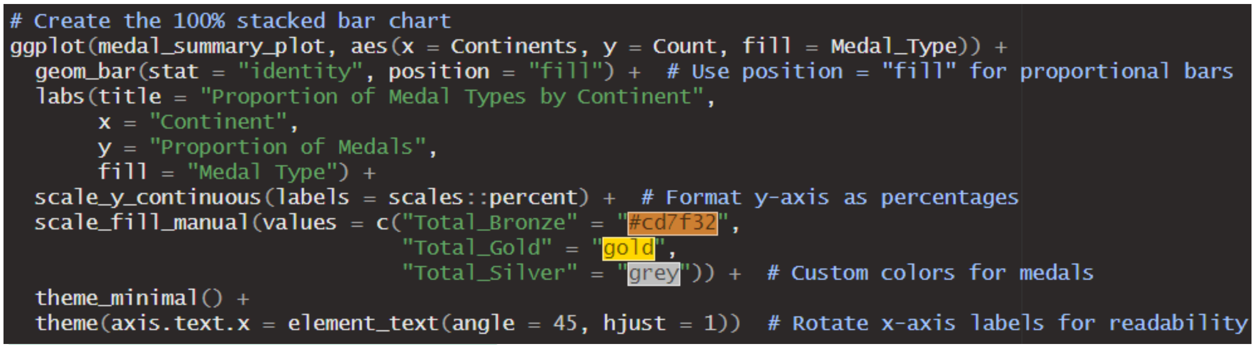


- correlation between only numeric variables. None of those are under Blue which means variables are Highly corelating.

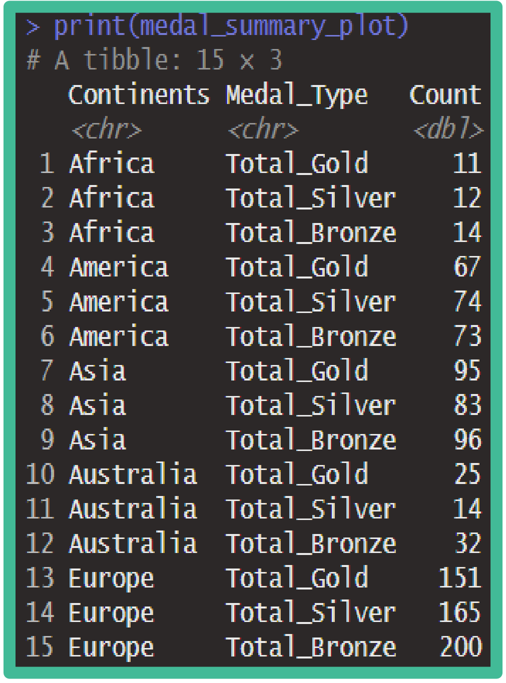
III.



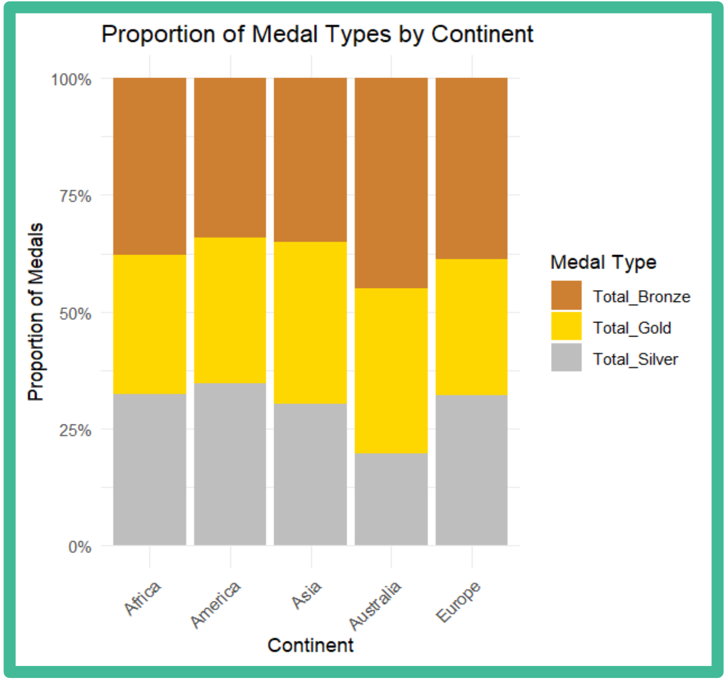
```
# Create the desired output using dplyr
medal_summary_plot <- olympics_data %>%
  # Select the relevant columns
  group_by(Continents) %>%
  # Summarize total medals by type (Gold, Silver, Bronze)
  summarise(
    Total_Gold = sum(Gold, na.rm = TRUE),
    Total_Silver = sum(Silver, na.rm = TRUE),
    Total_Bronze = sum(Bronze, na.rm = TRUE)
  ) %>%
  # Reshape to long format using pivot_longer
  pivot_longer(cols = starts_with("Total"), names_to = "Medal_Type", values_to = "Count")
```



```
# Create the 100% stacked bar chart
ggplot(medal_summary_plot, aes(x = Continents, y = Count, fill = Medal_Type)) +
  geom_bar(stat = "identity", position = "fill") +  # Use position = "fill" for proportional bars
  labs(title = "Proportion of Medal Types by Continent",
       x = "Continent",
       y = "Proportion of Medals",
       fill = "Medal Type") +
  scale_y_continuous(labels = scales::percent) +  # Format y-axis as percentages
  scale_fill_manual(values = c("Total_Bronze" = "#cd7f32",
                               "Total_Gold" = "gold",
                               "Total_Silver" = "grey")) +  # Custom colors for medals
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels for readability
```

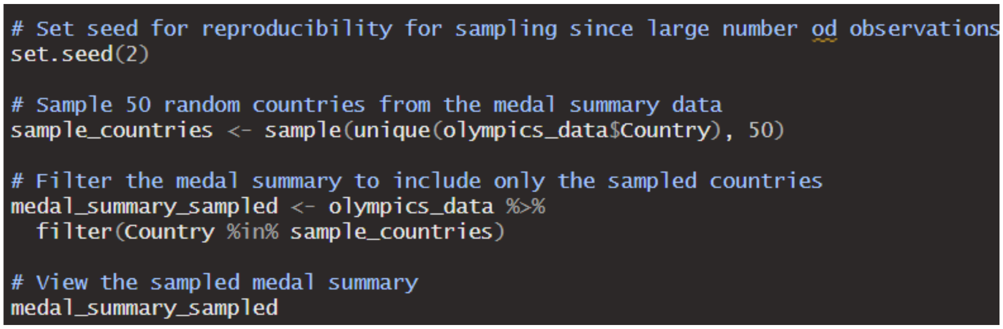


```
> print(medal_summary_plot)
# A tibble: 15 x 3
   Continents Medal_Type   Count
   <chr>      <chr>        <dbl>
 1 Africa     Total_Gold      11
 2 Africa     Total_Silver    12
 3 Africa     Total_Bronze    14
 4 America    Total_Gold      67
 5 America    Total_Silver    74
 6 America    Total_Bronze    73
 7 Asia       Total_Gold      95
 8 Asia       Total_Silver    83
 9 Asia       Total_Bronze    96
10 Australia  Total_Gold      25
11 Australia  Total_Silver    14
12 Australia  Total_Bronze    32
13 Europe     Total_Gold     151
14 Europe     Total_Silver   165
15 Europe     Total_Bronze   200
```

