My Portfolio

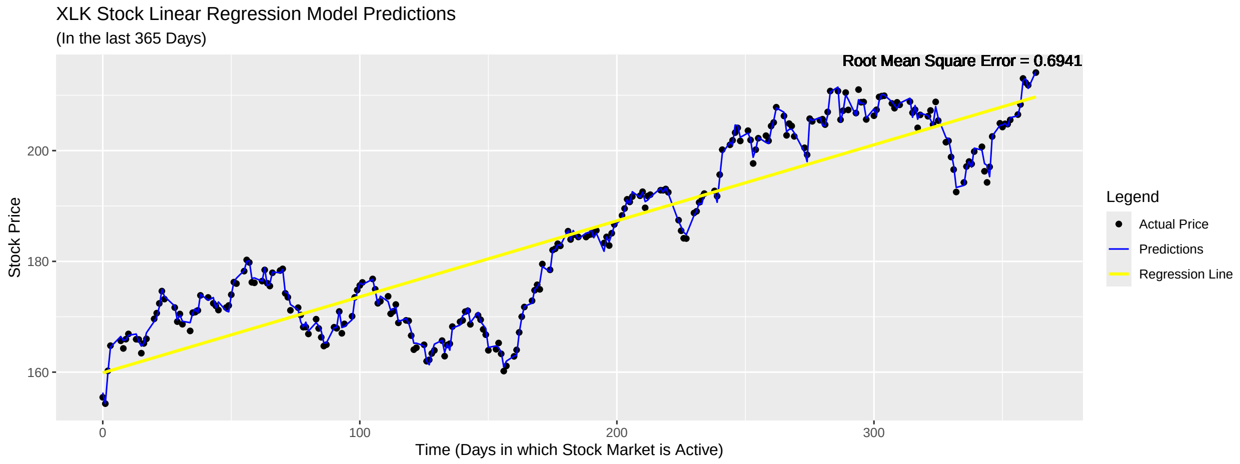
Andre Contreras

2023-2024

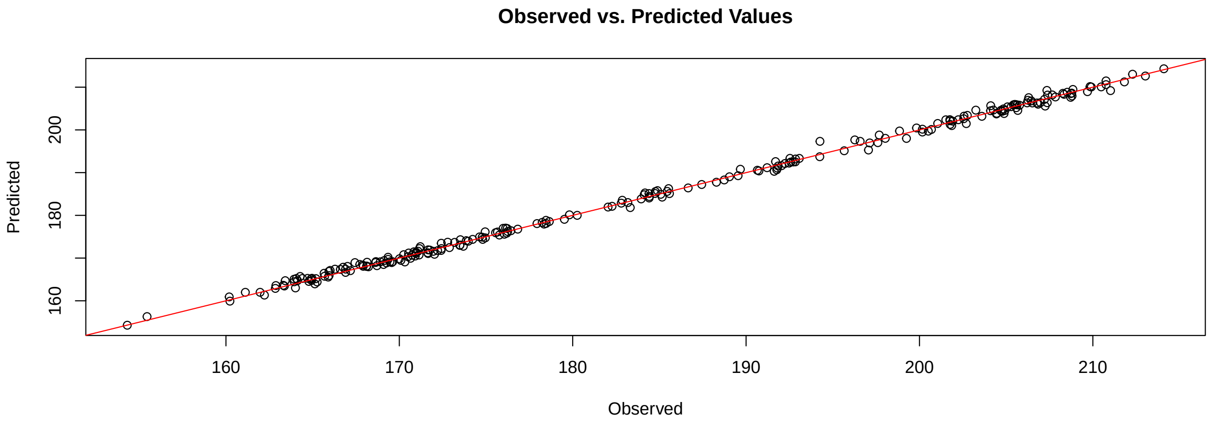
Linear Regression Machine Learning Model on XLK Stock

In the code below, I successfully train a Machine Learning model by gathering and manipulating the relevant data, splitting it into two sets (train & test), training a linear regression model, and testing it on the data set aside for testing. The resulting model displays a Root Mean Square Error of only 0.69, meaning that on average, my predictions are off by $0.69. I also included a regression line to show that on average, the stock price has gone up in the past year as well as a perfect fit line in the other plot.

```r
# 1. Read the CSV files into data frames
data <- read_csv("XLK dataset.csv")

data$Date <- as.Date(data$Date, format("%m/%d/%Y"))
data$Date <- as.numeric(data$Date)

# 2. Split the dataset into training and testing subsets
training_data <- data[1:5884, ]
testing_data <- data[5885:6134, ]

# Train Model using linear regression model ('.' indicates all other variables will be used as predicto
model <- lm(Close ~ ., data = training_data)

predictions <- predict(model, newdata = testing_data)

# Evaluate performance using Root Mean Square Error
rmse <- sqrt(mean((testing_data$Close - predictions)^2))

# Plot
ggplot(testing_data, aes(Date-19500, Close)) +
  geom_point(aes(col = "Actual Price")) +  # Testing Data points (Actual price values)
  geom_line(aes(y = predictions, col = "Predictions")) +  # Add a line for model predictions
  geom_smooth(method = "lm", aes(col = "Regression Line"), se = FALSE) + # Regression Line
  labs(x = "Time (Days in which Stock Market is Active)", y = "Stock Price", title = "XLK Stock Linear 
  geom_text(aes(label = paste("Root Mean Square Error =", round(rmse, 4)), x = Inf, y = Inf), hjust = 1
  scale_color_manual(name = "Legend", values = c("black", "blue", "yellow"))
```



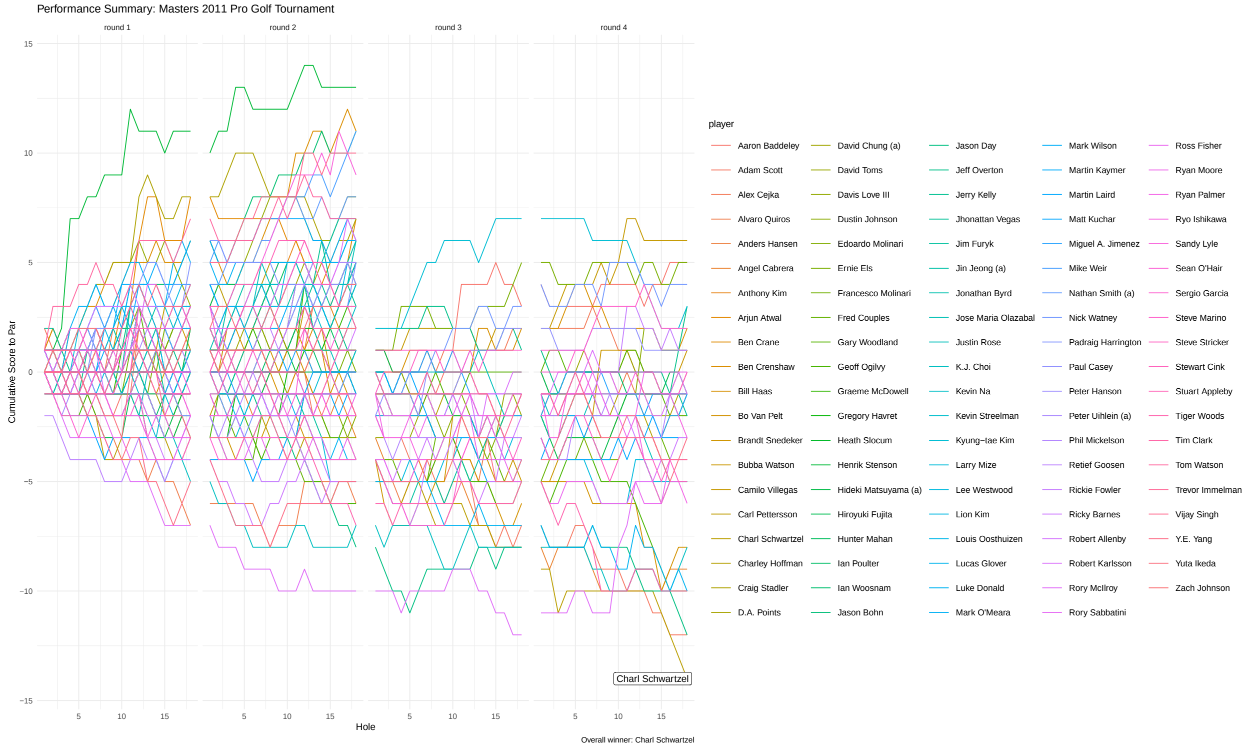
```r
plot(testing_data$Close, predictions,
     xlab = "Observed", ylab = "Predicted",
     main = "Observed vs. Predicted Values",
     col = "black")
# Red "Perfect" Line
abline(0, 1, col = "red")
```



