```r
# Set up
setwd("~/Desktop/STAT 38191 Project")
library(tidyverse)

# Data import
olympic = read_csv("athlete_events.csv")
```

서론

1. 자료 개괄 설명

```r
olympic
```

```
## # A tibble: 271,116 x 15
##       ID Name  Sex     Age Height Weight Team  NOC   Games  Year Season
##    <int> <chr> <fct> <int>  <dbl>  <dbl> <chr> <chr> <chr> <int> <fct>
##  1     1 A Di… M        24    180     80 China CHN   1992…  1992 Summer
##  2     2 A La… M        23    170     60 China CHN   2012…  2012 Summer
##  3     3 Gunn… M        24     NA     NA Denm… DEN   1920…  1920 Summer
##  4     4 Edga… M        34     NA     NA Denm… DEN   1900…  1900 Summer
##  5     5 Chri… F        21    185     82 Neth… NED   1988…  1988 Winter
##  6     5 Chri… F        21    185     82 Neth… NED   1988…  1988 Winter
##  7     5 Chri… F        25    185     82 Neth… NED   1992…  1992 Winter
##  8     5 Chri… F        25    185     82 Neth… NED   1992…  1992 Winter
##  9     5 Chri… F        27    185     82 Neth… NED   1994…  1994 Winter
## 10     5 Chri… F        27    185     82 Neth… NED   1994…  1994 Winter
## # … with 271,106 more rows, and 4 more variables: City <chr>, Sport <chr>,
## #   Event <chr>, Medal <fct>
```

```r
summary(olympic)
```

```
##        ID            Name           Sex             Age
##  Min.   :     1   Length:271116      M:196594   Min.   :10.00
##  1st Qu.: 34643   Class :character   F: 74522   1st Qu.:21.00
##  Median : 68205   Mode  :character              Median :24.00
##  Mean   : 68249                                 Mean   :25.56
##  3rd Qu.:102097                                 3rd Qu.:28.00
##  Max.   :135571                                 Max.   :97.00
##                                                 NA's   :9474
##      Height         Weight         Team              NOC
##  Min.   :127.0   Min.   : 25.0   Length:271116      Length:271116
##  1st Qu.:168.0   1st Qu.: 60.0   Class :character   Class :character
##  Median :175.0   Median : 70.0   Mode  :character   Mode  :character
##  Mean   :175.3   Mean   : 70.7
##  3rd Qu.:183.0   3rd Qu.: 79.0
##  Max.   :226.0   Max.   :214.0
##  NA's   :60171   NA's   :62875
##     Games               Year         Season          City
##  Length:271116      Min.   :1896   Summer:222552   Length:271116
```

```
##  Class :character   1st Qu.:1960   Winter: 48564   Class :character
##  Mode  :character   Median :1988                   Mode  :character
##                     Mean   :1978
##                     3rd Qu.:2002
##                     Max.   :2016
##
##     Sport              Event              Medal
##  Length:271116      Length:271116      Gold  : 13372
##  Class :character   Class :character   Bronze: 13295
##  Mode  :character   Mode  :character   Silver: 13116
##                                        NA's  :231333
```

1896 년 아테네 올림픽부터 2016 년 리오 올림픽까지 올림픽의 역사를 담고 있는 자료이다.[1] 1896 년 첫 올림픽이 개최된 이래 1992 년까지는 하계와 동계 올림픽 모두 4 년 주기로 같은 연도에 개최되었지만 그 이후 동계 올림픽은 1994 년부터 4 년 주기로, 하계 올림픽은 1996 년을 시작으로 4 년 주기로 열리게 된다.

자료는 총 271116 개의 관측값과 15 개의 변수를 가지며 ID, Age, Year 는 이산형 변수, Sex, Season, Medal 은 범주형 변수, 나머지 변수들은 문자형 변수로 파싱했다. 해당 자료에서 각 관측값은 한 종목에 출전한 선수 개개인에 해당한다. 즉, 한 선수가 서로 다른 종목에 출전하였다면 서로 다른 관측값을 가진다. 위 자료에서 총 출전 인원은 ID 의 count 값으로 보아야 한다. 위의 summary 를 참고하면, 지금까지 총 135571 명이 경기에 출전했음을 알 수 있다.

아래는 자료의 각 변수를 설명하는 표이다.

column	value	column2	value2
ID	선수의 고유 숫자	**Games**	연도 및 계절
Name	선수명	**Year**	연도
Sex	성별 (M, F)	**Season**	계절 (Summer, Winter)
Age	나이	**City**	개최 도시
Height	키 (단위 cm)	**Sport**	종목
Weight	몸무게 (단위 kg)	**Event**	세부종목
Team	출전 팀명	**Medal**	메달 (Gold, Silver, Bronze, NA)
NOC	IOC 국가 코드		

[1] 자료 출처: *https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results*

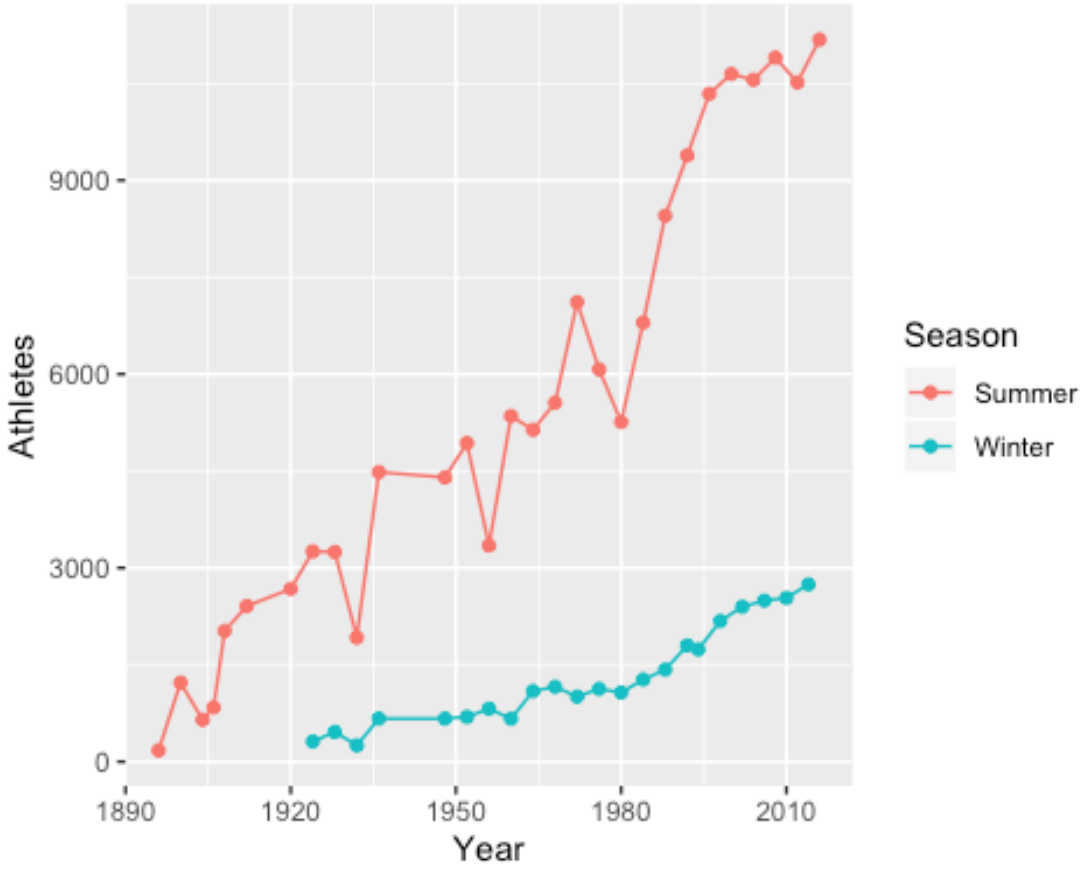
```
# check missing values
olympic %>% is.na() %>% colSums()

##     ID   Name    Sex    Age Height Weight   Team    NOC  Games   Year
##      0      0      0   9474  60171  62875      0      0      0      0
## Season   City  Sport  Event  Medal
##      0      0      0      0 231333
```

본격적인 분석에 앞서 데이터의 결측값을 살펴보면 Age, Height, Weight, Medal 총 4 개의 feature 에 결측값이 존재함을 알 수 있다. 이때 Medal 의 대부분은 결측값임을 알 수 있는데 각 경기 당 메달의 개수가 제한되어 있기 때문으로 추측할 수 있다.

2. 전반적인 시계열 분석 및 역사적 뒷배경

a) 연도별 출전 선수

```
olympics %>% group_by(Year, Season) %>% summarize(Athletes = n_distinct(ID)) %
>% ggplot(aes(Year, Athletes, color = Season)) + geom_point() + geom_line()
```



그래프를 살펴보면 가장 눈에 띄는 것은 출전 선수의 수가 대폭 감소하는 1935 년경, 1955 년경과 1980 년경의 점들과 1920 년 직전과 1935 ~ 1950 년경

연속한 두 점을 잇는 선의 길이가 다른 선들에 비해 길다는 것이다. 이 두 가지 두드러지는 특징의 원인은 역사적 사실을 감안하여 이해할 수 있다.

```
olympics %>% count(Year) %>% filter(Year %in% c(1925:1940))
```

```
## # A tibble: 3 x 2
##    Year     n
##   <int> <int>
## 1  1928  5574
## 2  1932  3321
## 3  1936  7401
```

```
olympics %>% count(Year) %>% filter(Year %in% c(1950:1965))
```

```
## # A tibble: 4 x 2
##    Year     n
##   <int> <int>
## 1  1952  9358
## 2  1956  6434
## 3  1960  9235
## 4  1964  9480
```

```
olympics %>% count(Year) %>% filter(Year %in% c(1970:1990))
```

```
## # A tibble: 5 x 2
##    Year     n
##   <int> <int>
## 1  1972 11959
## 2  1976 10502
## 3  1980  8937
## 4  1984 11588
## 5  1988 14676
```

첫번째로 세 감소점의 구체적인 연도와 수치를 살펴보기 위해 적당한 시기의 자료만을 추출해보면 1932 년, 1956 년과 1976 ~ 1984 년에 감소점이 나오는 것을 확인할 수 있다. 1929 년을 시작으로 10 년간 지속된 미국 경제 대공황은 1930 년대 초반에 가장 악화된 상태였고 따라서 1932 년 엘에이에서 개최된 올림픽은 당시 불경기의 타격을 받아 출전할 수 있었던 선수가 많지 않았음을 짐작할 수 있다. 다음으로 1956 년도 직전과 직후 연도에 비해 출전 선수가 현저히 적은데 당시 제 2 차 중동 전쟁으로 인해 이집트, 이라크와 레바논이 불참하였고 헝가리 혁명으로 인해 소련이 올림픽에 불참함으로써 네덜란드, 스페인 그리고 스위스의 불참을 이끌었고 대만의 독립적인 출전이 인정됨에 불만을 품은 중국이 보이콧을 함으로써 일어난 일이라는 것을 알 수 있다. 마지막으로 1976 년부터 1984 년까지 12 년동안 세 차례의 올림픽 경기 참여도 급감 또한 흥미로운 역사적 뒷배경을 지니고 있다.

1976 년은 당시 인종 차별 정책을 실시하고 있던 남아프리카 공화국의 참가가 허용되자 아프리카 28 개국이 불참을 선언한 사건을 통해 이해할 수 있다. 1980 년은 소련의 아프카니스탄 침공을 항의하기 위해 미국, 캐나다, 서독, 일본과 우리나라를 포함한 서방 진영 45 ~ 50 개국이 불참하였고 중소국경분쟁으로 인해 중국도 불참하면서 출전 선수의 수가 대폭 감소하였다. 1984 년은 이러한 서방 진영의 1980 년 올림픽 보이콧에 대한 보복으로 소련, 독일 민주 공화국(당시 동독), 알바니아 등 동구권 15 개국이 불참하였다.

```
olympics %>% count(Year, Season) %>% filter(Year %in% c(1910:1920))
```

```
## # A tibble: 2 x 3
##    Year Season     n
##   <int> <fct>  <int>
## 1  1912 Summer  4040
## 2  1920 Summer  4292
```

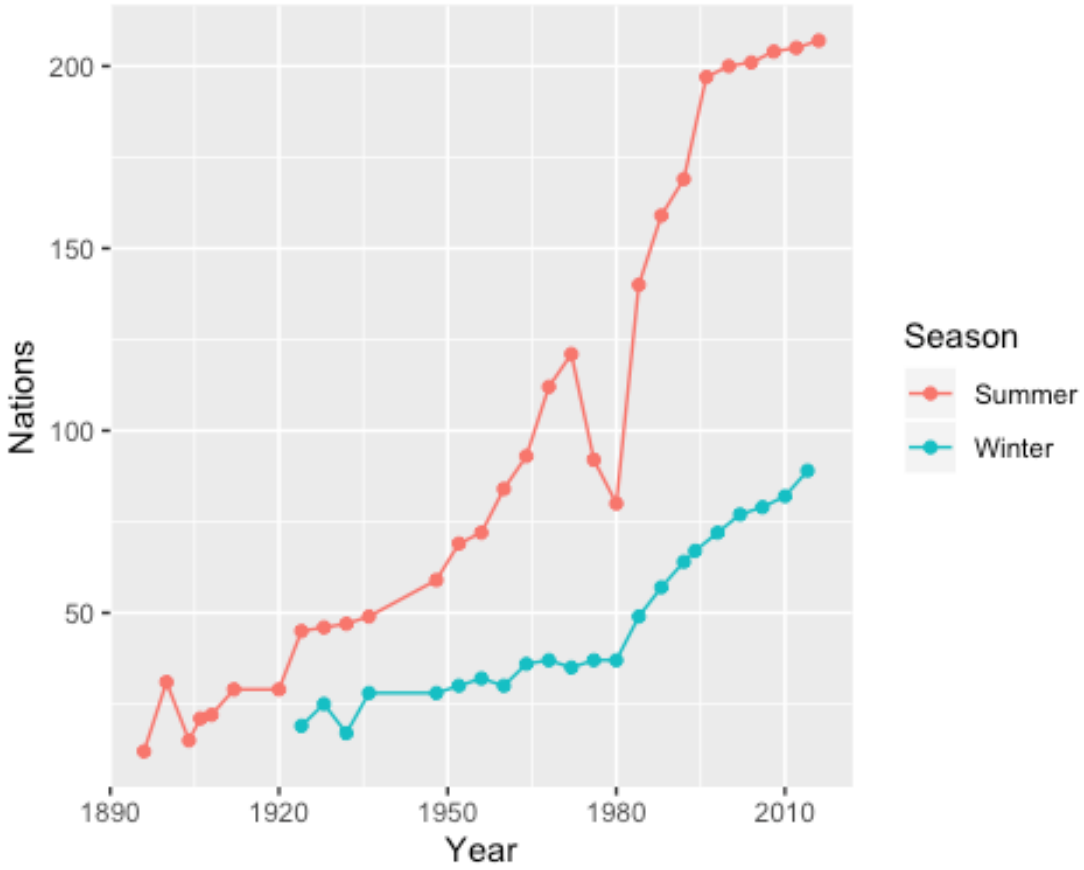
```
olympics %>% count(Year, Season) %>% filter(Year %in% c(1935:1950))
```

```
## # A tibble: 4 x 3
##    Year Season     n
##   <int> <fct>  <int>
## 1  1936 Summer  6506
## 2  1936 Winter   895
## 3  1948 Summer  6405
## 4  1948 Winter  1075
```

두번째로 그래프를 통해 1910 년도와 1930 년대 후반에 한동안 올림픽이 중단되었음을 유추할 수 있다. 이에 따라 1910 ~ 1920 년과 1935 ~ 1950 년 사이의 자료만을 추출해보면 1916 년과 1940, 1944 년에 열렸어야 할 올림픽이 개최되지 않았음을 알 수 있다. 이는 실제로 세계 1 차 대전과 세계 2 차 대전이 있었던 시기라는 역사적 사실을 감안한다면 납득할 수 있는 사실이다.

b) 연도별 참가국

```
olympics %>% group_by(Year, Season) %>% summarize(Nations = n_distinct(NOC)) %
>% ggplot(aes(Year, Nations, color = Season)) + geom_point() + geom_line()
```




```
olympics %>% group_by(Year, Season) %>% summarize(Nations = n_distinct(NOC)) %
>% filter(Season == "Winter", Year %in% c(1920:1940))
```

```
## # A tibble: 4 x 3
## # Groups:   Year [4]
##    Year Season Nations
##   <int> <fct>    <int>
## 1  1924 Winter      19
## 2  1928 Winter      25
## 3  1932 Winter      17
## 4  1936 Winter      28
```

```
olympics %>% group_by(Year, Season) %>% summarize(Nations = n_distinct(NOC)) %
>% filter(Season == "Summer", Year %in% c(1965:1990))
```

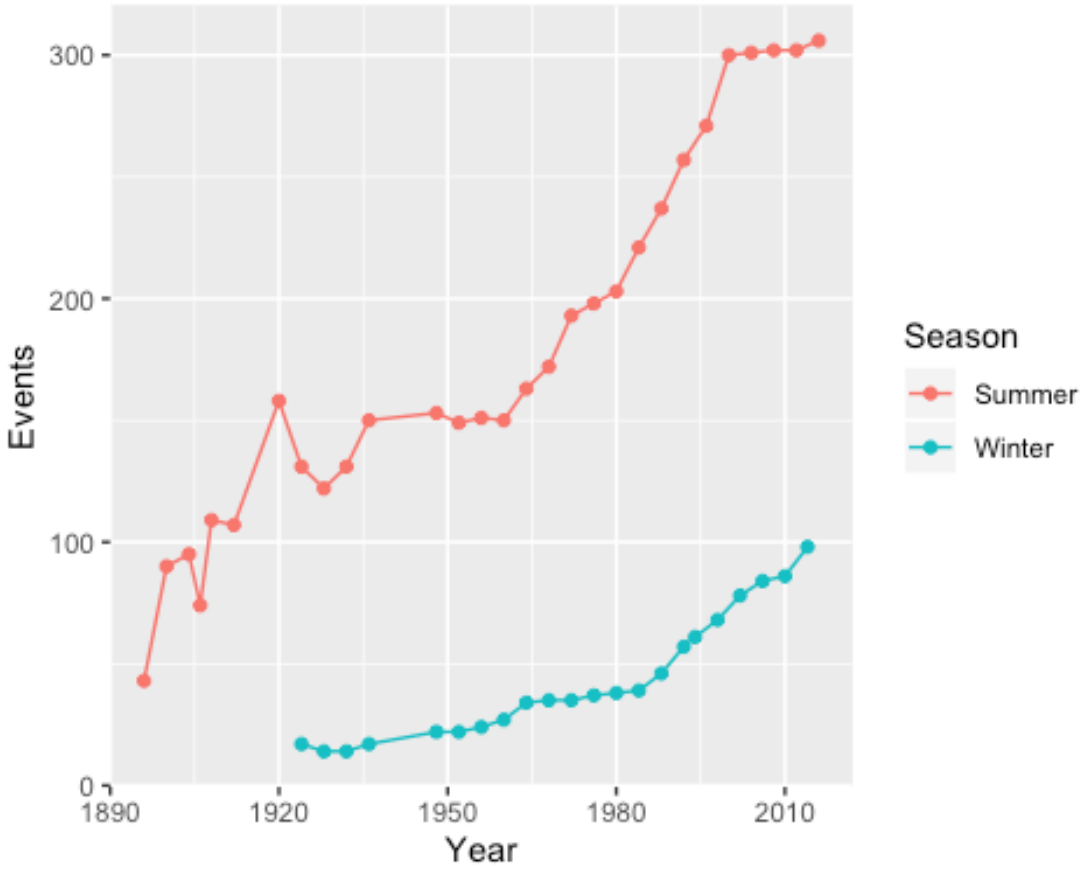
```
## # A tibble: 6 x 3
## # Groups:   Year [6]
##    Year Season Nations
##   <int> <fct>    <int>
```

```
## 1  1968 Summer     112
## 2  1972 Summer     121
## 3  1976 Summer      92
## 4  1980 Summer      80
## 5  1984 Summer     140
## 6  1988 Summer     159
```

일단 시간이 흐를수록 매 올림픽에 참가하는 나라의 수는 증가하는 추세임을 볼 수 있다. 2000 년대에 들어서서는 큰 변화가 없고 실제로 세상에 존재하는 나라의 수가 200 에 가깝다(국가를 규정하는 기준에 따라 약간씩 다른 수치를 가짐)는 사실을 감안할 때 현재는 포화상태에 이르렀다고 할 수 있다. 연도별 참가국의 수는 위에서 살펴본 연도별 출전 선수와 다르게 1932 년과 1956 년에 급격한 감소가 보이지 않는다. 1932 년의 출전 선수 대폭 감소는 경제 대공황으로 인해 올림픽 참가비를 낼 수 없었던 선수가 많은 것이 주원인이었고 계속해서 참가국의 수가 증가하고 있었음을 염두해두면 실제 참가국의 수에는 변화가 없었을 것이다. 1956 년 또한 상대적으로 인구가 많은 소련과 중국이 모두 불참하였기 때문에 선수의 수에는 타격이 컸지만 참가국의 수에는 큰 변화가 없었다. 하지만 1976 년과 1980 년의 경우 각각 아프리카 28 개국과 서방 진영 45 ~ 50 개국이 불참하였기 때문에 참가국 수에도 큰 타격이 있었음을 유추할 수 있다.

c) 연도별 종목 수

```
olympics %>% group_by(Year, Season) %>% summarize(Events = n_distinct(Event))
 %>% ggplot(aes(Year, Events, color = Season)) + geom_point() + geom_line()
```



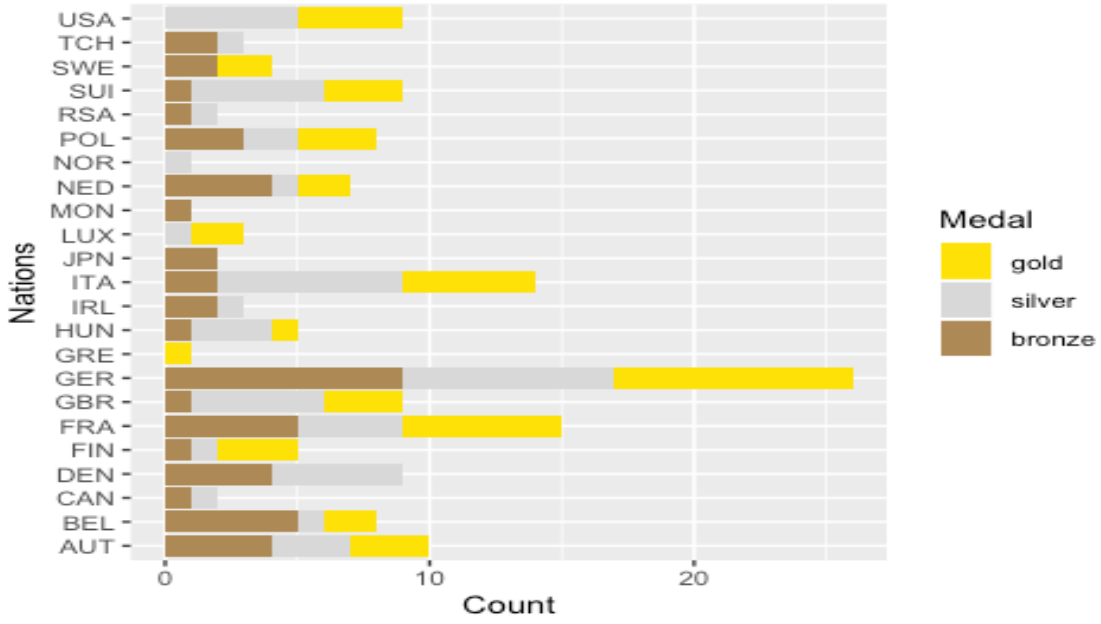
연도별 종목 수도 하계와 동계 모두 시간에 따라 꾸준히 증가해온 추세이다. 그러나 종목 수의 경우 무한정으로 늘어날 수 없고 하계 종목은 2000 년대 이후로 거의 변화가 없음으로 보아 어느 정도 포화상태에 이르렀음을 알 수 있다.

미술 종목

```
art = filter(olympics, Sport == "Art Competitions")
arrange(art, desc(Year))
```

```
## # A tibble: 3,578 x 15
##       ID Name  Sex     Age Height Weight Team  NOC   Games  Year Season
##    <int> <chr> <fct> <int>  <dbl>  <dbl> <chr> <chr> <chr> <int> <fct>
##  1    19 Win … M        54     NA     NA Finl… FIN   1948…  1948 Summer
##  2  1337 Olof… M        71     NA     NA Swed… SWE   1948…  1948 Summer
##  3  4072 "Cor… M        34     NA     NA Neth… NED   1948…  1948 Summer
##  4  4952 Mari… M        40     NA     NA Fran… FRA   1948…  1948 Summer
##  5  5146 Geor… M        84     NA     NA Grea… GBR   1948…  1948 Summer
##  6  5623 Kons… M        NA     NA     NA Gree… GRE   1948…  1948 Summer
##  7  5623 Kons… M        NA     NA     NA Gree… GRE   1948…  1948 Summer
##  8  5665 Hans… M        26     NA     NA Swed… SWE   1948…  1948 Summer
##  9  5728 Uuno… M        44     NA     NA Finl… FIN   1948…  1948 Summer
## 10  5981 John… M        34     NA     NA Grea… GBR   1948…  1948 Summer
```

```
art_gold = art %>% filter(!is.na(Medal), Medal == "Gold") %>% group_by(NOC) %
>% summarize(gold = n()) %>% arrange(desc(gold))
art_silver = filter(art, !is.na(Medal), Medal == "Silver") %>% group_by(NOC) %
>% summarize(silver = n()) %>% arrange(desc(silver))
art_bronze = filter(art, !is.na(Medal), Medal == "Bronze") %>% group_by(NOC) %
>% summarize(bronze = n()) %>% arrange(desc(bronze))
art_gold %>% full_join(art_silver, by = "NOC") %>% full_join(art_bronze, by =
"NOC") %>% gather("gold", "silver", "bronze", key = medal, value = count) %>%
ggplot(aes(NOC, y = count, fill = ordered(medal, levels = c("gold", "silver",
"bronze")))) + geom_bar(stat = "identity") + coord_flip() + labs(x = "Nations
", y = "Count", fill = "Medal") + scale_fill_manual(values = c("#fee101", "#d7
d7d7", "#ad8a56"))
```



```
art %>% filter(NOC == "GER", !is.na(Medal)) %>% group_by(Year) %>% summarize(n
 = n())
```

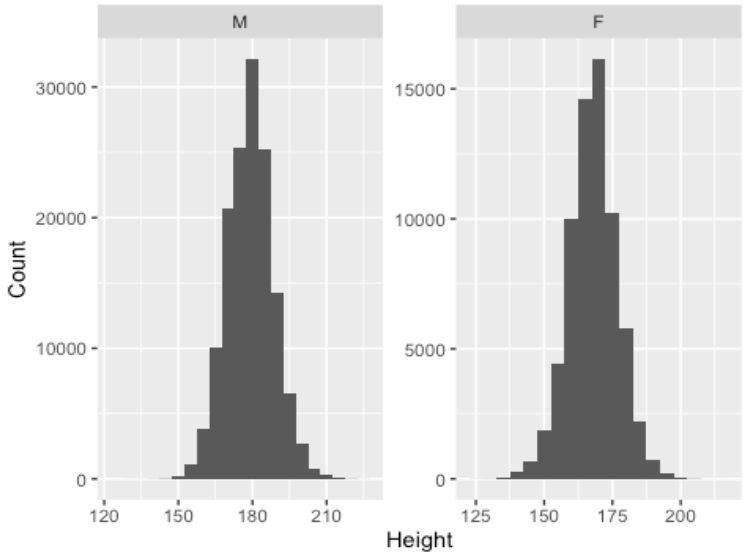
```
## # A tibble: 4 x 2
##    Year     n
##   <int> <int>
## 1  1912     1
## 2  1928     8
## 3  1932     3
## 4  1936    14
```

오늘날에는 없지만 올림픽 초기에는 존재했던 종목 중 대표적인 것은 미술 종목이다. 미술 종목은 1948 년을 마지막으로 시행되고 그 이후 사라졌는데 흥미로운 점은 독일이 1920 년, 1924 년 그리고 1948 년 올림픽에 참가가 금지되었음에도 불구하고 독보적으로 가장 많은 메달을 땄다는 사실이다. 당시 역사적 뒷배경을 찾아보면 독일이 1936 년 베를린에서 개최된 올림픽의 미술 종목을 나치 선전 목적으로 악용했음을 알 수 있다. 따라서 독일의 자료만을 추출하여 메달 수를 비교해본 결과 전 연도들에 비해 1936 년에 급격히 많은 메달을 땄음을 확인할 수 있다.

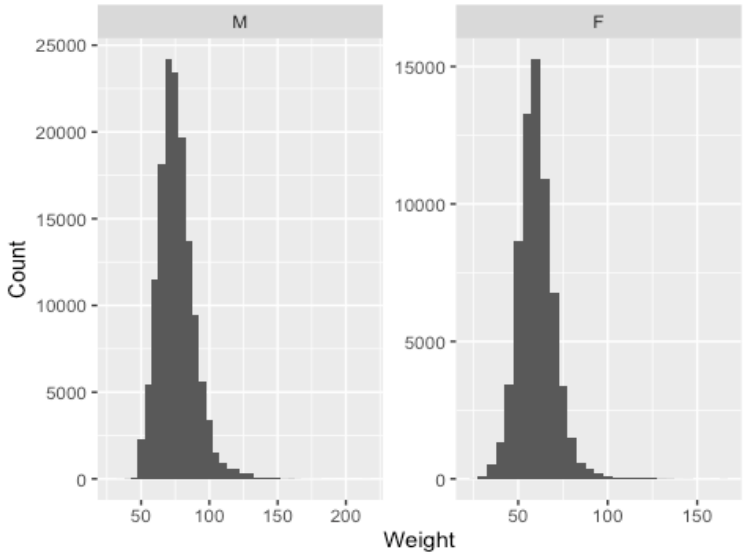
1. Height & Weight

1. 전체 선수들의 키와 몸무게 분포

```
olympics %>% filter(!is.na(Height)) %>% group_by(Sex) %>% ggplot(aes(Height))
 + geom_histogram(binwidth = 5) + facet_wrap(~Sex, scales = "free") + ylab("Co
unt")
```




```
olympics %>% filter(!is.na(Weight)) %>% group_by(Sex) %>% ggplot(aes(Weight))
 + geom_histogram(binwidth = 5) + facet_wrap(~Sex, scales = "free") + ylab("Co
unt")
```



키와 몸무게는 여성이 상대적으로 남성에 비해 값이 작기 때문에 분포를 더 잘 살펴보기 위해 scales = "free"를 지정하였다. 키의 분포는 눈에 띄지 않지만 약간

왼쪽으로 치우쳐 있고 특히 여자 선수들의 분포가 그러하다. 반면 몸무게의 경우 남녀 모두 봉우리가 그래프의 중간이 아닌 왼쪽으로 많이 치우쳐 있다는 것을 보면 수가 적어 그래프에는 잘 표현되지 않았지만 분포가 왼쪽으로 많이 치우쳐져 있음을 확인할 수 있다.