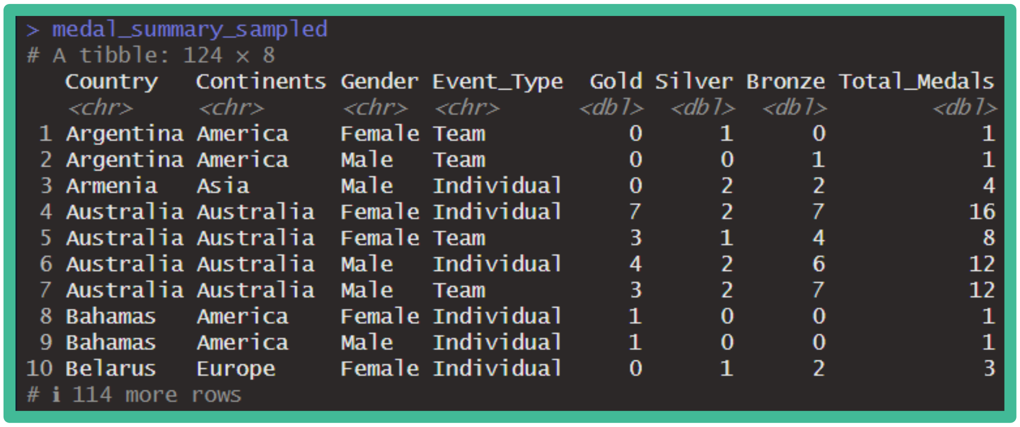


Stacked bar chart of Medal type percentage grouped by Continents. Africa and Europe have equally shared the medal type counts and where Australia taken less number of Silver than other medals.

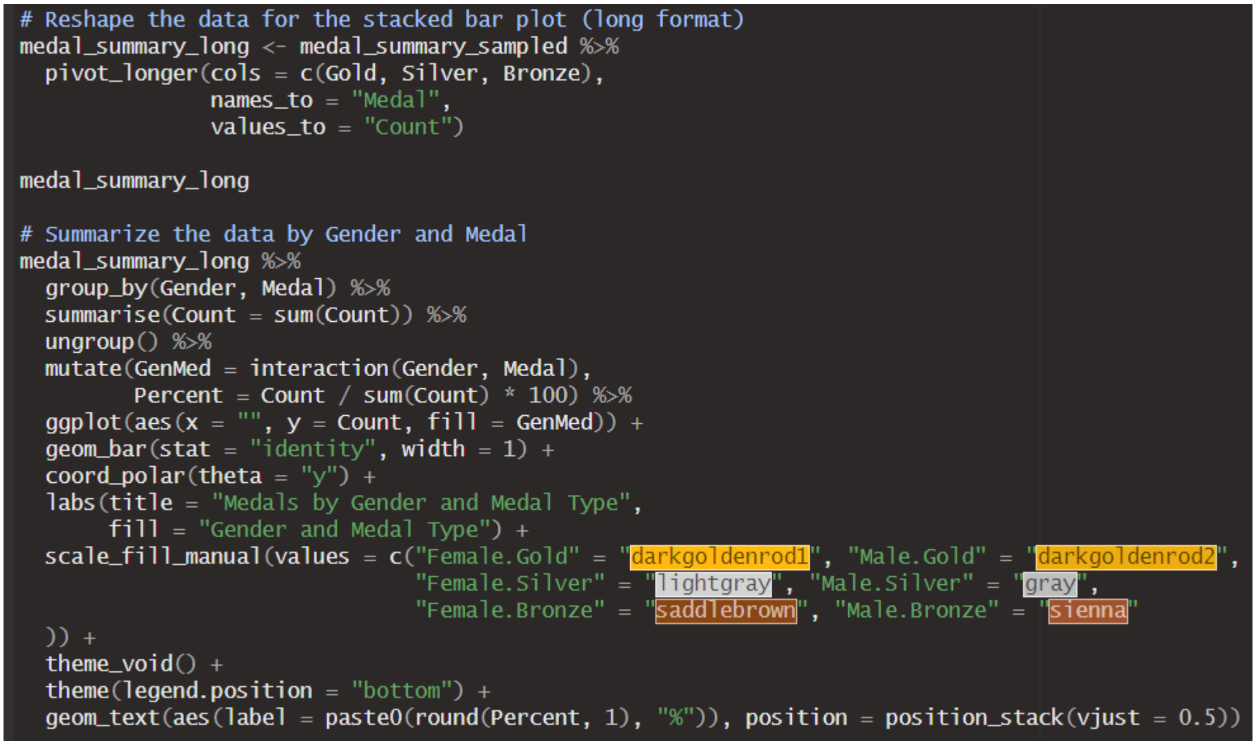
IV. sampling dataset for easy visualisation, grouping country, gender and eventype