2011 Masters Golf Tournament Project

The plot below is a line graph I created visualizing the summary of the Masters 2011 Pro Golf Tournament, along with the performance of each golfer and the overall winner of the comptetition (Charl Schwartzel).

```r
# Binds the rows of round1, round2,round3, and round4; specifies the name of the new column that will b
rounds <- bind_rows(round1, round2, round3, round4, .id = "round")


scorecard <- rounds %>%
  pivot_longer(cols = "1":"18", names_to = "hole", values_to = "score") %>%
  mutate(round = as.integer(round), hole = as.integer(hole), score = as.integer(score))


performance <- scorecard %>%
  left_join(course, by = "hole") %>%
  mutate(difference_to_par = score - par) %>%
  group_by(player) %>%
  mutate(cumulative_to_par = cumsum(difference_to_par)) %>%
  ungroup() %>%
  select(player, round, hole, difference_to_par, cumulative_to_par)

winner <- performance %>%
  filter(round == 4) %>%
  top_n(1, wt = -cumulative_to_par)

ggplot(performance) +
  geom_line(aes(x = hole, y = cumulative_to_par, col = player)) +
  facet_grid(. ~ round, labeller = labeller(round = c("1" = "round 1", "2" = "round 2", "3" = "round 3"
  geom_label(aes(x = hole - 4, y = cumulative_to_par, label = player), data = winner) +
  labs(title = "Performance Summary: Masters 2011 Pro Golf Tournament", x = "Hole", y = "Cumulative Scor
  theme_minimal() +
  theme(legend.position = "right", legend.key.size = unit(c(1, 1), "cm"), legend.text = element_text(si
```

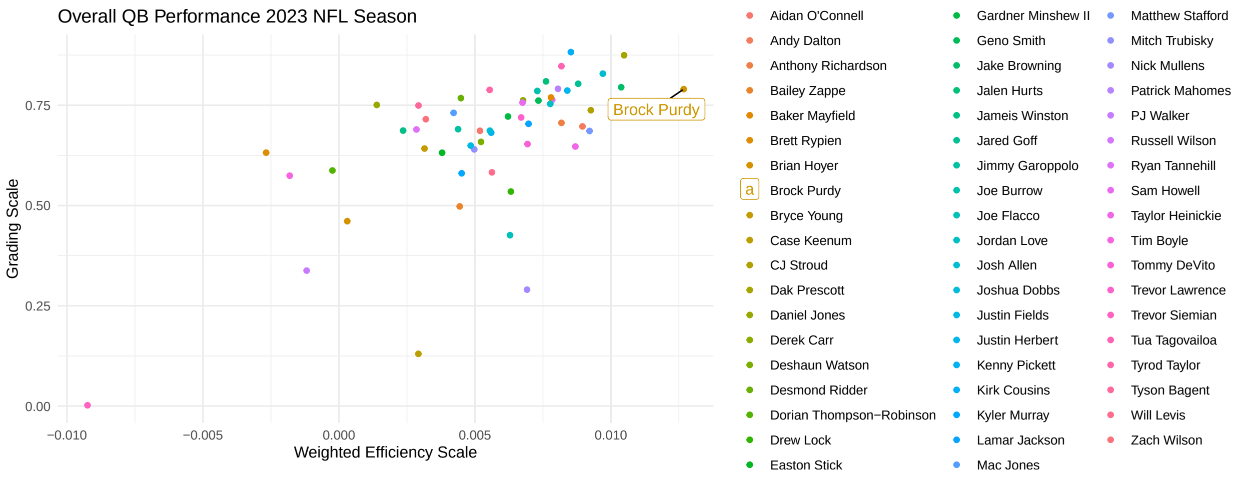


NFL 2023 QB Performance Project

Overall QB Performance:

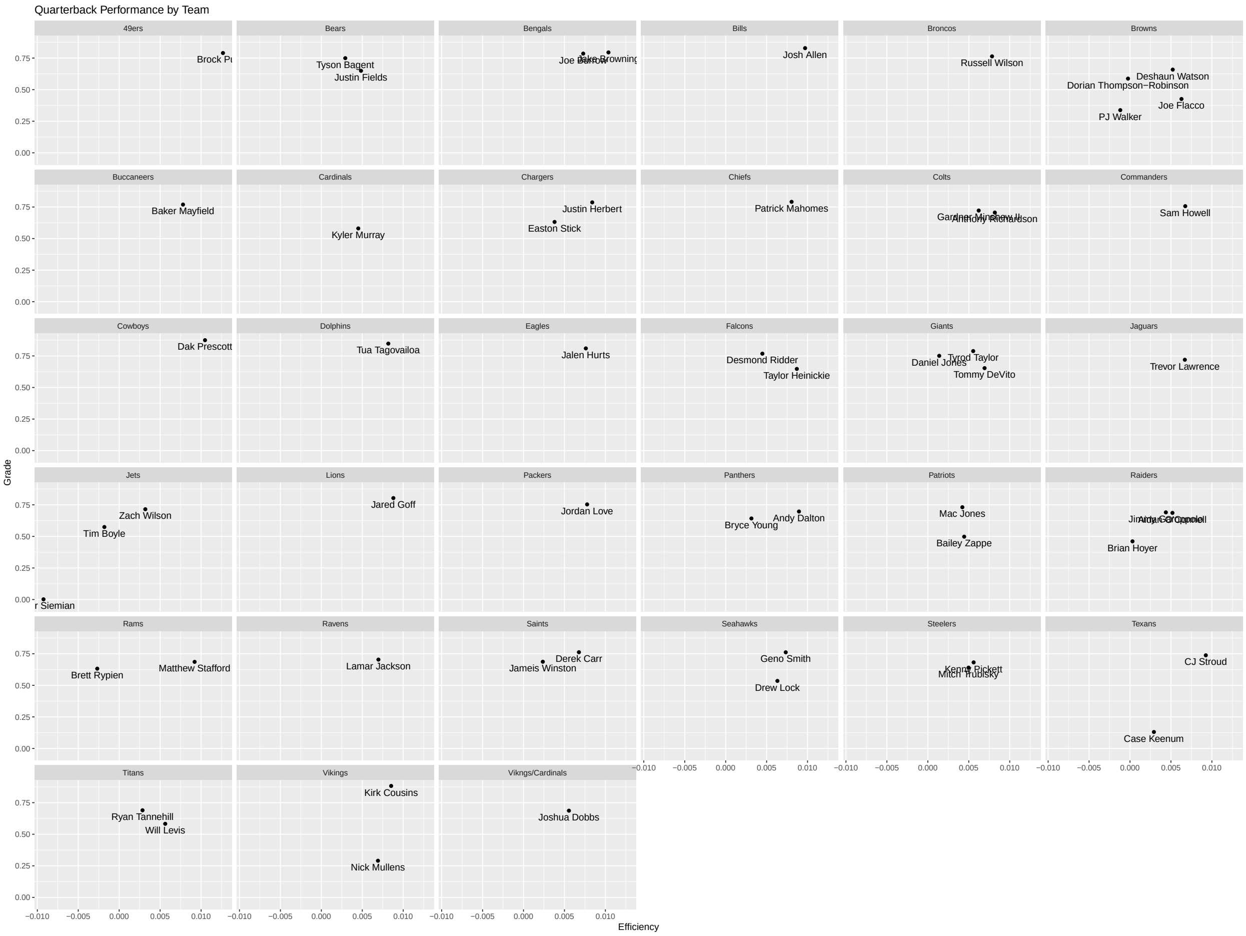
I created a spreadsheet of all the NFL Quarterbacks that played a minimum of one game in the 2023-2024 season and compiled all of their advanced statistics. I then developed a formula in Excel aiming to grade their efficiency, MVP rating, and overall grade, and transferred the file into R to visualize my findings. The name displayed is my MVP according to my grading system and efficiency formula

```
qbdata <- read_excel("C:\\Users\\ajcon\\Downloads\\Portfolio\\qbdata.xlsx")

mvp <- qbdata %>%
  top_n(1, qbdata$Grade*qbdata$Efficiency)

ggplot(qbdata, aes(x = Efficiency, y = Grade, col = QB)) +
  geom_point() +
  geom_label_repel(data = mvp, aes(label = QB), nudge_x = -0.001, nudge_y = -0.05, segment.color = "bla
  labs(x = "Weighted Efficiency Scale", y = "Grading Scale", title = "Overall QB Performance 2023 NFL Se
  theme_minimal() +
  theme(legend.position = "right")
```



QB Performance by Team:

```
ggplot(qbdata, aes(x = Efficiency, y = Grade)) +
  geom_point() +
  geom_text(aes(label = QB), nudge_x = 0, nudge_y = -0.05) +
  facet_wrap(. ~ Team) +
  labs(x = "Efficiency", y = "Grade", title = "Quarterback Performance by Team")
```



A Data-led Look into the History of the Olympics

After gaining access to mulitple datasets of the Olympics containing every instance throughout every competition since the inaugural season back in 1896 (Greece) up until the 2016 Games in Brazil, I decided my free time would be well spent answering a couple of questions I, like many others (I think), have been wondering:

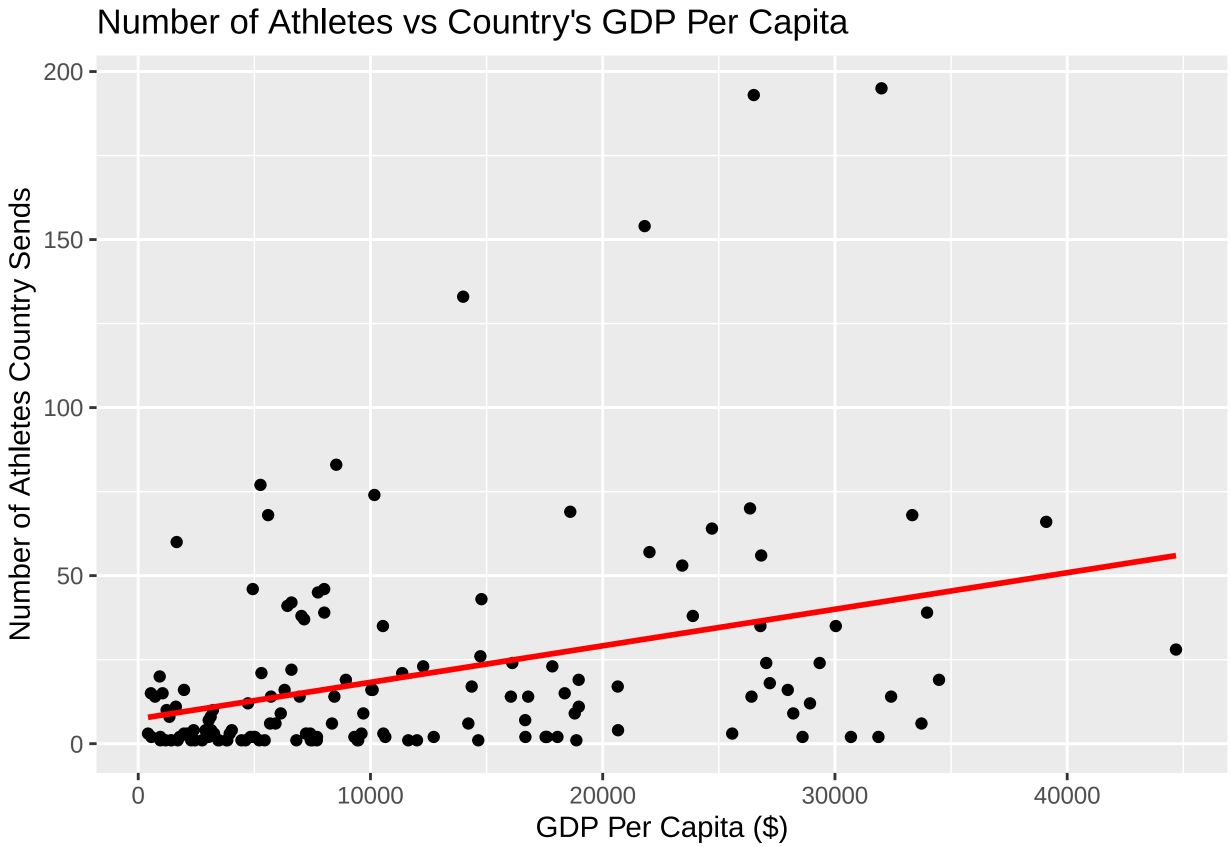
1. Does the economical stability of a country affect the number of athletes it sends to the olympics and the number of medals it wins?

2. Does hosting the olympics correlate to winning more medals that year?

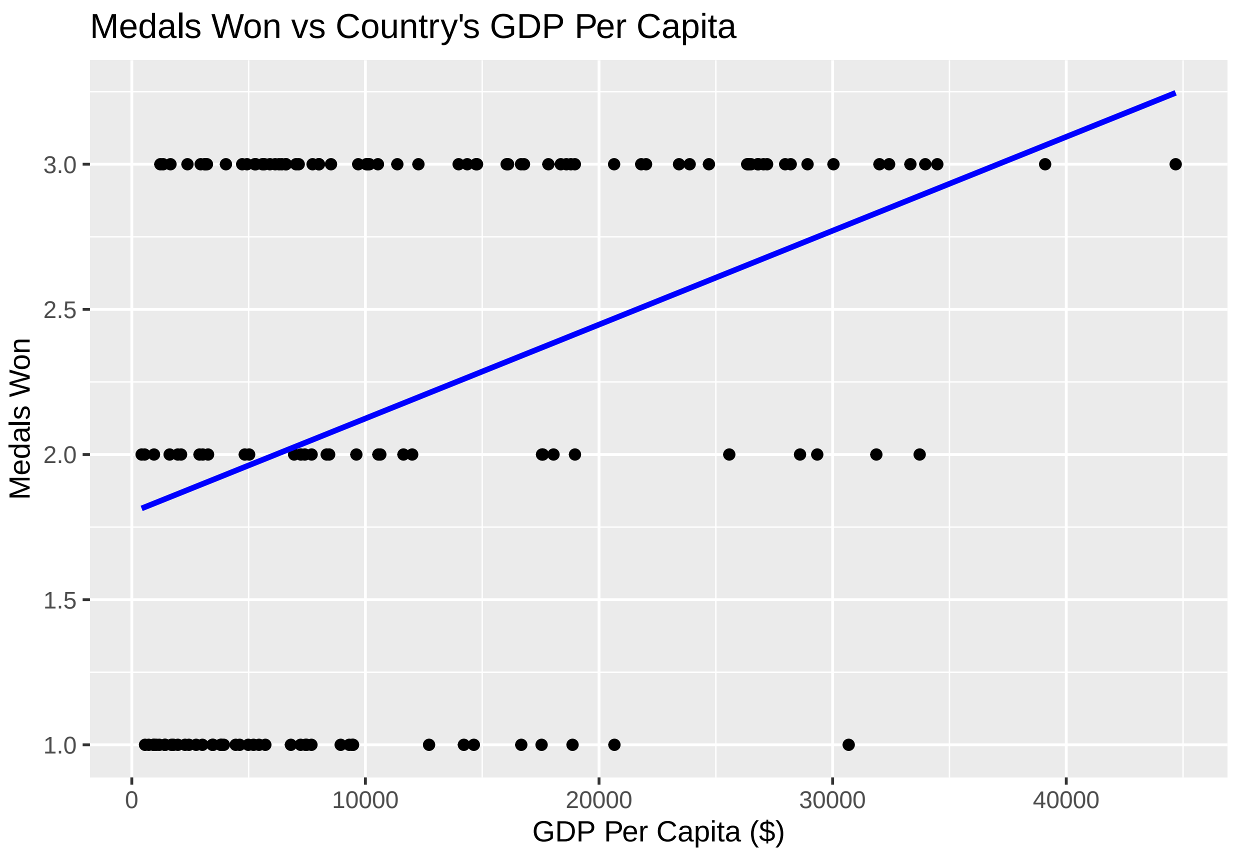
Part I

Below are the results I found for the first question, along with the code I wrote to filter and manipulate the data so I can visualize it in a more effective manner.

```r
athlete_events <- read_csv(
   file = 'athlete_events.csv',
   col_types = cols(ID = 'i', Age = 'i', Height = 'i', Year = 'i')
)

nearest_year <- function(olympics_year) {
  gapminder_year <- seq(1952, 2007, by = 5)
  nearest_year <- gapminder_year[which.min(abs(olympics_year - gapminder_year))]
  return(nearest_year)
}

olympics_data <-
  athlete_events %>%
  filter(!is.na(Medal)) %>%
  count(Games, Event, NOC, Medal, Team, Year, Name) %>%
  mutate(year = nearest_year(Year))

country_money <- gapminder %>%
  group_by(country) %>%
  select(country, year, gdpPercap)


athletes_by_country_year <- olympics_data %>%
  group_by(Team, Year) %>%
  summarise(Total_Athletes = n_distinct(Name), .groups = 'drop')

medals_by_country_year <- olympics_data %>%
  group_by(Team, Year) %>%
  summarise(Total_Medals = n_distinct(Medal), .groups = 'drop')

joined_data_athletes <- inner_join(athletes_by_country_year, country_money, by = c("Year" = "year", "Tea
  filter(!is.na(gdpPercap))

joined_data_medals <- inner_join(medals_by_country_year, country_money, by = c("Year" = "year", "Team" =
  filter(!is.na(gdpPercap))

ggplot(joined_data_athletes, aes(x = gdpPercap , y = Total_Athletes)) +
  geom_point() +
  geom_smooth(method = "lm", color = "red", se = FALSE) + #Plotting the athlete correlation
  labs(title = "Number of Athletes vs Country's GDP Per Capita", x = "GDP Per Capita ($)", y = "Number
```



```r
ggplot(joined_data_medals, aes(x = gdpPercap , y = Total_Medals)) +
  geom_point() +
  geom_smooth(method = "lm", color = "blue", se = FALSE) + #lm creates a smooth line to show a clear re
    labs(title = "Medals Won vs Country's GDP Per Capita", x = "GDP Per Capita ($)", y = "Medals Won")
```