2. 연도별 출전 선수들의 키와 몸무게

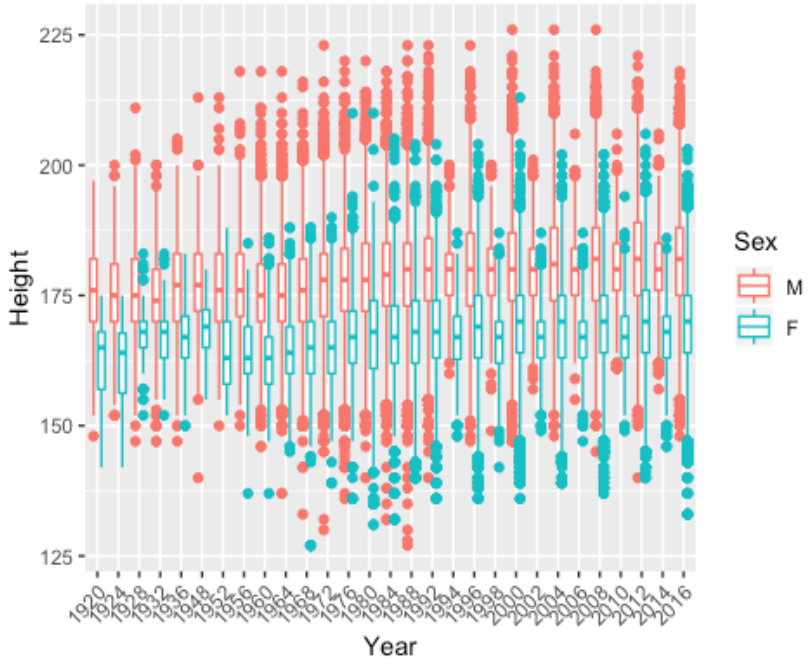
```
olympics %>% filter(Sex == "M", !is.na(Height), !is.na(Weight)) %>% arrange(Ye
ar)
```

```
## # A tibble: 140,124 x 15
##       ID Name  Sex     Age Height Weight Team  NOC   Games  Year Season
##    <int> <chr> <fct> <int>  <dbl>  <dbl> <chr> <chr> <chr> <int> <fct>
##  1 16616 "Tho… M        21    183     66 Unit… USA   1896…  1896 Summer
##  2 16616 "Tho… M        21    183     66 Unit… USA   1896…  1896 Summer
##  3 22700 Jame… M        27    175     72 Unit… USA   1896…  1896 Summer
##  4 22700 Jame… M        27    175     72 Unit… USA   1896…  1896 Summer
##  5 22700 Jame… M        27    175     72 Unit… USA   1896…  1896 Summer
##  6 24423 "Tho… M        23    176     66 Unit… USA   1896…  1896 Summer
##  7 24423 "Tho… M        23    176     66 Unit… USA   1896…  1896 Summer
##  8 27318 Dimi… M        NA    185    106 Gree… GRE   1896…  1896 Summer
##  9 29084 Kurt… M        21    179     73 Germ… GER   1896…  1896 Summer
## 10 29084 Kurt… M        21    179     73 Germ… GER   1896…  1896 Summer
## # … with 140,114 more rows, and 4 more variables: City <chr>, Sport <chr>,
## #   Event <chr>, Medal <fct>
```

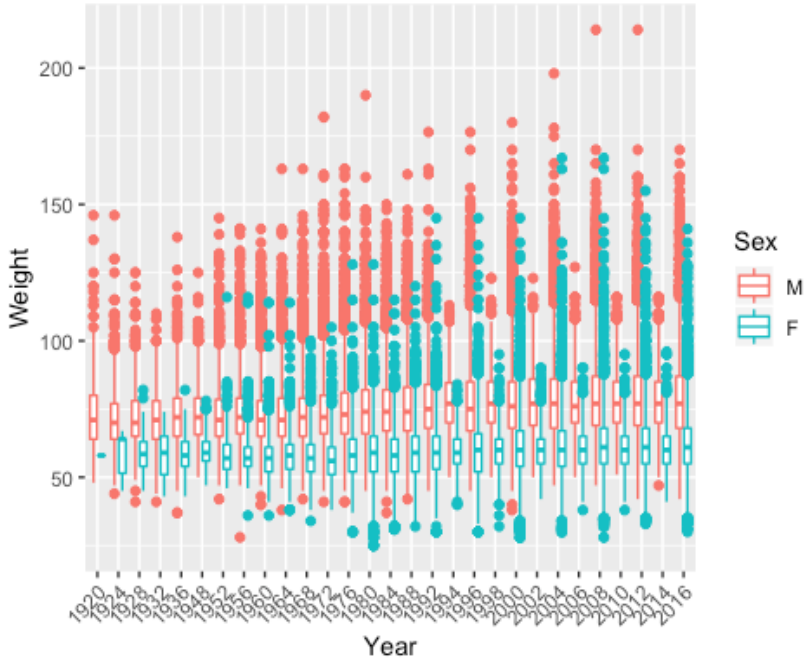
```
olympics %>% filter(Sex == "F", !is.na(Height), !is.na(Weight)) %>% arrange(Ye
ar)
```

```
## # A tibble: 66,729 x 15
##       ID Name  Sex     Age Height Weight Team  NOC   Games  Year Season
##    <int> <chr> <fct> <int>  <dbl>  <dbl> <chr> <chr> <chr> <int> <fct>
##  1 53409 Ludo… F        35    165     58 Finl… FIN   1920…  1920 Summer
##  2 31394 "Ger… F        18    167     64 Unit… USA   1924…  1924 Summer
##  3 31394 "Ger… F        18    167     64 Unit… USA   1924…  1924 Summer
##  4 31394 "Ger… F        18    167     64 Unit… USA   1924…  1924 Summer
##  5 39348 "Ade… F        36    169     67 Unit… USA   1924…  1924 Summer
##  6 47618 Sonj… F        11    155     45 Norw… NOR   1924…  1924 Winter
##  7 53409 Ludo… F        39    165     58 Finl… FIN   1924…  1924 Winter
##  8 98258 Suza… F        19    160     45 Fran… FRA   1924…  1924 Summer
##  9  9639 "Flo… F        17    169     58 Cana… CAN   1928…  1928 Summer
## 10  9639 "Flo… F        17    169     58 Cana… CAN   1928…  1928 Summer
## # … with 66,719 more rows, and 4 more variables: City <chr>, Sport <chr>,
## #   Event <chr>, Medal <fct>
```

```
olympics %>% filter(!is.na(Height), Year >= 1920) %>% ggplot(aes(factor(Year),
 Height, color = Sex)) + geom_boxplot() + labs(x = "Year") + theme(axis.text.x
 = element_text(angle=45, hjust=1))
```




```
olympics %>% filter(!is.na(Weight), Year >= 1920) %>% ggplot(aes(factor(Year),
 Weight, color = Sex)) + geom_boxplot() + labs(x = "Year") + theme(axis.text.x
 = element_text(angle=45, hjust=1))
```

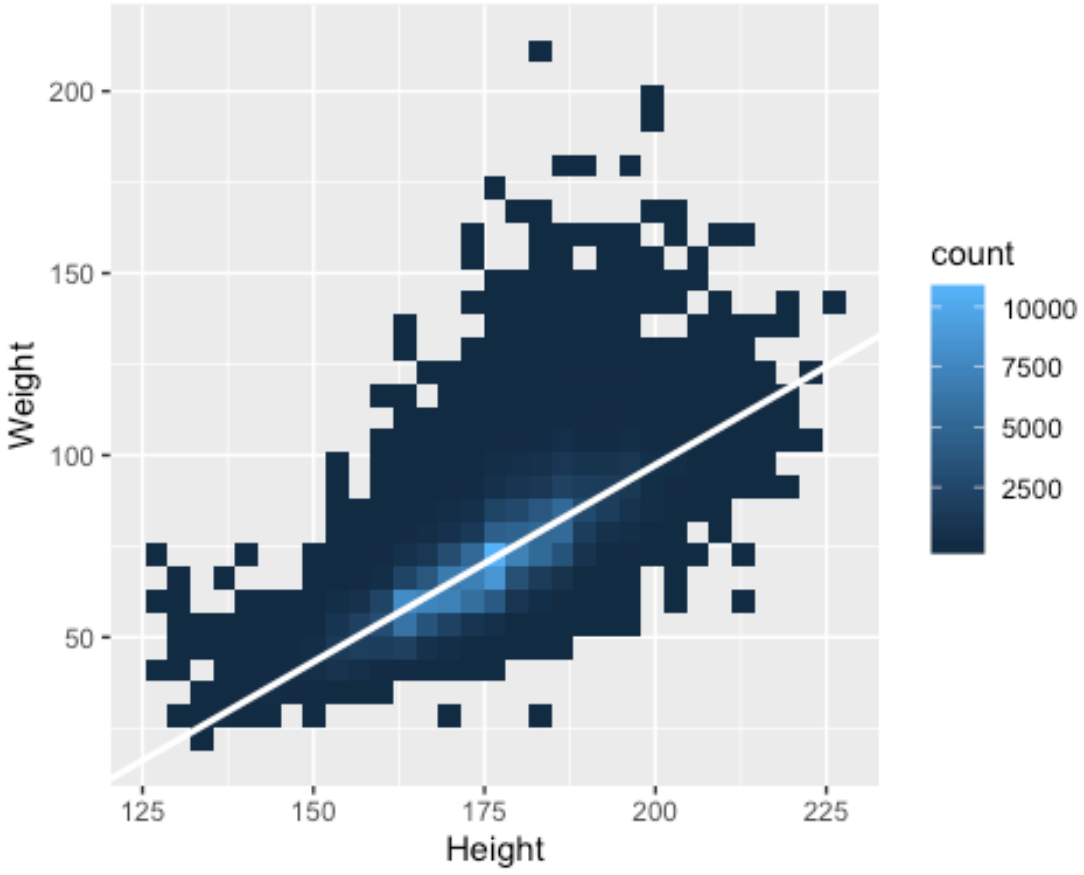


일단 모든 성별의 키와 몸무게 자료가 기록되어 있기 시작한 1920 년 이후의 자료만을 추출하였다. 연도별 선수들의 키와 몸무게 상자 그림을 보면 예전과 지금 선수들의 키와 몸무게가 남녀 모두 아주 약간 증가하였지만 크게 변화하지는

않았다. 키의 분포는 상자그림의 상자를 제외한 위아래로도 많다는 것으로 보아 분포가 퍼져 있음을 알 수 있다. 몸무게의 분포는 이상점이 상자의 아래보다 위가 압도적으로 많은 것을 볼 수 있는데 이것은 위에서 전체적인 몸무게의 분포가 오른쪽으로 많이 치우쳐져 있다는 점을 명확히 보여준다.

3. 키와 몸무게 상관관계 분석

```
reg_coef = coef(lm(Weight ~ Height, data = olympics))
olympics %>% filter(Year >= 1920) %>% ggplot(aes(Height, Weight)) + geom_bin2d
(na.rm = TRUE) + geom_abline(intercept = reg_coef[1], slope = reg_coef[2], col
or = "white", size = 1)
```



키와 몸무게 사이에 분명한 양의 상관관계가 존재함을 볼 수 있다. 점들의 분포가 회귀 직선 아래보다 위쪽에 더 많은 것으로 보아 잔차의 분포가 오른쪽으로 치우쳐져 있다고 유추할 수 있다

4. 종목별 키와 몸무게 분석

a) 중간값

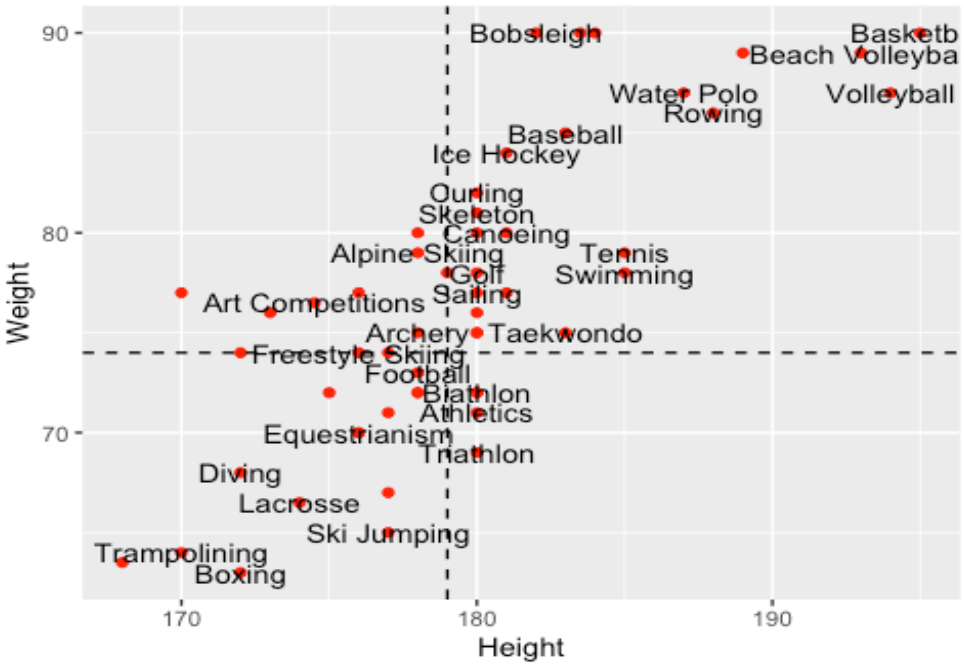
(i) 남자 선수

```
olympics %>% filter(Sex == "M", !is.na(Height), !is.na(Weight)) %>% summarize
(median_h = median(Height), median_w = median(Weight))
```

```
## # A tibble: 1 x 2
##   median_h median_w
##      <dbl>    <dbl>
## 1      179       74
```

남자 선수들의 키와 몸무게 중간값은 각각 179cm 와 74kg 이다.

```
olympics %>% filter(Sex == "M", !is.na(Height), !is.na(Weight)) %>% group_by(S
port) %>% summarize(m_height = median(Height), m_weight = median(Weight)) %>%
ggplot(aes(m_height, m_weight)) + geom_point(color = "red") + geom_text(aes(la
bel = Sport), check_overlap = TRUE) + geom_vline(xintercept = 179, linetype =
2) + geom_hline(yintercept = 74, linetype = 2) + labs(x = "Height", y = "Weigh
t")
```



종목별 키와 몸무게의 중간값을 전체 선수들의 중간값과 비교해보기 위하여 각 종목의 위치를 종목명과 함께 나타내고 전체 중간값은 점선으로 표현한 그림이다. 농구와 배구의 점이 오른쪽 위에 위치한 것으로 보아 선수들이 키도 월등히 크고 몸무게도 많이 나감을 알 수 있다. 봅슬레이의 경우 중간 위에 위치하고 따라서

키는 타종목 선수들과 비슷하지만 몸무게는 그에 비해 매우 많이 나가는 선수가 많음을 의미한다. 반대로 다이빙과 체조 등의 종목은 키와 몸무게 모두 전반적인 선수들에 비해 그 값이 작다.

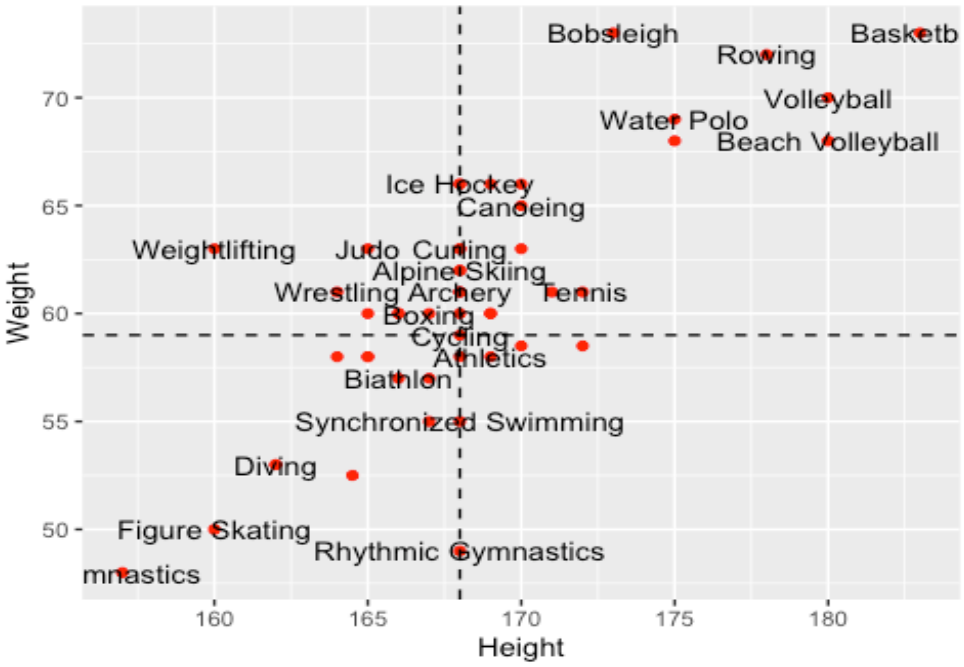
(ii) 여자 선수

```
olympics %>% filter(Sex == "F", !is.na(Height), !is.na(Weight)) %>% summarize
(median_h = median(Height), median_w = median(Weight))
```

```
## # A tibble: 1 x 2
##   median_h median_w
##      <dbl>    <dbl>
## 1      168       59
```

여자 선수들의 키와 몸무게 중간값은 각각 168cm 와 59kg 이다.

```
olympics %>% filter(Sex == "F", !is.na(Height), !is.na(Weight)) %>% group_by(S
port) %>% summarize(m_height = median(Height), m_weight = median(Weight)) %>%
ggplot(aes(m_height, m_weight)) + geom_point(color = "red") + geom_text(aes(la
bel = Sport), check_overlap = TRUE) + geom_vline(xintercept = 168, linetype =
2) + geom_hline(yintercept = 59, linetype = 2) + labs(x = "Height", y = "Weigh
t")
```



여자 선수도 남자 선수와 종목별 양상이 비슷하다. 한 가지 주목할 점은 앞서 보인 키와 몸무게의 양의 상관관계를 고려하면 역도 선수들은 키는 작은 편에 속하면서 키에 비해 몸무게가 굉장히 많이 나가는 선수들이 많음을 확인할 수 있다. 또한

남자 선수의 그림에서 check_overlap = TRUE 옵션 때문에 볼 수 없었던 종목 중 피겨 스케이팅을 볼 수 있고 이 종목 또한 전반적인 타종목 선수들에 비해 선수들이 키도 작고 몸무게도 덜 나감을 알 수 있다.

b) 평균

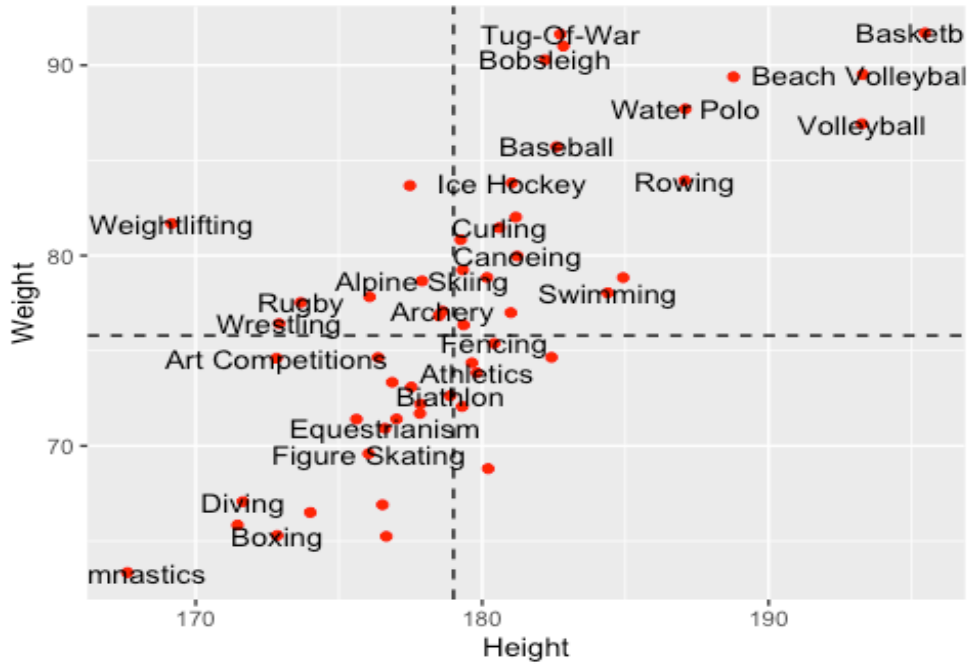
(i) 남자 선수

```
olympics %>% filter(Sex == "M", !is.na(Height), !is.na(Weight)) %>% summarize
(mean_h = mean(Height), mean_w = mean(Weight))
```

```
## # A tibble: 1 x 2
##   mean_h mean_w
##    <dbl>  <dbl>
## 1   179.   75.8
```

남자 선수들의 키와 몸무게 평균은 178.93cm 와 75.75kg 이다. 남자 선수의 몸무게 분포는 오른쪽으로 치우쳐져 있었으므로 평균이 앞서 구한 중간값 74kg 보다 약간 크다. 키의 경우는 중간값인 179cm 와 거의 일치한다.

```
olympics %>% filter(Sex == "M", !is.na(Height), !is.na(Weight)) %>% group_by(S
port) %>% summarize(m_height = mean(Height), m_weight = mean(Weight)) %>% ggpl
ot(aes(m_height, m_weight)) + geom_point(color = "red") + geom_text(aes(label
= Sport), check_overlap = TRUE) + geom_vline(xintercept = 179, linetype = 2) +
 geom_hline(yintercept = 75.8, linetype = 2) + labs(x = "Height", y = "Weight
")
```



앞서 중간값으로 표현한 그림과 매우 유사한 양상을 보인다. 위에서 보지 못했던 종목 중 하나로 줄다리기(Tug-of-War)가 봅슬레이 근처에 위치하고 있는데 이에 따라 줄다리기 선수들 역시 키는 평균과 비슷하지만 몸무게는 많이 나가는 선수들이 대부분임을 의미한다.

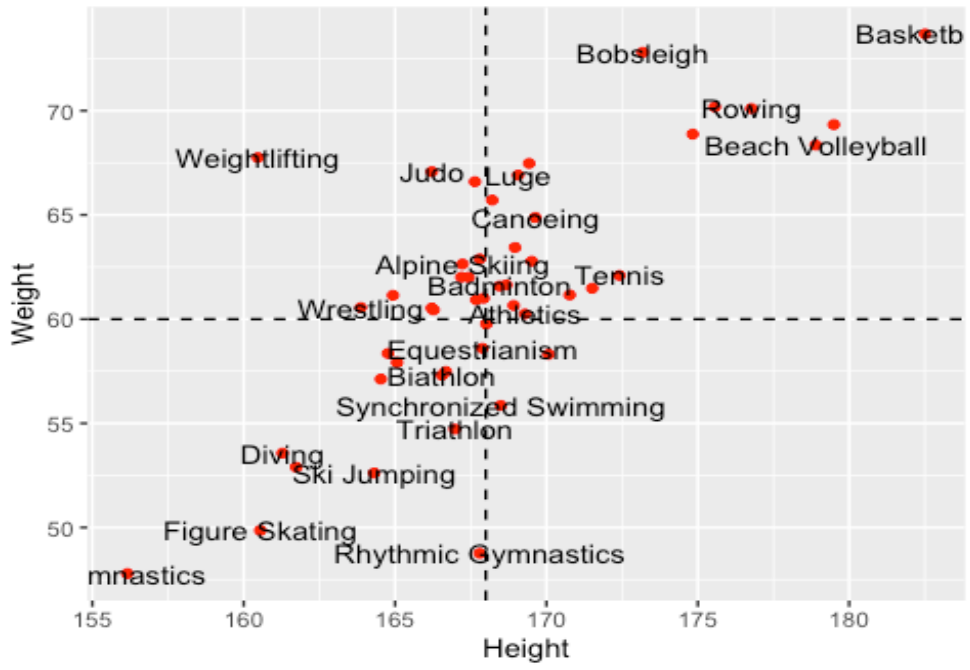
(ii) 여자 선수

```
olympics %>% filter(Sex == "F", !is.na(Height), !is.na(Weight)) %>% summarize
(mean_h = mean(Height), mean_w = mean(Weight))
```

```
## # A tibble: 1 x 2
##   mean_h mean_w
##    <dbl>  <dbl>
## 1   168.   60.0
```

여자 선수들의 키와 몸무게 평균은 167.85cm 와 60.02kg 이다. 남자 선수와 마찬가지로 몸무게의 평균이 중간값인 59kg 보다 약간 크지만 키의 경우 별로 다르지 않다.

```
olympics %>% filter(Sex == "F", !is.na(Height), !is.na(Weight)) %>% group_by(S
port) %>% summarize(m_height = mean(Height), m_weight = mean(Weight)) %>% ggpl
ot(aes(m_height, m_weight)) + geom_point(color = "red") + geom_text(aes(label
= Sport), check_overlap = TRUE) + geom_vline(xintercept = 168, linetype = 2) +
 geom_hline(yintercept = 60, linetype = 2) + labs(x = "Height", y = "Weight")
```

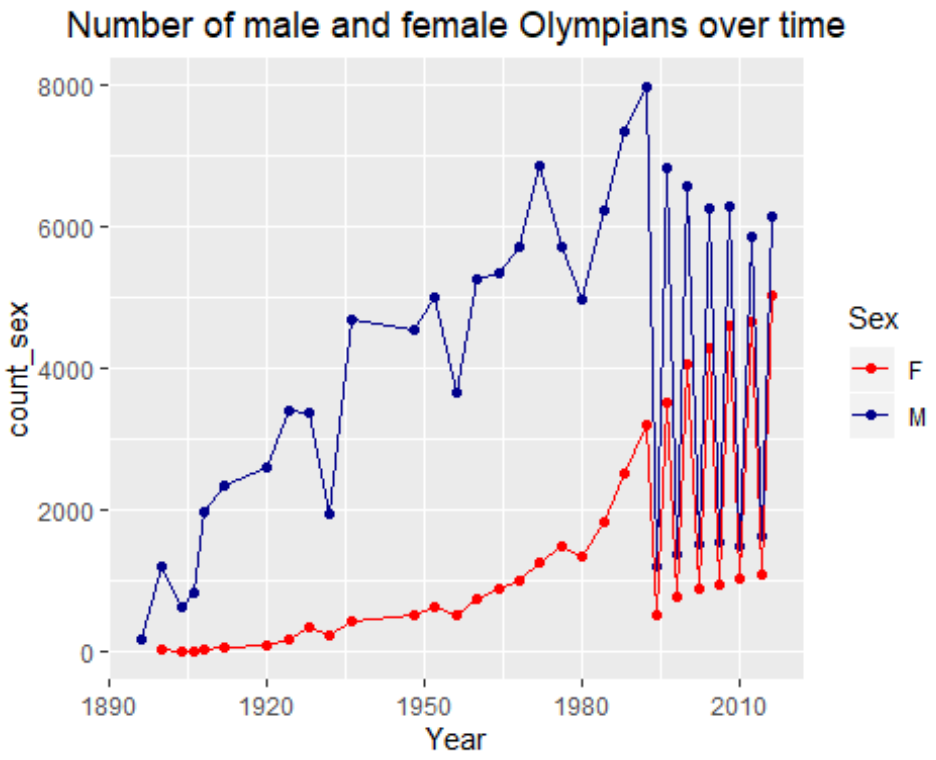


마찬가지로 왼쪽 위에 위치하는 역도와 오른쪽 위에 위치하는 농구가 두드러진다.

2. Gender

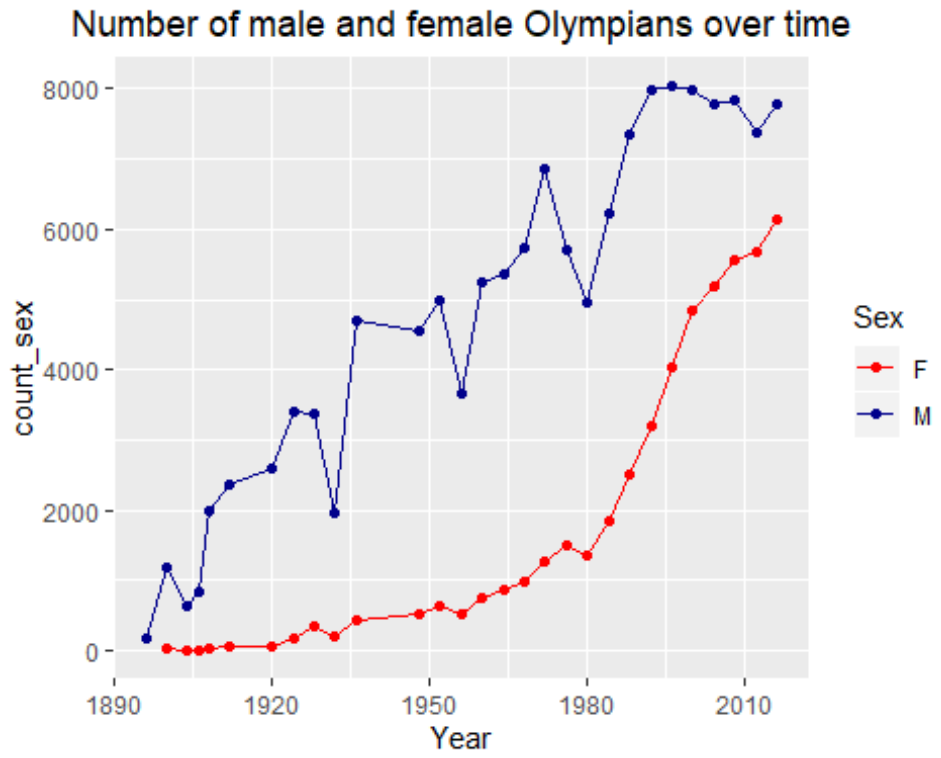
성별에 따른 데이터 분석

```
#시간에 따른 남성과 여성 참가선수의 수 비교 그래프
(athlete_sex <- athlete %>%
  group_by(Year, Sex) %>%
  summarize(count = length(unique(ID))))

athlete_sex <- olympic %>%
  group_by(Year, Sex) %>%
  summarize(count_sex = length(unique(ID)))

ggplot(athlete_sex, aes(x = Year, y = count_sex, group = Sex, color = Sex)) +
  geom_point()+geom_line() +
  scale_color_manual(values = c('red','darkblue')) + labs(title = 'Number of m
ale and female Olympians over time') +
  theme(plot.title=element_text(hjust=0.5))
```



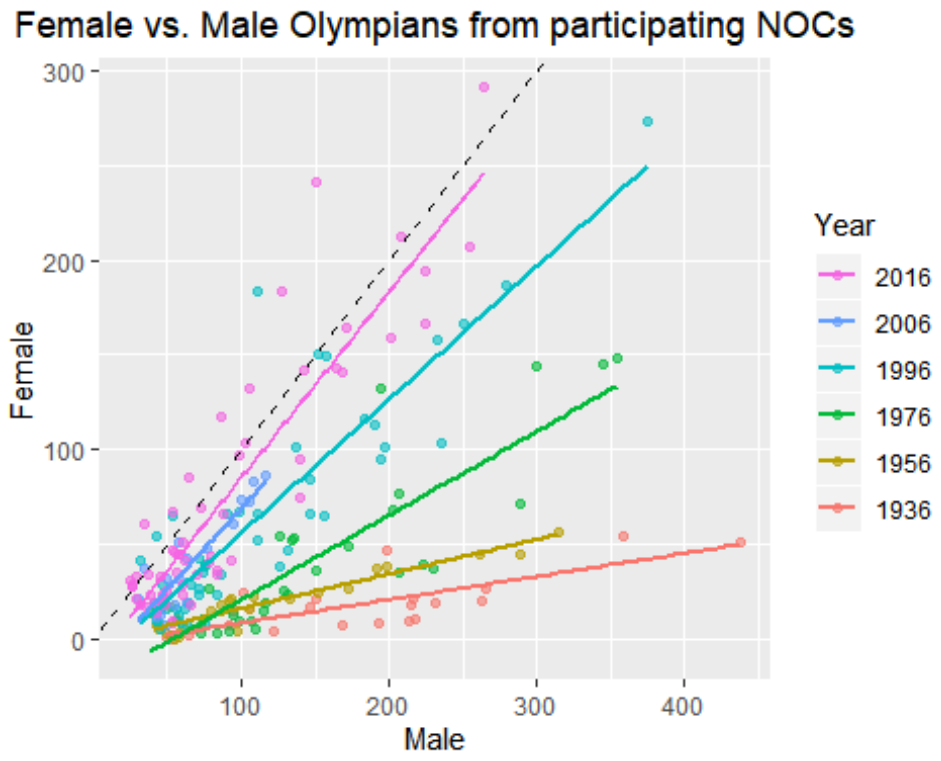
>위를 통해 여성 참여가 1900 년부터 일어났다는 것을 알 수 있다. 다만, 1994 년 이후 동계올림픽과 하계올림픽이 2 년 단위로 번갈아 가며 개최되어 지그재그로 나타나는 패턴이 보이기 때문에 전체적으로 파악하기 힘들었다. 따라서 동계올림픽과 하계올림픽을 하나로 합쳐 아래와 같이 그래프를 그려보았다.

```r
year<-athlete_sex$Year
for(i in 1:69){ #동계올림픽을 하계올림픽 연도에 맞추기
  if((year[i]>=1994) && ((year[i])%%4==2)){
    year[i]=year[i]+2
    }
}
athlete_sex$Year <-year
athlete_sex2<-athlete_sex%>%
  group_by(Year,Sex)%>%
  summarise(count_sex=sum(count_sex))

ggplot(athlete_sex2, aes(x = Year, y = count_sex, group = Sex, color = Sex)) +
  geom_point()+geom_line() +
  scale_color_manual(values = c('red','darkblue')) + labs(title = 'Number of m
ale and female Olympians over time') +
  theme(plot.title=element_text(hjust=0.5))
```



이를 통해 남성과 여성의 참여 비율 차이가 점점 줄어들고 있다는 것 또한 알 수 있었다.

```r
counts_NOC <- athlete %>% filter(Year %in% c(1936,1956,1976,1996,2006,2016)) %
>%
  group_by(Year, NOC, Sex) %>%
  summarize(Count = length(unique(ID))) %>%
  spread(Sex, Count) %>%
  mutate(Total = sum(M,F,na.rm=TRUE)) %>%
  filter(Total > 49)
names(counts_NOC)[3:4] <- c("Male","Female")
counts_NOC$Male[is.na(counts_NOC$Male)] <- 0
counts_NOC$Female[is.na(counts_NOC$Female)] <- 0
counts_NOC$Year <- as.factor(counts_NOC$Year)

ggplot(counts_NOC, aes(x=Male, y=Female, group=Year, color=Year)) +
  geom_point(alpha=0.6) +
  geom_abline(intercept=0, slope=1, linetype="dashed") +
  geom_smooth(method="lm", se=FALSE) +
  labs(title = "Female vs. Male Olympians from participating NOCs") +
  theme(plot.title = element_text(hjust = 0.5)) +
  guides(color=guide_legend(reverse=TRUE))
```



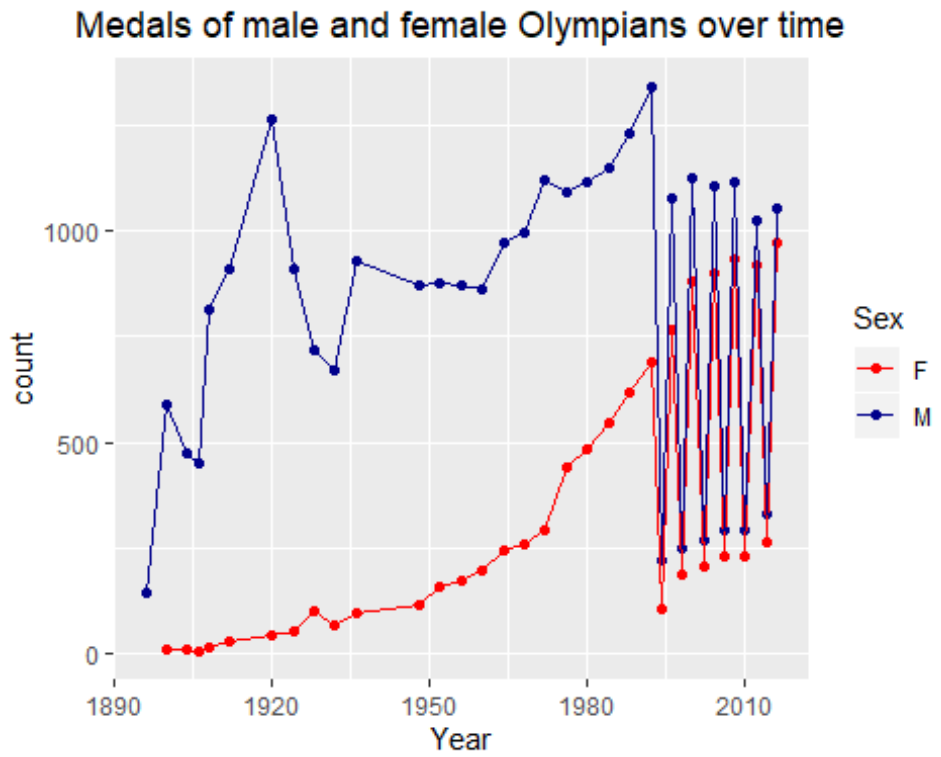
>여전히 여성 선수와 남성 선수간의 출전자 수 차이가 있지만, 10 년 주기로 비교하였을 때 여성 선수 참여율이 점차 늘어나는 것을 확인할 수 있다. (기울기가 점차 커지는 것을 확인할 수 있기 때문이다.)

