Assignment 4 - Data Visualization

Name:Penca Matei

Studentnumber: s4039696

Visualization

Packages & Data

This is a historical dataset on the modern Olympic Games, including all the Games from Athens 1896 to Rio 2016. The data was retrieved from here.

```
library(ggplot2)
library(tidyverse)
data <- read.csv("https://stulp.gmw.rug.nl/dataviz/athlete_events.csv",
  header = TRUE
)
```

Code to create visualization

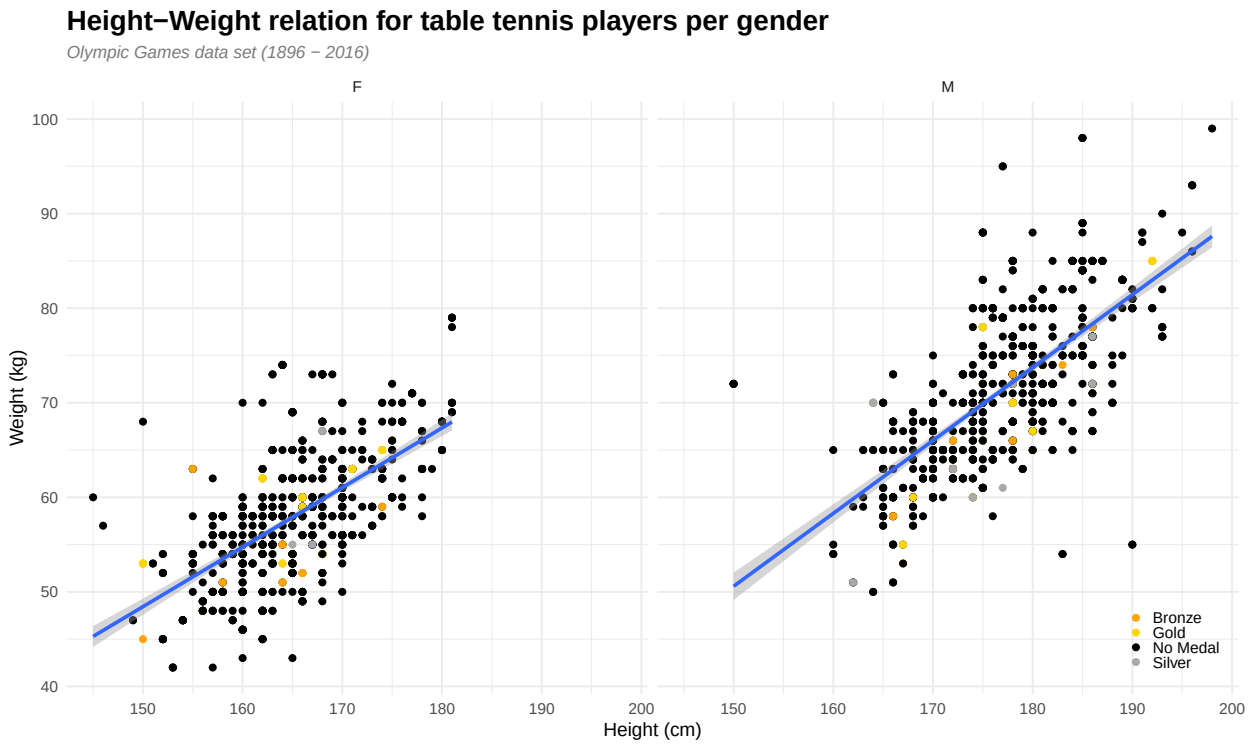
```
# Because the data set is very big, I focus on table tennis (ping pong), and I
# clean out any NAs in weight or height
pingpong_data <- data %>% filter(
  Sport == "Table Tennis", !is.na(Height),
  !is.na(Weight)
)

# I transform the NAs into 'No medal' as it's easier to read
pingpong_data <- pingpong_data %>% mutate(Medal = case_when(
  is.na(Medal) ~ "No Medal",
  !is.na(Medal) ~ as.character(Medal)
))

# Create custom theme for labels and legend position
custom_theme <- theme(
  plot.title = element_text(size = 16, face = "bold", colour = "black"),
  plot.subtitle = element_text(size = 10, face = "italic", colour = "gray48"),
  legend.position = c(0.945, .1),
  legend.key.size = unit(0.3, "cm"),
  legend.title = element_blank()
)

ggplot(pingpong_data, aes(x = Height, y = Weight, color = Medal)) +
  geom_point(size = 1.5) +
```

```
facet_grid(~Sex) +
geom_smooth(aes(x = Height, y = Weight), method = "lm", inherit.aes = FALSE
) +
scale_color_manual(values = c("orange", "gold", "black", "darkgrey")) +
scale_y_continuous(breaks = seq(0, 100, by = 10)) +
labs(
  x = "Height (cm)", y = "Weight (kg)",
  title = "Height-Weight relation for table tennis players per gender",
  subtitle = "Olympic Games data set (1896 - 2016)"
) +
theme_minimal() +
custom_theme
```



Description of visualization

In the graph above I have decided to split the data between Male (M) and Female (F) to better showcase the distribution of values of values. In both cases we can see that there is a clear relationship between the weight and the height of a professional table tennis player. The general rule seems to be that heavier people are also taller, with a few exception (the outliers on the graph). This relation can also be seen by the linear regression line that has the confidence intervals larger at the margins as it encounters a few outliers.

From the graph we can also see that males seem to have higher values for both height and weight then females which makes sense. With this thought in mind, I also added to graph the distribution of medals to see if there is any relation between the 2 variables plotted and the chances of getting a medal. We can see that the medals are also quite uniformly distributed, and it is rather hard to say that being a smaller person can give you an advantage.

We can notice that all the medals are close to the regression line which tells us that in order to perform well you need good proportions between height and weight.

Because I am using a dot plot, some of the points might overlap so it is hard to get a real idea of the population size. The colors that I chose might also not be the best options, but I thought it would be nice to correlate them to the real life color of the medals.