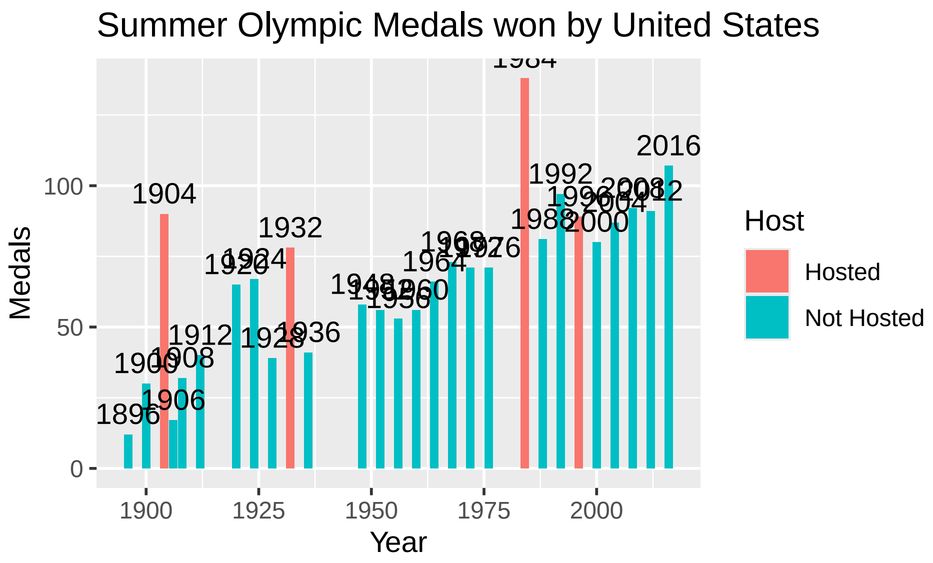


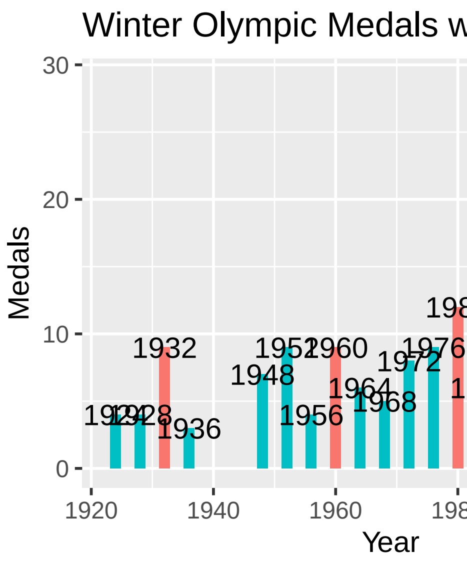
As we can see, there is in fact a positive correlation between a country's gdp per capita and the number of medals and athletes a country has. This means that the higher the gdp is, the more medals it wins and more athletes it sends to the Olympics.

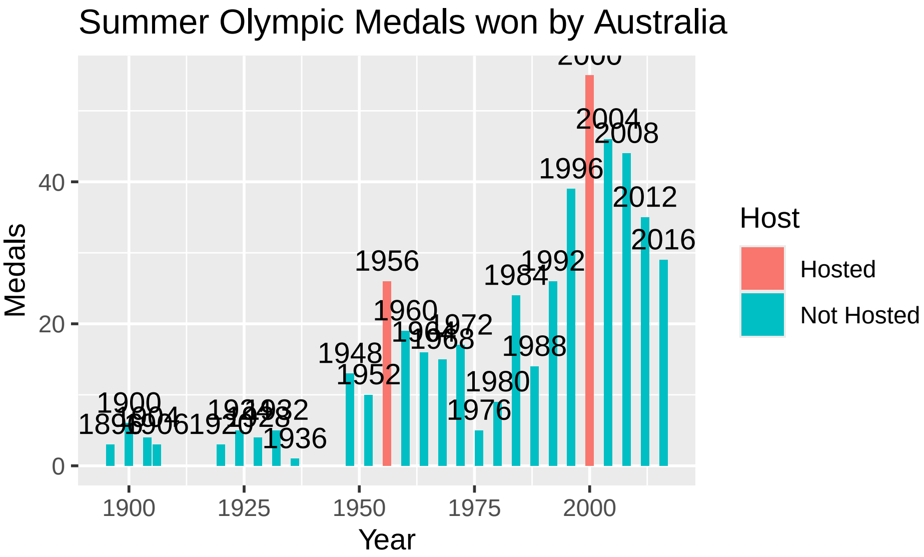
Part II

For the second question… I started by joining data sets together and creating a function that will filter the joint dataset for each country and in each of the seasons: determine whether they hosted or not. The function also displays a plot to compare the amount of medals that country won when they hosted vs when they did not. We will then compare and draw reasonable conclusions by creating a histogram containing the average number of medals all countries combined have won when they host vs in the competitions before.

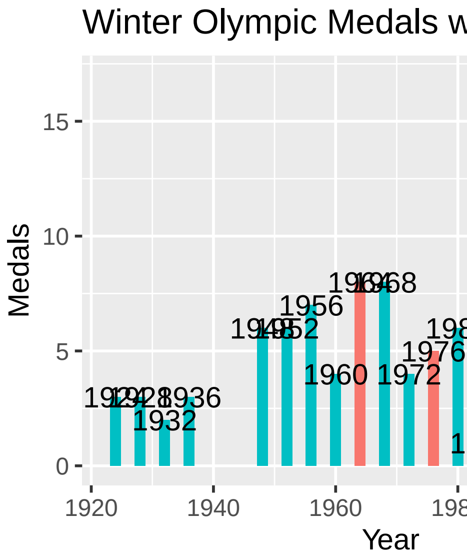
```
generate_country_medals_plot <- function(country_code, country_name, summer_hosts, winter_hosts) {

  # SUMMER
  summer_plot <- NULL

  if (length(summer_hosts) > 0) {
    summer_medals <- data %>%
      filter(NOC == country_code & !is.na(Medal) & Season == "Summer" & Year %in% c(1896:2016)) %>%
      distinct(Year, Event) %>%
      group_by(Year) %>%
      summarise(Medal_Count = n())

    summer_medals$Host <- ifelse(summer_medals$Year %in% summer_hosts, "Hosted", "Not Hosted")
    summer_plot <- ggplot(summer_medals, aes(x = Year, y = Medal_Count, fill = Host)) +
      geom_bar(stat = "identity", position = "dodge") +
      geom_text(aes(label = Year), vjust = -0.5, position = position_dodge(width = 0.9)) +
      labs(x = "Year", y = "Medals", title = paste("Summer Olympic Medals won by", country_name))
  }

  # WINTER
  winter_plot <- NULL

  if (length(winter_hosts) > 0) {
    winter_medals <- data %>%
      filter(NOC == country_code & !is.na(Medal) & Season == "Winter" & Year %in% c(1896:2016)) %>%
      distinct(Year, Event) %>%
      group_by(Year) %>%
      summarise(Medal_Count = n())

    winter_medals$Host <- ifelse(winter_medals$Year %in% winter_hosts, "Hosted", "Not Hosted")

    winter_plot <- ggplot(winter_medals, aes(x = Year, y = Medal_Count, fill = Host)) +
      geom_bar(stat = "identity", position = "dodge") +
      geom_text(aes(label = Year), position = position_dodge(width = 0.9)) +
      labs(x = "Year", y = "Medals", title = paste("Winter Olympic Medals won by", country_name))
  }

  list(summer_plot = summer_plot, winter_plot = winter_plot)
}
```