남성 선수 수를 x 축 변수로, 여성 선수 수를 y 축 변수로 두는 그래프를 그려 확인해보면 시간이 흐를 수록 점차 그 기울기가 1 에 가까워지는 것을 확인할 수 있다.

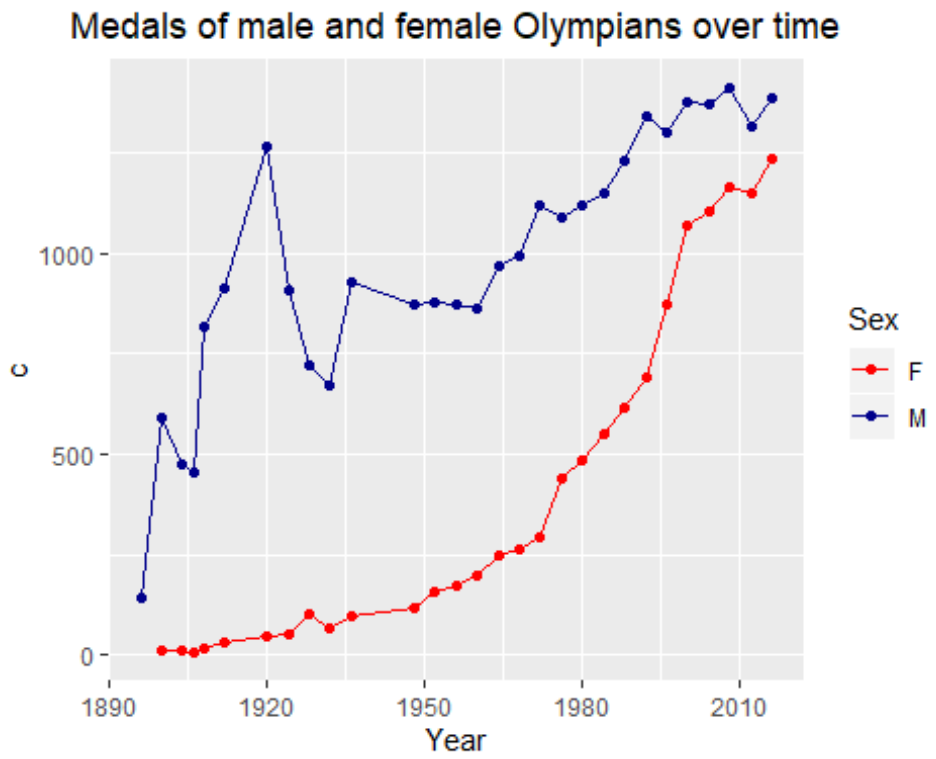
초반에 비해 여성의 참여율이 높아지기 때문에 직선 그래프가 시간이 지날수록 기울기가 증가하는 것을 확인할 수 있다. (1946 년은 세계 2 차 대전으로 인해 중단되었고, 1966 년은 올림픽 게임 협약 관련으로 인해 게임이 개최되지 않았다.)

연도와 성별에 따른 메달수 비교

```
athlete_sex_medal <- olympic %>%
  group_by(Year, Sex) %>%
  summarize(count = sum(!is.na(Medal)))

ggplot(athlete_sex_medal, aes(x = Year, y = count, group = Sex, color = Sex))
+ geom_point() +
  geom_line() +
  scale_color_manual(values = c('red','darkblue')) +
  labs(title = 'Medals of male and female Olympians over time') +
  theme(plot.title=element_text(hjust=0.5))
```

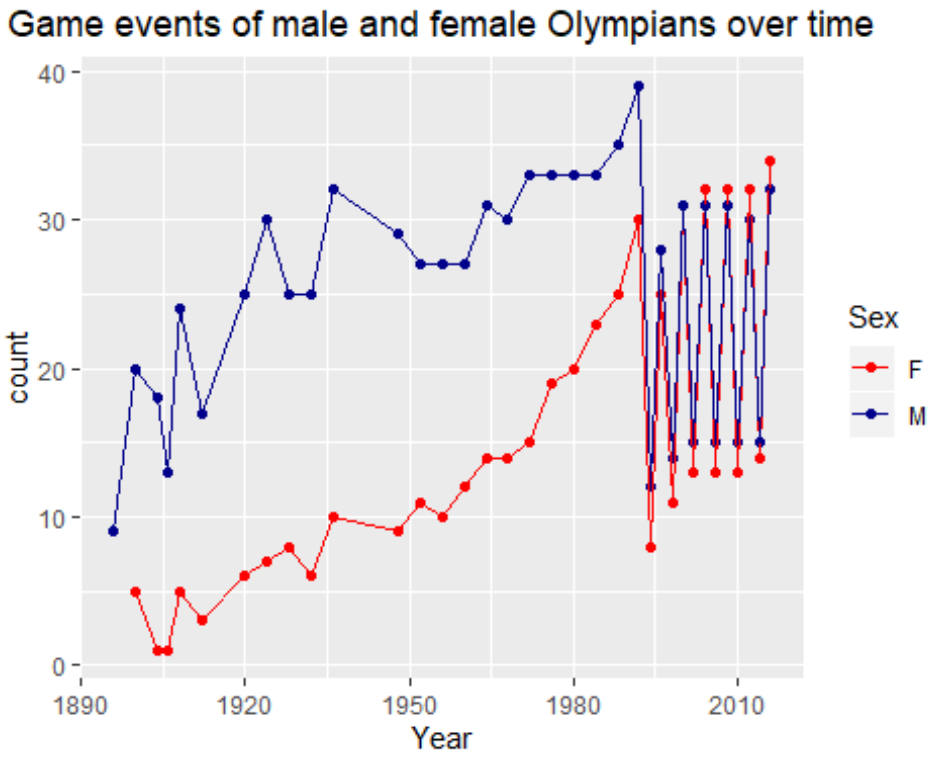


```r
year<-athlete_sex_medal$Year
for(i in 1:69){
  if( (year[i]>=1994) && ((year[i])%%4==2)){
    year[i]=year[i]+2
    }
}
athlete_sex_medal$Year <-year

athlete_sex_medal2<-athlete_sex_medal%>%
  group_by(Year,Sex)%>%
  summarise(c=sum(count))

ggplot(athlete_sex_medal2, aes(x = Year, y = c, group = Sex, color = Sex)) + g
eom_point() +
  geom_line() +
  scale_color_manual(values = c('red','darkblue')) +
  labs(title = 'Medals of male and female Olympians over time') +
  theme(plot.title=element_text(hjust=0.5))
```

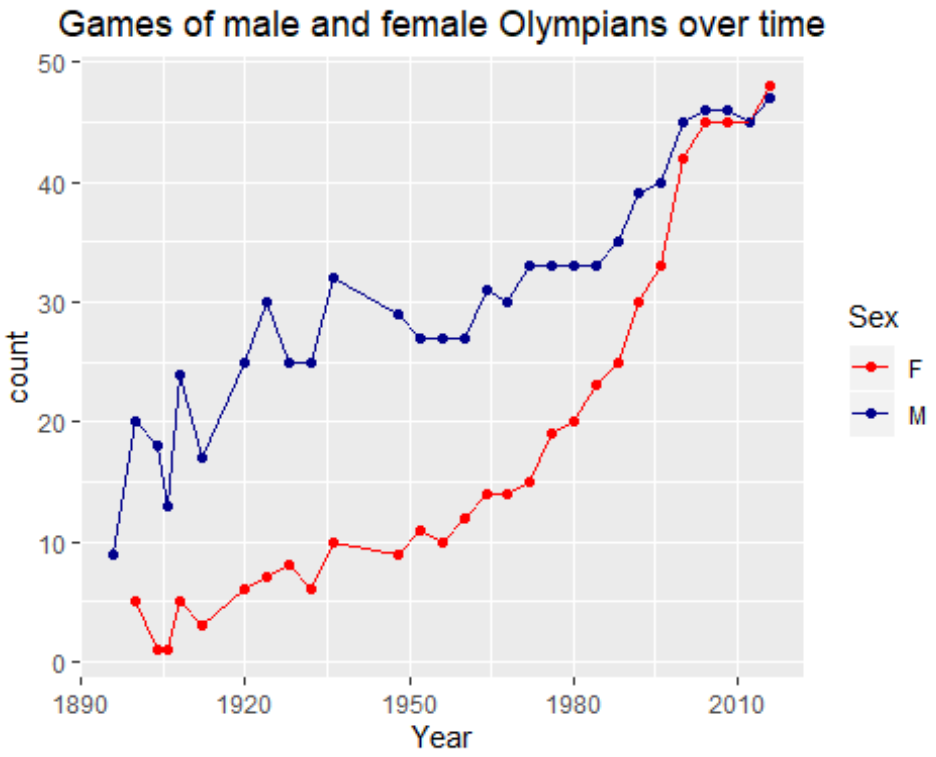


메달 수는 종목수와 비례하므로 여성 종목 자체가 적었다는 것을 확인할 수 있다. 시간이 흐를수록 여성과 남성의 차이가 줄어든 것을 확인할 수 있었다.

```r
athlete_sex_games <-  olympic%>%
  group_by(Year, Sex) %>%
  summarize(count = n_distinct(Sport))

ggplot(athlete_sex_games, aes(x = Year, y = count, group = Sex, color = Sex))
+
  geom_point() +
  geom_line() +
  scale_color_manual(values = c('red','darkblue')) +
  labs(title = 'Game events of male and female Olympians over time') +
  theme(plot.title=element_text(hjust=0.5))
```




```r
athlete_sex_games$Year<-year
athlete_sex_games2<-athlete_sex_games%>%
  group_by(Year,Sex)%>%
  summarize(count=sum(count))

ggplot(athlete_sex_games2, aes(x = Year, y = count, group = Sex, color = Sex))
 +
  geom_point() +
  geom_line() +
  scale_color_manual(values = c('red','darkblue')) +
  labs(title = 'Games of male and female Olympians over time') +
  theme(plot.title=element_text(hjust=0.5))
```



이 그래프가 앞서 본 메달 수와 매우 유사함을 확인할 수 있다.

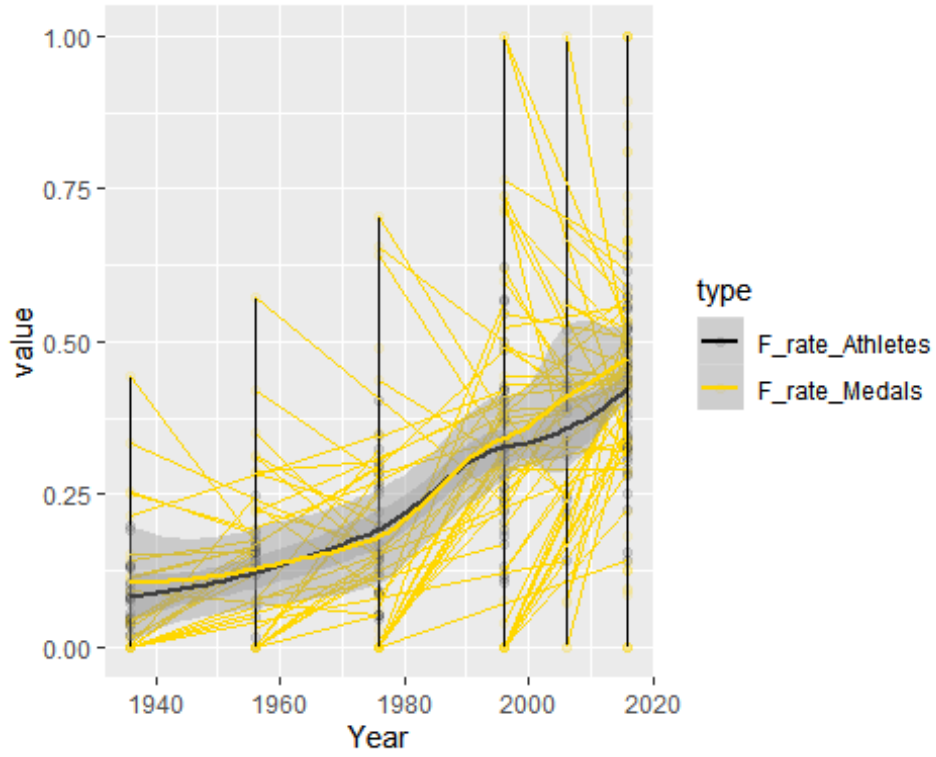
여성의 연도별 참가율과 메달 비율

여성의 메달 비율 또한 10 년 주기로 비교해볼 수 있다.

```
athlete_pre <- athlete %>%
  filter(Year %in% c(1936, 1956, 1976, 1996, 2006, 2016)) %>%
  group_by(Year, NOC, Sex) %>%
  summarise(Athletes = length(unique(ID)), Medals = sum(!is.na(Medal)))

athelte_number <- athlete_pre %>%
  select(- Medals) %>%
  spread(key = Sex, value = Athletes) %>%
  rename(F_Athletes = F,
         M_Athletes = M)

athelte_medals <- athlete_pre %>%
  select(- Athletes) %>%
  spread(key = Sex, value = Medals) %>%
  rename(F_Medals = F,
         M_Medals = M)
```

```
props <- full_join(athelte_number, athelte_medals, by = c('Year','NOC'))
props$F_Athletes[is.na(props$F_Athletes)] <- 0
props$M_Athletes[is.na(props$M_Athletes)] <- 0
props$F_Medals[is.na(props$F_Medals)] <- 0
props$M_Medals[is.na(props$M_Medals)] <- 0

props <- props %>%
  mutate(F_rate_Athletes = F_Athletes/(F_Athletes + M_Athletes),
         F_rate_Medals = F_Medals/(F_Medals + M_Medals)) %>%
  filter((F_Athletes + M_Athletes) > 49)


years <- props %>%
  gather(F_rate_Athletes, F_rate_Medals, key = 'type', value = 'value')

ggplot(years, aes(x=Year, y= value, colour = type)) +
  geom_point(na.rm = TRUE, alpha = 0.1) +
  geom_line(aes(group = NOC)) +
  geom_smooth()+
  scale_color_manual(values = c("black", "gold"))+
  ylim(0,1)
```

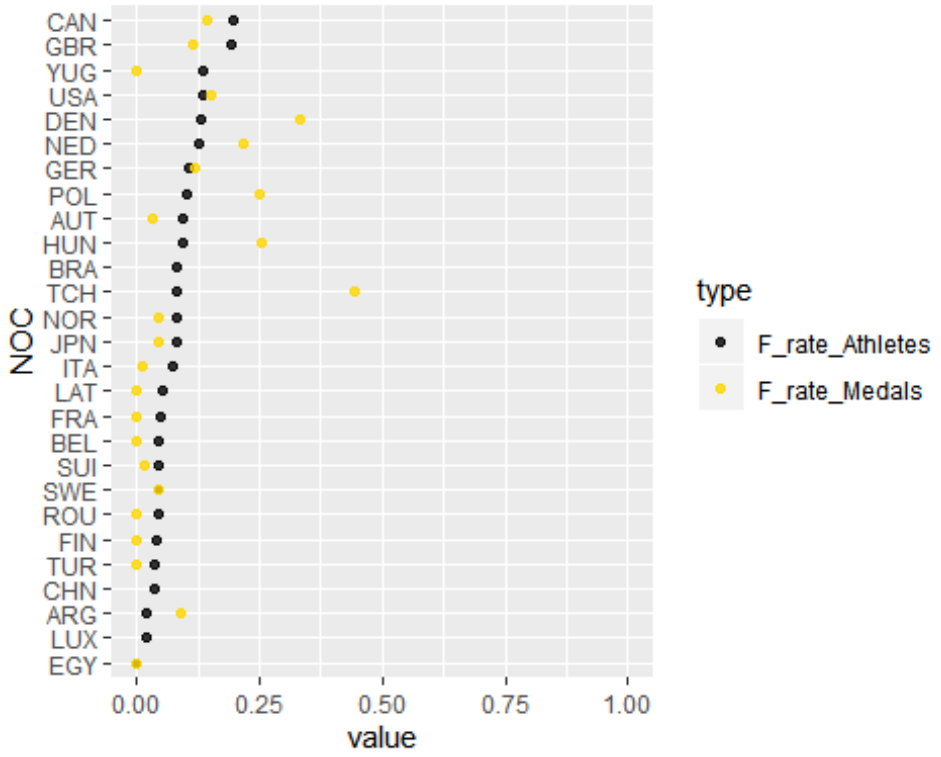
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



여성의 연도별 참가율과 메달 획득율 모두 증가하는 패턴을 보이는 것을 확인할 수 있다.

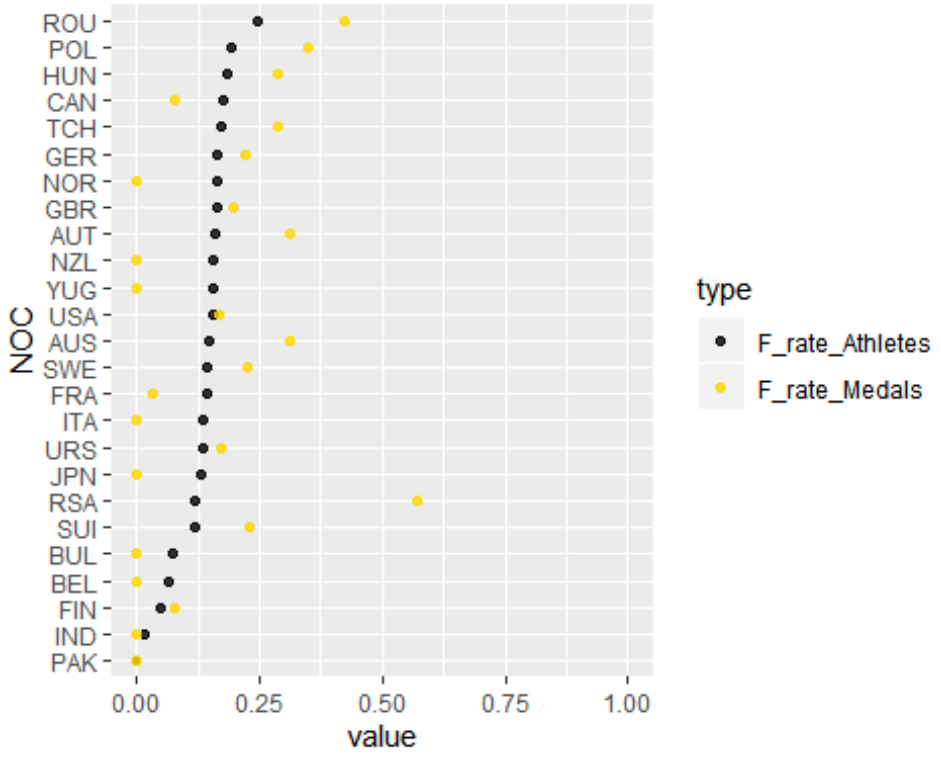
연도별 시대 배경과 대조 및 확인

```
years_1936 <- years %>%
  filter(Year == 1936)

lev_1936 <- years_1936 %>%
  filter(type == 'F_rate_Athletes') %>%
  arrange(value) %>%
  select(NOC)

ggplot(years_1936, aes(x= value, y= NOC, color=type)) +
  geom_point(na.rm = TRUE, alpha = 0.8) +
  scale_color_manual(values = c("black", "gold")) +
  theme(plot.title = element_text(hjust = 'center')) +
  xlim(0,1)
```



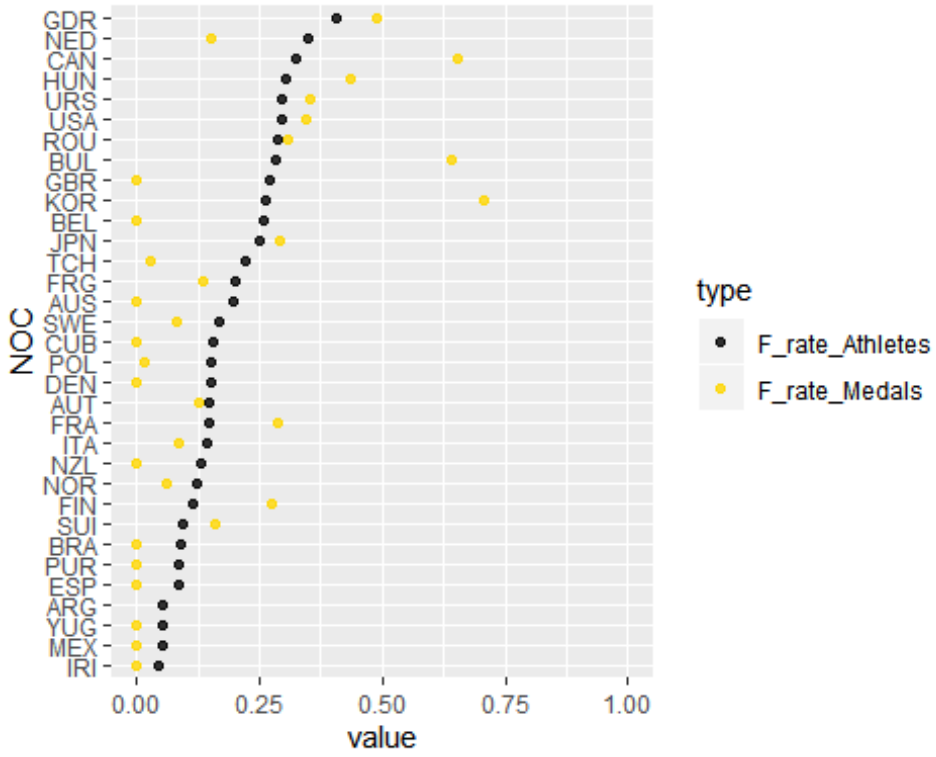
1936 년 여성 선수 참가 비율은 25%를 넘기지 않았다. 그러나 일부 국가의 여성 선수들이 두드러지는 성과를 보여 해당 국가의 여성 메달 획득 비율 자체는 높았던 것을 확인할 수 있다.

```r
years_1956 <- years %>%
  filter(Year == 1956)

lev_1956 <- years_1956 %>%
  filter(type == 'F_rate_Athletes') %>%
  arrange(value) %>%
  select(NOC)

ggplot(years_1956, aes(x= value, y= NOC, color=type)) +
  geom_point(na.rm = TRUE, alpha = 0.8) +
  scale_color_manual(values = c("black", "gold")) +
  theme(plot.title = element_text(hjust = 'center')) +
  xlim(0,1)
```



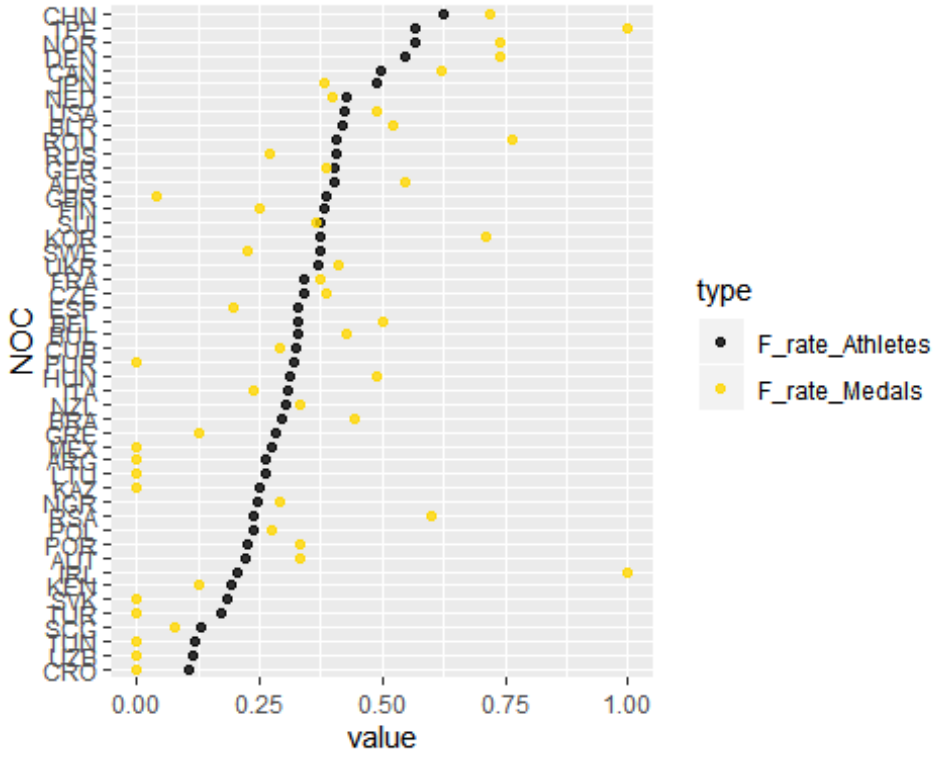
1936 년에 비해 여성 선수 출전 국가가 많아진 것을 확인할 수 있다. 대체적으로 공산주의 국가들(루마니아, 폴란드, 헝가리, 러시아)의 여성 선수 메달 획득 비율이 높은 것을 확인할 수 있다. 조사 결과 서구권 국가들의 보이콧이 있었던 해였다. 여성 선수 출전 국가가 많아짐에 따라 각 나라별 메달 획득 비율은 전반적으로 떨어진 것을 확인할 수 있다.

```r
years_1976 <- years %>%
  filter(Year == 1976)

lev_1976 <- years_1976 %>%
  filter(type == 'F_rate_Athletes') %>%
  arrange(value) %>%
  select(NOC)
ggplot(years_1976, aes(x= value, y= NOC, color=type)) +
  geom_point(na.rm = TRUE, alpha = 0.8) +
  scale_color_manual(values = c("black", "gold")) +
  theme(plot.title = element_text(hjust = 'center')) +
  xlim(0,1)
```



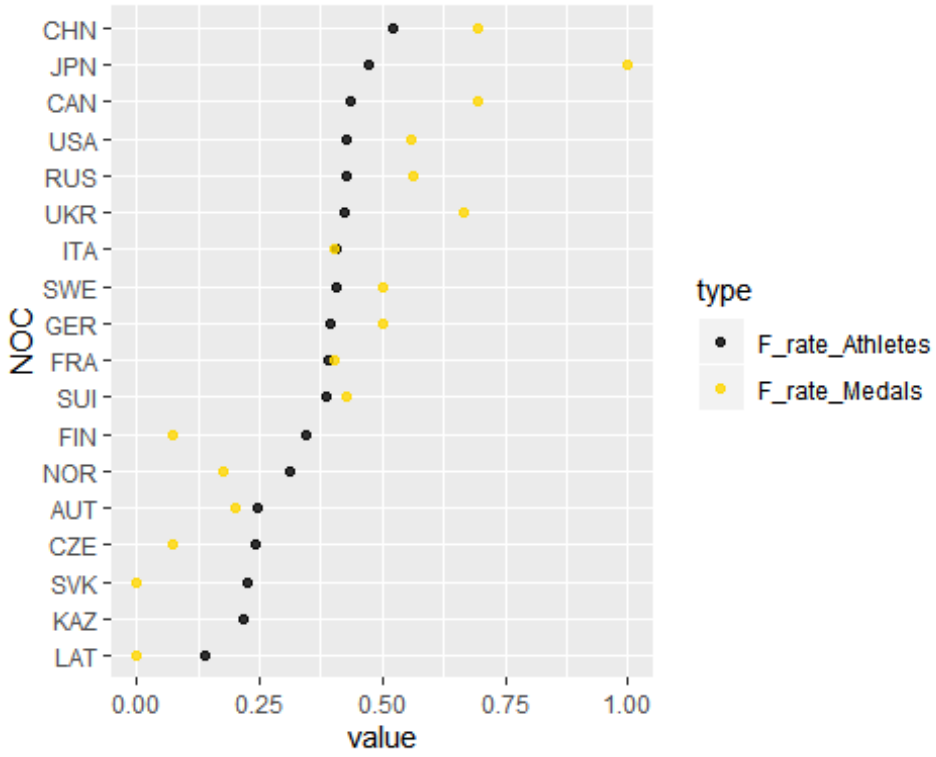
1956 년에 비해서 여성 선수 출전 국가가 는 것을 확인할 수 있다. 1956 년에 보이콧했던 나라들이 참가하면서 상위에 자본주의 서구권 국가들의 이름이 보이는 것을 확인할 수 있다. 이 리스트를 통해 1956 년에 목록 자체에서 없던 대한민국 또한 이름을 올려 여성 참가율과 메달 획득 비율이 높아진 것을 확인할 수 있다.

```r
years_1996 <- years %>%
  filter(Year == 1996)

lev_1996 <- years_1996 %>%
  filter(type == 'F_rate_Athletes') %>%
  arrange(value) %>%
  select(NOC)
ggplot(years_1996, aes(x= value, y= NOC, color=type)) +
  geom_point(na.rm = TRUE, alpha = 0.8) +
  scale_color_manual(values = c("black", "gold")) +
  theme(plot.title = element_text(hjust = 'center')) +
  xlim(0,1)
```