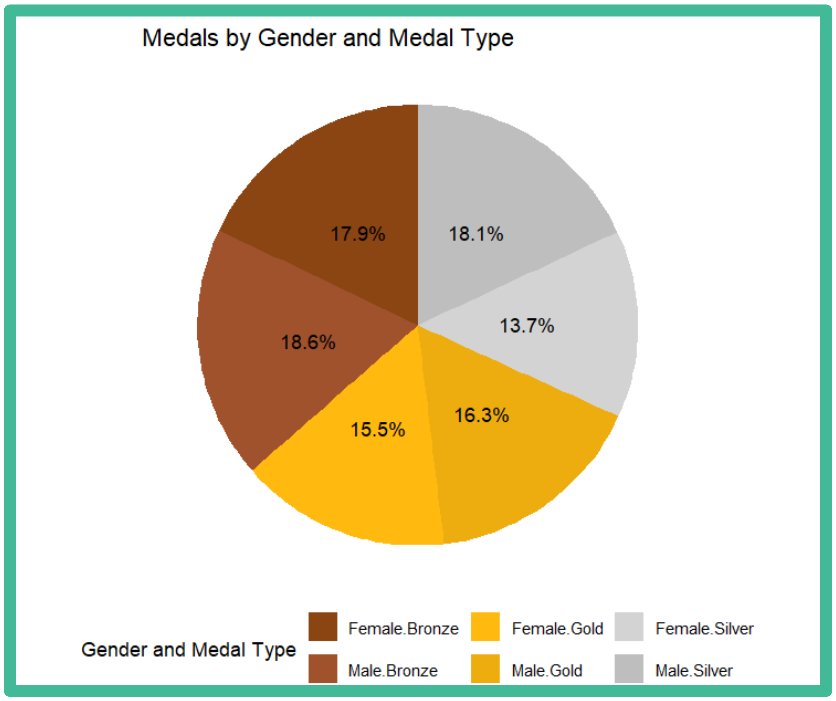
```
# Set seed for reproducibility for sampling since large number od observations
set.seed(2)

# Sample 50 random countries from the medal summary data
sample_countries <- sample(unique(olympics_data$Country), 50)

# Filter the medal summary to include only the sampled countries
medal_summary_sampled <- olympics_data %>%
  filter(Country %in% sample_countries)

# View the sampled medal summary
medal_summary_sampled
```



```
> medal_summary_sampled
# A tibble: 124 × 8
   Country   Continents Gender Event_Type  Gold Silver Bronze Total_Medals
   <chr>     <chr>      <chr>  <chr>      <dbl>  <dbl>  <dbl>        <dbl>
 1 Argentina America    Female Team           0      1      0            1
 2 Argentina America    Male   Team           0      0      1            1
 3 Armenia   Asia       Male   Individual     0      2      2            4
 4 Australia Australia  Female Individual     7      2      7           16
 5 Australia Australia  Female Team           3      1      4            8
 6 Australia Australia  Male   Individual     4      2      6           12
 7 Australia Australia  Male   Team           3      2      7           12
 8 Bahamas   America    Female Individual     1      0      0            1
 9 Bahamas   America    Male   Individual     1      0      0            1
10 Belarus   Europe     Female Individual     0      1      2            3
# ℹ 114 more rows
```

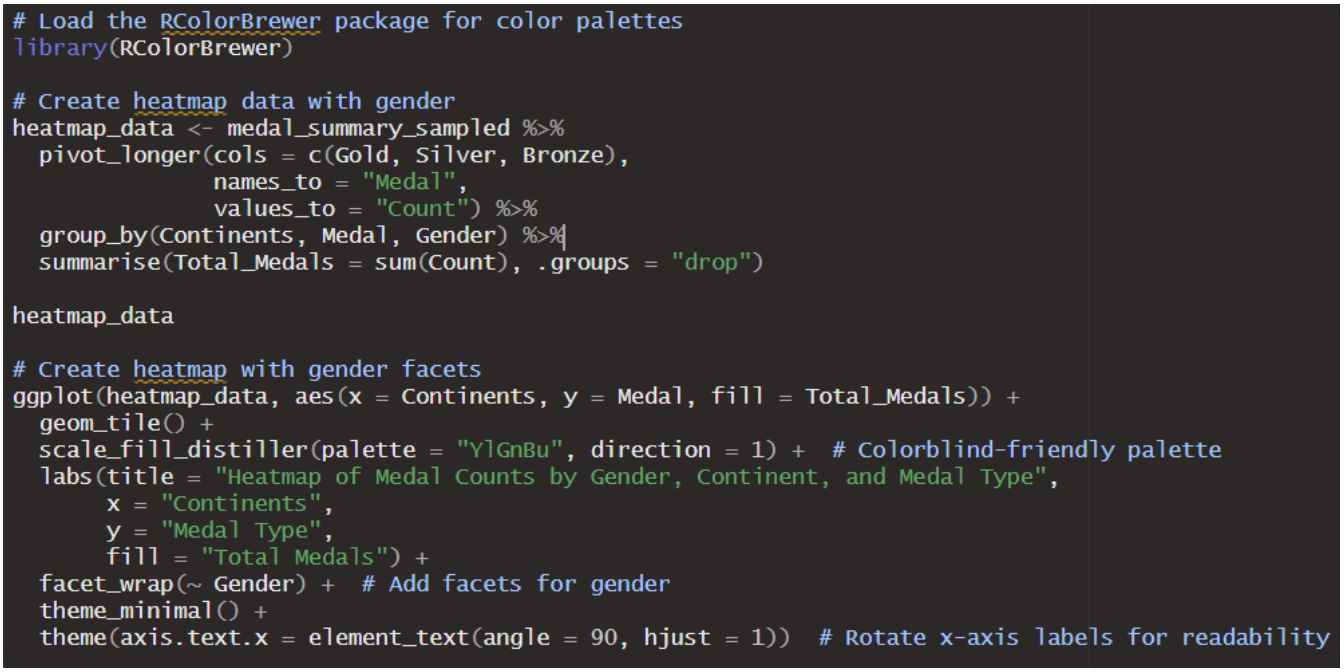


```
# Reshape the data for the stacked bar plot (long format)
medal_summary_long <- medal_summary_sampled %>%
  pivot_longer(cols = c(Gold, Silver, Bronze),
               names_to = "Medal",
               values_to = "Count")

medal_summary_long

# Summarize the data by Gender and Medal
medal_summary_long %>%
  group_by(Gender, Medal) %>%
  summarise(Count = sum(Count)) %>%
  ungroup() %>%
  mutate(GenMed = interaction(Gender, Medal),
         Percent = Count / sum(Count) * 100) %>%
  ggplot(aes(x = "", y = Count, fill = GenMed)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  labs(title = "Medals by Gender and Medal Type",
       fill = "Gender and Medal Type") +
  scale_fill_manual(values = c("Female.Gold" = "darkgoldenrod1", "Male.Gold" = "darkgoldenrod2",
                               "Female.Silver" = "lightgray", "Male.Silver" = "gray",
                               "Female.Bronze" = "saddlebrown", "Male.Bronze" = "sienna"
  )) +
  theme_void() +
  theme(legend.position = "bottom") +
  geom_text(aes(label = paste0(round(Percent, 1), "%")), position = position_stack(vjust = 0.5))
```