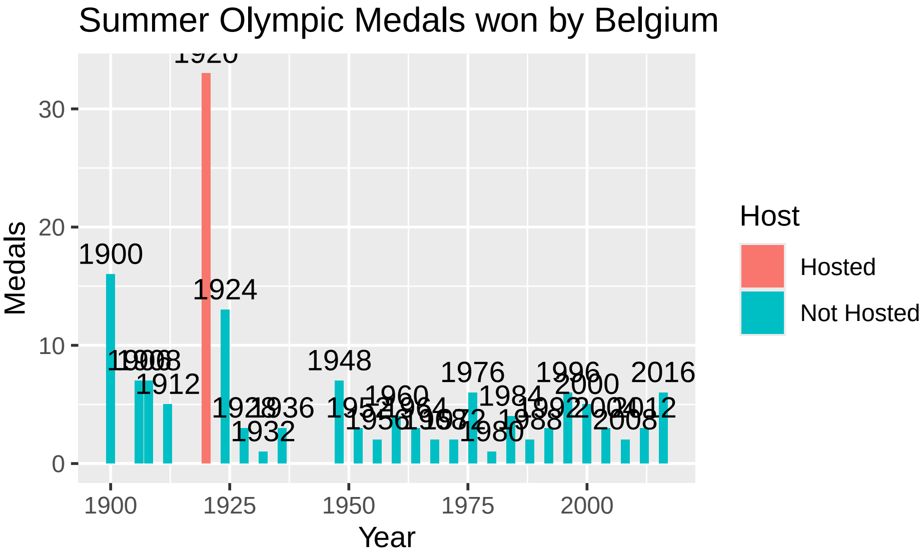




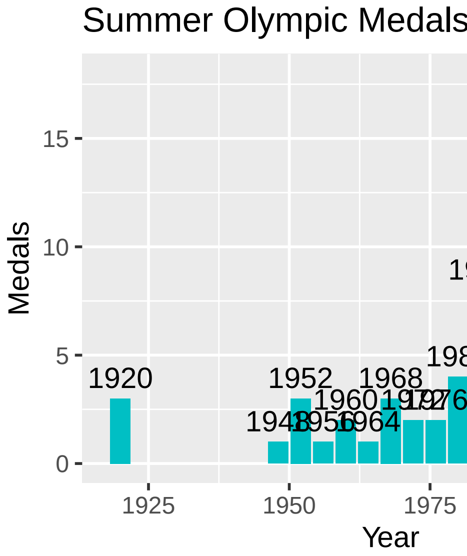
Summer Olympic Medals won by Australia



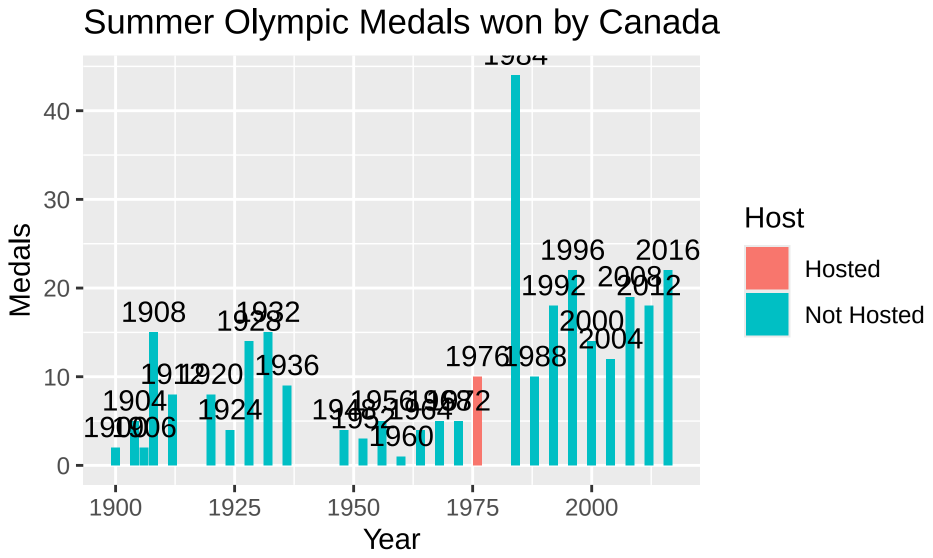
Winter Olympic Medals w

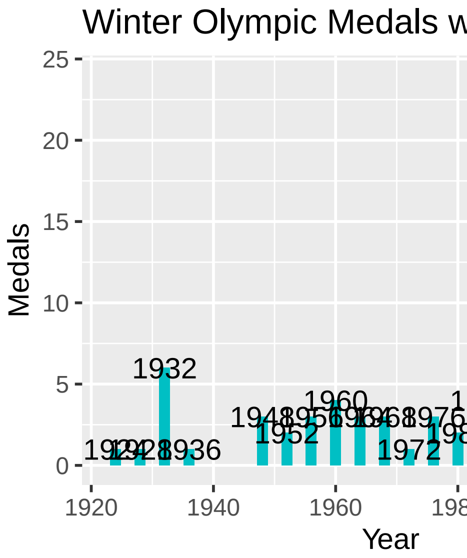


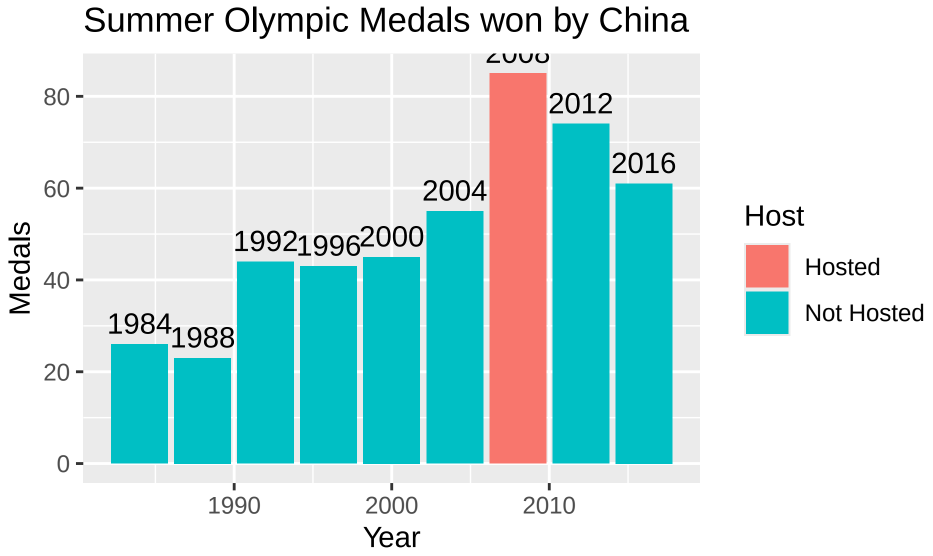
Summer Olympic Medals won by Belgium

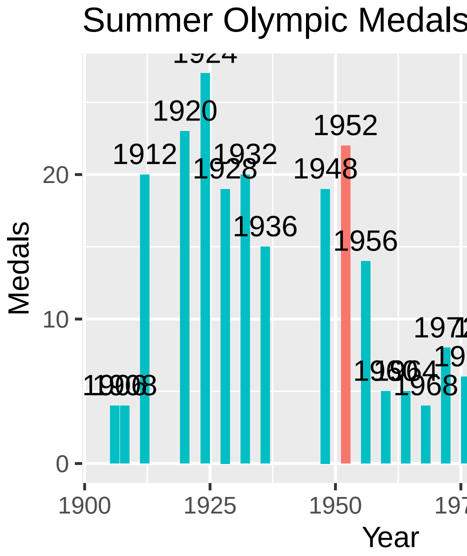


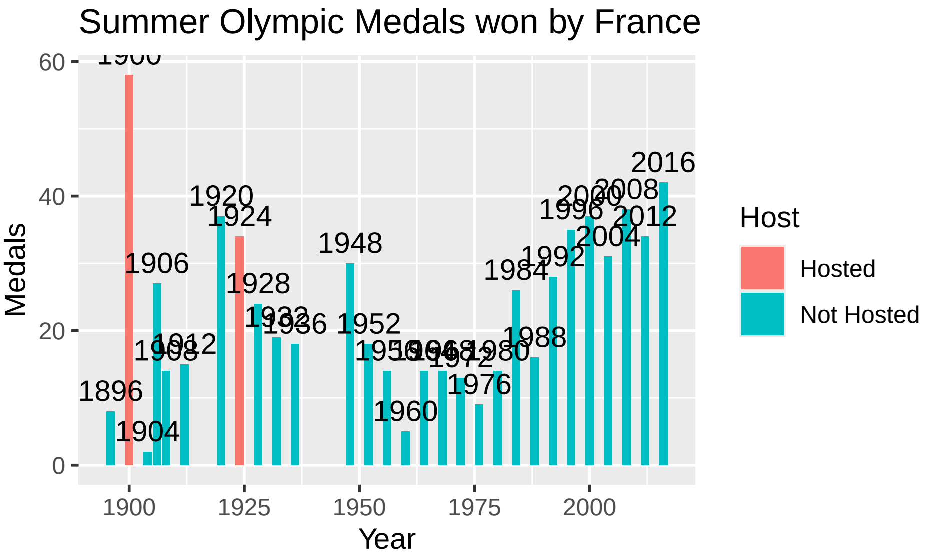
Summer Olympic Medals

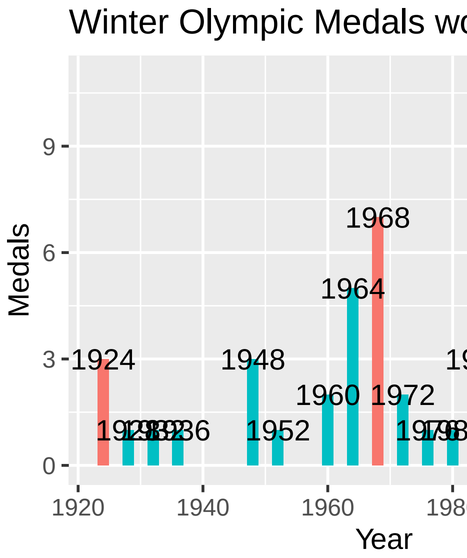

Summer Olympic Medals won by Canada
Medals
Year
0
10
20
30
40
1900
1925
1950
1975
2000
1900
1904
1906
1908
1912
1920
1924
1928
1932
1936
1948
1952
1956
1960
1964
1968
1972
1976
1984
1988
1992
1996
2000
2004
2008
2012
2016
Host
Hosted
Not Hosted

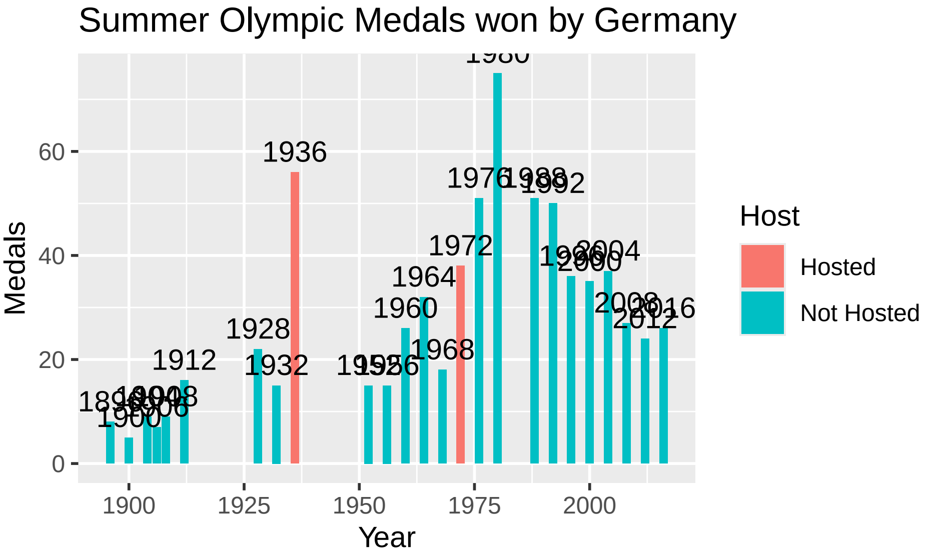

Winter Olympic Medals w
Medals
Year
0
5
10
15
20
25
1920
1940
1960
198
1924
1928
1932
1936
1948
1952
1956
1960
1964
1968
1972
1976