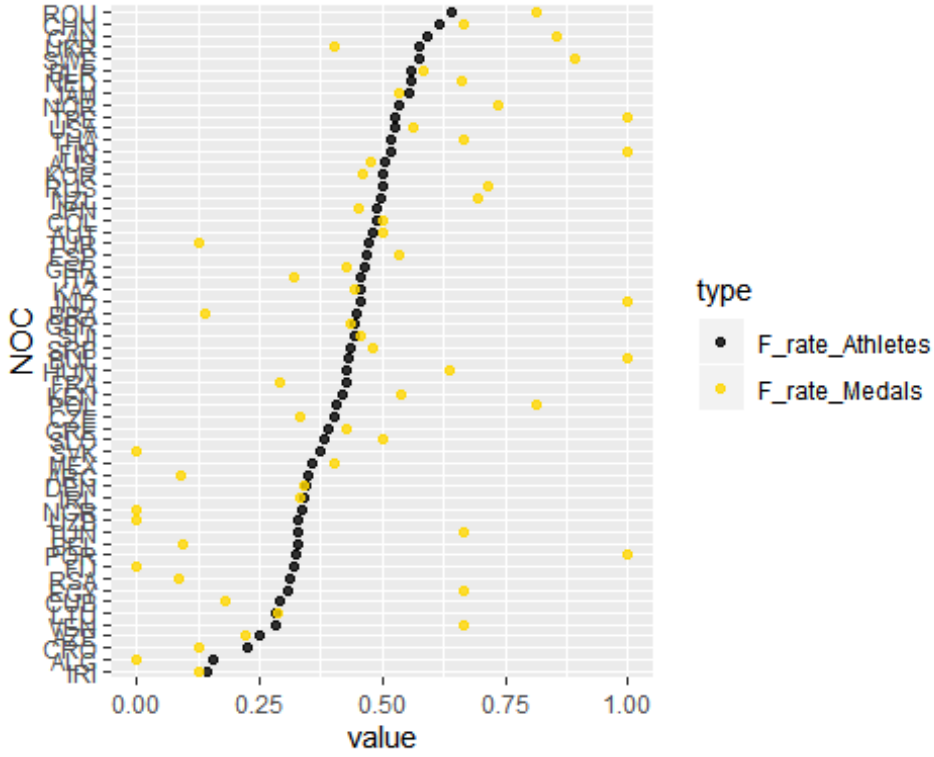
여성 참가 국가 자체가 점차 느는 것을 확인할 수 있다. 여성 선수 참여비율이 50%가 넘는 국가들도 나타난다. 1984 년 대한민국 여성 최초 금메달 획득 이후, 여성 메달 획득 비율 (60% 이상) 또한 늘었음을 확인할 수 있다. 일부 국가에서는 (타이완, 아일랜드) 여성 선수가 그 해의 메달을 모두 딴 경우도 확인할 수 있다.

```r
years_2006 <- years %>%
  filter(Year == 2006)

lev_2006 <- years_2006 %>%
  filter(type == 'F_rate_Athletes') %>%
  arrange(value) %>%
  select(NOC)
ggplot(years_2006, aes(x= value, y= NOC, color=type)) +
  geom_point(na.rm = TRUE, alpha = 0.8) +
  scale_color_manual(values = c("black", "gold")) +
  theme(plot.title = element_text(hjust = 'center')) +
  xlim(0,1)
```



2006 년 올림픽에는 출전국가 자체가 줄어 여성 선수 출전 국가 자체가 1996 년에 비해 줄었다. (1996 년 197 개국 출전, 2006 년 80 개국 출전, 동계 올림픽 참여 국가가 하계 올림픽 참여 국가보다 적다) 일부 국가에서는 (일본) 여성 선수가 그 해의 메달을 모두 딴 경우를 확인할 수 있다.

```r
years_2016 <- years %>%
  filter(Year == 2016)

lev_2016 <- years_2016 %>%
  filter(type == 'F_rate_Athletes') %>%
  arrange(value) %>%
  select(NOC)
ggplot(years_2016, aes(x= value, y= NOC, color=type)) +
  geom_point(na.rm = TRUE, alpha = 0.8) +
  scale_color_manual(values = c("black", "gold")) +
  theme(plot.title = element_text(hjust = 'center')) +
  xlim(0,1)
```



2016 년 여성 선수 참여 국가는 기존보다 확 늘었다는 것을 확인할 수 있다. 여성 선수 메달 획득 비율 또한 기존보다 크게 늘었다는 것을 확인할 수 있다. 대한민국 또한 15 위권에 이름을 올리는 것을 확인할 수 있다.

3. Age

```
Sys.setlocale('LC_ALL','C')

olympic$Age%>%
  summary

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   10.00   21.00   24.00   25.56   28.00   97.00    9474
```

Age 변수의 Summary 를 통해 분포를 대략적으로 살펴본 결과 최연소 참가자의 나이는 10 세, 최고령 참가자가 97 세임을 알 수 있었다. 또한 NA 값이 대략 9474 개가 있는 것으로 보아 올림픽 초기의 나이 기록일 것이라고 추측해볼 수 있다. IQR 의 값이 7 정도로 되는 것을 보아서 Age 의 대략적인 분포는 중앙 쪽에 몰려있는 굉장히 좁은 분포로 추측할 수 있다.

나이 missing value 살펴보기

```
olympic %>% filter(is.na(Age)) %>%
  summarise(id = n_distinct(ID))

## # A tibble: 1 x 1
##      id
##   <int>
## 1  6368
```

총 6368 명의 선수가 나이 정보가 존재하지 않음을 알 수 있다.

1) NA 값은 올림픽 초기의 나이 기록일 것

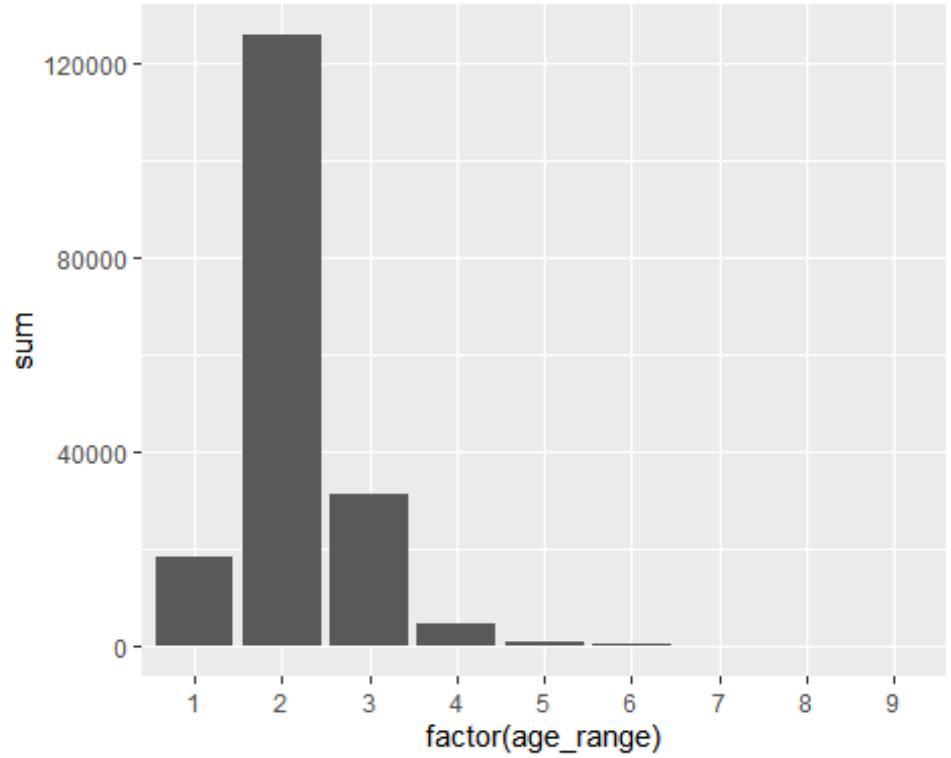
```
no_age <- olympic %>% filter(is.na(Age))
no_age %>% group_by(Year) %>% summarise(unique = n_distinct(ID)) %>%
  arrange(desc(unique))

## # A tibble: 28 x 2
##     Year unique
##    <dbl>  <int>
##  1  1948    881
##  2  1924    748
##  3  1928    718
##  4  1900    586
##  5  1920    553
##  6  1908    487
##  7  1956    476
##  8  1906    449
##  9  1932    261
```

```
## 10  1952    194
## # ... with 18 more rows
```

현재보다는 과거 게임의 데이터에서 나이에 대한 missing value 가 많이 일어났음을 알 수 있다. 그러나 가설이 참이라고 결론 지을 만큼 뚜렷한 패턴이 존재하지는 않는다. 혹은 해당 가설 이외에도, 1948 년 혹은 1920 년 등이 세계대전 직후 등임을 고려했을 때 데이터 손실과 전쟁의 연관성에 대해 추후 생각해볼 수 있을 듯하다.

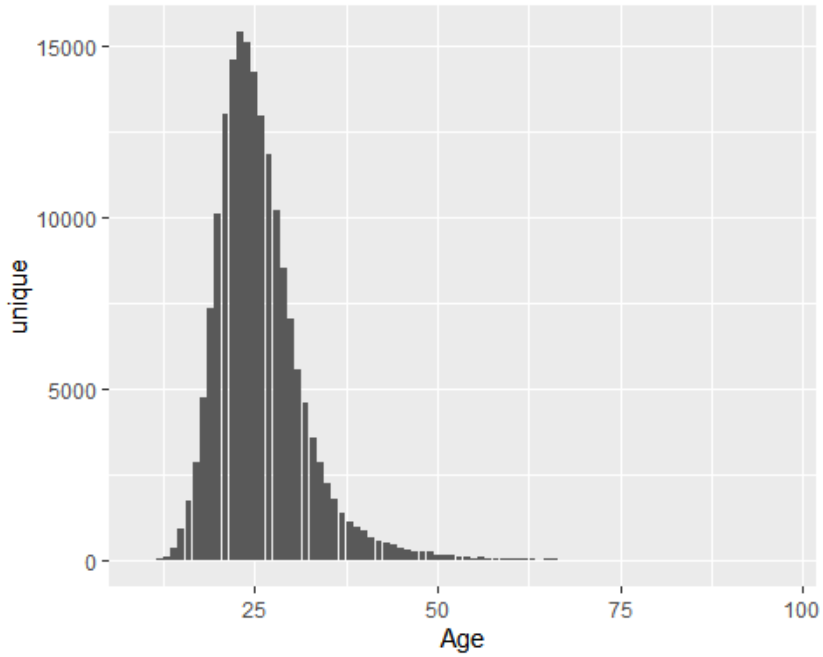
Age distribution

```
olympic %>% filter(!is.na(Age)) %>%
  group_by(Age) %>% summarise(unique = n_distinct(ID)) %>%
  mutate(age_range = Age %/% 10) %>%
  group_by(age_range) %>% summarise(sum = sum(unique)) %>%
  ggplot(aes(x = factor(age_range), y = sum)) +
  geom_bar(stat = 'identity')
```



연령대별 분포는 20, 30, 10, ... 순으로 많음을 알 수 있다.

```
olympic %>% filter(!is.na(Age)) %>%
  group_by(Age) %>% summarise(unique = n_distinct(ID)) %>%
  ggplot(aes(x = Age, y = unique)) +
  geom_bar(stat = 'identity')
```



Age 에 대한 histogram 이다. 오른쪽으로 skewed 되어있는 형태이며, 대략 25 세 부근에 데이터가 집중되어 있음을 알 수 있다. 아까 살펴보았듯, 97 세의 고령 선수의 데이터 때문에 x range 가 100 정도까지 표현이 되어 있음을 알 수 있다. 여기서 올림픽 고령 참가자들이 존재하고 있음을 알 수 있는데, 이러한 고령 참가선수 데이터에 대해 더 살펴볼 필요가 있을 듯하다.

고령 선수들의 종목 살펴보기

```
olympic %>% summarise(n_distinct(Sport))
```

```
## # A tibble: 1 x 1
##   `n_distinct(Sport)`
##                 <int>
## 1                  66
```

올림픽 종목으로 선정되었던 종목들은 총 66 개이다.

```
# 선수별
olympic %>% filter(Age >= 60) %>%
  group_by(Sport) %>%
  summarise(unique = n_distinct(ID)) %>%
  arrange(desc(unique))
```

```
## # A tibble: 8 x 2
##   Sport           unique
##   <chr>            <int>
```

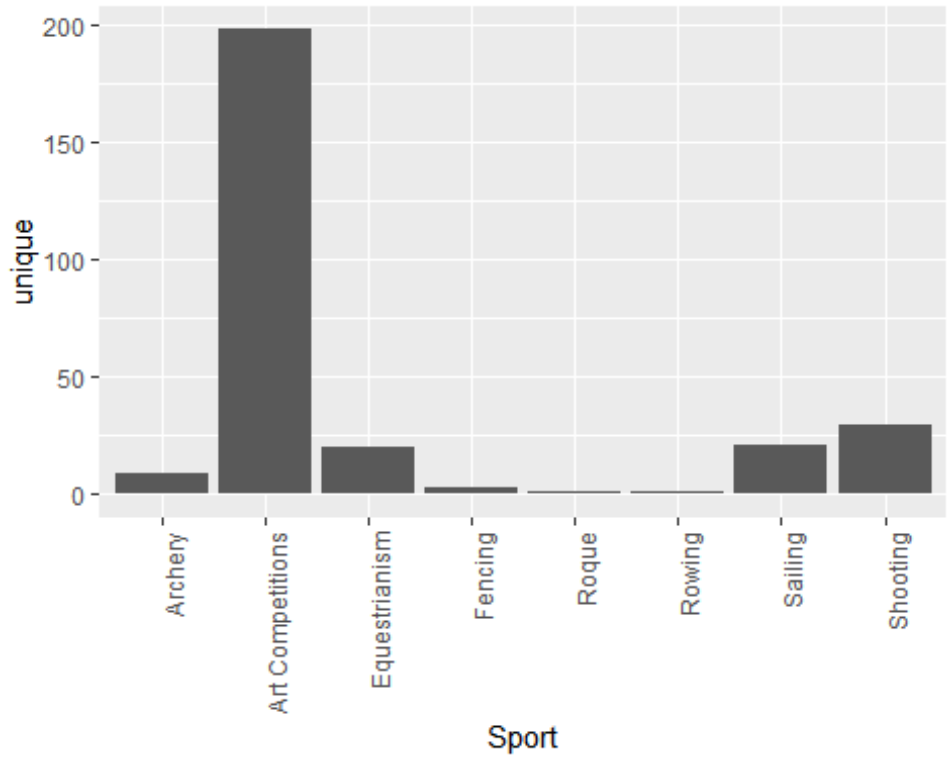
```
## 1 Art Competitions    198
## 2 Shooting             30
## 3 Sailing              21
## 4 Equestrianism        20
## 5 Archery               9
## 6 Fencing               3
## 7 Roque                 1
## 8 Rowing                1
```

```
# 경기별
olympic %>% filter(Age >= 60) %>%
  group_by(Sport) %>%
  count() %>% arrange(desc(n))
```

```
## # A tibble: 8 x 2
## # Groups:   Sport [8]
##   Sport              n
##   <chr>          <int>
## 1 Art Competitions   508
## 2 Shooting            57
## 3 Equestrianism       41
## 4 Sailing             28
## 5 Archery             19
## 6 Fencing              3
## 7 Roque                1
## 8 Rowing               1
```

고령자가 많이 출전한 종목의 결과를 선수별로 살펴보았을 때의 순위는 'Art -> Shooting -> Sailing -> Equestrianism -> Archery -> Fencing -> Roque -> Rowing' 이고, 고령자 출전 경기가 많은 종목의 순위는 'Art -> Shooting -> Equestrianism -> Sailing -> Archery -> Fencing -> Roque -> Rowing' 으로 비슷하나, Sailing 과 Equestrianism(승마) 간의 순위가 뒤바뀌었다. 승마는 한 명의 선수가 많은 게임(혹은 다양한 세부 종목)에 참가한 것으로 이해할 수 있다.

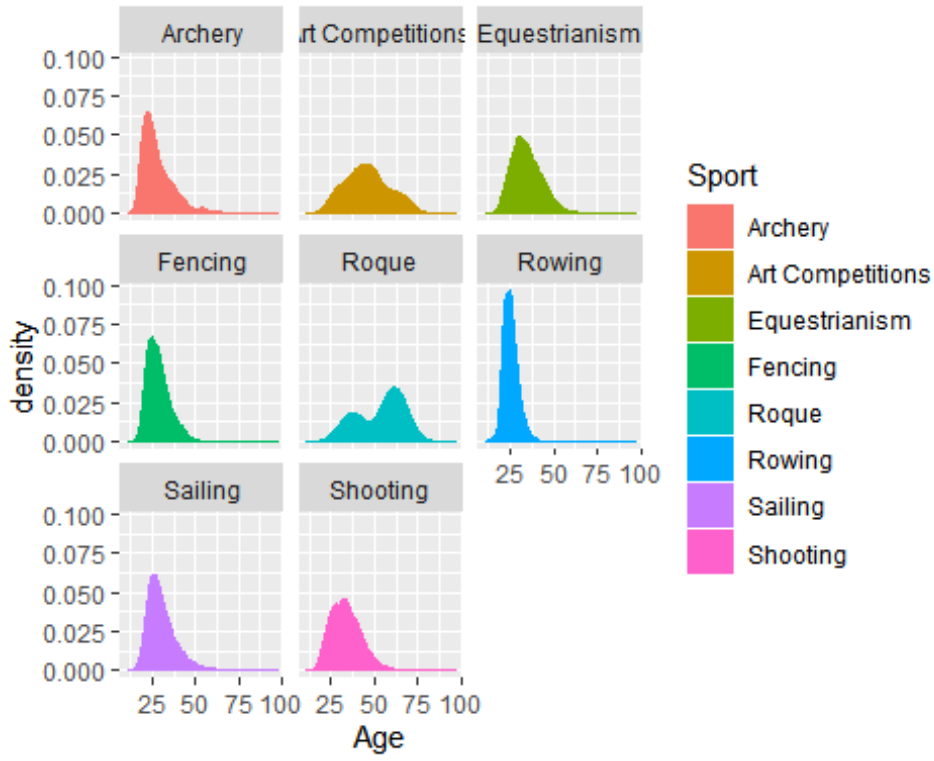
```
olympic %>% filter(Age >= 60) %>%
  group_by(Sport) %>%
  summarise(unique = n_distinct(ID)) %>%
  ggplot(aes(x = Sport, y = unique)) +
  geom_bar(stat = 'identity') +
  theme(axis.text.x=element_text(angle=90, hjust=1))
```



각 종목별 60 세 이상의 고령 선수들의 수를 나타낸 자료이다. 이때 60 세 이상의 고령 선수들이 출전하는 종목은 66 개 중 총 8 개로 굉장히 제한되어 있음을 알 수 있다. 또한 고령 출전자는 대부분 Art Competitions 에 출전함을 알 수 있으며 그 다음은 사격, 승마 등의 순서로 많았다. 아주 적은 참가자지만 펜싱, 조정, 로크 등에도 출전했음을 알 수 있다.

Art Competitions

```
elder_sports <- olympic %>% filter(Age >= 60) %>%
  group_by(Sport) %>%
  summarise(unique = n_distinct(ID)) %>% pull(Sport)

olympic %>% filter(Sport %in% elder_sports) %>%
  filter(!is.na(Age)) %>%
  ggplot(aes(x = Age, fill = Sport, color = Sport)) +
  geom_density() +
  facet_wrap(~Sport)
```



고령 출전 선수들의 종목의 분포를 살펴보았을 때 Art Competitions 와 Roque 의 경우 다른 종목들과 비교해서 출전 선수들의 나이가 상대적으로 많음을 알 수 있다. Roque 의 경우에는 애초에 두 개의 peak 값이 있고, 특히 가장 높은 peak 는 대략 60 대임을 알 수 있다.

저연령 선수 출전 종목

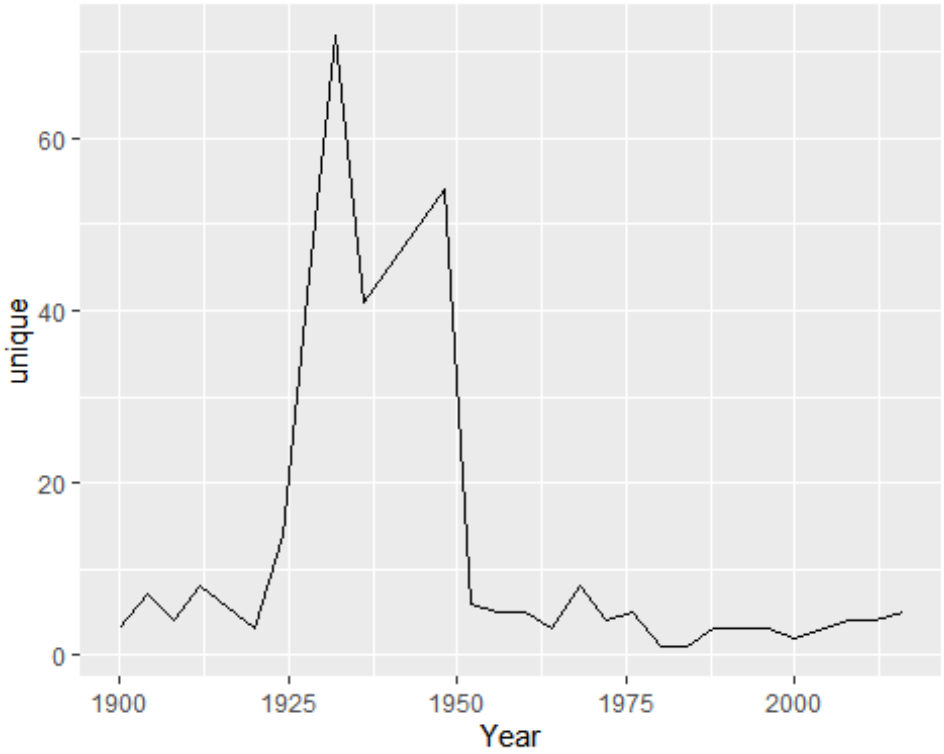
```
olympic %>% filter(Age <= 12) %>%
  group_by(Sport) %>%
  summarise(unique = n_distinct(ID)) %>% arrange(desc(unique))
```

```
## # A tibble: 6 x 2
##   Sport          unique
##   <chr>           <int>
## 1 Figure Skating     15
## 2 Swimming           12
## 3 Gymnastics          5
## 4 Rowing              4
## 5 Athletics           1
## 6 Diving              1
```

12 세 이하의 나이가 어린 선수들의 올림픽 출전 종목은 피겨스케이팅, 수영, 체조, 조정 등이 있었다.

고령 선수들이 나온 데에는 trend 가 있을까?

```
olympic %>% filter(Age >= 60) %>%
  group_by(Year) %>% summarise(unique = n_distinct(ID)) %>%
  ggplot(aes(x = Year, y = unique)) +
  geom_line()
```



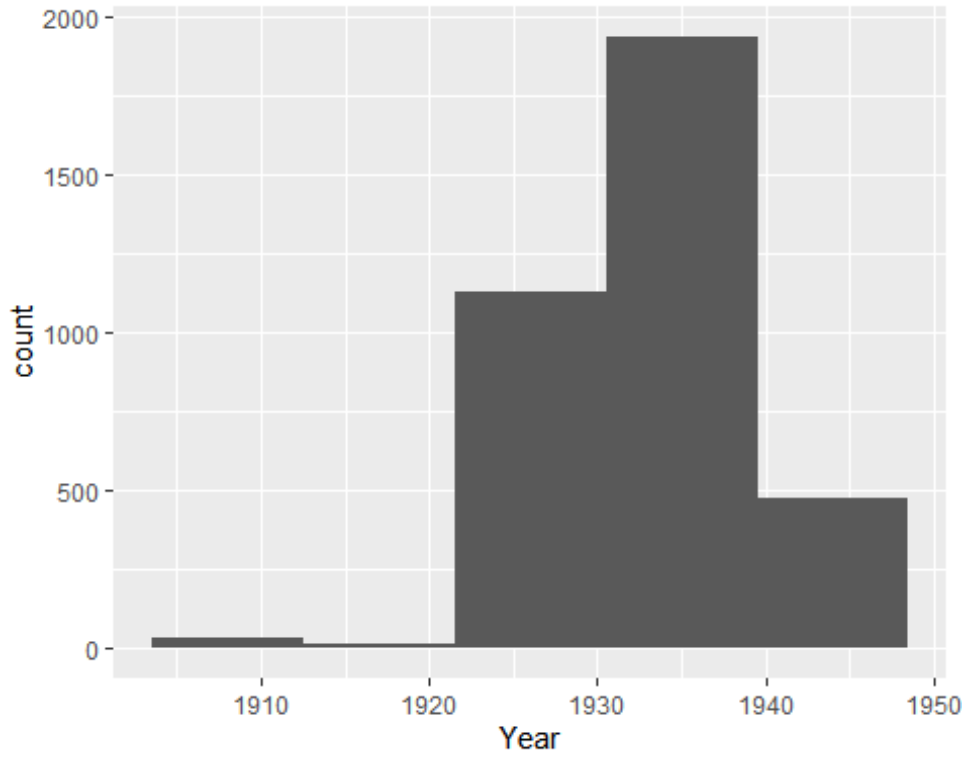
고령 선수들이 출전하는 데에 증가 혹은 감소하는 trend 가 보이지는 않는다. 다만, 1925 ~ 1950 년 사이에 많이 출전을 했음을 알 수 있고 이후 급감하고 별 다른 증감추세를 보이지 않고 있다.

2) 고령 선수들의 출전이 25 년부터 50 년 사이인 이유: 최다 출전 종목인 'Art Competitions'가 1950 년 이후에는 사라졌을 것이다.

```
olympic %>% filter(Sport == 'Art Competitions') %>%
  group_by(Year) %>%
  count() %>% arrange(desc(Year)) %>% head(1)
```

```
## # A tibble: 1 x 2
## # Groups:   Year [1]
##    Year     n
##   <dbl> <int>
## 1  1948   471
```

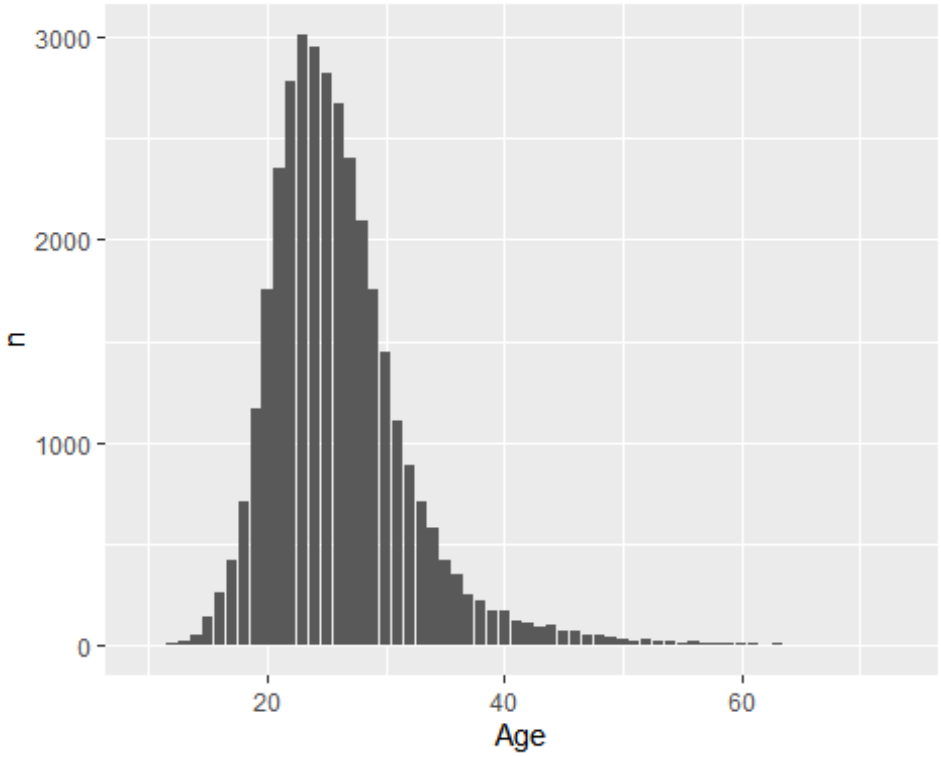
```r
olympic %>% filter(Sport == 'Art Competitions') %>%
  ggplot(aes(x = Year)) +
  geom_histogram(bins = 5)
```



실제로 Art Competitions 종목은 1925 년에서 1950 년 사이에 거의 대다수가 분포해 있으며, 1948 년이 마지막임을 알 수 있다. 따라서 가설은 참이라고 결론 지을 수 있다.

Age 별 메달의 분포

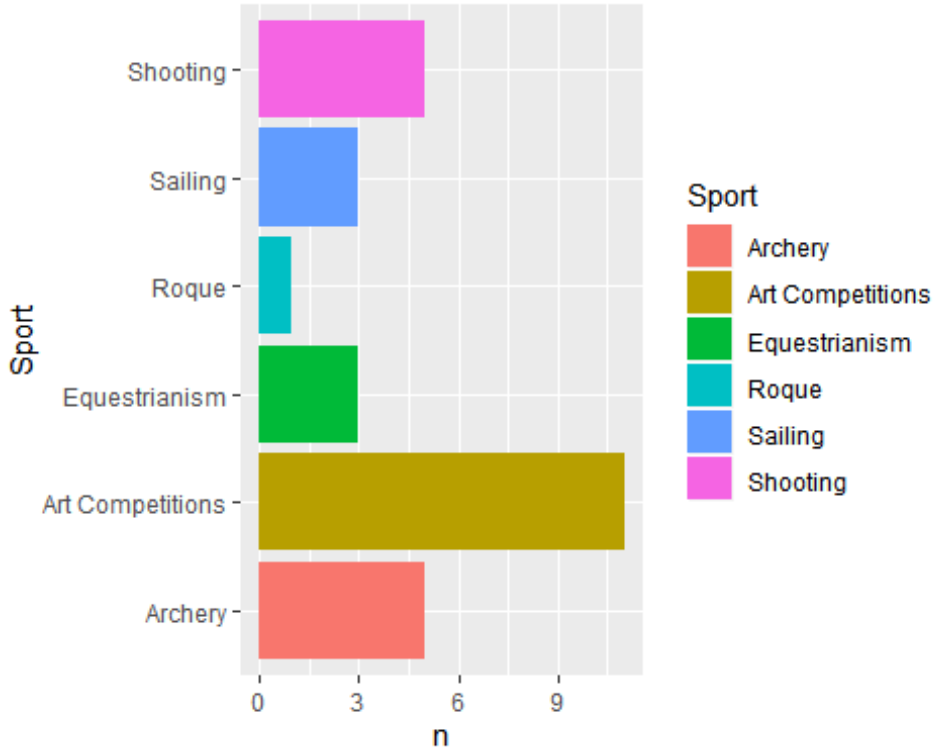
```r
olympic %>% filter(!is.na(Medal)) %>% filter(!is.na(Age)) %>%
  group_by(Age) %>% summarise(n = n_distinct(ID)) %>%
  ggplot(aes(x = Age, y = n)) +
  geom_bar(stat = 'identity')
```



기존 Age 의 분포 자체와 큰 변화는 없다. 다만 주목할 만한 것은, 여전히 skewed 되어 있다는 점에서 고령 출전 선수도 메달을 땄음을 알 수 있다. 메달을 딴 종목이 무엇인지 보고, 이때 메달을 딴 것이 이례적인 것인지 혹은 선수들의 분포 자체가 고령이었는지를 보고자 한다.