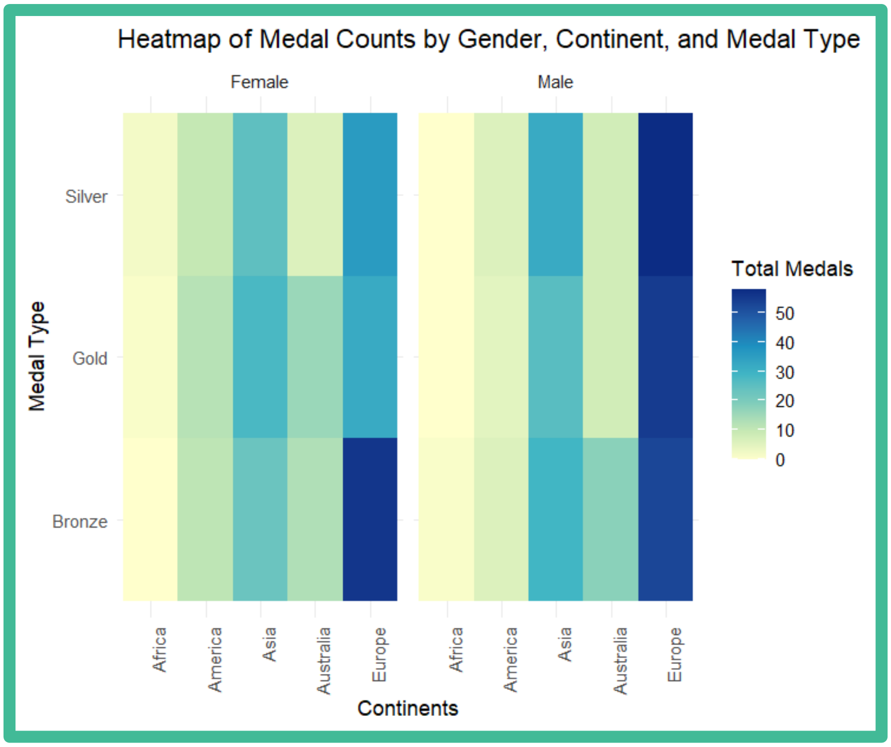


Pie chart representing medal types grouped by gender. Where all are seems to be equally shared. But Female Silver Medals are lower than all other diversifying.

V.



```
# Load the RColorBrewer package for color palettes
library(RColorBrewer)

# Create heatmap data with gender
heatmap_data <- medal_summary_sampled %>%
  pivot_longer(cols = c(Gold, Silver, Bronze),
               names_to = "Medal",
               values_to = "Count") %>%
  group_by(Continents, Medal, Gender) %>%
  summarise(Total_Medals = sum(Count), .groups = "drop")

heatmap_data

# Create heatmap with gender facets
ggplot(heatmap_data, aes(x = Continents, y = Medal, fill = Total_Medals)) +
  geom_tile() +
  scale_fill_distiller(palette = "YlGnBu", direction = 1) +  # Colorblind-friendly palette
  labs(title = "Heatmap of Medal Counts by Gender, Continent, and Medal Type",
       x = "Continents",
       y = "Medal Type",
       fill = "Total Medals") +
  facet_wrap(~ Gender) +  # Add facets for gender
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))  # Rotate x-axis labels for readability
```

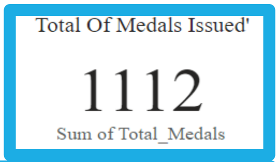


- Colorblind friendly Heatmap represents Medal types received majorly that are divided by gender, grouped by Continents. Europe highlighted as most received and Africa as less in all genders. Using this graph can further analyze data since it contains many variables and grouped by continents that give value to my RQ. For example, But Australia Female gold winners are comparatively higher than Australia Male Gold winners

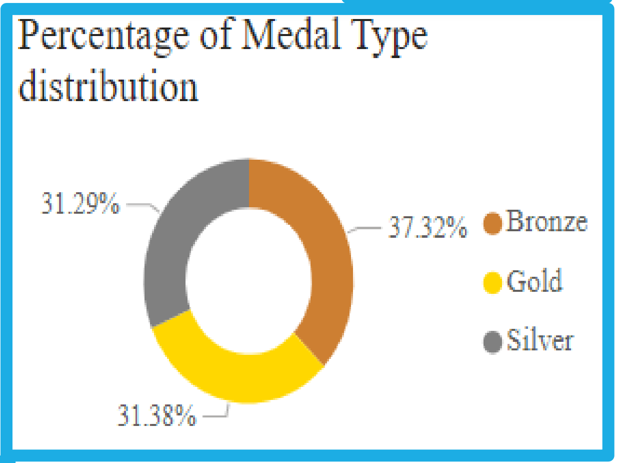
Data storytelling

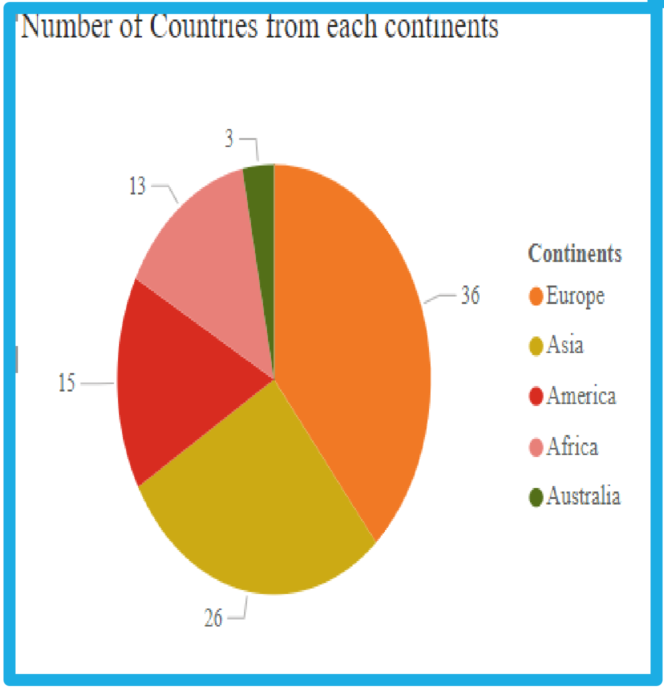
Comprehensive Analysis of Global Medal Distribution of Olympics 2020.

The Tokyo 2020 Olympics brought together athletes from across the global under tragic circumstances following the COVID – 19 pandemics. Though many countries stepped forward to bring their athletes to be a part of the event, that records with over 200 countries participation along with diverse level success across continents.



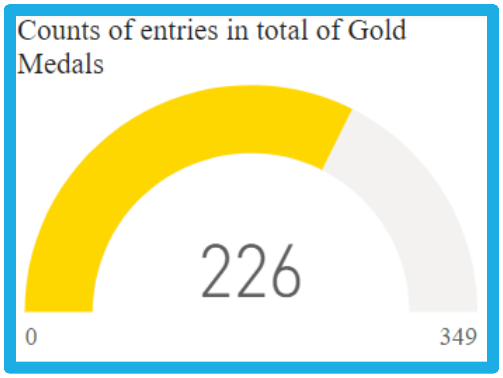
Total Number of Medals issued by Tokyo 2020 were 1112. Yes, the highest percentage of issues is Bronze at 37%, Which is contrary to popular belief. Following that Gold and Silver at nearly equal proportions.

