Project 4

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

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(ggplot2)
library(knitr)

athletes_data <- read_excel("~/Dropbox/Courses/EPFL Extension School/Rstudio course/Final Projects/Githu
    sheet = "athletes")
country_data<-read_excel("~/Dropbox/Courses/EPFL Extension School/Rstudio course/Final Projects/Github/
    sheet = "country")
games_data<-read_excel("~/Dropbox/Courses/EPFL Extension School/Rstudio course/Final Projects/Github/Fi
    sheet = "games")
medals_data<-read_excel("~/Dropbox/Courses/EPFL Extension School/Rstudio course/Final Projects/Github/F
    sheet = "medals")
```

Part 1

some athletes competed for different countries over time

```
athlete_participated_different_country<-country_data %>%
  select(-Age,-Games)%>%
  group_by(athlete_id)%>%
  mutate(participation=n()) %>%
  filter(participation!=1) %>%
  count()

athlete_participated_different_country_table<-athlete_participated_different_country %>%
  top_n(20) %>%
  knitr::kable()
```

Yes, many athletes (37121) have participated for different countries.

Part 2

ten athletes that took part in most games

```
Participation_number_of_games<-medals_data%>%
  select(athlete_id,Games) %>%
  group_by(athlete_id) %>%
  summarise(number_of_games=n())%>%
  arrange(desc(number_of_games))%>%
  top_n(10)

Participation_number_of_games_table<-Participation_number_of_games %>%
  kable()
```

ten athletes that took part in most games are :

athlete_id	number_of_games
77710	58
106296	39
115354	38
119591	36
44875	32
53240	32
89187	32
119590	32
129196	32
55047	31
76437	31
83312	31
106156	31

Part 3

athlete(s) kept a Gold medal for the longest time

```
Athlete_gold_longest_time<-medals_data%>%
  select(athlete_id,Games,Sport,Medal)%>%
  mutate(Year=as.numeric(str_sub(Games,start = 1L,end = 4L)))%>%
  filter(Medal=="Gold")%>%
  group_by(athlete_id)%>%
  mutate(number_of_years=max(Year)-min(Year))%>%
  arrange(desc(number_of_years))

  Athlete_gold_longest_time_by_sport<-Athlete_gold_longest_time %>%
    select(Sport,number_of_years) %>%
    group_by(Sport) %>%
    top_n(1)

Athlete_gold_longest_time_table<-Athlete_gold_longest_time %>%
  kable()
```

Fencing athlete(s) kept a Gold medal for the longest time.

Part 4

country(ies) kept a Gold medal for the longest time

```
Countries_gold_longest_time<-medals_data%>%
  select(athlete_id,Games,Team,Sport,Medal)%>%
  mutate(Year=as.numeric(str_sub(Games,start = 1L,end = 4L)))%>%
  filter(Medal=="Gold")%>%
  group_by(Team,Sport)%>%
  mutate(number_of_years=max(Year)-min(Year))%>%
  ungroup(Sport) %>%
  arrange(desc(number_of_years))


Countries_gold_longest_time_table<-Countries_gold_longest_time %>%
  select(Team,Sport,number_of_years) %>%
  filter(number_of_years==120) %>%
  group_by(Team) %>%
  count() %>%
  kable()
```

country(ies) kept a Gold medal for the longest time are :

Team	n
France	129
Germany	46
Great Britain	14
Greece	20
Hungary	36
United States	654

Part 5

ten athletes that competed in the most events (some athletes take part in more than one event during games)

```
Athletes_participated_most_events<-medals_data%>%
  select(athlete_id,Event)%>%
  group_by(athlete_id) %>%
  summarise(number_of_events=n())%>%
  arrange(desc(number_of_events))%>%
  top_n(10)    #Not sure why answer for Part2 and Part 5 are same???

Athletes_participated_most_events_table<-Athletes_participated_most_events %>%
  kable()
```

ten athletes that competed in the most events :

athlete_id	number_of_events
77710	58
106296	39
115354	38
119591	36
44875	32
53240	32
89187	32
119590	32
129196	32
55047	31
76437	31
83312	31
106156	31

Part 6

number of medals per country (rows) and per year (column)

```
Number_of_medals_per_country<-medals_data%>%
  select(Games,Team,Medal)%>%
  mutate(Year=as.numeric(str_sub(Games,start = 1L,end = 4L)))%>%
  select(-Games)%>%
  group_by(Team,Year)%>%
  summarise(total_number_of_medals_per_country=n())%>%
  ungroup() %>%
  top_n(15)