고령 선수들의 메달 종목

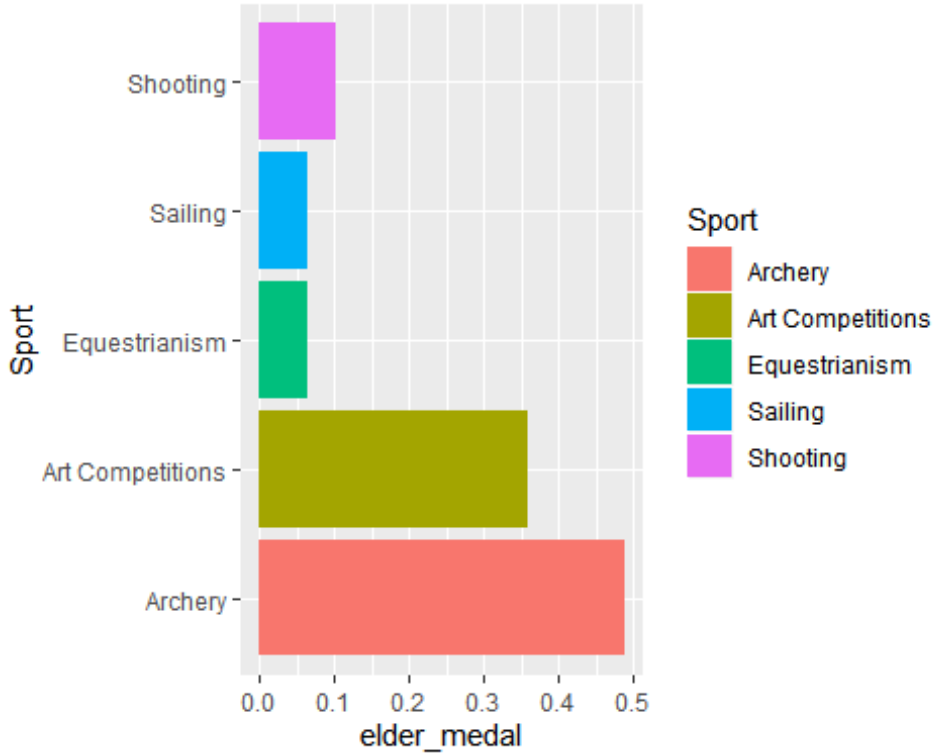
```
olympic %>% filter(Age >= 60) %>%
  filter(!is.na(Medal)) %>%
  group_by(Sport) %>%
  summarise(n = n_distinct(ID)) %>%
  ggplot(aes(x = Sport, y = n)) +
  geom_bar(stat = 'identity', aes(fill = Sport)) +
  coord_flip()
```



앞서 살펴 본 결과 고령 선수들의 출전 종목은 Art, Shooting, Equestrianism, Sailing, Archery, Fencing, Roque, Rowing 등이 있었다. 여기서 펜싱과 조정을 제외하고 Sport 종목에서는 다 메달을 획득했음을 알 수 있다.

종목별 고연령자들이 메달을 차지한 비율 (경기별)

```
olympic %>% filter(!is.na(Age)) %>%
  group_by(Sport) %>%
  summarise(elder_medal = 100*mean(Age >= 60 & !is.na(Medal))) %>%
  filter(elder_medal != 0) %>%
  arrange(desc(elder_medal)) %>%
  filter(Sport != 'Roque') %>%
  ggplot(aes(x = Sport, y = elder_medal, fill = Sport)) +
  geom_bar(stat = 'identity') +
  coord_flip()
```



종목별로 메달 입상자 중 고령자의 비율을 살펴본 결과이다. Roque 의 경우가 가장 메달을 딴 비율(%)이 높았으나, 해당 종목은 게임 한 번 만의 기록이므로 의미 있는 기록은 아니므로 제외하고 그림을 그린 결과, Archery(양궁)가 가장 높고 Art Competitions 가 그 뒤를 이었다. 비율 자체가 크진 않으나 Archery 의 경우 노령 선수들의 활약이 상대적으로 가장 두드러졌음을 알 수 있다.

4. Nations

```
gdp<-read.csv('c:/temp/gdp.csv', stringsAsFactors = TRUE)
gdp<-as_tibble(gdp)
```

1) 메달 수가 많은 국가와 GDP 변화의 연관성

1. 총 메달 수 많은 국가

```
olympicTot<-olympic%>%
  filter(Medal!=0)%>%
  group_by(NOC)%>%
  summarise(medal_tot=n(),na.rm=TRUE)%>%
  select(-na.rm)%>%
  arrange(desc(medal_tot))%>%
  filter(rank(desc(medal_tot))<=10)
olympicTot
```

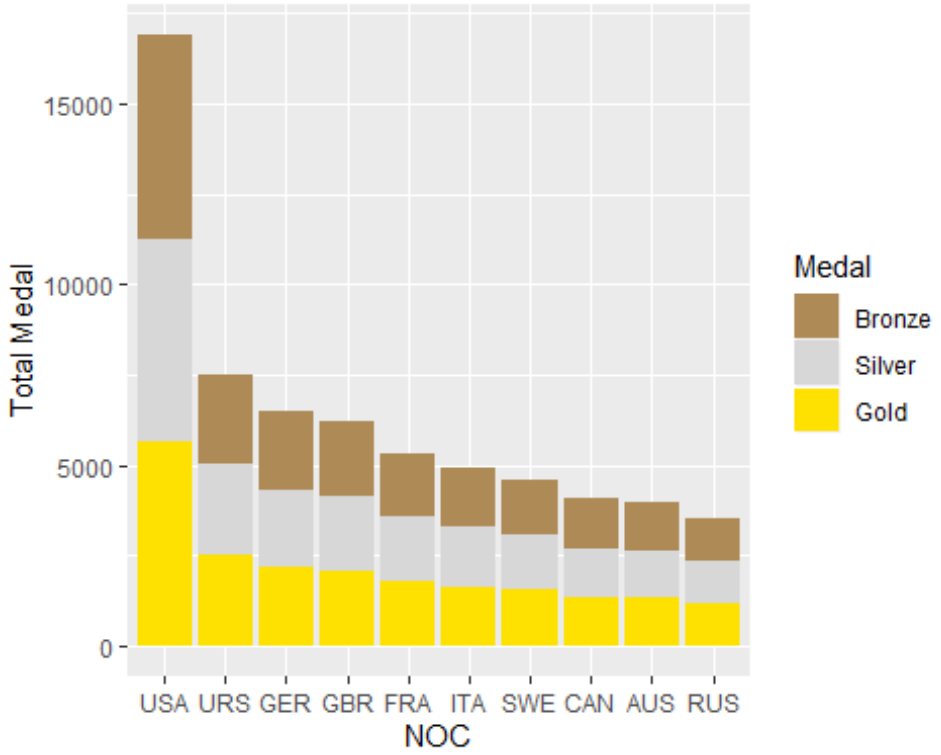
```
##    NOC   medal_tot
##    <fct>     <int>
##  1 USA        5637
##  2 URS        2503
##  3 GER        2165
##  4 GBR        2068
##  5 FRA        1777
##  6 ITA        1637
##  7 SWE        1536
##  8 CAN        1352
##  9 AUS        1320
## 10 RUS        1165
```

2. 추출 국가의 금은동 비율

```
left_join(olympicTot,olympic,by="NOC")%>%
  group_by(NOC,medal_tot,Medal)%>%
  summarise(medal_sum=n())%>%
  spread(key=Medal,value=medal_sum)
## # Groups:   NOC, medal_tot [230]
##    NOC   medal_tot   `0` Bronze  Gold Silver
##    <fct>     <int> <int>  <int> <int>  <int>
##  1 AUS        1320  6318    517   348    455
##  2 CAN        1352  8381    451   463    438
##  3 FRA        1777 10981    666   501    610
##  4 GBR        2068 10188    651   678    739
##  5 GER        2165  7665    746   745    674
##  6 ITA        1637  9078    531   575    531
##  7 RUS        1165  3978    408   390    367
```

```
##  8 SWE        1536  6803    535   479    522
##  9 URS        2503  3182    689  1082    732
## 10 USA        5637 13216   1358  2638   1641
```

```
olympicMedal<-left_join(olympicTot,olympic,by="NOC")%>%
  group_by(NOC,medal_tot,Medal)%>%
  summarise(medal_sum=n())%>%
  spread(key=Medal,value=medal_sum)%>%
  gather('Bronze':'Silver',key="Medal",value="medal_sum")%>%
  arrange(NOC,Medal)

ggplot(olympicMedal,aes(x=reorder(NOC,-medal_tot,sum),medal_tot,fill=ordered
(Medal,levels=c("Bronze","Silver","Gold"))))+geom_bar(stat="identity")+xlab("
NOC")+ylab("Total Medal")+labs(fill="Medal")+scale_fill_manual(values=c("#ad8
a56","#d7d7d7","#fee101"))
```

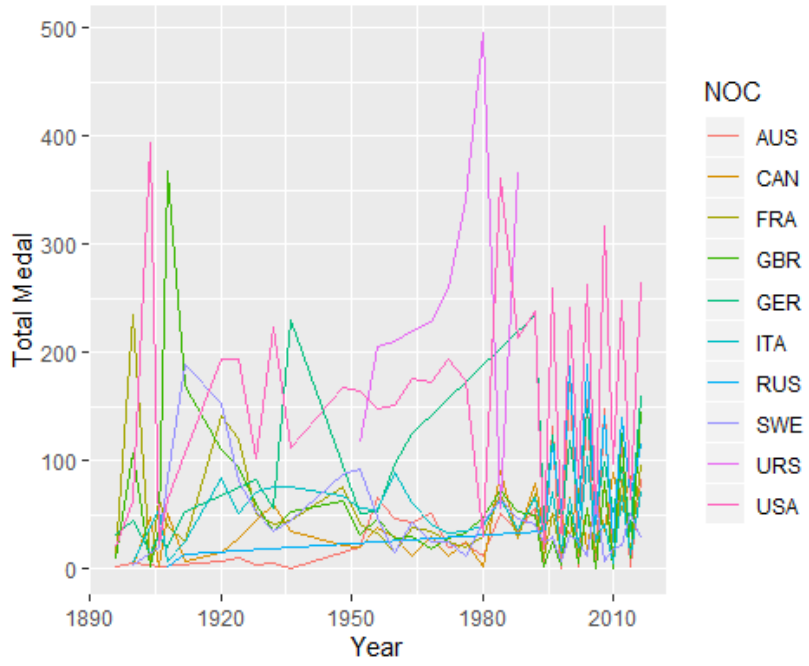


USA	URS	GER	GBR	FRA	ITA	SWE	CAN	AUS	RUS
미국	소련	독일	영국	프랑스	이탈리아	스웨덴	캐나다	오스트레일리아	러시아

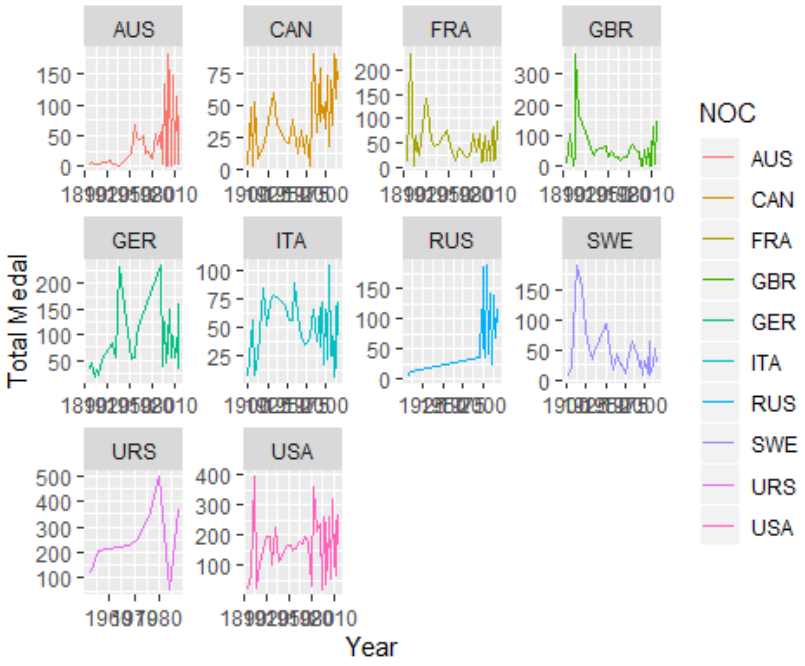
추출된 국가가 획득한 메달의 비율에 대한 그림을 그려보았다. 국가 내에서 봤을 때, 획득한 금, 은, 동메달의 비율은 비슷하고, 그에 따라서 국가 간 비교를 해보면 전체 메달이 많은 국가가 각 메달의 수도 많다는 것을 알 수 있다.

3-1. 추출 국가의 메달 수 변화

```
olympicTot<-olympic%>%
  filter(Medal!=0)%>%
  group_by(NOC)%>%
  summarise(medal_tot=n(),na.rm=TRUE)%>%
  arrange(desc(medal_tot))%>%
  filter(rank(desc(medal_tot))<=10)%>%
  select(NOC)
olympicYear<-olympic%>%
  filter(Medal!=0)%>%
  group_by(NOC,Year)%>%
  summarise(medal_tot=n(),na.rm=TRUE)%>%
  arrange(NOC,Year)


inner_join(olympicTot,olympicYear,by="NOC")%>%
  select(-na.rm)%>%
  ggplot(aes(Year,medal_tot,group=NOC))+geom_line(aes(color=NOC))+ylab("Total
 Medal")
```

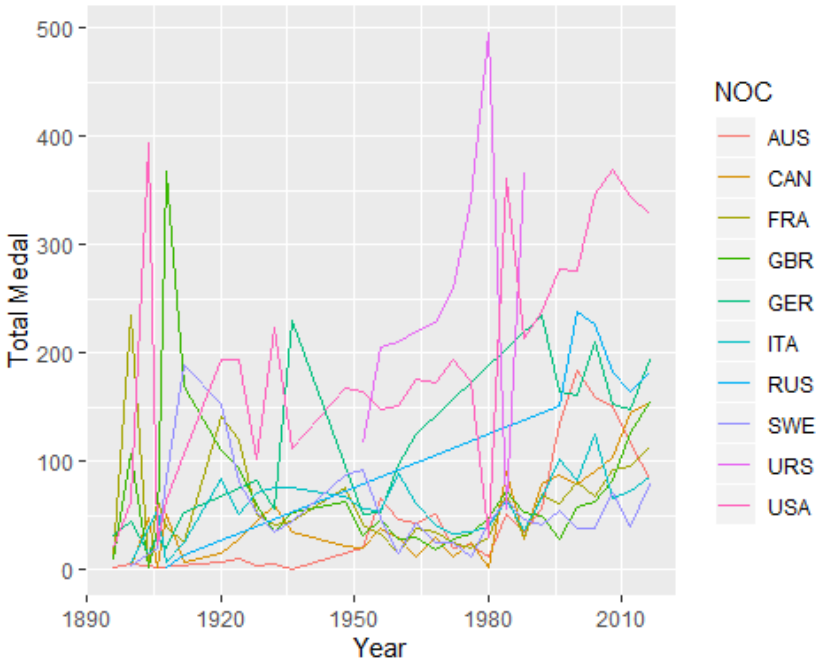



```
inner_join(olympicTot,olympicYear,by="NOC")%>%
  select(-na.rm)%>%
  ggplot(aes(Year,medal_tot,group=NOC))+geom_line(aes(color=NOC))+facet_wrap
(~NOC,scales="free")+ylab("Total Medal")
```

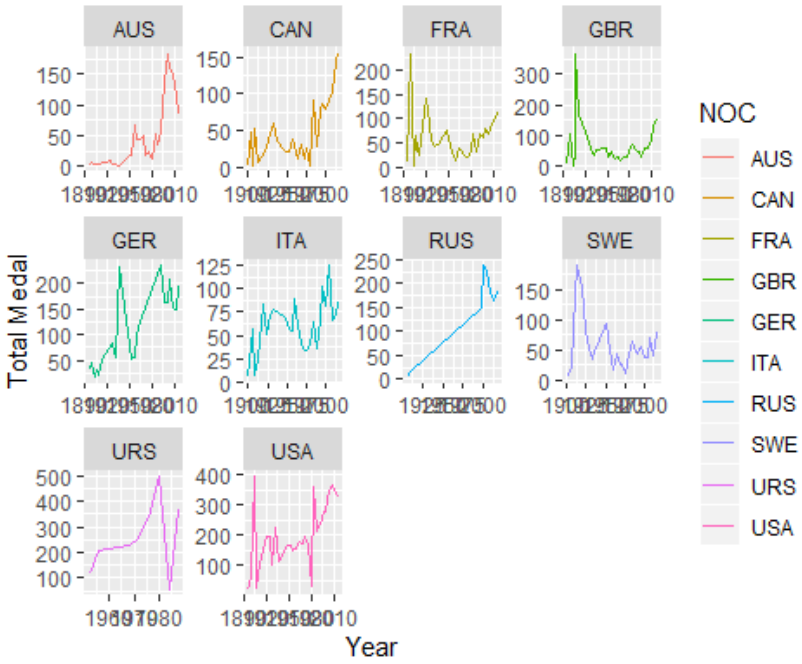


10 개 국가에 대해 전체 메달 수의 변화 그림을 그려보았고, 전반적으로 오스트레일리아, 캐나다, 러시아, 미국은 메달 수가 증가했고, 스웨덴은 감소했으며 프랑스, 영국은 감소 후 증가, 이탈리아, 독일은 증가 후 감소하는 것처럼 보이는 것을 알 수 있다. 약 1990 년대부터 메달 수 그래프의 변화가 급격해 진 것이 보이는데, 1994 년부터 동계, 하계 올림픽을 나눠서 개최했기 때문이다. 이렇게 나타난 지그재그 패턴으로 인해 전체적인 파악이 힘들어 나누어 개최된 것을 하나로 합쳐 그림을 그려보았다.

```
olympicTot2 <- olympicTot%>%
  mutate(Year=stringr::str_replace(Year, "1994", "1996"))%>%
  mutate(Year=stringr::str_replace(Year, "1998", "2000"))%>%
  mutate(Year=stringr::str_replace(Year, "2002", "2004"))%>%
  mutate(Year=stringr::str_replace(Year, "2006", "2008"))%>%
  mutate(Year=stringr::str_replace(Year, "2010", "2012"))%>%
  mutate(Year=stringr::str_replace(Year, "2014", "2016"))
olympicTot2$Year<-as.integer(olympicTot2$Year)

olympicTot2<-olympicTot2%>%
  group_by(NOC,Year)%>%
  summarise(medal_tot=sum(medal_tot),na.rm=TRUE)%>%
  arrange(medal_tot)



ggplot(olympicTot2,aes(Year,medal_tot,group=NOC))+geom_line(aes(color=NOC))+y
lab("Total Medal")
```

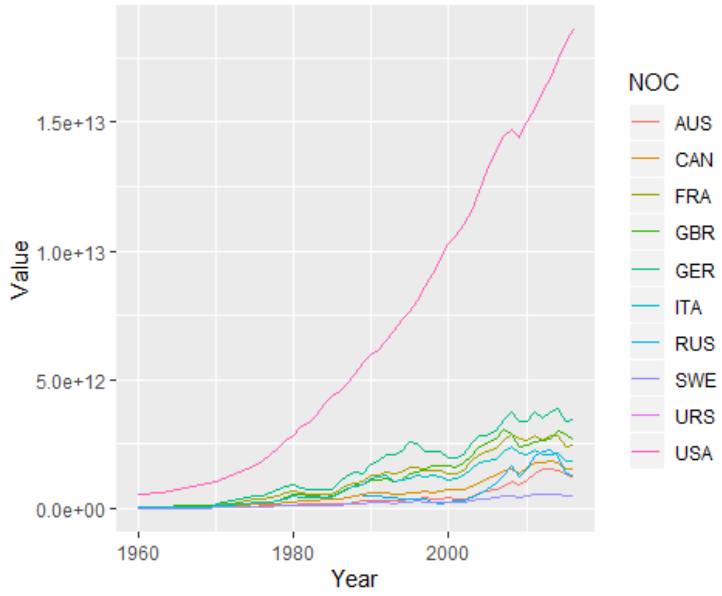



```
ggplot(olympicTot22,aes(Year,medal_tot,group=NOC))+geom_line(aes(color=NOC))+
ylab("Total Medal")+facet_wrap(~NOC,scales="free")+ylab("Total Medal")
```



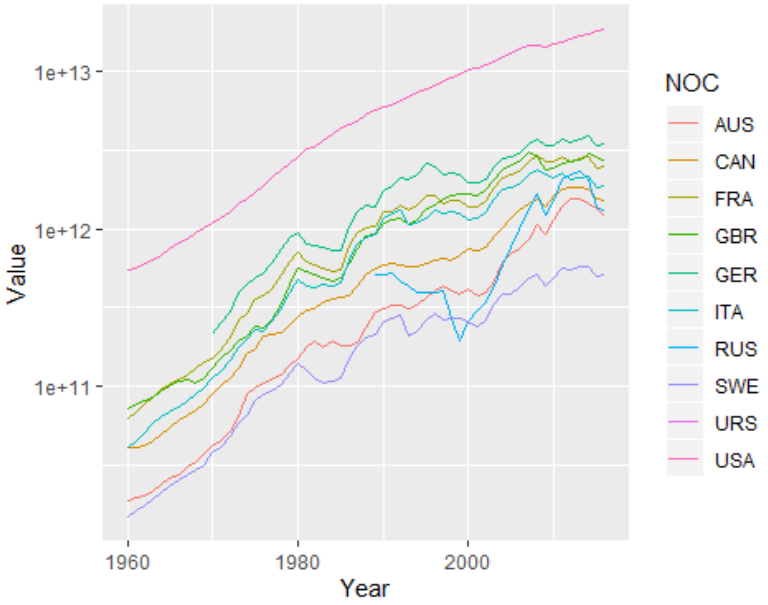
3-2. 추출 국가 의 GDP 변화

```
left_join(olympicTot,gdp,by="NOC")%>%
 ggplot(aes(Year,Value,group=NOC))+geom_line(aes(color=NOC))
```



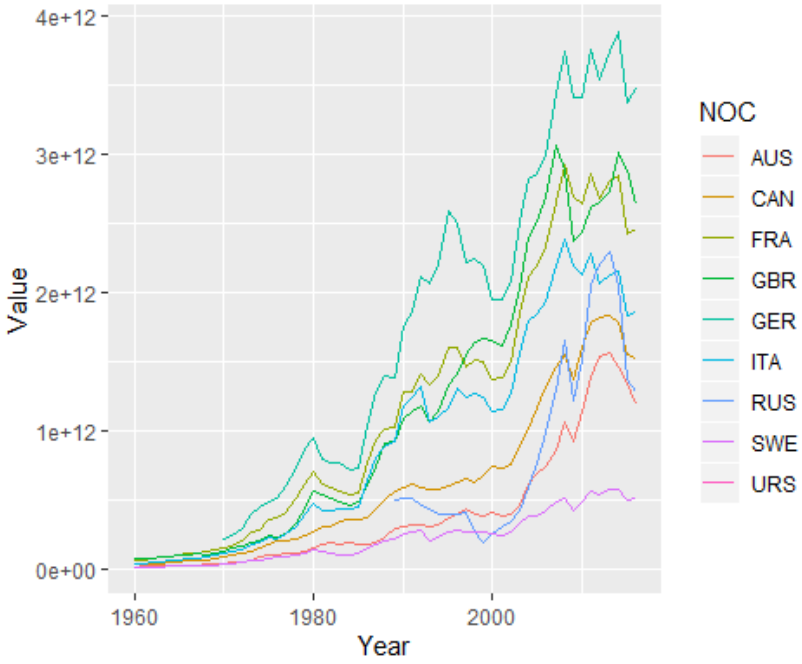
GDP 증가율과 메달 수 간에 상관관계 유무를 파악하고자 메달 수가 가장 많은 10 개 국가의 GDP 변화에 대한 그래프를 그려본 결과, 미국과 다른 국가 간의 데이터 스케일의 차이로 인해 국가별 GDP 의 변화를 파악하기 힘들었다. 따라서 GDP 값에 로그 변환을 해준 그래프와 미국의 자료를 제외한 그래프, 총 2 개의 그래프를 그려보았다.

```
left_join(olympicTot,gdp,by="NOC")%>%
  ggplot(aes(Year,Value,group=NOC))+geom_line(aes(color=NOC))+scale_y_log10()
```

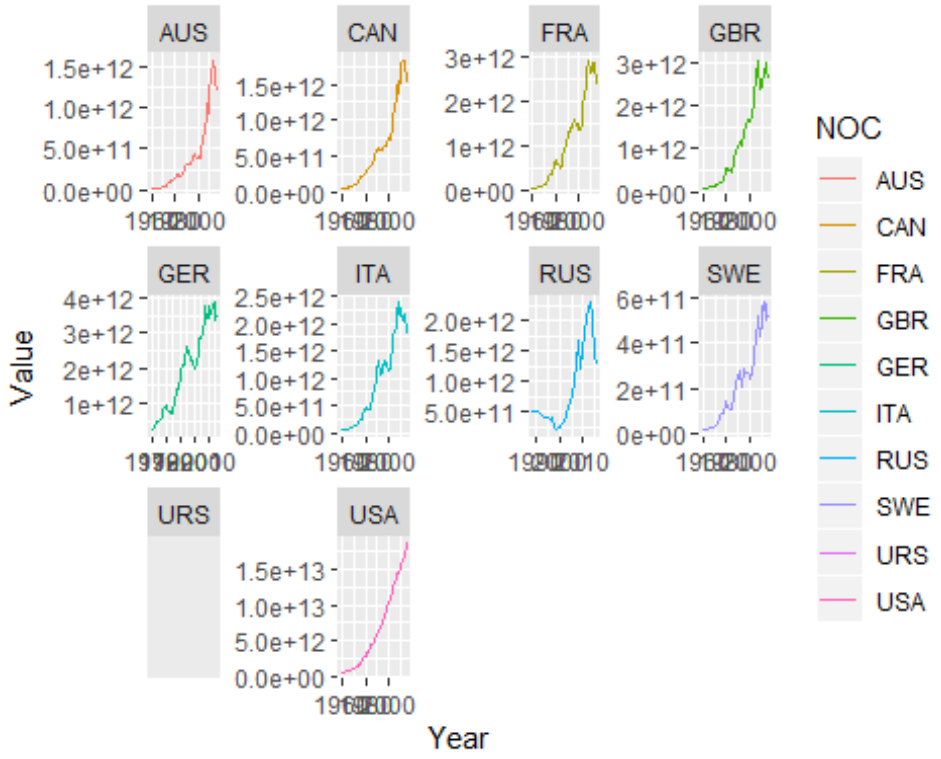


그래프의 기울기가 미국, 러시아, 캐나다, 오스트레일리아는 상대적으로 커 보이고, 스웨덴은 작아 보이며, 영국은 점점 작아지는 것처럼 보인다.

```
left_join(olympicTot,gdp,by="NOC")%>%
  filter(NOC!="USA")%>%
  ggplot(aes(Year,Value,group=NOC))+geom_line(aes(color=NOC))
```




```
left_join(olympicTot,gdp,by="NOC")%>%
  ggplot(aes(Year,Value,group=NOC))+geom_line(aes(color=NOC))+facet_wrap(~NO
C,scales="free")
```



GDP 변화가 큰 국가를 순서대로 나열해보면 미국, 독일, 영국, 프랑스, 이탈리아, 캐나다, 오스트레일리아, 러시아, 스웨덴 순서가 되는데 이는 앞선 메달 수가 많은 순서와 유사하게 나타난다. 따라서 메달 수가 많(적)으면 GDP 의 증가율이 크(작)다고 할 수 있다. 이때 소련의 경우 자료가 존재하지 않아 그림이 그려지지 않았다.

2) GDP 변화가 큰 국가의 메달 수 파악

1. GDP 변화 큰 국가

```
gdp1<-gdp %>% spread(Year, Value) %>%
  mutate(growth = (`2016` - `1960`) / `1960`) %>%
  select(Country.Name, NOC, growth) %>%
  semi_join(olympic, by = 'NOC') %>%
  arrange(desc(growth)) %>% head(10)

gdp1
```

```
## # A tibble: 10 x 3
##    Country.Name         NOC   growth
##    <fct>               <fct>  <dbl>
##  1 Singapore           SGP     421.
##  2 Korea, Rep.         KOR     355.
##  3 Hong Kong SAR, China HKG     242.
##  4 China               CHN     187.
##  5 Ireland             IRL     156.
##  6 Thailand            THA     146.
##  7 Israel              ISR     121.
##  8 Brazil              BRA     117.
##  9 Japan               JPN     111.
## 10 Dominican Republic  DOM     105.
```

위와는 반대로 GDP 의 변화가 큰 10 개국을 추출해 그 국가들의 메달 수 변화를 보려고 했다.