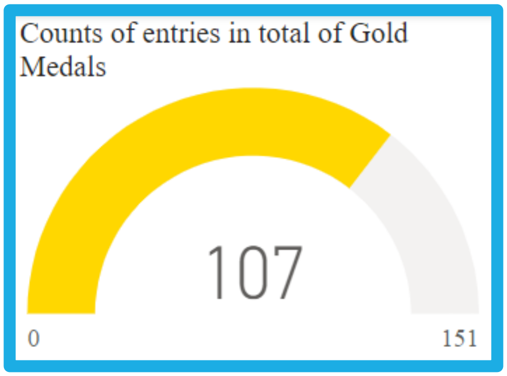




From that there were 36 countries that represent Europe that emerged as the dominant continent, securing most medals in the Olympics 2020. In order that Asia, America, Africa and Australia were also exceptional continents that celebrated victory. Number of countries participated and won on behalf of Australia continents is very low.

Let's focus more on other factors along with continents. For example,





For all Continents

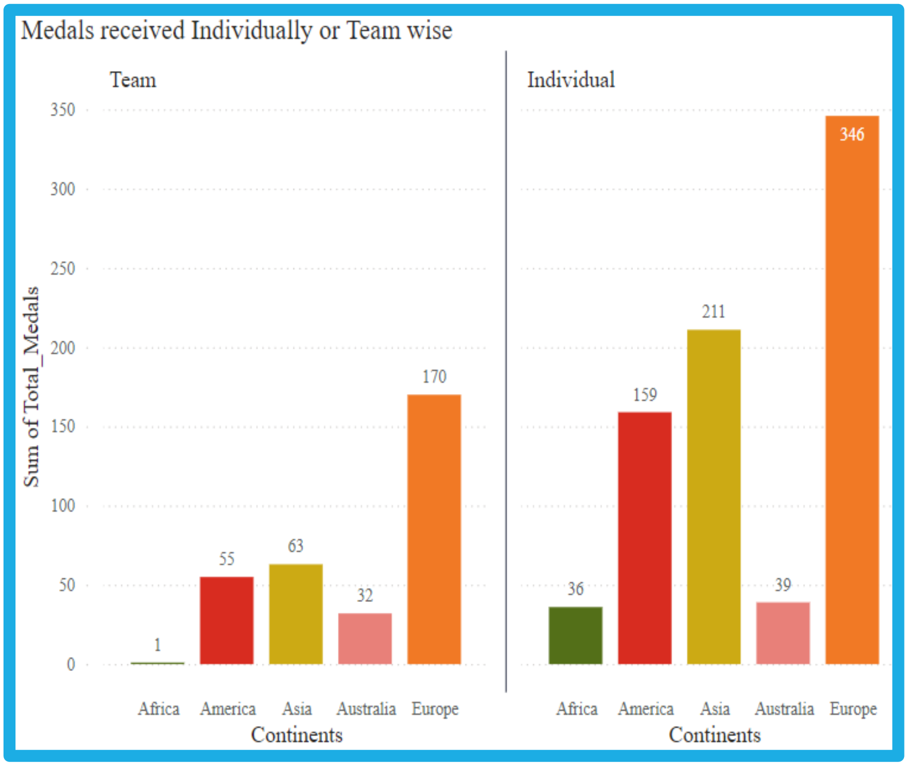
Entirely gold Medals issued were 349 Maximum, where 226 entries

After selecting Europe Continent

Total of Gold medals for Europe was 151, 107 entries

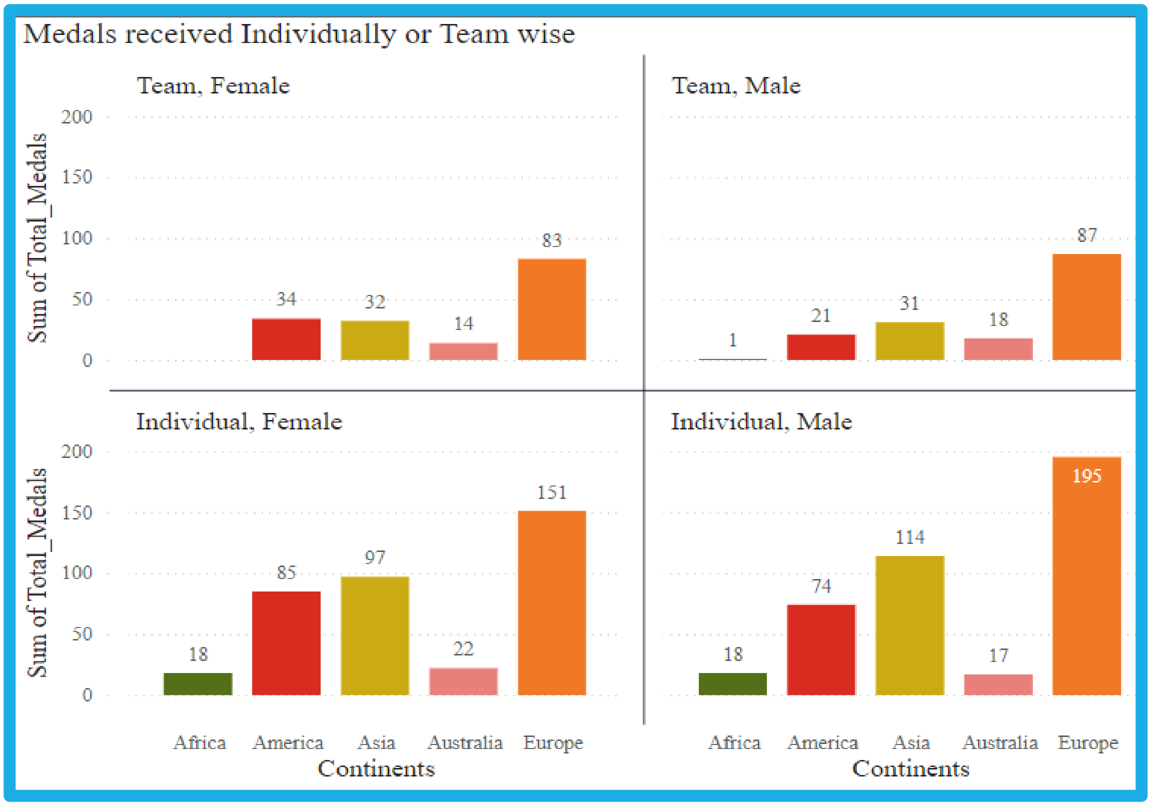
But why?

Countries were broken down by gender such as Male and female. While Event type separated further as Individual and Team games.