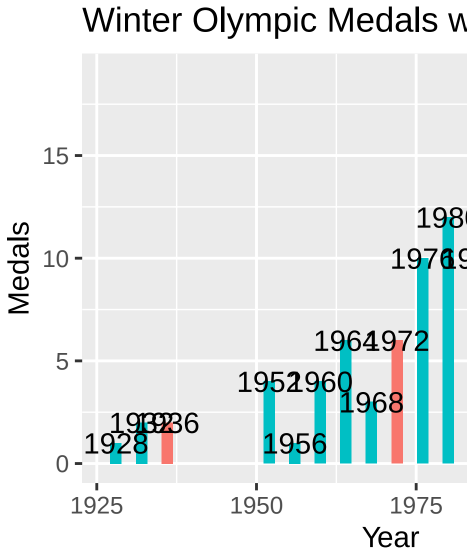

Summer Olympic Medals won by China
Medals
Year
0
20
40
60
80
1990
2000
2010
1984
1988
1992
1996
2000
2004
2008
2012
2016
Host
Hosted
Not Hosted

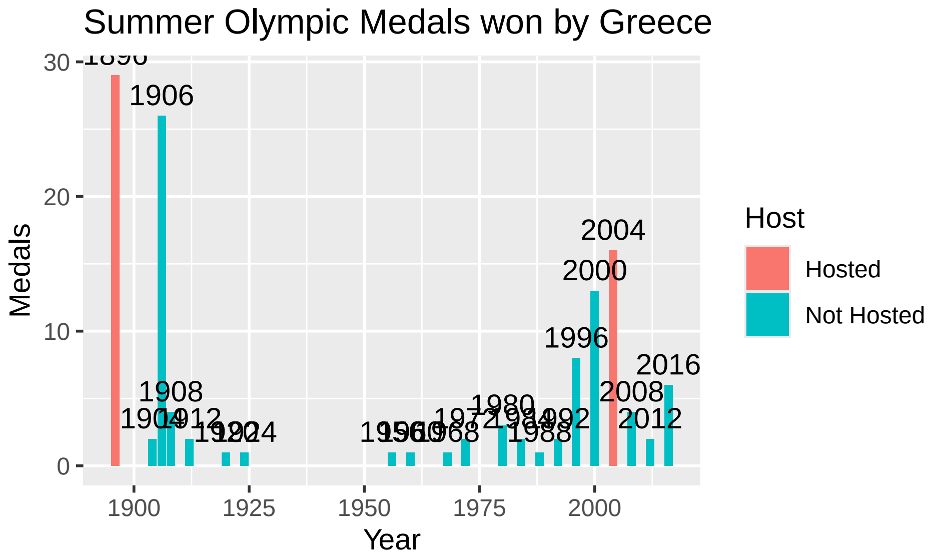

Summer Olympic Medals
Medals
Year
0
10
20
1900
1925
1950
197
1906
1908
1912
1920
1924
1928
1932
1936
1948
1952
1956
1960
1964
1968
1972

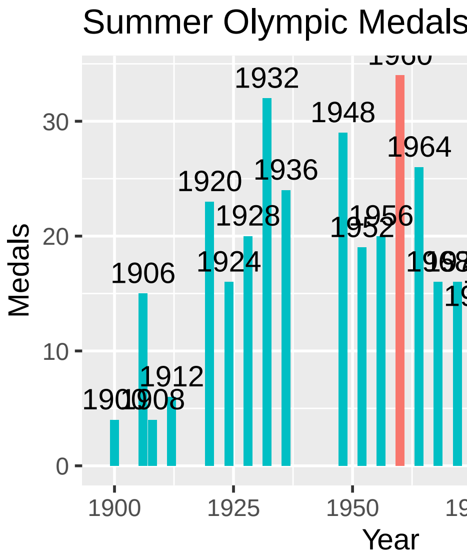

Summer Olympic Medals won by France
Medals
Year
0
20
40
60
1900
1925
1950
1975
2000
1896
1900
1904
1906
1908
1912
1920
1924
1928
1932
1936
1948
1952
1956
1960
1964
1968
1972
1976
1980
1984
1988
1992
1996
2000
2004
2008
2012
2016
Host
Hosted
Not Hosted

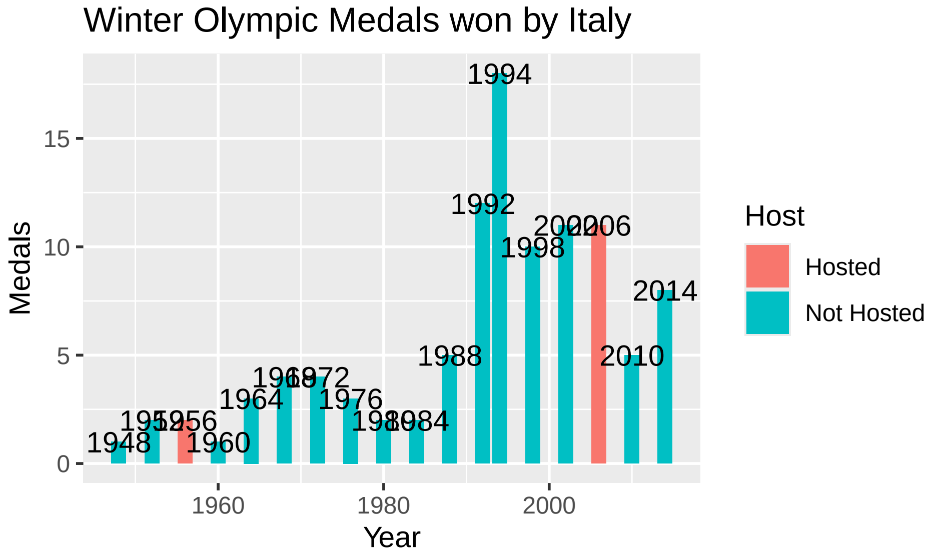

Winter Olympic Medals wo
Medals
Year
0
3
6
9
1920
1940
1960
198
1924
1928
1932
1936
1948
1952
1960
1964
1968
1972
1976
19

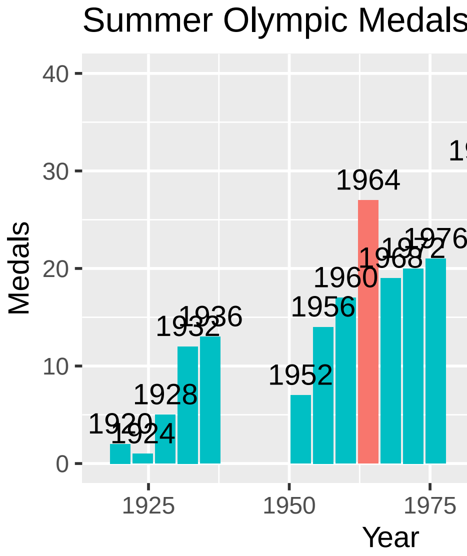

Summer Olympic Medals won by Germany
Medals
Year
0
20
40
60
1900
1925
1950
1975
2000
1896
1900
1904
1906
1908
1912
1928
1932
1936
1952
1956
1960
1964
1968
1972
1976
1980
1988
1992
1996
2000
2004
2008
2012
2016
Host
Hosted
Not Hosted

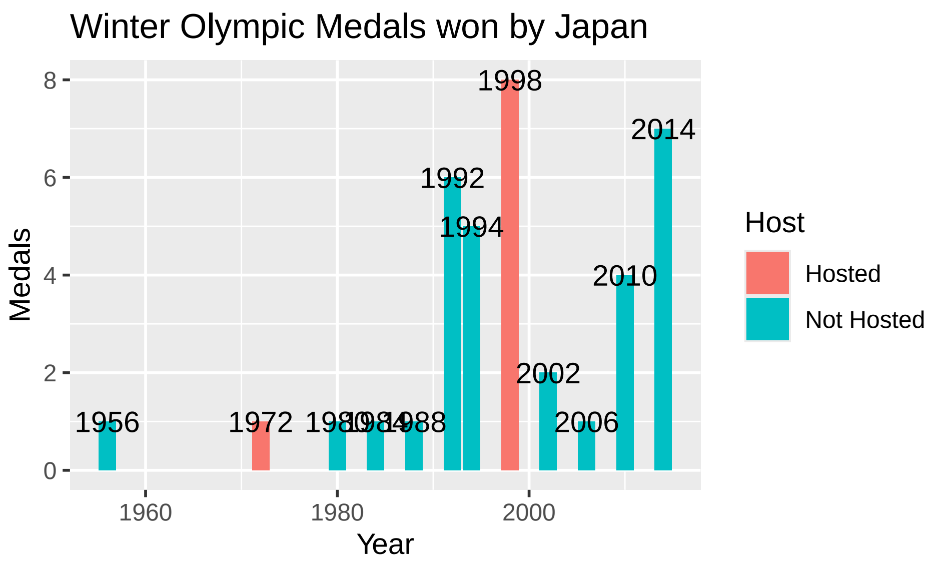

Winter Olympic Medals w
Medals
Year
0
5
10
15
1925
1950
1975
1928
1932
1936
1952
1956
1960
1964
1968
1972
1976
198