2. 추출 국가의 메달 수 비율

```
gdpTot<-olympic%>%
  semi_join(gdp1)%>%
  filter(Medal!=0)%>%
  select(Year,NOC,Medal)%>%
  group_by(NOC)%>%
  summarise(medal_tot=n())

## Joining, by = "NOC"

gdpTot
```

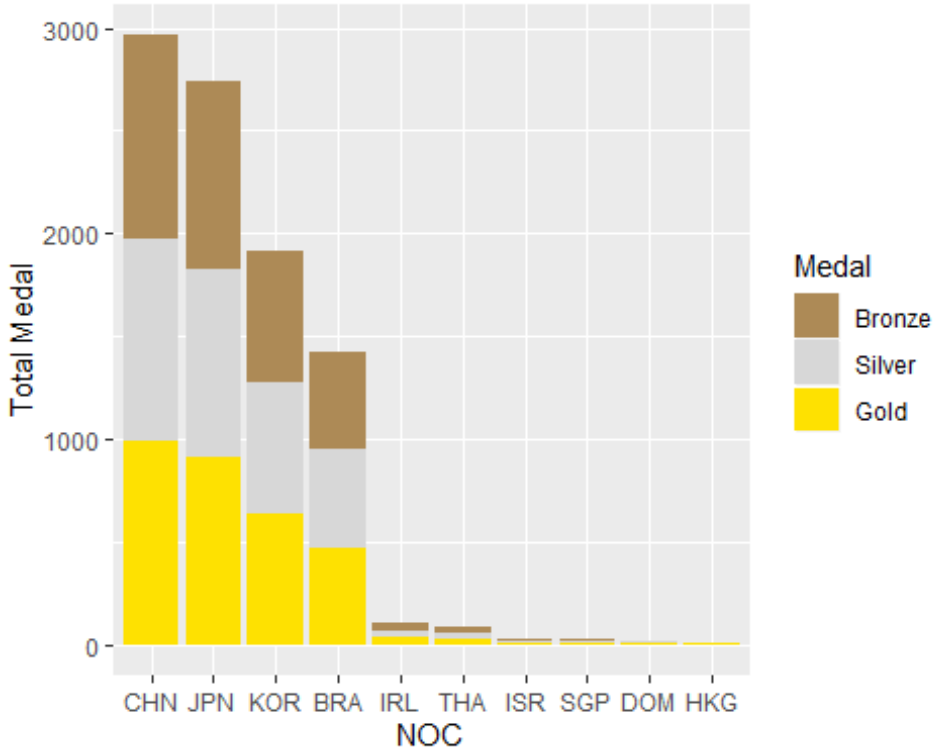
```
## # A tibble: 10 x 2
##    NOC   medal_tot
##    <fct>     <int>
##  1 BRA         475
##  2 CHN         989
##  3 DOM           7
##  4 HKG           4
```

```
##  5 IRL         35
##  6 ISR          9
##  7 JPN        913
##  8 KOR        638
##  9 SGP          9
## 10 THA         30
```

```
gdpMedal<-left_join(gdpTot,olympic,by="NOC")%>%
  group_by(NOC,medal_tot,Medal)%>%
  summarise(medal_sum=n())%>%
  spread(key=Medal,value=medal_sum)%>%
  gather('Bronze':'Silver',key="Medal",value="medal_sum")%>%
  arrange(NOC,Medal)
gdpMedal
```

```
## # A tibble: 30 x 5
## # Groups:   NOC, medal_tot [230]
##    NOC   medal_tot   `0` Medal  medal_sum
##    <fct>     <int> <int> <chr>      <int>
##  1 BRA         475  3373 Bronze       191
##  2 BRA         475  3373 Gold         109
##  3 BRA         475  3373 Silver       175
##  4 CHN         989  4152 Bronze       292
##  5 CHN         989  4152 Gold         350
##  6 CHN         989  4152 Silver       347
##  7 DOM           7   270 Bronze         2
##  8 DOM           7   270 Gold           3
##  9 DOM           7   270 Silver         2
## 10 HKG           4   681 Bronze         1
## # ... with 20 more rows
```

```
ggplot(gdpMedal,aes(x=reorder(NOC,-medal_tot,sum),medal_tot,fill=ordered(Meda
l,levels=c("Bronze","Silver","Gold"))))+geom_bar(stat="identity")+xlab("NOC")
+ylab("Total Medal")+labs(fill="Medal")+scale_fill_manual(values=c("#ad8a56",
"#d7d7d7","#fee101"))
```

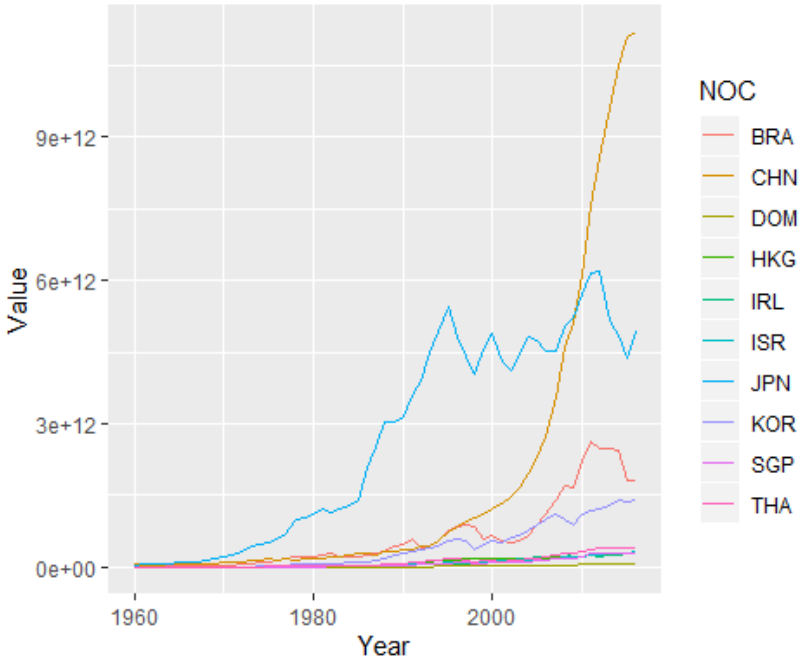


CHN	JPN	KOR	BRA	IRL	THA	ISR	SGP	DOM	HKG
중국	일본	한국	브라질	아일랜드	태국	이스라엘	싱가포르	도미니카	홍콩

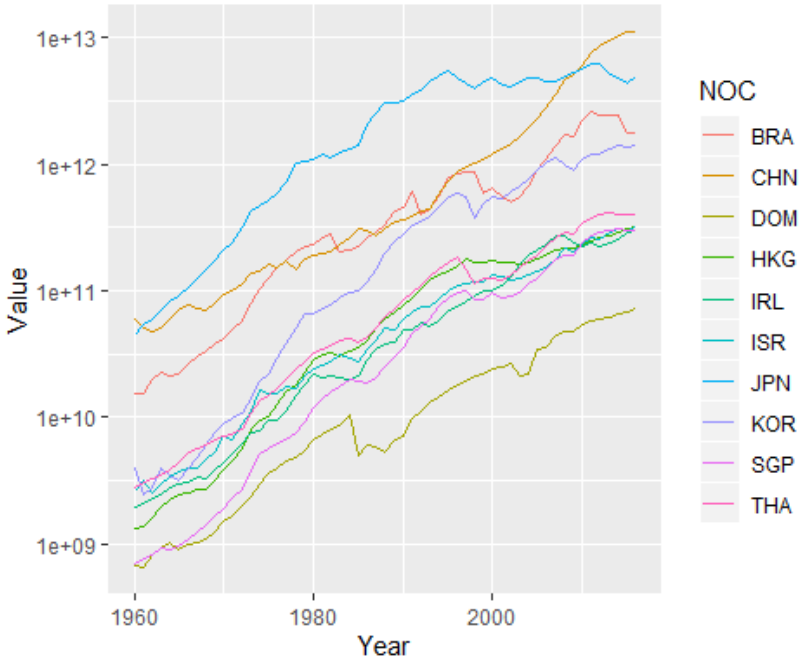
중국, 일본, 한국, 브라질 밑으로는 메달 수가 급격하게 줄어들기 때문에 이 네 국가를 중심으로 더 자세한 그림을 그릴 예정이다.

3-1. 추출 국가의 GDP 변화

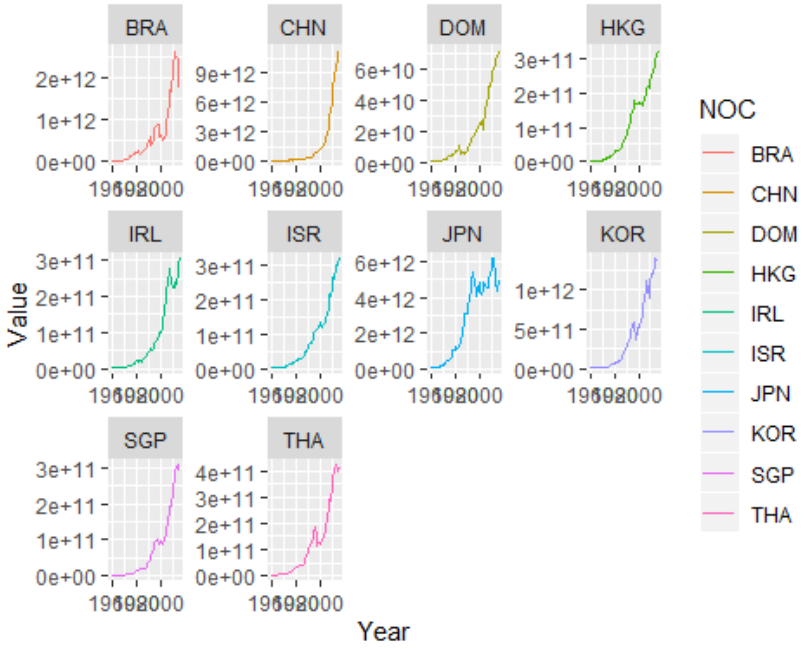
```
gdp%>%
  semi_join(gdp1)%>%
  ggplot(aes(Year,Value,group=NOC))+geom_line(aes(color=NOC))
```




```
gdp%>%
  semi_join(gdp1)%>%
  ggplot(aes(Year,Value,group=NOC))+geom_line(aes(color=NOC))+scale_y_log10()
```

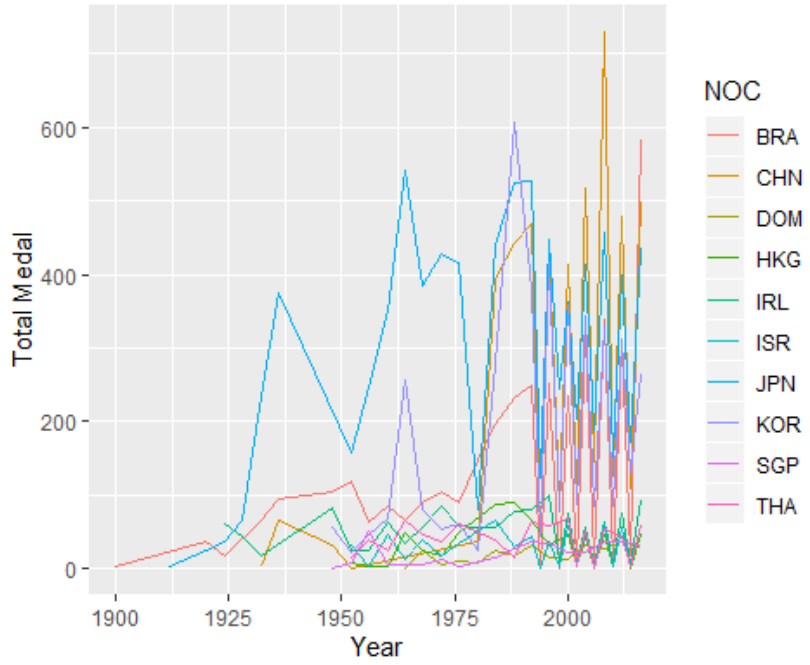


```
gdp%>%
  semi_join(gdp1)%>%
  ggplot(aes(Year,Value,group=NOC))+geom_line(aes(color=NOC))+facet_wrap(~NOC,scales="free")
```

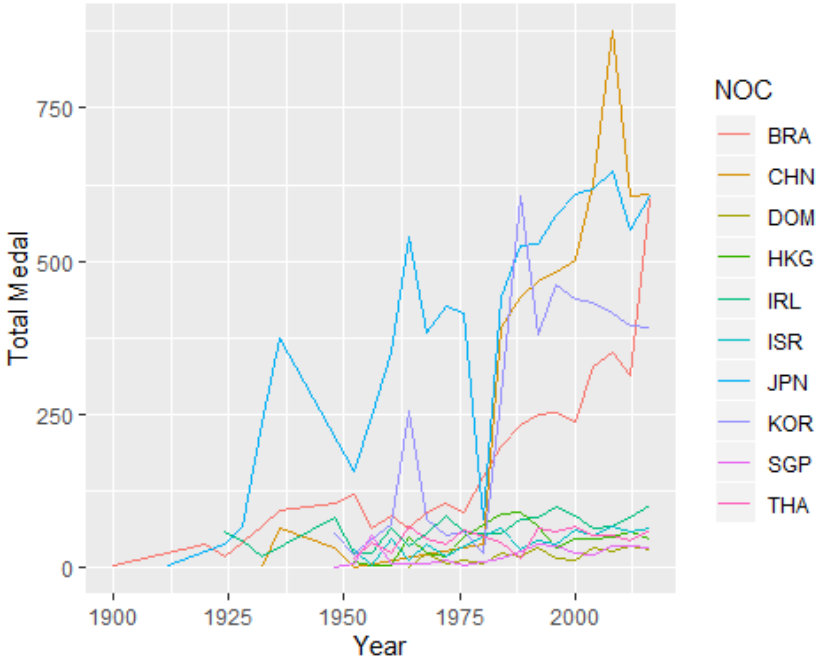


3-2. 추출 국가의 메달 수 변화

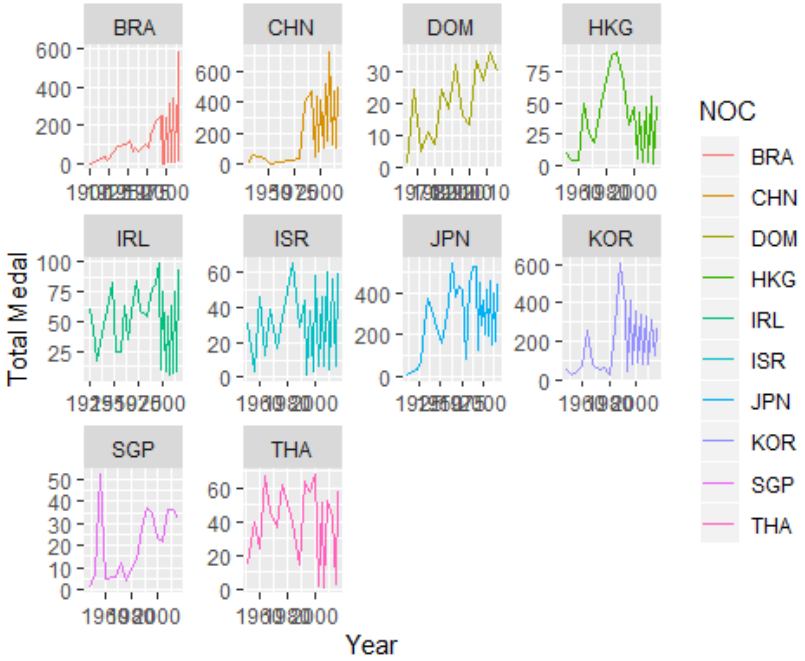
```
olympic%>%semi_join(gdp1)%>%group_by(NOC,Year)%>%
  summarise(medal_tot=n())%>%arrange(NOC,Year)%>%
  ggplot(aes(Year,medal_tot,group=NOC))+geom_line(aes(color=NOC))+ylab("Total
 Medal")
```



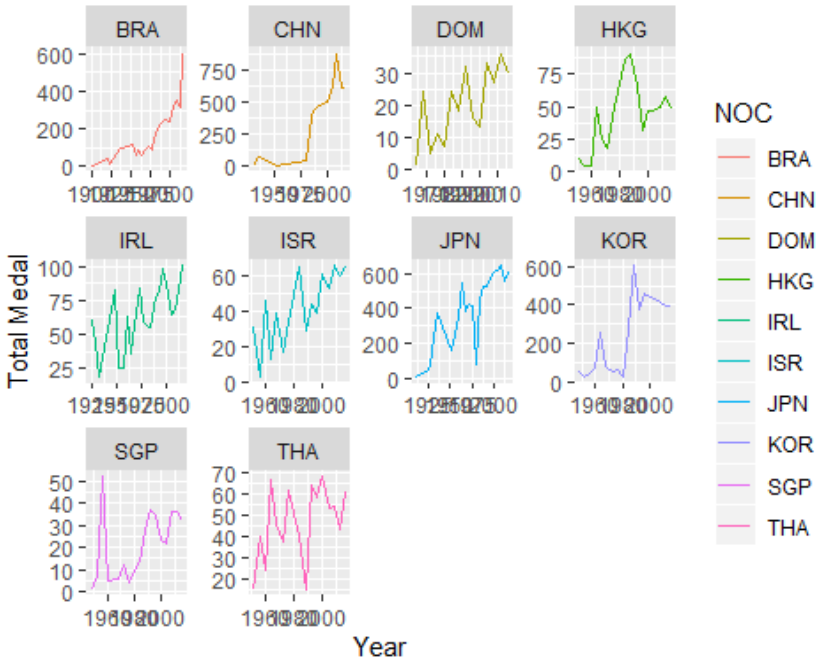

```
olympic2%>%semi_join(gdp1)%>%group_by(NOC,Year)%>%
  summarise(medal_tot=n())%>%arrange(NOC,Year)%>%
  ggplot(aes(Year,medal_tot,group=NOC))+geom_line(aes(color=NOC))+ylab("Total
 Medal")
```




```
olympic%>%
  semi_join(gdp1)%>%
  group_by(NOC,Year)%>%
  summarise(medal_tot=n())%>%
  arrange(NOC,Year)%>%
  ggplot(aes(Year,medal_tot,group=NOC))+geom_line(aes(color=NOC))+facet_wrap
(~NOC,scales="free")+ylab("Total Medal")
```



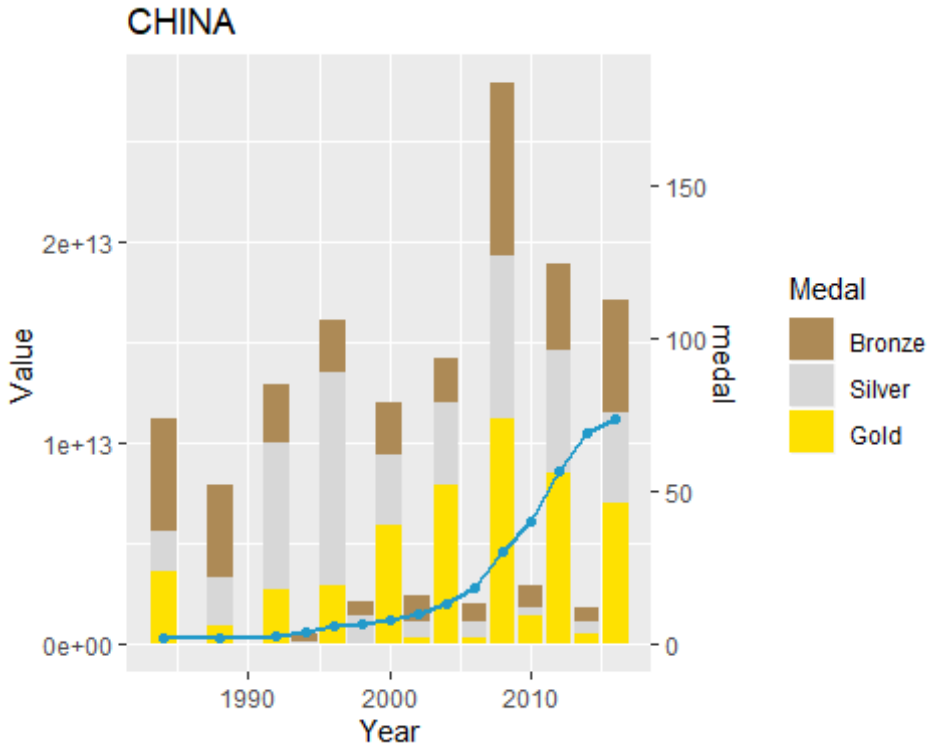
```
olympic2%>%
  semi_join(gdp1)%>%
  group_by(NOC,Year)%>%
  summarise(medal_tot=n())%>%
  arrange(NOC,Year)%>%
  ggplot(aes(Year,medal_tot,group=NOC))+geom_line(aes(color=NOC))+facet_wrap
(~NOC,scales="free")+ylab("Total Medal")
```



GDP 가 증가하는 국가 모두 특정 연도의 피크를 제외하고 보면 전반적으로 메달 수가 증가하는 경향성이 보인다.

1 중국의 GDP& 메달 수

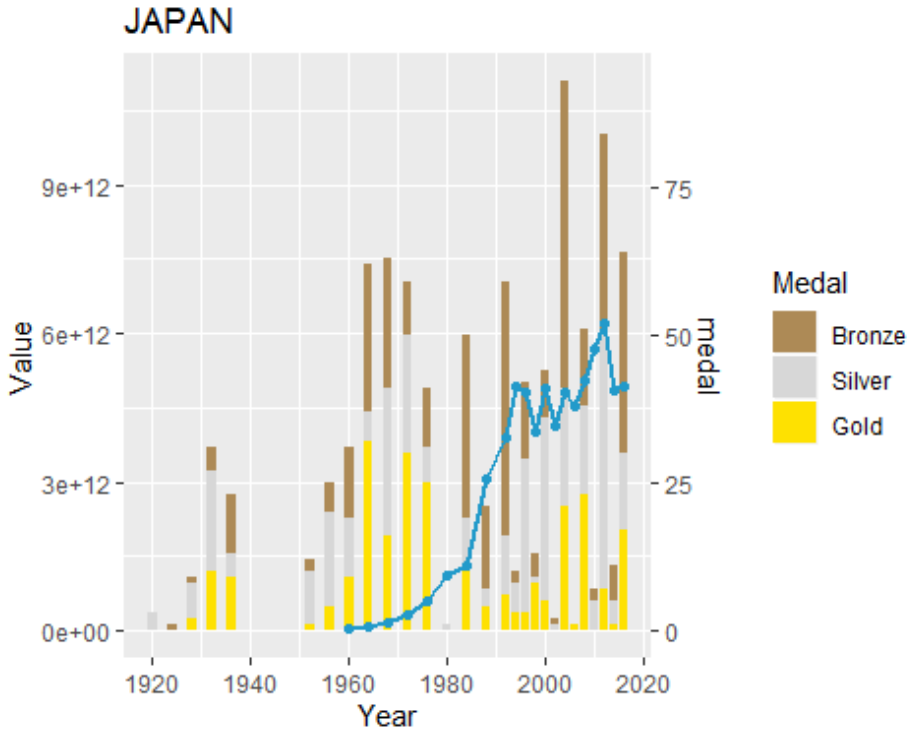
```
gdpCHN<-gdp%>%
  filter(NOC=="CHN")
china_gdp <- olympic%>%
  filter(NOC=="CHN"&Medal!=0)%>%
  inner_join(gdpCHN,by="Year")%>%
  group_by(Year,Value)%>%
  summarise(medal_tot=n(),na.rm=TRUE)%>%
  select(-na.rm)%>%
  arrange(Year)
china_medal2 <- olympic%>%
  filter(NOC=="CHN"&Medal!=0)%>%
  group_by(Year,Medal)%>%
  summarise(medal_tot=n())
max_ratio <- max(china_gdp$Value)/max(china_medal2$medal_tot)

ggplot(data = china_gdp, aes(x = Year, y = Value)) +
  geom_bar(data = china_medal2, aes(Year, y = medal_tot*max_ratio, fill=ordere
d(Medal,levels=c("Bronze","Silver","Gold"))), stat = 'identity') +
  labs(fill="Medal")+scale_fill_manual(values=c("#ad8a56","#d7d7d7","#fee101
")) +
  scale_y_continuous(sec.axis = sec_axis(~ ./max_ratio, name = "medal")) +
  geom_line(color = '#229aca', size = 1) +
  geom_point(color = '#229aca')+ggtitle("CHINA")
```



GDP 와 메달 수 둘 다 눈에 띄게 증가하는 경향이 보인다.

2 일본의 GDP& 메달 수

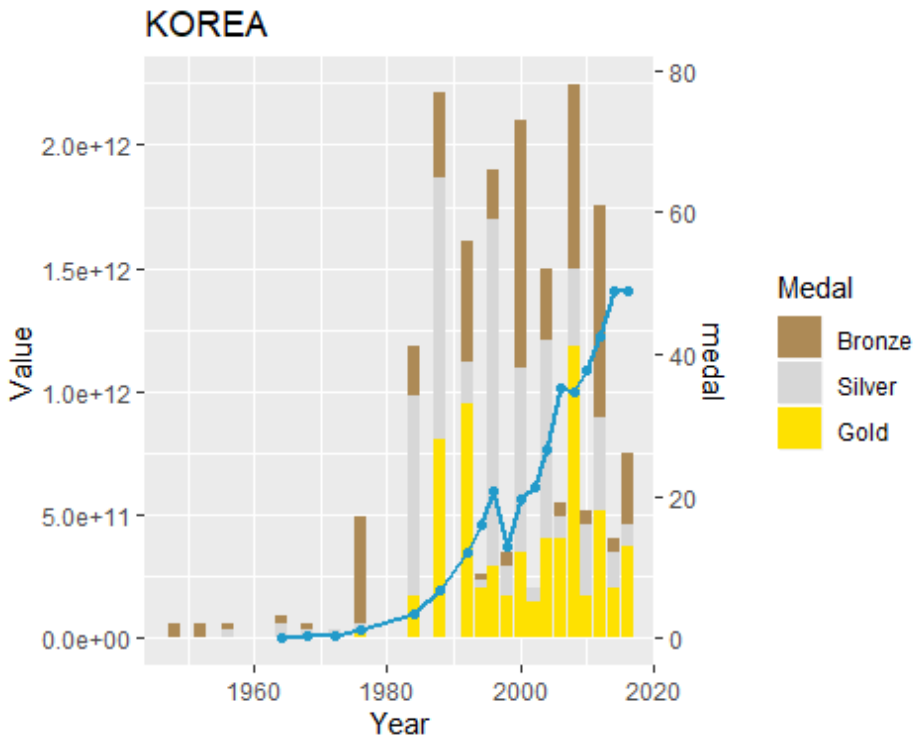
```
gdpJPN<-gdp%>%
  filter(NOC=="JPN")
japan_gdp <- olympic%>%
  filter(NOC=="JPN"&Medal!=0)%>%
  inner_join(gdpJPN,by="Year")%>%
  group_by(Year,Value)%>%
  summarise(medal_tot=n(),na.rm=TRUE)%>%
  select(-na.rm)%>%
  arrange(Year)
japan_medal2 <- olympic%>%
  filter(NOC=="JPN"&Medal!=0)%>%
  group_by(Year,Medal)%>%
  summarise(medal_tot=n())
max_ratio <- max(japan_gdp$Value)/max(japan_medal2$medal_tot)

ggplot(data = japan_gdp, aes(x = Year, y = Value)) +
  geom_bar(data = japan_medal2, aes(Year, y = medal_tot*max_ratio, fill=ordere
d(Medal,levels=c("Bronze","Silver","Gold"))), stat = 'identity') +
  labs(fill="Medal")+scale_fill_manual(values=c("#ad8a56","#d7d7d7","#fee101
")) +
  scale_y_continuous(sec.axis = sec_axis(~ ./max_ratio, name = "medal")) +
  geom_line(color = '#229aca', size = 1) +
  geom_point(color = '#229aca')+ggtitle("JAPAN")
```



GDP 와 메달 수 둘 다 눈에 띄게 증가하는 경향이 보인다.

3 한국의 GDP& 메달 수

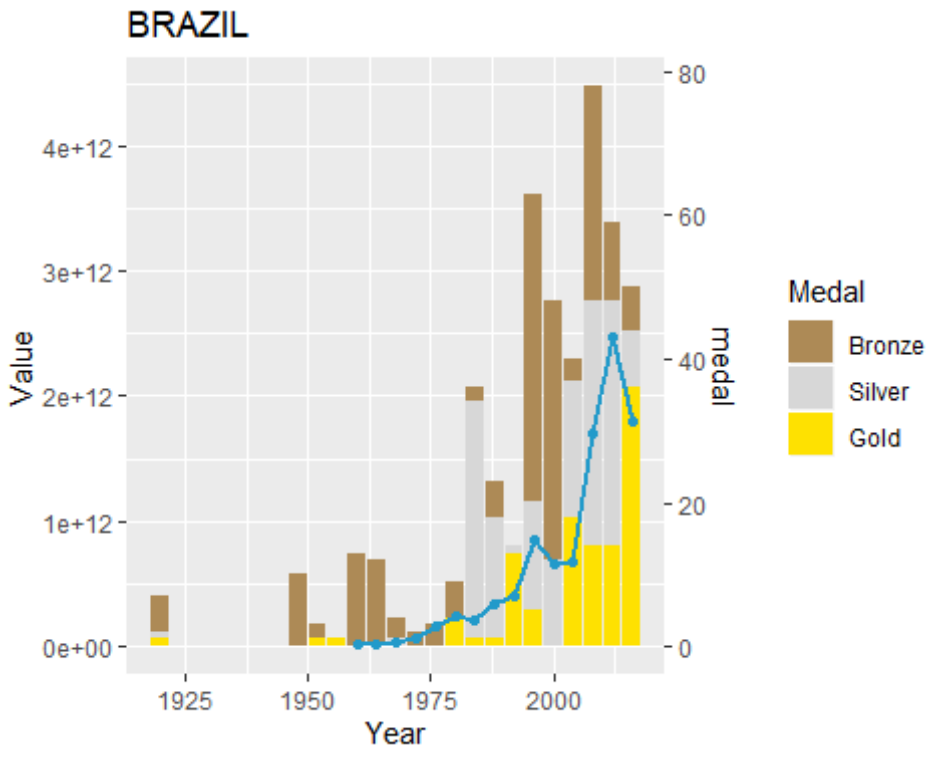
```
gdpKOR<-gdp%>%
  filter(NOC=="KOR")
korean_gdp <- olympic%>%
  filter(NOC=="KOR"&Medal!=0)%>%
  inner_join(gdpKOR,by="Year")%>%
  group_by(Year,Value)%>%
  summarise(medal_tot=n(),na.rm=TRUE)%>%
  select(-na.rm)%>%
  arrange(Year)
korean_medal2 <- olympic%>%
  filter(NOC=="KOR"&Medal!=0)%>%
  group_by(Year,Medal)%>%
  summarise(medal_tot=n())
max_ratio <- max(korean_gdp$Value)/max(korean_medal2$medal_tot)

ggplot(data = korean_gdp, aes(x = Year, y = Value)) +
  geom_bar(data = korean_medal2, aes(Year, y = medal_tot*max_ratio, fill=order
ed(Medal,levels=c("Bronze","Silver","Gold"))), stat = 'identity') +
  labs(fill="Medal")+scale_fill_manual(values=c("#ad8a56","#d7d7d7","#fee101
")) +
  scale_y_continuous(sec.axis = sec_axis(~ ./max_ratio, name = "medal")) +
  geom_line(color = '#229aca', size = 1) +
  geom_point(color = '#229aca')+ggtitle("KOREA")
```



GDP 와 메달 수 둘 다 눈에 띄게 증가하는 경향이 보인다.

4 브라질 의 GDP& 메달 수

```
gdpBRA<-gdp%>%
  filter(NOC=="BRA")
brazil_gdp <- olympic%>%
  filter(NOC=="BRA"&Medal!=0)%>%
  inner_join(gdpBRA,by="Year")%>%
  group_by(Year,Value)%>%
  summarise(medal_tot=n(),na.rm=TRUE)%>%
  select(-na.rm)%>%
  arrange(Year)
brazil_medal2 <- olympic%>%
  filter(NOC=="BRA"&Medal!=0)%>%
  group_by(Year,Medal)%>%
  summarise(medal_tot=n())
max_ratio <- max(brazil_gdp$Value)/max(brazil_medal2$medal_tot)

ggplot(data = brazil_gdp, aes(x = Year, y = Value)) +
  geom_bar(data = brazil_medal2, aes(Year, y = medal_tot*max_ratio, fill=order
ed(Medal,levels=c("Bronze","Silver","Gold"))), stat = 'identity') +
  labs(fill="Medal")+scale_fill_manual(values=c("#ad8a56","#d7d7d7","#fee101
")) +
  scale_y_continuous(sec.axis = sec_axis(~ ./max_ratio, name = "medal")) +
  geom_line(color = '#229aca', size = 1) +
  geom_point(color = '#229aca')+ggtitle("BRAZIL")
```



GDP 와 메달 수 둘 다 눈에 띄게 증가하는 경향이 보인다.

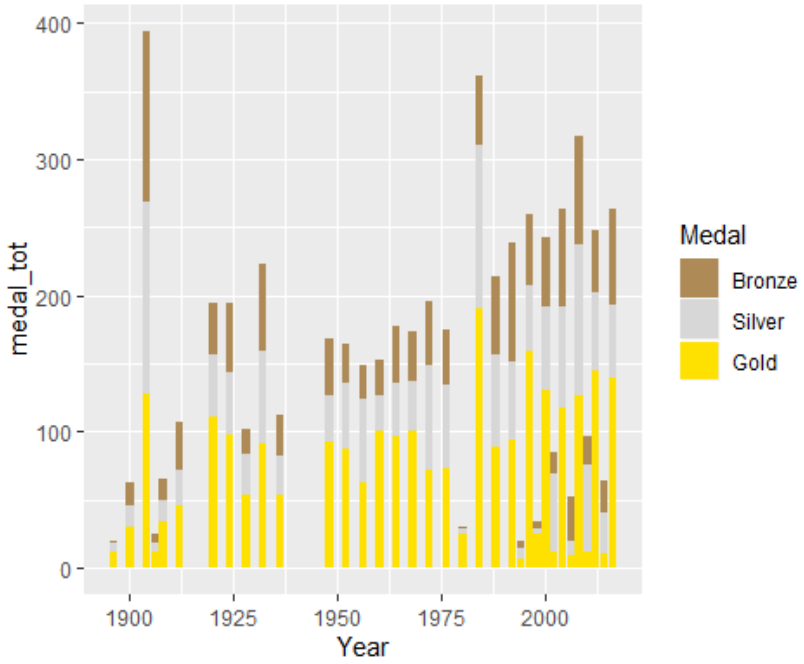
따라서 GDP 가 증가(감소)하면 메달 수도 증가(감소)한다고 할 수 있다. 결론적으로 GDP 와 메달 수는 서로 관계가 있다는 것을 알 수 있다.