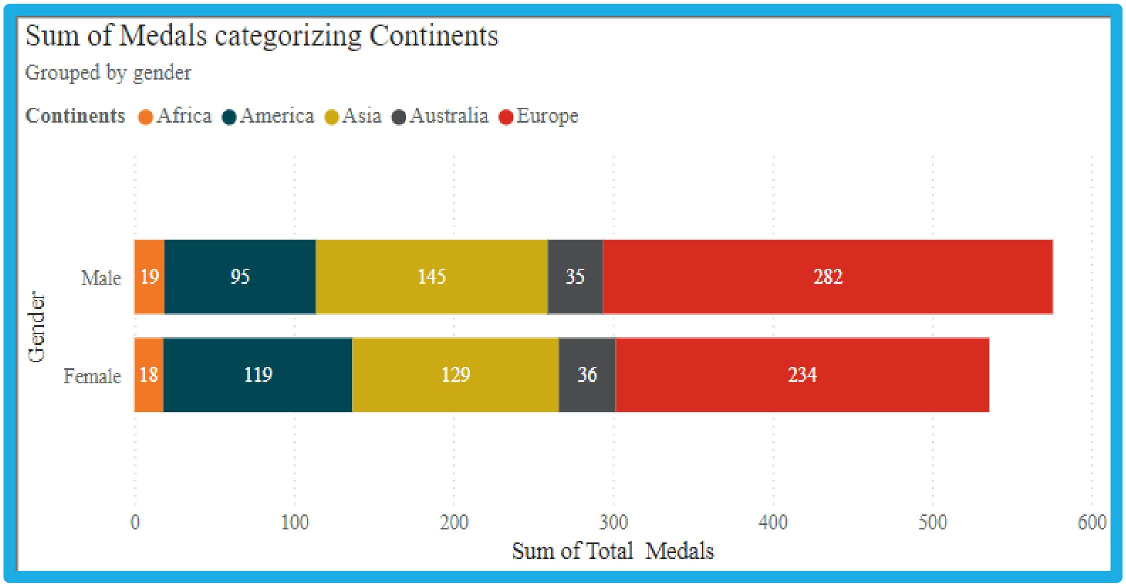
Going through colorblind friendly bar charts the event type involvement, continents wise, Individual games must have been very competitive than Teams. Analyzing deeply Asia has moderate medal victory and none other than Europe has highest reach in both types of games. Considering fair and near to equal frequency counts Australia has 32 : 39 for Team and Individual games.

Further dividing this by gender,

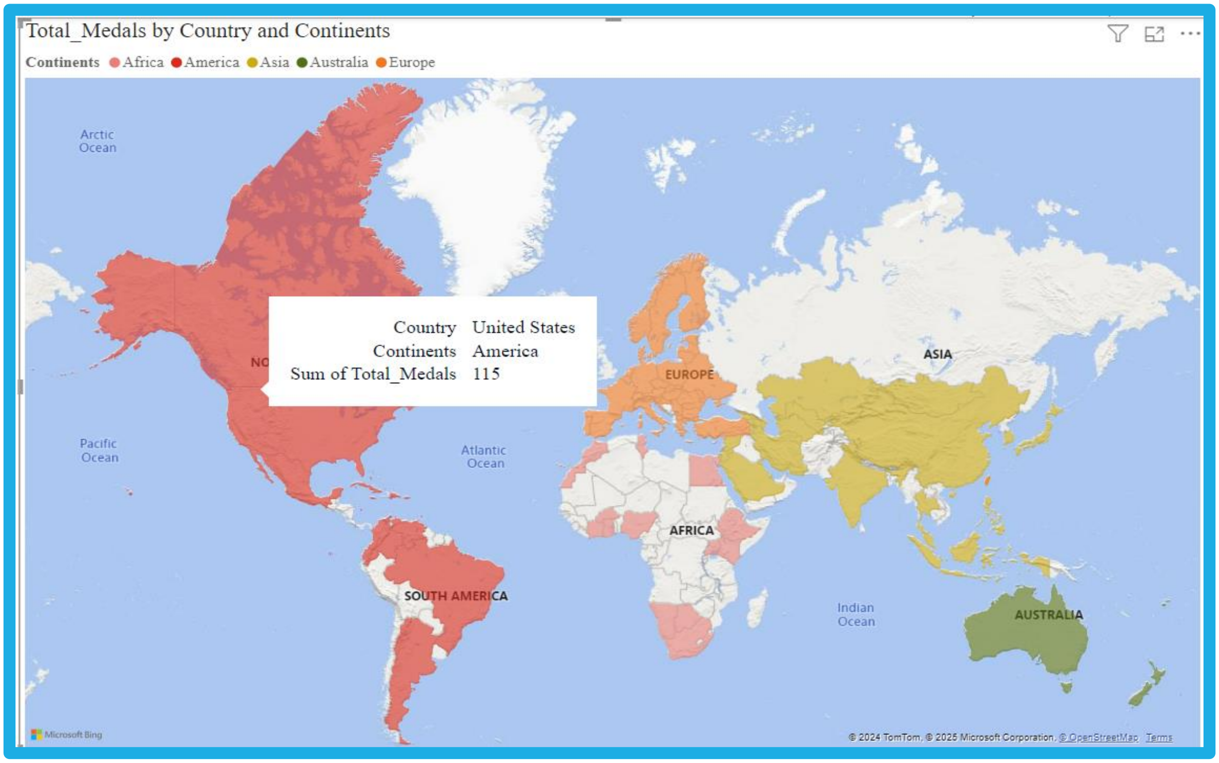


Male athletes outperformed female athletes in total medal, particularly in individual games. For instance, if you closely see Europe, medals mostly taken by Men at 195 and low counts of medals received at 151 by female. These small multiples ensure that individual games won many medals by Male and Female in general. An interesting fact here is Africa, in Both genders taken equal count of Medals for Individual games. In team games, gender participation seems equal, but it must be improved more and bring up to individual game standards.

In the race of Gender, over the discrimination, it is not too far to reach equality. Continents have given the best to equality but in total it's controversial. There must be more appreciation and facilities given for women empowerment here.