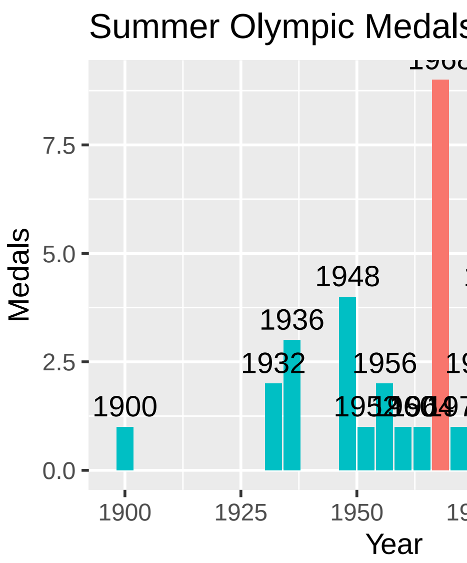

Summer Olympic Medals won by Greece

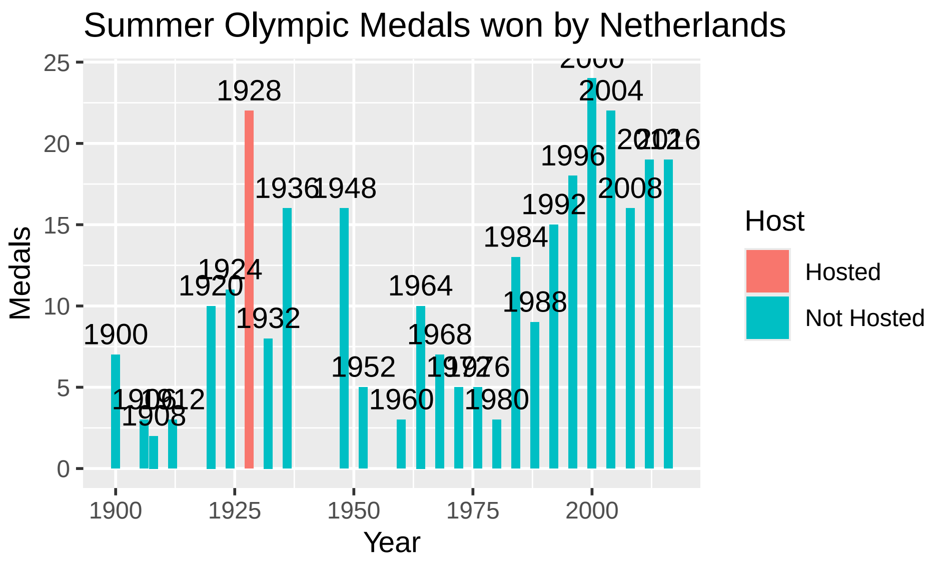

Summer Olympic Medals

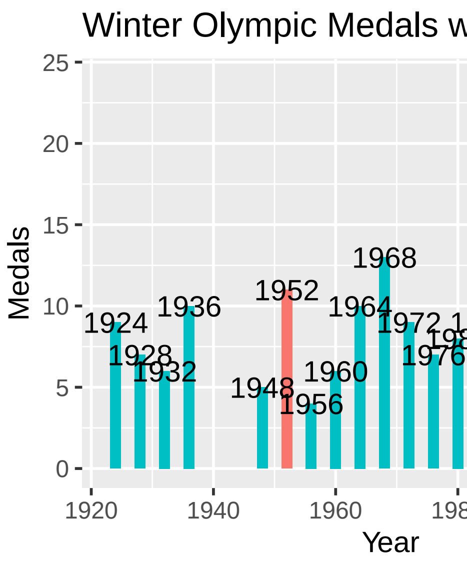

Winter Olympic Medals won by Italy

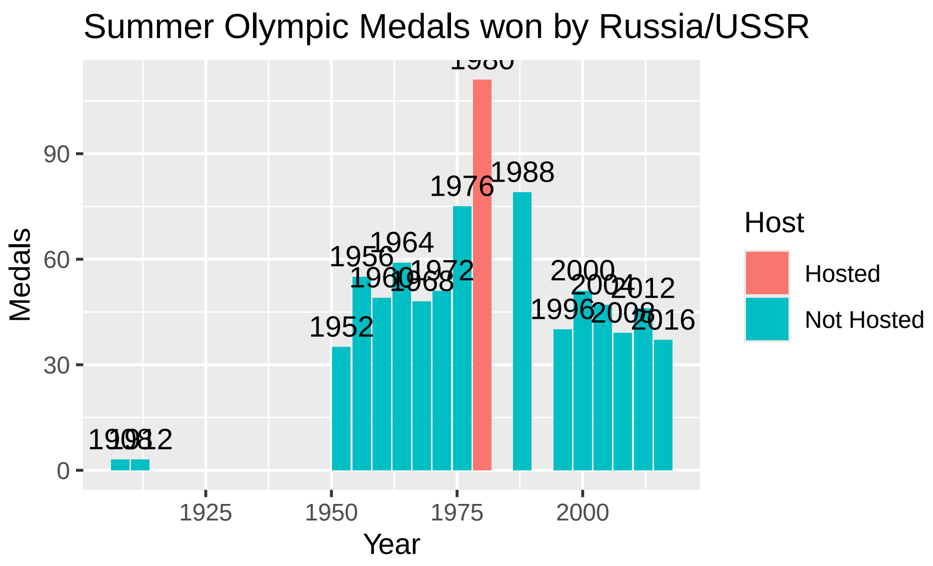

Summer Olympic Medals


Winter Olympic Medals won by Japan

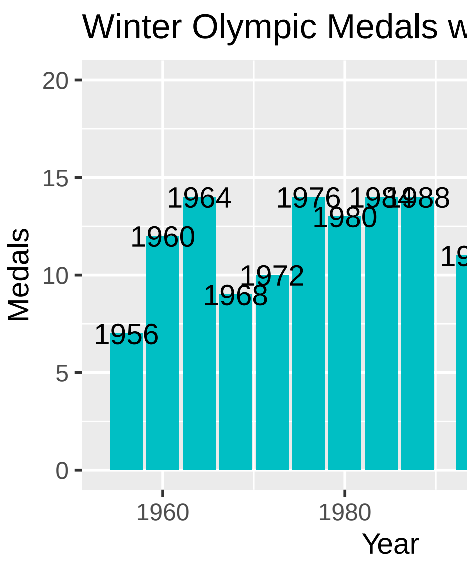

Summer Olympic Medals


Summer Olympic Medals won by Netherlands

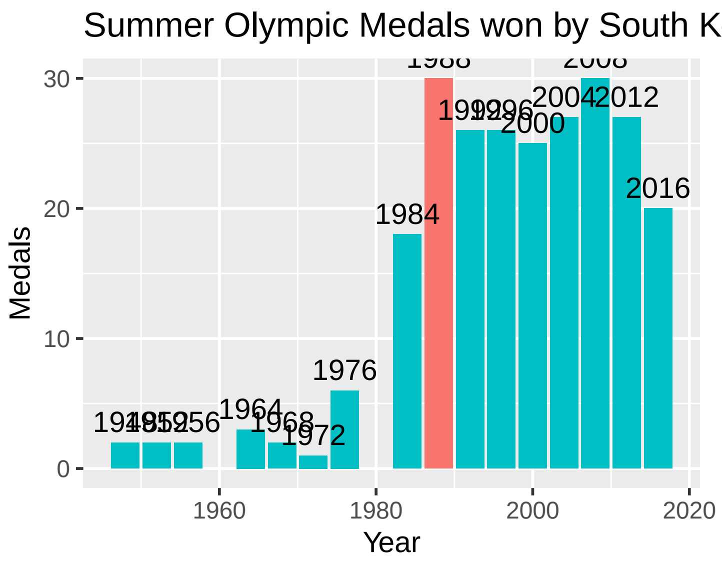

Winter Olympic Medals w



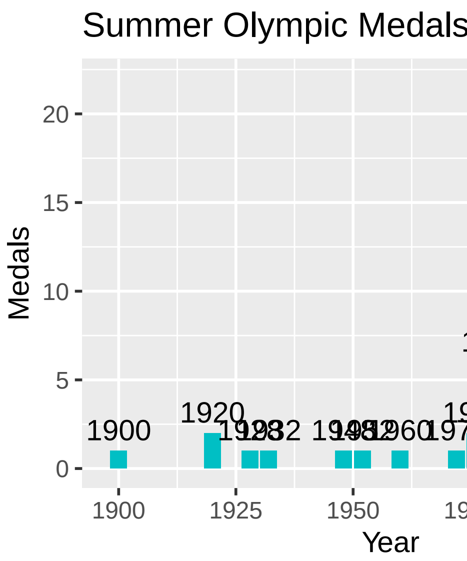
Summer Olympic Medals won by Russia/USSR