3) 일부 국가의 메달 분석

. 1916 년 베를린 올림픽과 1940 년 도쿄올림픽, 1944 년 헬싱키 올림픽은 각각 제 1,2 차 세계대전으로 취소되었기 때문에 결측값이다.

1. 미국

```
olympic%>%
  filter(NOC=="USA"&Medal!=0)%>%
  group_by(Year,Medal)%>%
  summarise(medal_tot=n())%>%
  ggplot(aes(Year,medal_tot,fill=ordered(Medal,levels=c("Bronze","Silver","Gold"))))+geom_bar(stat="identity")+labs(fill="Medal")+scale_fill_manual(values=c("#ad8a56","#d7d7d7","#fee101"))
```

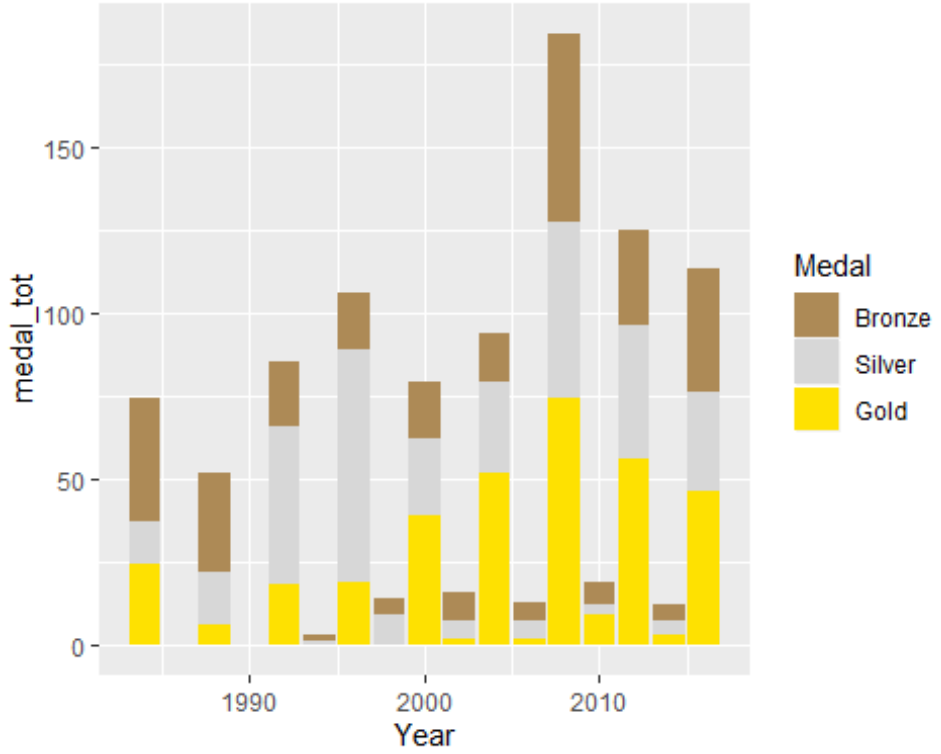


미국의 경우 1904 년에 메달 수가 갑자기 증가한 것을 알 수 있는데, 1904 년 하계 올림픽이 미국 미주리주 세인트 루이스에서 개최됐기 때문에 미국에게 유리했다고 생각해 볼 수 있다. 1980 년 모스크바 올림픽 땐 미국이 보이콧에 동조했기 때문에, 개인 자격의

선수만 파견해서 메달의 수가 급격히 줄었다. 1984 년에도 메달 수가 급격히 증가했고, 이것도 1904 년과 같은 이유로 캘리포니아주 로스앤젤레스에서 개최됐기 때문에 유리했을 것이라고 추측 가능하다.

2. 중국

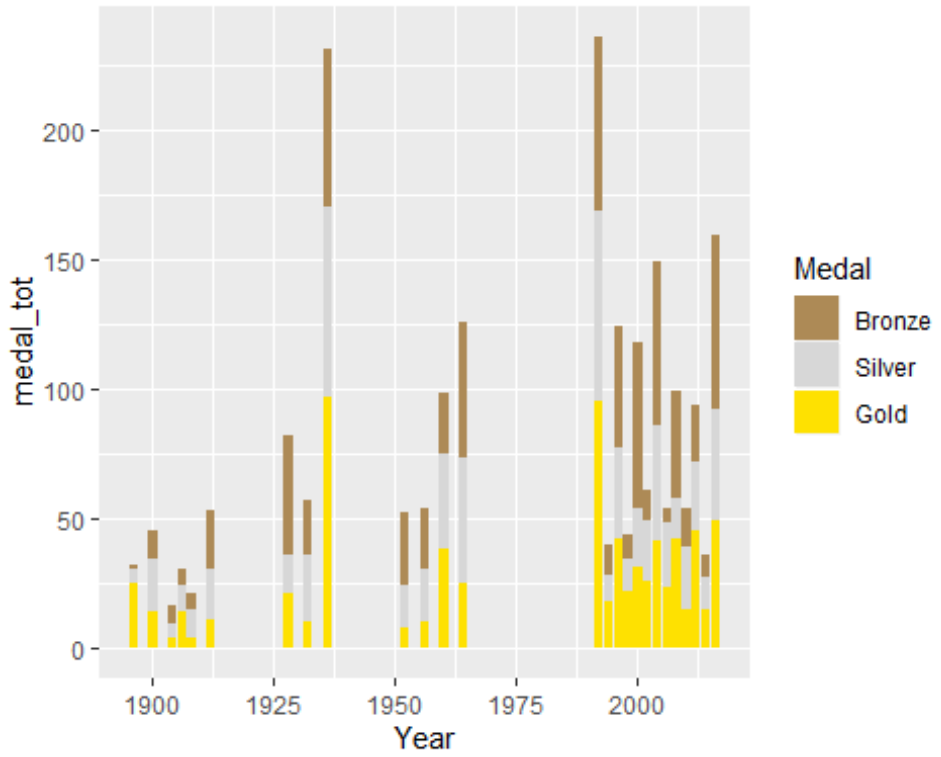
```
olympic%>%
  filter(NOC=="CHN"&Medal!=0)%>%
  group_by(Year,Medal)%>%
  summarise(medal_tot=n())%>%
  ggplot(aes(Year,medal_tot,fill=ordered(Medal,levels=c("Bronze","Silver","Gold"))))+geom_bar(stat="identity")+labs(fill="Medal")+scale_fill_manual(values=c("#ad8a56","#d7d7d7","#fee101"))
```



중국의 경우 1984 년부터 메달 자료가 존재하는데, 중화인민공화국이 수립한 이후 1956 년부터 1980 년까지 중화민국(타이완)의 참가에 항의하여 불참했기 때문이다. 참가 이후 메달이 증가해가는 것을 알 수 있고, 2008 년엔 자국의 베이징에서 올림픽이 개최됐기 때문에 다른 해에 비해 메달 수가 많다.

3. 독일

```
olympic%>%
  filter(NOC=="GER"&Medal!=0)%>%
  group_by(Year,Medal)%>%
  summarise(medal_tot=n())%>%
  ggplot(aes(Year,medal_tot,fill=ordered(Medal,levels=c("Bronze","Silver","Gold"))))+geom_bar(stat="identity")+labs(fill="Medal")+scale_fill_manual(values=c("#ad8a56","#d7d7d7","#fee101"))
```



1968 년~1988 년 까지의 그림이 없는데, 이 이유는 이 당시 독일이 동독과 서독으로 나눠졌었기 때문이다. 따라서 이 시기만 추출하여 동독과 서독을 나누어 그림을 그려보았다. 1936 년엔 자국의 베를린에서 올림픽을 개최했기 때문에, 메달 수가 많은 것이라고 추정할 수 있다.

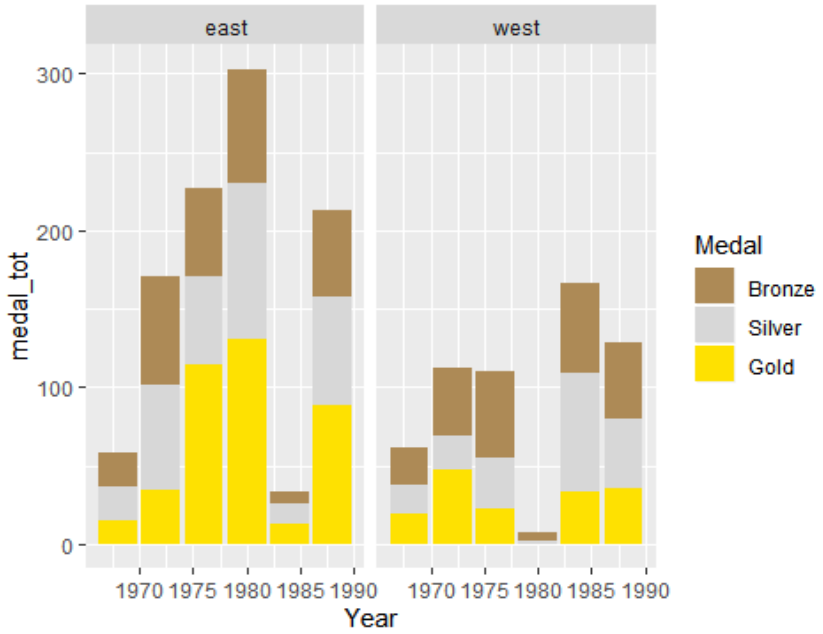
```
# 동독, 서독 분단 시기
olympic%>%
  filter(NOC=="GER"&Medal!=0) %>%
  group_by(Year) %>%
  count()
```

```
## # A tibble: 26 x 2
## # Groups:   Year [26]
##    Year     n
##   <int> <int>
```

```
##  1  1896    32
##  2  1900    45
##  3  1904    16
##  4  1906    30
##  5  1908    21
##  6  1912    53
##  7  1928    82
##  8  1932    57
##  9  1936   231
## 10  1952    52
## # ... with 16 more rows
```

독일의 연도별 메달 그래프를 보았을 때 1912 ~ 1928 년, 1936 ~ 1952, 1964 ~ 1992 년 사이에는 공백임을 알 수 있다. 1917 ~ 1925 년은 1 차 세계대전의 영향으로 보이고, 1936 년 ~ 1952 년 사이에는 2 차 세계대전의 영향임을 추측해볼 수 있다. 1964 년부터 1992 년의 공백은 동독, 서독으로 독일이 분단되었을 시기로 각각 동독과 서독으로 따로 출전함에 기인함을 알 수 있다.

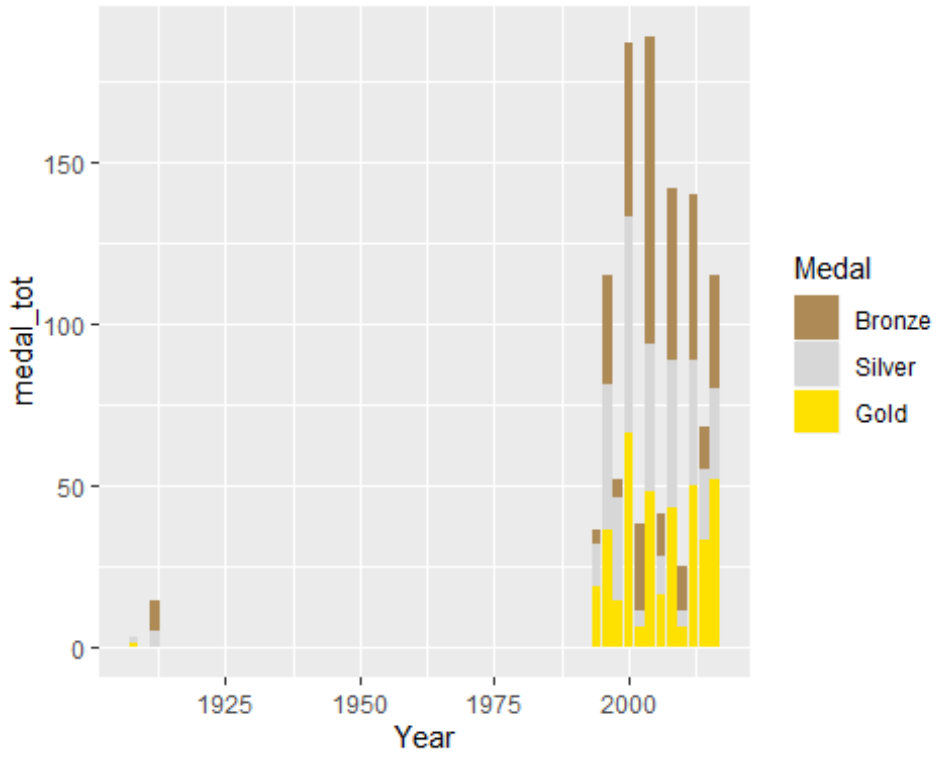
```
east_g = olympic%>%
  filter(NOC=="GDR"& Medal != 0)%>%
  group_by(Year,Medal)%>%
  summarise(medal_tot=n())

west_g = olympic%>%
  filter(NOC=="FRG" & Medal != 0)%>%
  group_by(Year,Medal)%>%
  summarise(medal_tot=n())

east_g %>% left_join(west_g, by = c('Year', 'Medal')) %>%
  replace_na(list(medal_tot.y = 0)) %>%
  rename(east = 'medal_tot.x', west = 'medal_tot.y') %>%
  gather(country, medal_tot, 'east':'west') %>%
  ggplot(aes(Year,medal_tot,fill=ordered(Medal,levels=c("Bronze","Silver","Go
ld"))))+
  geom_bar(stat="identity")+
  labs(fill="Medal")+scale_fill_manual(values=c("#ad8a56","#d7d7d7","#fee101
")) +
  facet_wrap(~country)
```



동독과 서독의 메달 분포를 살펴보면 전반적으로 동독이 서독보다 우수한 성적을 기록했음을 알 수 있다. 주목할 만한 점은 1980 년도에는 동독이 서독에 비해 압도적으로 메달 수가 많은 데 비해 그 다음 올림픽에서는 서독이 동독을 앞섰다는 점이다. 이는 앞서 서론에서 말했던 것과 같이 1980 년에는 미국, 서독 등의 자본주의 국가들이 참가하지 않았고, 그 다음 84 년 올림픽에서는 소련, 동독을 비롯한 사회주의 국가들이 참가하지 않았음에 영향을 받았을 것이라고 추측 가능하다.

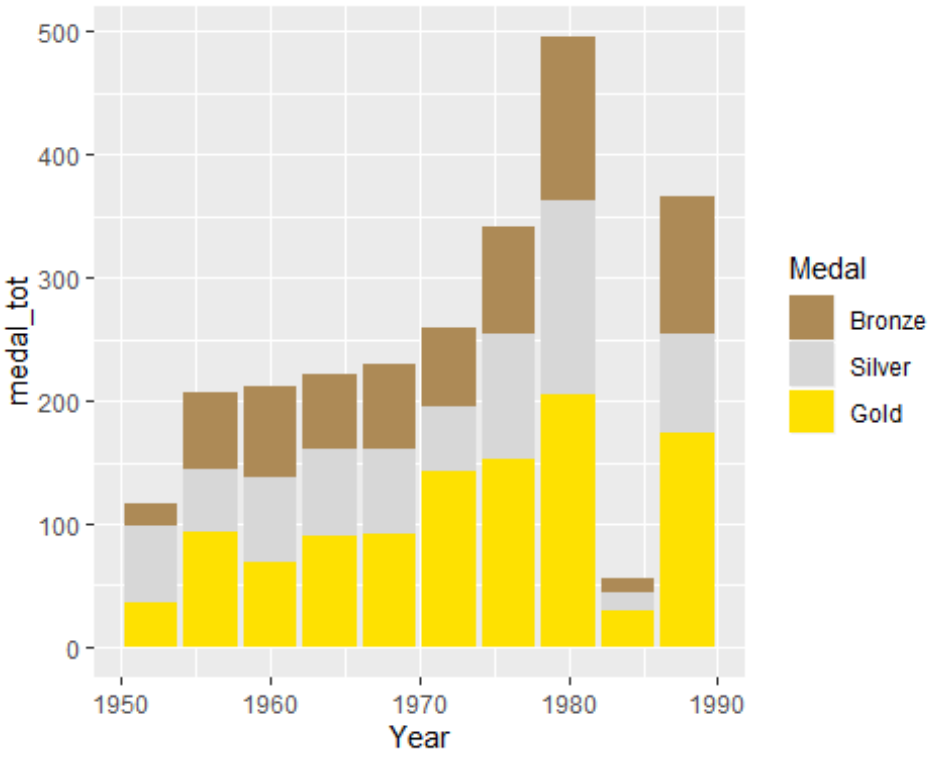
4. 러시아

```
olympic%>%
  filter(NOC=="RUS"&Medal!=0)%>%
  group_by(Year,Medal)%>%
  summarise(medal_tot=n())%>%
  ggplot(aes(Year,medal_tot,fill=ordered(Medal,levels=c("Bronze","Silver","Gold"))))+geom_bar(stat="identity")+labs(fill="Medal")+scale_fill_manual(values=c("#ad8a56","#d7d7d7","#fee101"))
```



1920 년 서방 국가들의 봉쇄조치로 올림픽에 불참했고, 1922 년 소비에트 연방(소련)의 결성으로 1991 년까지 측정 값이 존재하지 않는다. 따라서 이 시기의 소련의 메달 수 변화를 보기 위한 그림을 그려보았다.

```
olympic%>%
  filter(NOC=="URS"&Medal!=0)%>%
  group_by(Year,Medal)%>%
  summarise(medal_tot=n())%>%
  ggplot(aes(Year,medal_tot,fill=ordered(Medal,levels=c("Bronze","Silver","Go
ld"))))+geom_bar(stat="identity")+labs(fill="Medal")+scale_fill_manual(values
=c("#ad8a56","#d7d7d7","#fee101"))
```

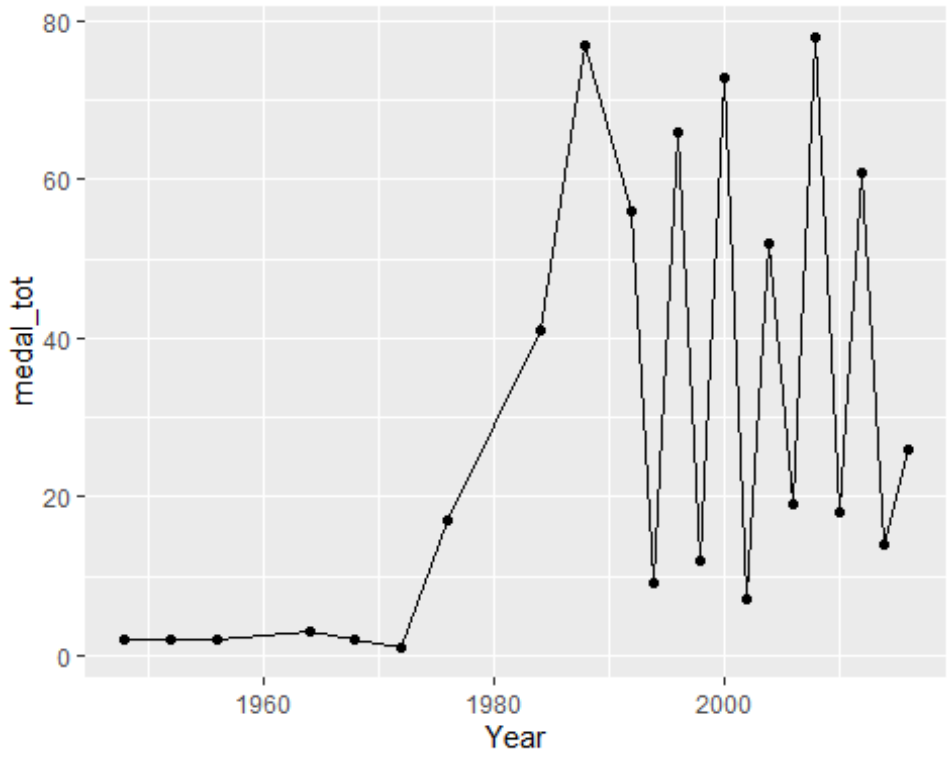


전반적인 메달 수가 증가하는 것을 알 수 있고, 1984 년 로스앤젤레스 올림픽 보이콧에 동조하면서 불참했기 때문에 메달 수가 급격히 줄었다.

5. 한국

(1)메달 수 변화

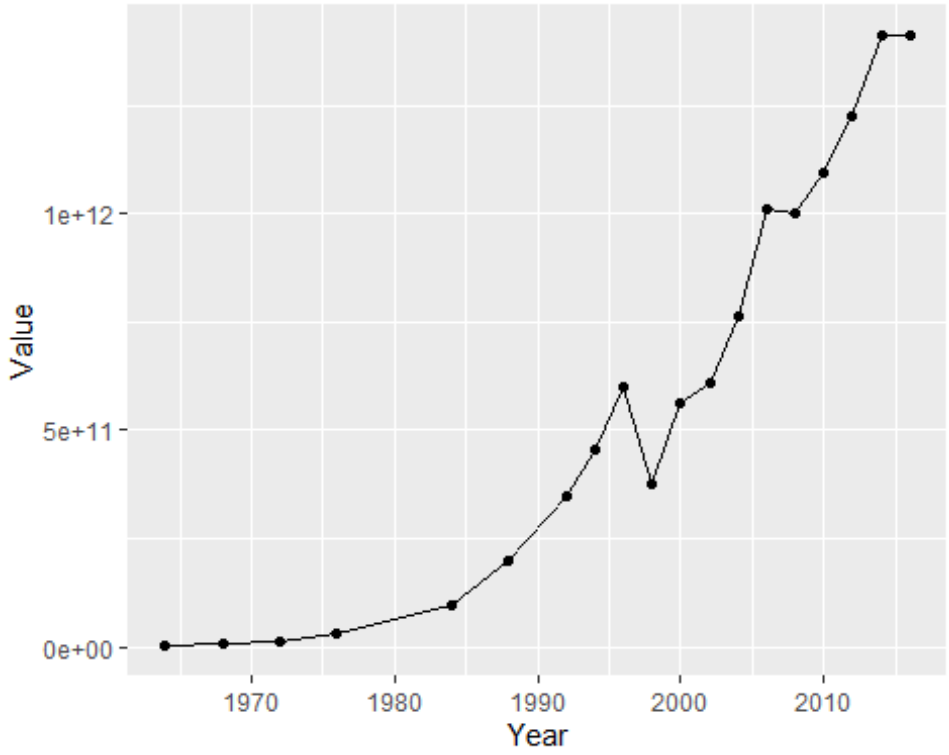
```
gdpKOR<-gdp%>%
  filter(NOC=="KOR")
olympic%>%
  filter(NOC=="KOR"&Medal!=0)%>%
  left_join(gdpKOR,by="Year")%>%
  group_by(NOC.x,Year)%>%
  summarise(medal_tot=n(),na.rm=TRUE)%>%
  select(-na.rm)%>%
  arrange(Year)%>%
  ggplot(aes(Year,medal_tot))+geom_point()+geom_line()
```



대체적으로 메달 수가 증가하는 것을 볼 수 있다. 다만 1994 년부터 하계올림픽과 동계올림픽을 분리하여 2 년주기로 개최되어 반복적으로 증가와 감소하는 패턴을 볼 수 있다.

(2) GDP 변화

```
olympic%>%
  filter(NOC=="KOR"&Medal!=0)%>%
  inner_join(gdpKOR,by="Year")%>%
  group_by(Year,Value)%>%
  summarise(medal_tot=n(),na.rm=TRUE)%>%
  select(-na.rm)%>%
  arrange(Year)%>%
  ggplot(aes(Year,Value))+geom_point()+geom_line()
```

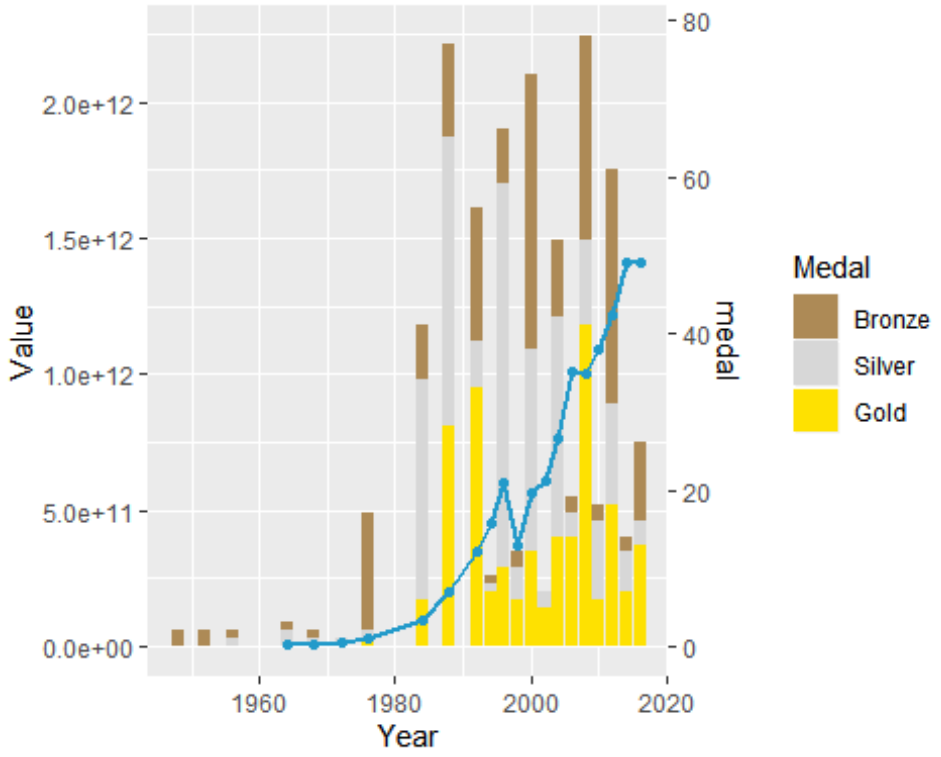


한국의 GDP 는 대체적으로 증가하지만, 1998 년경에 잠깐 감소한 것을 볼 수 있다. 이는 한국의 IMF 영향이라고 추측할 수 있다.

(3) GDP 변화 & 메달 수 별 변화

```
gdpKOR<-gdp%>%
  filter(NOC=="KOR")

korean_gdp <- olympic%>%
  filter(NOC=="KOR"&Medal!=0)%>%
  inner_join(gdpKOR,by="Year")%>%
  group_by(Year,Value)%>%
  summarise(medal_tot=n(),na.rm=TRUE)%>%
  select(-na.rm)%>%
  arrange(Year)

korean_medal2 <- olympic%>%
  filter(NOC=="KOR"&Medal!=0)%>%
  group_by(Year,Medal)%>%
  summarise(medal_tot=n())


max_ratio <- max(korean_gdp$Value)/max(korean_medal2$medal_tot); max_ratio

## [1] 28775510204
```

```
ggplot(data = korean_gdp, aes(x = Year, y = Value)) +
  geom_bar(data = korean_medal2, aes(Year, y = medal_tot*max_ratio, fill=order
ed(Medal,levels=c("Bronze","Silver","Gold"))), stat = 'identity') +
  labs(fill="Medal")+scale_fill_manual(values=c("#ad8a56","#d7d7d7","#fee101
")) +
  scale_y_continuous(sec.axis = sec_axis(~ ./max_ratio, name = "medal")) +
  geom_line(color = '#229aca', size = 1) +
  geom_point(color = '#229aca')
```



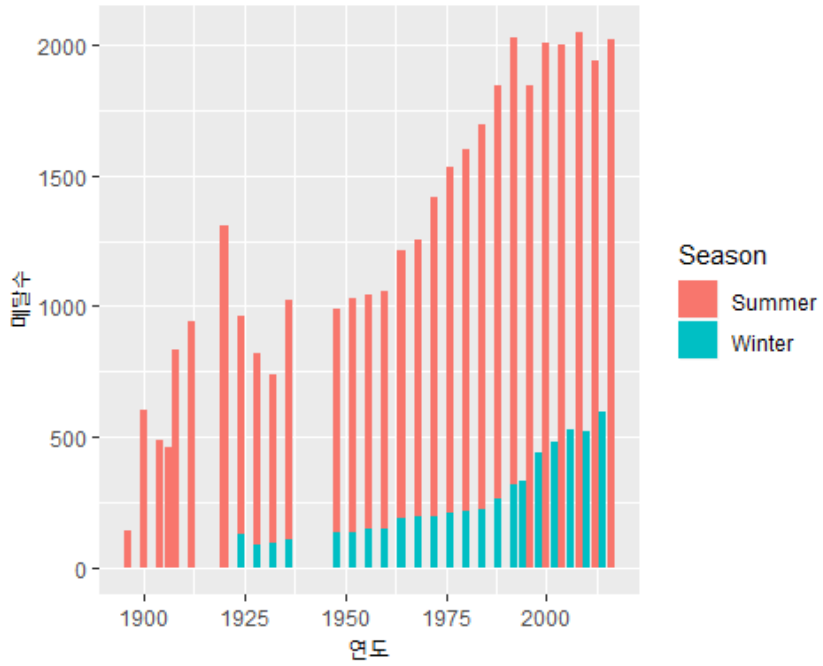
우선 메달 분포를 보면 1960 년 로마 올림픽 때는 성적이 부진하여 메달 기록이 없고, 1980 년 모스크바 올림픽 당시에는 미국에 동조하여 올림픽 참가를 거부하여 메달 기록이 없다. 또한 한국의 GDP 가 1960 년대 이후 증가하는 것을 볼 수 있는데, 이에 맞춰 메달 수도 대체적으로 증가한 것을 볼 수 있다. 그 중에서도 1988 년에 메달 수가 폭발적으로 증가하였는데, 이는 한국에서 개최된 서울올림픽의 영향, 즉 소위 말하는 '개최국 버프'임을 알 수 있다.