Number_of_medals_per_country_table<-Number_of_medals_per_country %>%
  kable()
```

number of medals per country (rows) and per year (column) are :

Team	Year	total_number_of_medals_per_country
Australia	2000	762
China	2008	708
Germany	1992	820
Great Britain	1908	752
Soviet Union	1980	773
Soviet Union	1988	771
Unified Team	1992	832
United States	1904	823
United States	1932	778
United States	1984	821
United States	1988	886
United States	1992	936
United States	1996	827
United States	2000	748
United States	2008	744

Part 7

Relationship between country and the probability of winning a medal

```
# Sorry I did not realize till recently that i did not finish Part 7, will try my best before the proje

#Total_number_of_medals_counted_by_year<-Number_of_medals_per_country %>%
#group_by(Year) %>%
#summarise(Total_number_of_medals_counted=sum(total_number_of_medals_per_country))
```

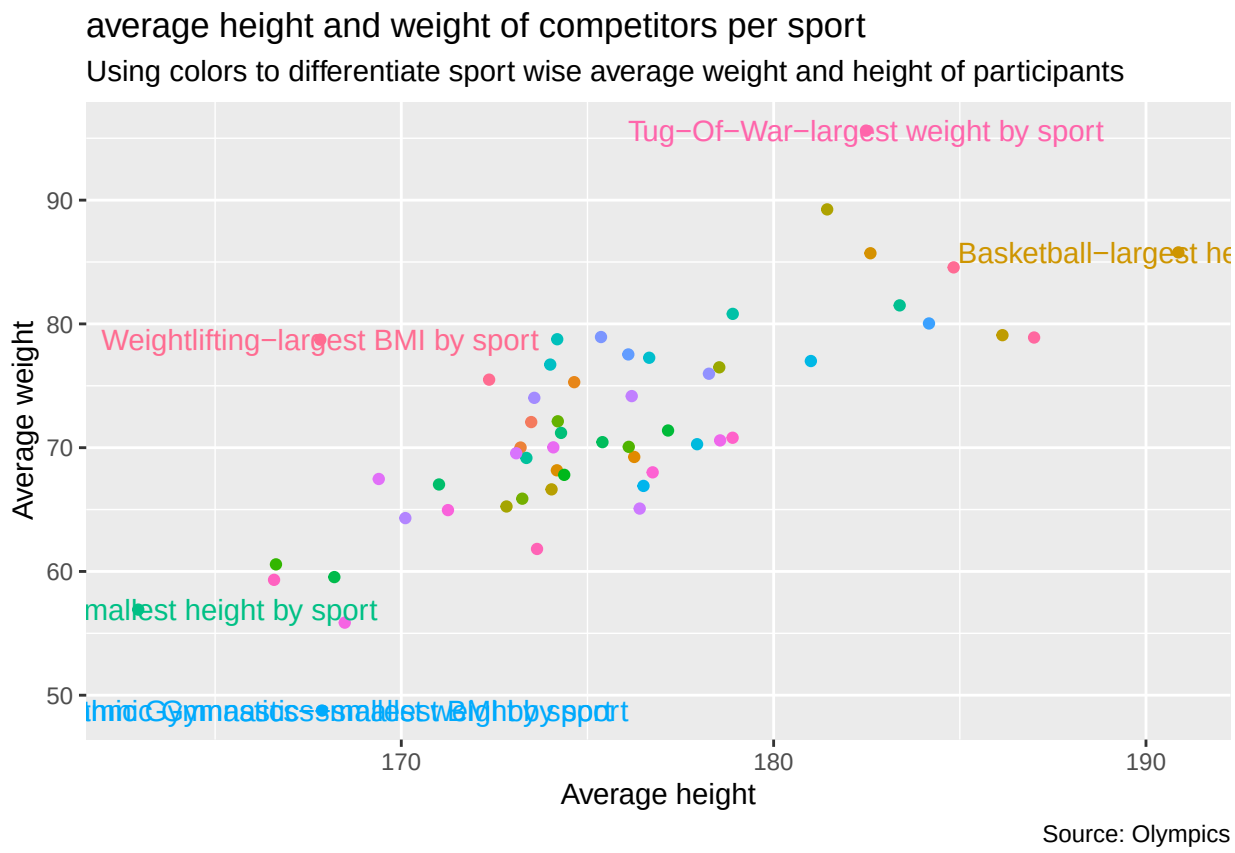
Part 8

Scatterplot showing the average height and weight of competitors per sport