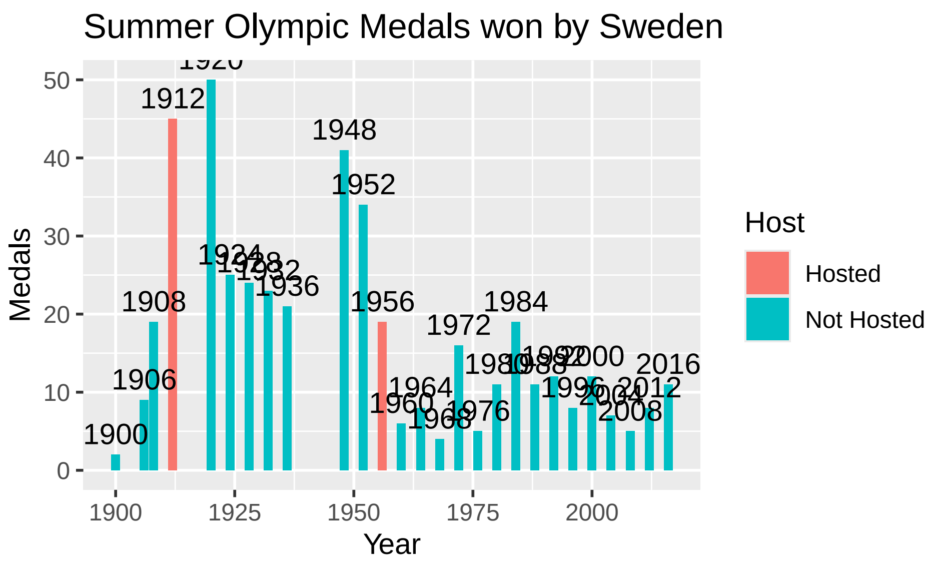
Winter Olympic Medals w



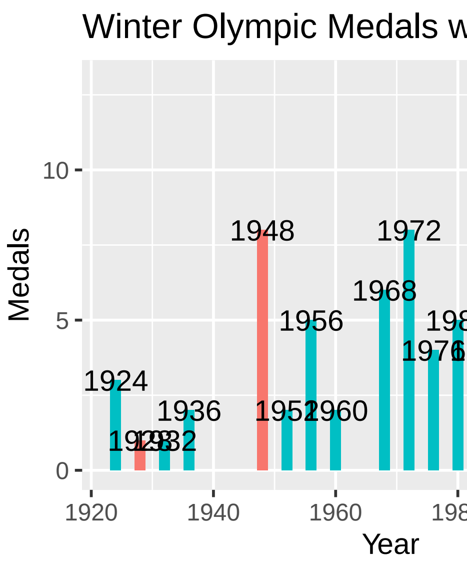
Summer Olympic Medals won by South Korea



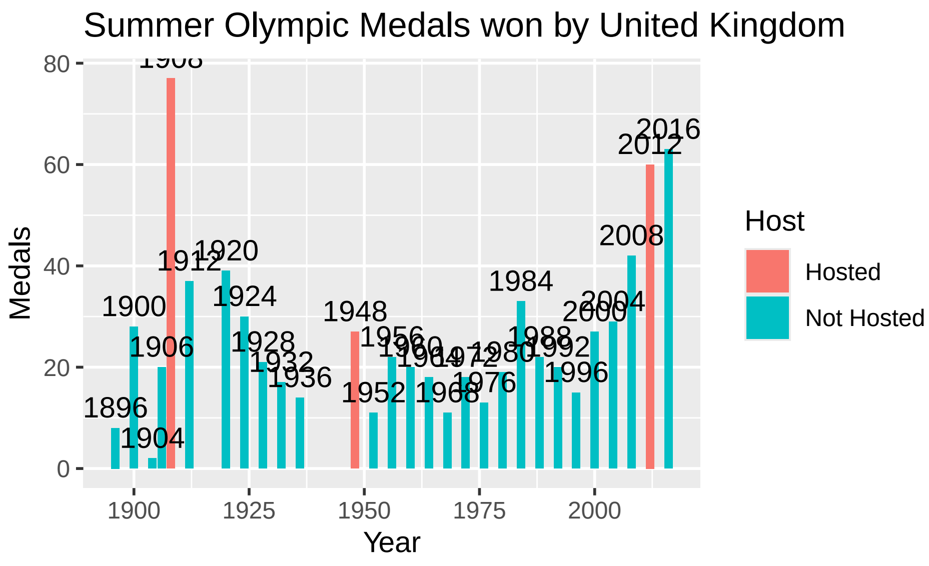
Summer Olympic Medals



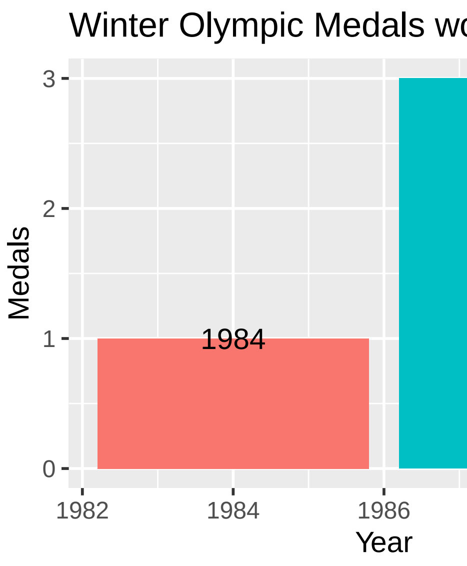
Summer Olympic Medals won by Sweden



Winter Olympic Medals w



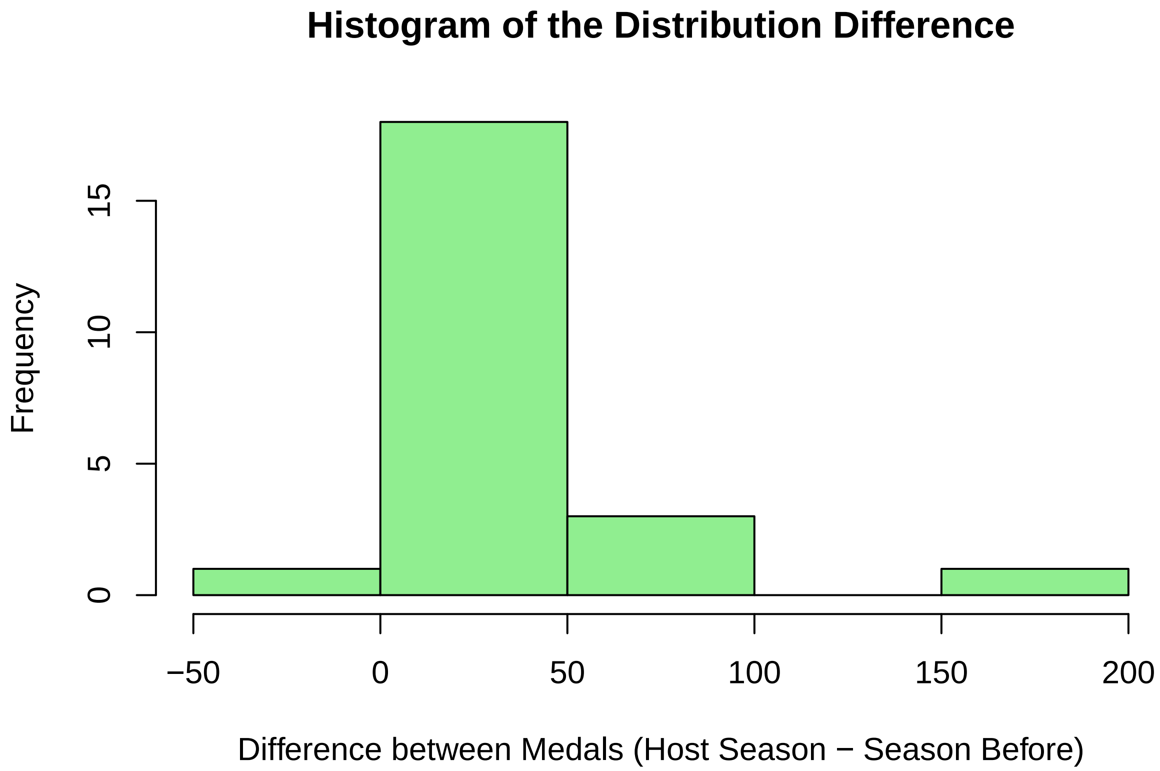
Summer Olympic Medals won by United Kingdom



Winter Olympic Medals wo

As stated earlier, I created a histogram of the difference of medals **(by subtracting the medals won when they host minus the medals won in the olympic season directly prior)** to draw a reasonable conclusion.

```
hist_data <- tibble(
  NOC = c("USA", "AUS", "AUT", "BEL", "BRA", "CAN", "CHN", "FIN", "FRA", "GER", "GRE", "ITA", "JPN", "ME
  Medals_Won_Host = c(451, 83, 13, 33, 18, 39, 85, 22, 109, 102, 45, 47, 36, 9, 22, 29, 130, 30, 22, 64
  Medals_Won_Year_Before = c(265, 49, 8, 5, 17, 29, 55, 19, 52, 39, 16, 33, 22, 1, 11, 18, 84, 18, 4, 53

new <- hist_data %>%
  mutate(Distribution_difference = Medals_Won_Host - Medals_Won_Year_Before)

hist(new$Distribution_difference, xlab = "Difference between Medals (Host Season - Season Before)", mair
```



We can see there is a positive host effect country on the amount of medals won when a country hosts the olympics vs when they don't because there is an overall positive difference.

I am also currently in the process of completing my fifth project - Analyzing the passing networks of the undefeated 2003-2004 Arsenal FC soccer team to use Graph Theory concepts and techniques in hopes of optimizing strategies that will lead to better results and more wins in present day soccer.