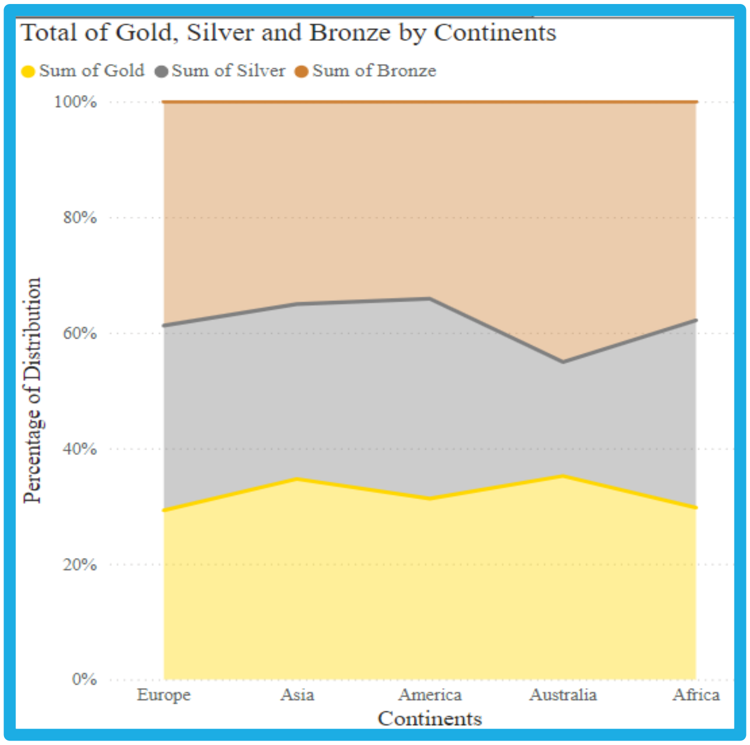


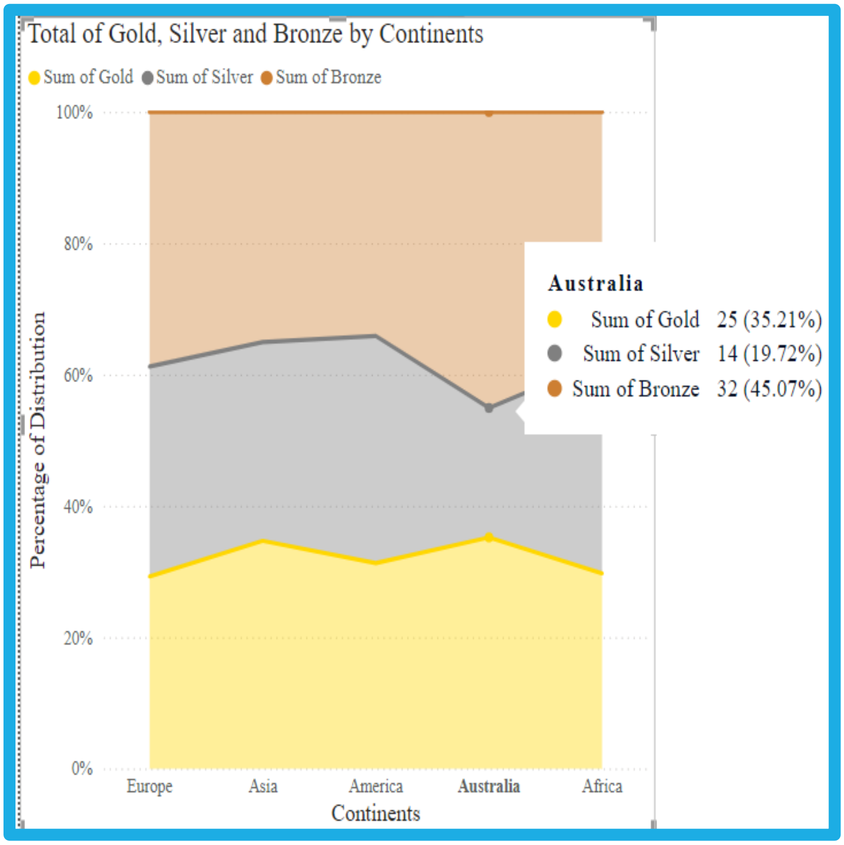
The world map represents all the countries won in Olympics 2020, differentiating Continents by diversifying colors. Though this Spatial level insight where population will differ continent to continent.


Total_Medals by Country and Continents
Continents ● Africa ● America ● Asia ● Australia ● Europe
Arctic Ocean
Pacific Ocean
Atlantic Ocean
Indian Ocean
NO
EUROPE
ASIA
AFRICA
SOUTH AMERICA
AUSTRALIA
Microsoft Bing
© 2024 TomTom, © 2025 Microsoft Corporation, © OpenStreetMap Terms

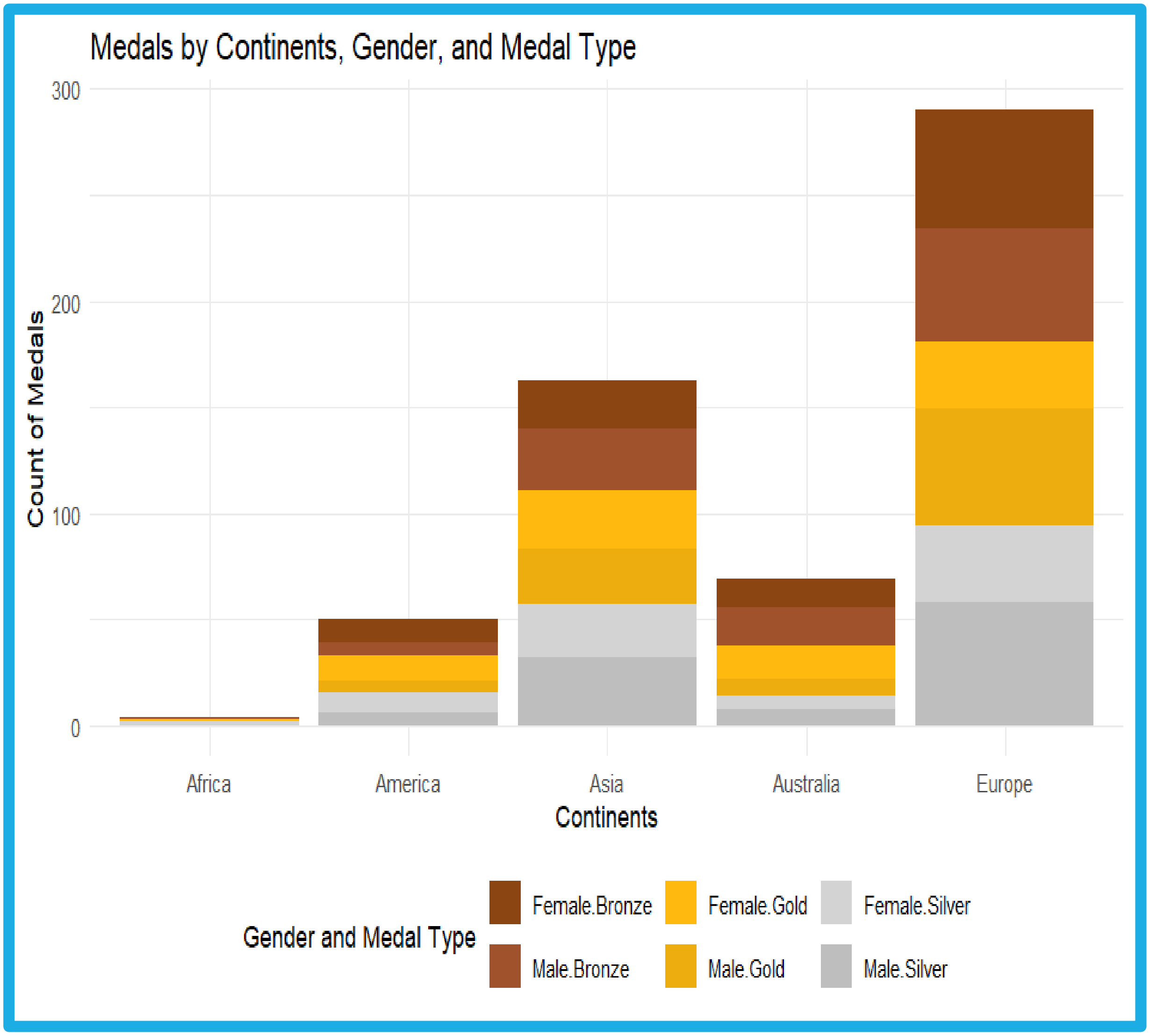
Country United States
Continents America
Sum of Total_Medals 115

Coming back to broader approach, by considering the percentages in Area chart where Bronze captures wide space, and Silver has narrowed space, and this ensures the Silver has the least medal distribution as well. Asia and Australia have same percentage of gold medal distribution. Africa has the least in all formats.