```
medals_atheltes_left_join<-medals_data%>%
  left_join(athletes_data,
            by = c("athlete_id" ="ID"))%>%
  mutate(BMI=Weight/(Height/100)^2)

Average_weight_by_sport<-medals_atheltes_left_join%>%
  group_by(Sport)%>%
  summarize(mean_weight = mean(Weight, na.rm = TRUE))
Average_height_by_sport<-medals_atheltes_left_join%>%
  group_by(Sport)%>%
  summarize(mean_height = mean(Height, na.rm = TRUE))

Average_BMI_by_sport<-medals_atheltes_left_join%>%
  group_by(Sport)%>%
  summarize(mean_BMI = mean(BMI, na.rm = TRUE))

largest_weight_by_sport <- Average_weight_by_sport %>%
  slice(which.max(mean_weight))
```

```r
smallest_weight_by_sport <- Average_weight_by_sport %>%
  slice(which.min(mean_weight))
largest_height_by_sport <- Average_height_by_sport %>%
  slice(which.max(mean_height))
smallest_height_by_sport <- Average_height_by_sport %>%
  slice(which.min(mean_height))
largest_BMI_by_sport <- Average_BMI_by_sport %>%
  slice(which.max(mean_BMI))
smallest_BMI_by_sport <- Average_BMI_by_sport %>%
  slice(which.min(mean_BMI))

weight_height<-left_join(Average_weight_by_sport,Average_height_by_sport, by=c("Sport"))
weight_height_BMI<-left_join(weight_height,Average_BMI_by_sport, by=c("Sport"))

ggplot2::ggplot(data= weight_height_BMI,
                aes(y=mean_weight,
                    x=mean_height,
                    color=Sport)) + geom_point() +
  theme(legend.position="none")+
  geom_text(data= weight_height_BMI%>%
              slice(which.max(mean_weight)),
            aes(label=stringr::str_c(largest_weight_by_sport$Sport,
                                     "largest weight by sport",sep="-")), check_overlap = T) +
  geom_text(data= weight_height_BMI%>%
              slice(which.min(mean_weight)),
            aes(label=stringr::str_c(smallest_weight_by_sport$Sport,
                                     "smallest weight by sport",sep="-")), check_overlap = T) +

  geom_text(data= weight_height_BMI%>%
              slice(which.max(mean_height)),
            aes(label=stringr::str_c(largest_height_by_sport$Sport,
                                     "largest height by sport",sep="-")), check_overlap = T) +

  geom_text(data= weight_height_BMI%>%
              slice(which.min(mean_height)),
            aes(label=stringr::str_c(smallest_height_by_sport$Sport,
                                     "smallest height by sport",sep="-")), check_overlap = T) +

  geom_text(data= weight_height_BMI%>%
              slice(which.max(mean_BMI)),
            aes(label=stringr::str_c(largest_BMI_by_sport$Sport,
                                     "largest BMI by sport",sep="-")), check_overlap = T) +

  geom_text(data= weight_height_BMI%>%
              slice(which.min(mean_BMI)),
            aes(label=stringr::str_c(smallest_BMI_by_sport$Sport,
                                     "smallest BMI by sport",sep="-")), check_overlap = T) +

  labs(title = "average height and weight of competitors per sport",
       subtitle = "Using colors to differentiate sport wise average weight and height of participants",
       caption = "Source: Olympics",
       x = "Average height",
       y = "Average weight",
```

```
       fill = "Sport")
```



Part 9

Number of medals (gold, silver and bronze given per year, facet chart: summer and winter)

```
Medals_category_year_session<-medals_data%>%
  select(Games,Medal)%>%
  mutate(Year=as.numeric(str_sub(Games,start = 1L,end = 4L)))%>%
  mutate(Session=str_sub(Games,start = 5L))%>%
  select(-Games)%>%
  arrange(desc(Year))

medals_per_year <- Medals_category_year_session%>%
  group_by(Year, Medal,Session)%>%
  summarise(number_of_medals=n())%>%
  ungroup()

# Line plot for Part 9

medals_per_year %>% drop_na() %>%
  ggplot(mapping=aes(x=Year,
                     y=number_of_medals)) +
```

```
geom_line(aes(group=Medal, color=Medal)) +
theme(legend.position="bottom", legend.box = "horizontal")+
facet_wrap(vars(Session))+
labs (title = "Evolution of total number of medals per year",
      subtitle = "Using colors to differentiate medal categories",
      caption = "Source: Olympics",
      x = "Year",
      y = "Total number of medals")
```