5. Season

1. 하계올림픽, 동계올림픽 규모 비교

a) 연도에 따른 메달 수 하계/동계 비교

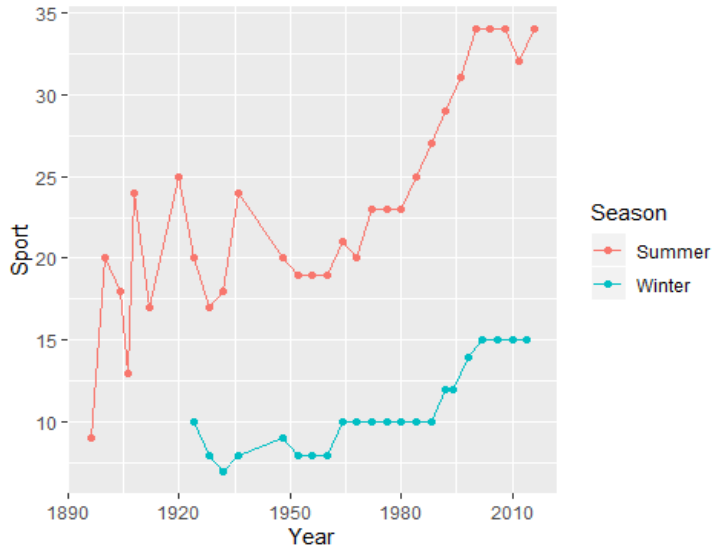
```
olympic%>%
  filter(!is.na(Medal))%>%
  group_by(Year,Season)%>%
  summarise(medal_count = n())%>%
  ggplot(aes(x=Year,y=medal_count,fill=Season))+geom_histogram(binwidth=0.5,s
tat='identity')+xlab("연도")+ylab("메달수")
```



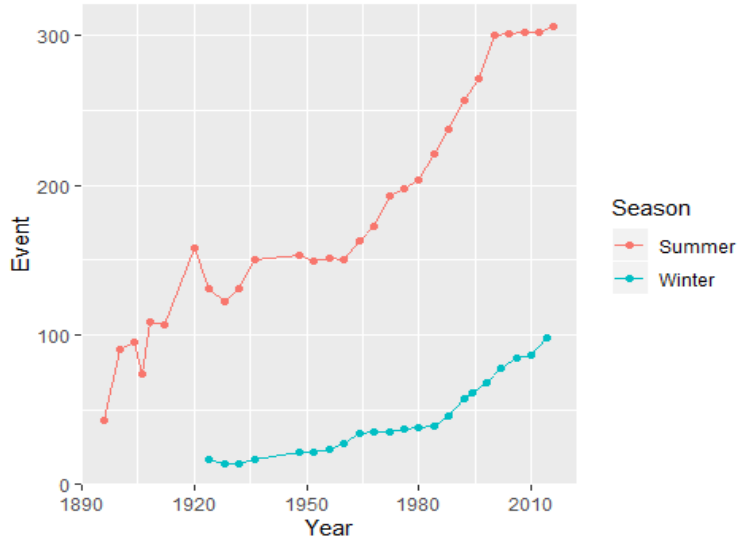
1800 년대 하계올림픽이 먼저 개최되었고, 이후 1924 년경 동계올림픽이 개최되었다는 것을 알 수 있다. 하계올림픽과 동계올림픽이 같은 연도에 개최되다가, 1994 년부터 동계올림픽, 하계올림픽이 2 년단위로 번갈아 개최되었다는 것 또한 알 수 있다. 이외에도, 하계올림픽의 메달수가 동계올림픽의 메달 수보다 월등히 많음을 알 수 있다.

b) 연도에 따른 종목 수 하계/동계 비교

```
olympic%>%
  group_by(Year,Season)%>%
  summarise(Sport=n_distinct(Sport))%>%
  ggplot(aes(x=Year,y=Sport))+
  geom_point(aes(color=Season))+geom_line(aes(color=Season))
```




```
olympic%>%
  group_by(Year,Season)%>%
  summarise(Event=n_distinct(Event))%>%
  ggplot(aes(x=Year,y=Event))+
  geom_point(aes(color=Season))+geom_line(aes(color=Season))
```

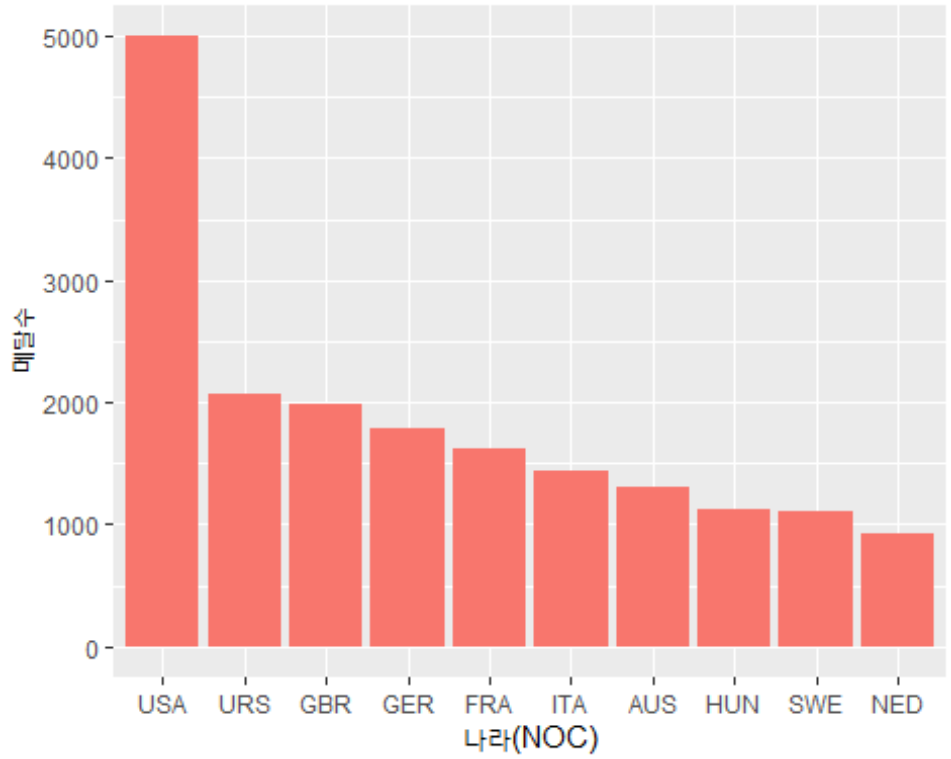


하계, 동계 올림픽의 Sport(종목)와 Event(세부 종목) 수를 연도별로 살펴본 결과, 하계올림픽이 동계올림픽에 비해 종목 수/세부 종목 수가 월등히 많다. 앞서 메달 수가 동계 대비 하계가 더 많았던 이유도 이러한 종목 수와 세부 종목 수의 차이에서 기인함을 알 수 있다. 위 두 결과를 종합하면, 하계올림픽이 동계올림픽보다 메달 수, 종목 수, 세부 종목수가 많은 것을 볼 수 있었다. 따라서 하계올림픽이 동계올림픽보다 더 규모가 크고 인기있다고 볼 수 있다.

2. 하계올림픽/동계올림픽 메달수와 나라의 기후간의 연관성

하계올림픽 메달 획득 수 상위 10 개국

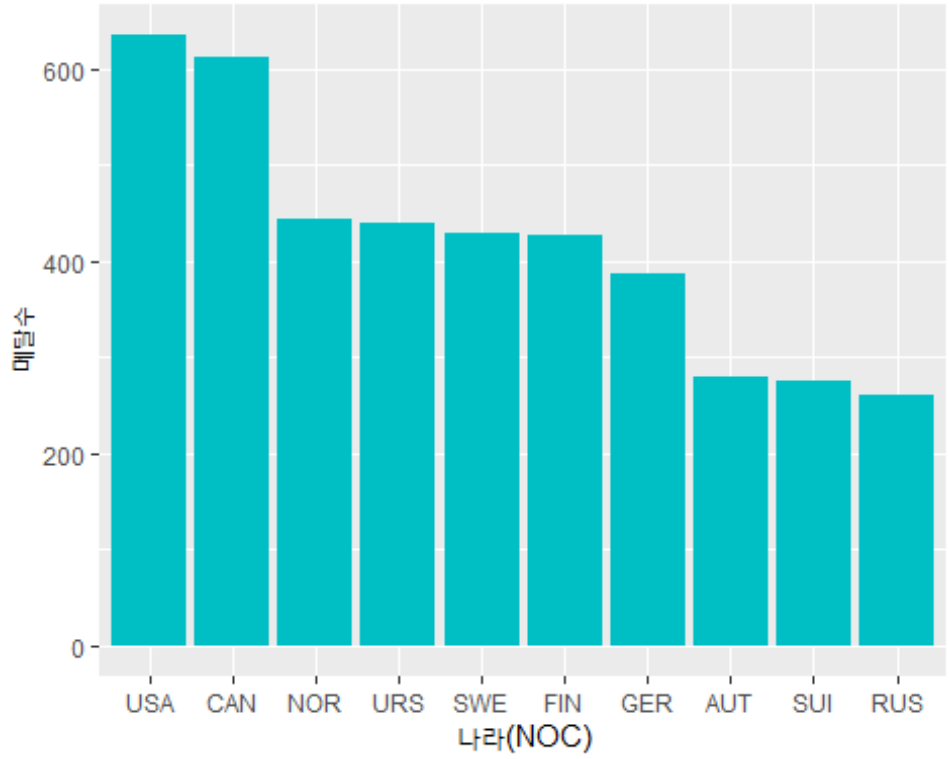
```
summer <- olympic %>%
  filter(Season=='Summer',!is.na(Medal))

summer %>%
  group_by(NOC)%>%
  summarise(medal_count=n())%>%
  filter(rank(desc(medal_count))<=10)%>%
  arrange(desc(medal_count))%>%
  ggplot(aes(x=reorder(NOC,-medal_count,sum),y=medal_count))+geom_bar(stat="i
dentity",fill="#F8766D",show.legend=FALSE)+xlab("나라(NOC)")+ylab("메달수")
```



하계올림픽에는 미국, 소련, UK, 독일, 프랑스, 이탈리아, 호주, 헝가리, 스위스, 네덜란드 순으로 메달수가 높았다. 이러한 결과는 하계올림픽에서 좋은 메달 성적을 보여준 국가들이 더운 나라일 것이라는 가설에 반대되는 결과이다. 하계올림픽 특성상 실내에서 진행되는 경기가 많기도 하며 특별한 자연환경의 구애가 없기 때문이라고 추측할 수 있다.

동계올림픽 메달 획득 수 상위 10 개국

```
winter <- olympic %>%
  filter(Season=='Winter',!is.na(Medal))

winter %>%
  group_by(NOC)%>%
  summarise(medal_count=n())%>%
  filter(rank(desc(medal_count))<=10)%>%
  arrange(desc(medal_count))%>%
  ggplot(aes(x=reorder(NOC,-medal_count),y=medal_count))+geom_bar(stat="ident
ity",fill="#00BFC4")+xlab("나라(NOC)")+ylab("메달수")
```



동계올림픽에서는 미국, 캐나다, 노르웨이, 소련, 스웨덴, 핀란드, 독일, 호주, 스위스, 러시아 순으로 메달 획득을 많이 했다. 대체적으로 추운 나라가 메달을 많이 딸 것이라는 가설을 뒷받침하는 결과임을 알 수 있다. 하계올림픽과 달리 동계올림픽 경기는 대부분 스키장, 아이스경기장, 봅슬레이 경기장 등 특별한 환경이 필요하기 때문이라고 추측할 수 있다. 특히, 노르웨이는 하계올림픽에 메달 상위 10 개 국에 들지 않았는데, 동계올림픽에서 우수한 성적을 가진 것을 볼 수 있었다.

하계올림픽/동계올림픽 국가별 메달 수 워드클라우드

```
summer_country_medals <- summer%>%
  group_by(NOC) %>%
  summarise(medals = n())
summer_country_medals <- na.omit(summer_country_medals)
wordcloud(summer_country_medals$NOC,summer_country_medals$medals,colors=brewe
r.pal(6,"Reds"),random.order=FALSE)

winter_country_medals <- winter%>%
  group_by(NOC) %>%
  summarise(medals = n())
winter_country_medals <- na.omit(winter_country_medals)
wordcloud(winter_country_medals$NOC,winter_country_medals$medals,colors=brewe
r.pal(6,"Blues"),random.order=FALSE)
```

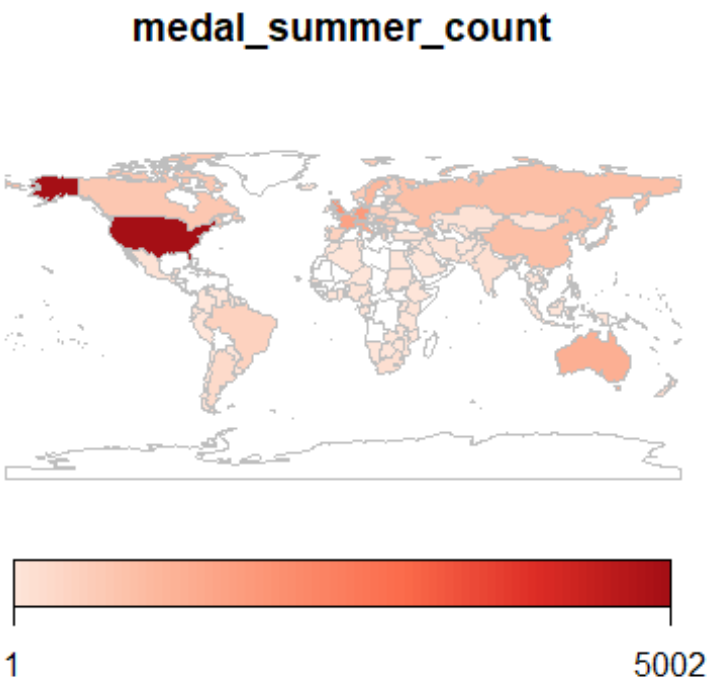
메달 획득 수 상위 10 개국을 표현한 워드클라우드이다. (각각 하계, 동계) 이 역시 앞선 막대 그래프와 같은 결과를 보여줌을 알 수 있다.

World Maps (set up)

```
library(plyr)

library(rworldmap)

library(repr)
options(repr.plot.width=6, repr.plot.height=6)
world <- map_data(map="world")
world <- world[world$region != "Antarctica",]

country_code <- read_csv('c:/Users/KimMinyoung/Documents/country_codes.txt')

country_code <- country_code %>% select(Country, IOC, ISO)
olympic2 <- olympic %>% left_join(country_code, by = c('NOC' = 'IOC'))

ISO_Medal<-olympic2%>%
  filter(!is.na(Medal))%>%
  group_by(ISO)%>%
  dplyr::summarise(medal_count=n())

ISO_Medal_winter<-olympic2%>%
  filter(Season=="Winter")%>%
  filter(!is.na(Medal))%>%
  group_by(ISO)%>%
  dplyr::summarise(medal_winter_count = n())

ISO_Medal_summer<-olympic2%>%
  filter(Season=="Summer")%>%
  filter(!is.na(Medal))%>%
  group_by(ISO)%>%
  dplyr::summarise(medal_summer_count=n())

winter_prop2 <- ISO_Medal%>%left_join(ISO_Medal_winter,by="ISO")
winter_prop2[is.na(winter_prop2)]<-0

winter_prop2<-winter_prop2%>%
  mutate(w_prop=medal_winter_count/medal_count*100)

summer_prop2 <- ISO_Medal%>%left_join(ISO_Medal_summer,by="ISO")
summer_prop2[is.na(summer_prop2)]<-0

summer_prop2<-summer_prop2%>%
  mutate(s_prop=medal_summer_count/medal_count*100)
```

하계올림픽 나라별 메달 수 지도

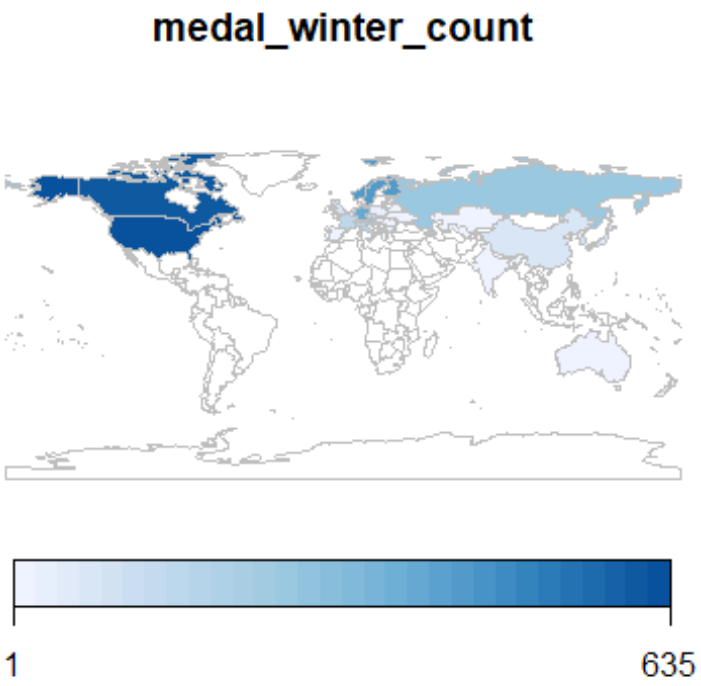
```
options(repr.plot.width=25, repr.plot.height=25)

world <- world[world$region != "Antarctica",]
colourPalette <- RColorBrewer::brewer.pal(6,'Reds')
sPDF1 <- joinCountryData2Map(summer_prop2,joinCode = "ISO3",nameJoinColumn =
"ISO")

## 135 codes from your data successfully matched countries in the map
## 2 codes from your data failed to match with a country code in the map
## 108 codes from the map weren't represented in your data

mapCountryData(sPDF1,nameColumnToPlot='medal_summer_count',colourPalette=colo
urPalette, catMethod ='fixedWidth',numCats = length(table(sPDF1$medal_summer_
count)))
```



세계지도에 표현하여 위도상으로 살펴본 결과 또한 마찬가지로, 따뜻한 나라일수록 메달 수가 많다는 것을 보여주지는 않았음을 알 수 있다. 오히려 아프리카 등 너무 더운 나라들의 메달 수가 적고, 중위도 지역에 위치한 온화한 나라들의 메달 수가 높았다.

동계올림픽 국가별 메달 수 지도

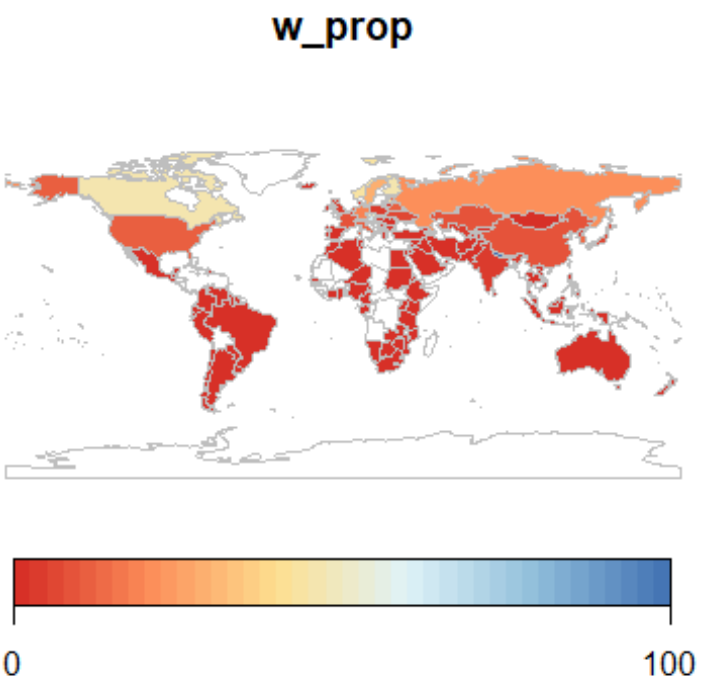
```
options(repr.plot.width=25, repr.plot.height=25)

world <- world[world$region != "Antarctica",]
colourPalette <- RColorBrewer::brewer.pal(6,'Blues')
sPDF2 <- joinCountryData2Map(winter_prop2,joinCode = "ISO3",nameJoinColumn =
"ISO")

## 135 codes from your data successfully matched countries in the map
## 2 codes from your data failed to match with a country code in the map
## 108 codes from the map weren't represented in your data

mapCountryData(sPDF2,nameColumnToPlot='medal_winter_count',colourPalette=colo
urPalette, catMethod ='fixedWidth',numCats = length(table(sPDF2$medal_winter_
count)))
```



지도에 표현하여 살펴본 결과, 북반구의 위도가 높은 지역에서 동계올림픽 메달 수가 많은 것을 확인할 수 있었다. 즉, 추운 지역에서 동계올림픽 메달 수가 많다고 할 수 있다.

나라별 하계올림픽, 동계올림픽 메달 비율

```
options(repr.plot.width=25, repr.plot.height=25)

world <- world[world$region != "Antarctica",]
colourPalette <- RColorBrewer::brewer.pal(6,'RdYlBu')
sPDF3 <- joinCountryData2Map(winter_prop2,joinCode = "ISO3",nameJoinColumn =
"ISO")

mapCountryData(sPDF3,nameColumnToPlot='w_prop',colourPalette=colourPalette, c
atMethod ='fixedWidth',numCats = length(table(sPDF3$w_prop)))
```



출전 인원이 많은 나라일수록 medal count 가 높게 잡힐 가능성이 크므로, 출전 인원의 영향을 배제하기 위해 나라별 하계올림픽, 동계올림픽 메달 비율에 대해 살펴보았다.

저위도지역에서는 동계올림픽보다 하계올림픽에 메달을 훨씬 많이 따는 것을 볼 수 있었고 노르웨이, 캐나다 등 고위도 지역, 히말라야 산맥 근처의 산지인 네팔 등 에서는 동계올림픽 메달 비율이 높아지는 것을 볼 수 있다. 다만, 대부분의 나라가 하계올림픽 메달 수가 많기 때문에 0~50 사이의 비중이 높다.

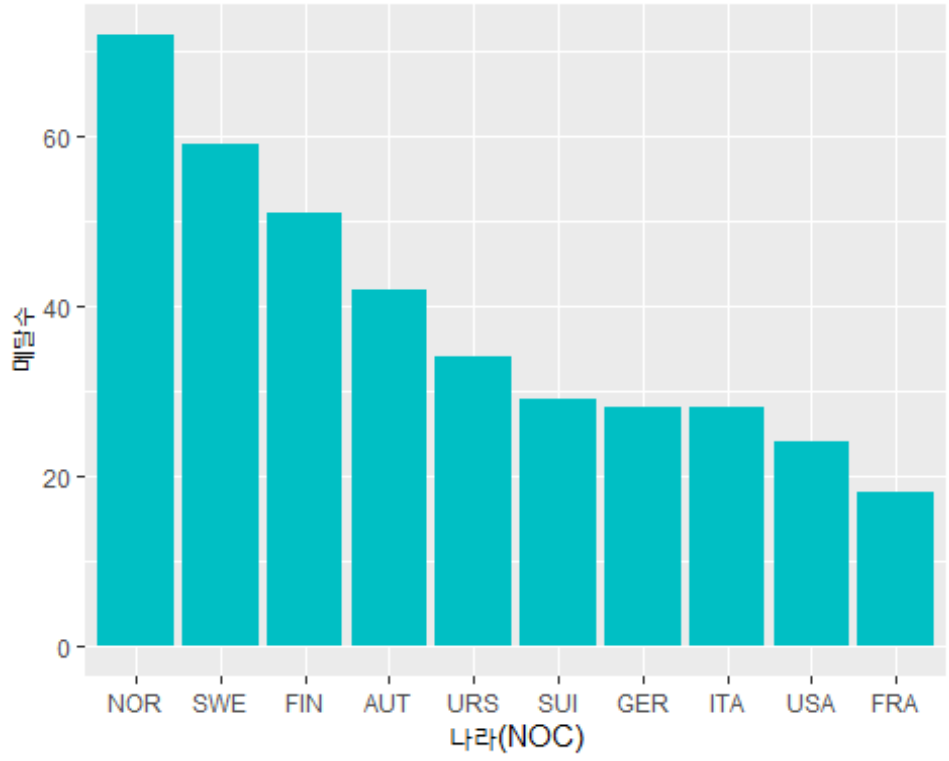
동계올림픽 나라별 메달수 자세히 살펴보기

앞서 추운나라가 동계올림픽에서 메달을 많이 따는 경향을 확인할 수 있었다.

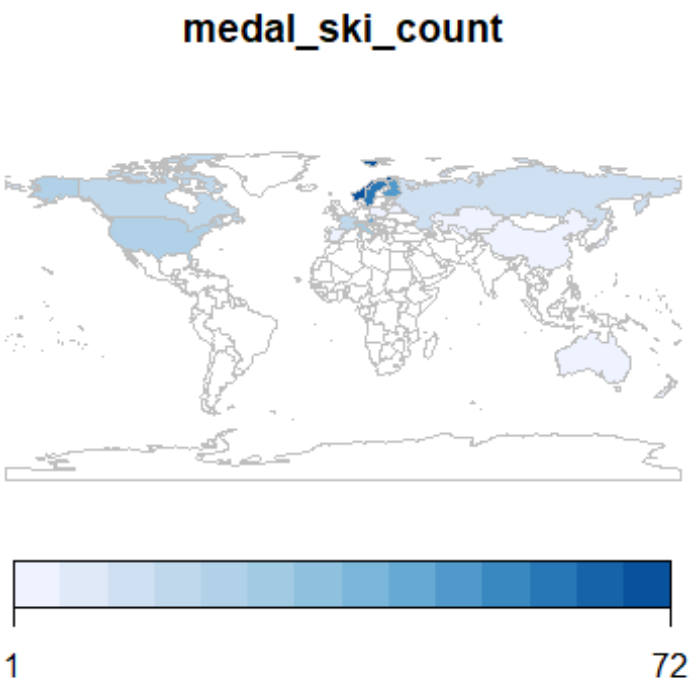
그중에서도 설상종목과 실내종목 간 차이가 있을거라 생각하고, 이를 확인하기 위해 그림을 그려보았다.

스키 종목(설상 종목)

```
winter_ski<-olympic%>%
  filter(!is.na(Medal),Season=="Winter",Sport==c("Alpine Skiing","Cross Country Skiing","Freestyle Skiing"))%>%
  group_by(NOC)%>%
  dplyr::summarise(medal_ski_count = n())
winter_ski%>%
  filter(rank(desc(medal_ski_count))<=10)%>%
  ggplot(aes(x=reorder(NOC,-medal_ski_count),y=medal_ski_count))+geom_bar(stat="identity",fill="#00BFC4")+xlab("나라(NOC)")+ylab("메달수")
```



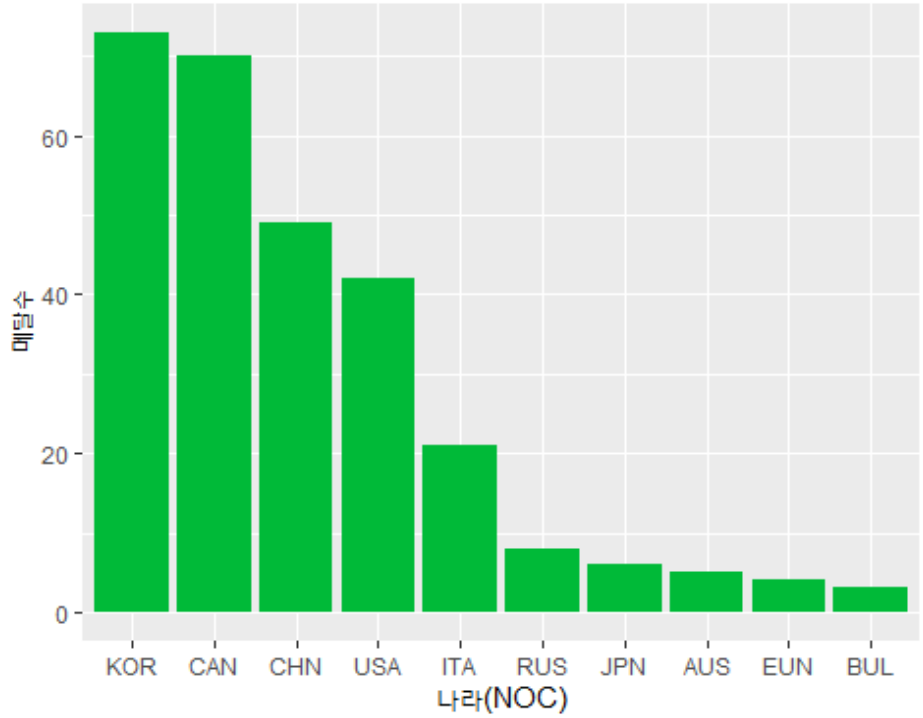
```
colourPalette <- RColorBrewer::brewer.pal(6,'Blues')
sPDF4 <- joinCountryData2Map(winter_ski,joinCode = "ISO3",nameJoinColumn = "N
OC")

mapCountryData(sPDF4,nameColumnToPlot='medal_ski_count',colourPalette=colourP
alette, catMethod ='fixedWidth',numCats = length(table(sPDF4$medal_ski_coun
t)))
```



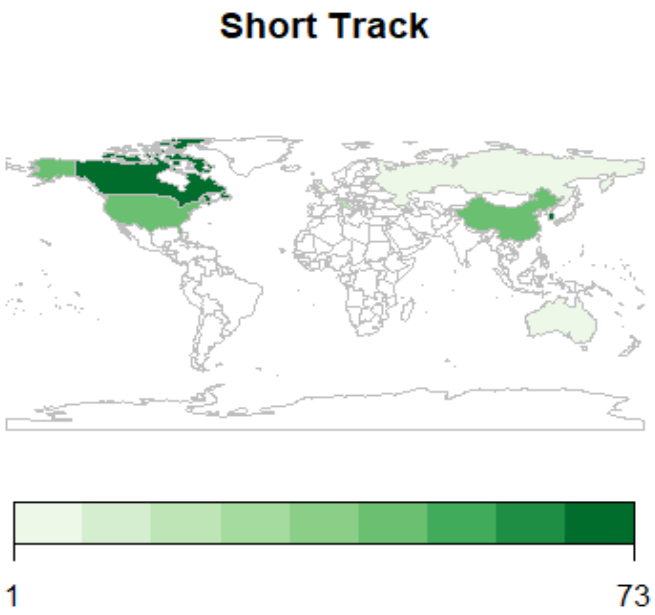
먼저 자연환경의 영향을 많이 받는 스키 종목에 대해 살펴보았다. 노르웨이, 스웨덴, 핀란드, 호주, 러시아, 독일, 이탈리아, 미국, 프랑스 순으로 높았다. 이를 통해 스키종목에서 추운나라가 우세하다는 것을 알 수 있었다. 스키경기를 하기 위해서는 특이한 지형, 즉 눈덮인 산지가 필요하기 때문에 위와 같은 결과가 나왔음을 추측할 수 있다.

쇼트트랙 종목(실내 종목)

```
winter_st<-olympic%>%
  filter(!is.na(Medal),Season=="Winter",Sport==c("Short Track Speed Skating
"))%>%
  group_by(NOC)%>%
  dplyr::summarise(medal_st_count = n())
winter_st%>%
  filter(rank(desc(medal_st_count))<=10)%>%
  ggplot(aes(x=reorder(NOC,-medal_st_count),y=medal_st_count))+geom_bar(stat=
"identity",fill="#00BA38")+xlab("나라(NOC)")+ylab("메달수")
```




```
colourPalette <- RColorBrewer::brewer.pal(6,'Greens')
sPDF5<- joinCountryData2Map(winter_st,joinCode = "ISO3",nameJoinColumn = "NOC
")

mapCountryData(sPDF5,nameColumnToPlot='medal_st_count',mapTitle="Short Track
",colourPalette=colourPalette, catMethod ='fixedWidth',numCats = length(table
(sPDF5$medal_st_count)))
```



위에서와 반대로 자연환경의 영향을 적게 받는 쇼트트랙 종목에 대해 살펴보았다. 한국, 캐나다, 중국, 미국, 이탈리아, 러시아, 일본, 호주 순으로 높았다. 쇼트트랙 경기는 실내에서 진행되기 때문에 설상종목보다 영향이 적을 것이라고 생각해볼 수 있다.

결론

1. Height&Weight : 종목에 따라 출전선수들의 전반적인 키와 몸무게가 상이하다.

특히 농구, volley ball 선수들은 키와 몸무게가 컸고, 역도 선수들은 키가 작고 몸무게가 큰 경향이 있었다.

2. Gender : 1900 년 여성 출전 허가 이후 점차 여성 출전 비율과 메달 획득률이 높아지고 있다

. 남성과 여성의 종목수, 메달수, 출전선수 차이가 컸으나, 남성과 여성의 차이가 줄어들고 있다.

3. Age : 출전선수들의 나이는 20 대 위주였으나, 97 세 등의 고령선수들의 존재도 돋보인다. 특히, 미술, 양궁 등에서 성과를 보인다.

4. Nations : GDP 와 메달 수는 어느정도 연관성이 있다.

1. gdp 가 큰 나라들은 메달수가 높았다.

2. gdp 가 크게 성장한 나라들은 메달수도 크게 증가했다.

+ 분단국가, 전쟁, 보이콧 등에 의해 각 나라별로 특이점이 존재했다.

5. Season : 하계올림픽/동계올림픽 메달수와 나라의 기후간의 관계가 있다.

1. 하계올림픽에는 너무 더운나라의 메달 수가 적었고, 중위도 지역의 온화한 기후를 가진 나라의 메달 수가 높았다.

2. 동계올림픽에는 추운나라의 메달 수가 높았다. 특히 스키 등 설상경기는 그런 경향성이 짙었다.

3. 비율상 더운 나라는 동계올림픽보다 하계올림픽에 메달을 많이 따는 경향이 있었다. 반대로 추운 나라는 동계올림픽 메달 비율이 다른나라에 비해 높았다.