Pinpoint and looking for Australia Continent that shows basic statistics of all three medals, where we can see Silver is 14 which is comparatively lower than other medals and the continents.



Finally, the chart says Europe has the highest medals and Africa is the lowest by considering only the factors Gender, Event type, Counts of Countries.

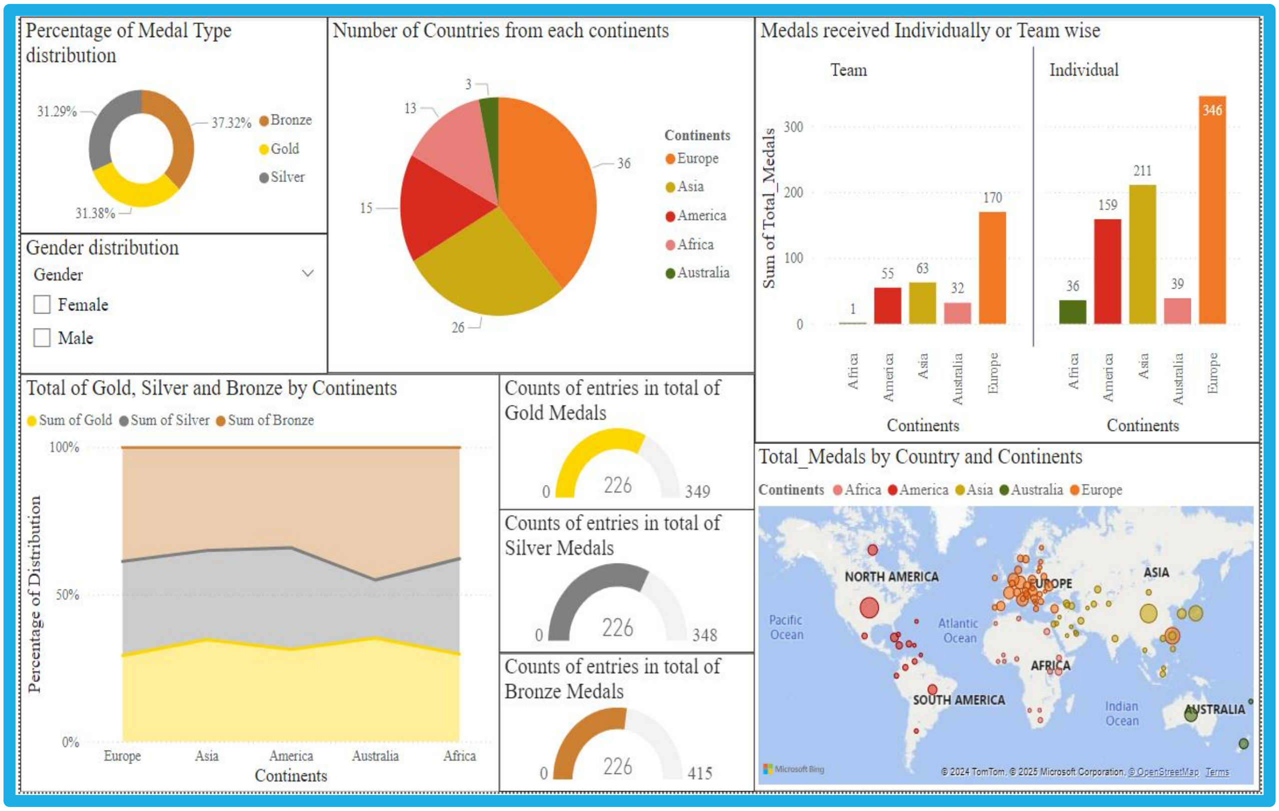
There are many economic and demographic factors that influence the Olympics but here we focused basically on terms of equity in sports infrastructure, especially underrepresented regions. This research led to find Majority and Minority percentage of medal winnings continents wise. Let's emphasize the need for inclusive growth in international sports by improving access to training facilities, nurturing talents in low-income regions, and promoting gender equity.

Key takeaways

- Europe has a high performance and inclusive positive impact on all factors.
- Asia and America perform well but trail behind Europe.
- The underappreciation of Africa and Australia suggests improving opportunities
- The balanced distribution of medal types and prominence of individual games explained us to enhance sportsmanship and humanity by involving in group events.
- As a negative result, the Olympics Refugees Team did not receive any medals at all.

Future Olympics can celebrate a more inclusive and diverse field of champions

Take a quick look at the preview of dashboard


Percentage of Medal Type distribution
31.29%
37.32%
31.38%
● Bronze
● Gold
● Silver
Number of Countries from each continents
3
13
36
15
26
Continents
● Europe
● Asia
● America
● Africa
● Australia
Medals received Individually or Team wise
Team
Individual
Sum of Total_Medals
300
200
100
0
1
55
63
32
170
36
159
211
39
346
Africa
America
Asia
Australia
Europe
Continents
Continents
Gender distribution
Gender
□ Female
□ Male
Total of Gold, Silver and Bronze by Continents
● Sum of Gold ● Sum of Silver ● Sum of Bronze
100%
50%
0%
Percentage of Distribution
Europe
Asia
America
Australia
Africa
Continents
Counts of entries in total of Gold Medals
0
226
349
Counts of entries in total of Silver Medals
0
226
348
Counts of entries in total of Bronze Medals
0
226
415
Total_Medals by Country and Continents
Continents ● Africa ● America ● Asia ● Australia ● Europe
NORTH AMERICA
EUROPE
ASIA
Pacific Ocean
Atlantic Ocean
AFRICA
SOUTH AMERICA
Indian Ocean
AUSTRALIA
Microsoft Bing
© 2024 TomTom, © 2025 Microsoft Corporation, © OpenStreetMap Terms

Refer Power BI dashboards here - https://app.powerbi.com/links/FNHeCjMeBe?ctid=9a5b5691-a451-49e7-93de-9c61cb04328b&pbi_source=linkShare