올림픽 운동선수 데이터 분석

graphics-Team5

Data Loading and Preprocessing

```
# 필요한 라이브러리 로드
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────────────── tidy
verse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.2     ✓ tibble    3.2.1
## ✓ lubridate 1.9.4     ✓ tidyr     1.3.1
## ✓ purrr     1.1.0
## ── Conflicts ────────────────────────────────────────────────────────
──── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to bec
ome errors
```

```
library(ggplot2)
library(tibble)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(dplyr)

# 데이터 불러오기
data <- read.csv("C:/graphics/athlete_events.csv")

# 데이터 구조 확인
head(data)
```

	ID	Name	S..	A..	Height	Wei...	Team	N..	Games
	<int>	<chr>	<chr>	<int>	<int>	<dbl>	<chr>	<chr>	<chr>
1	1	A Dijiang	M	24	180	80	China	CHN	1992 Summer
2	2	A Lamusi	M	23	170	60	China	CHN	2012 Summer
3	3	Gunnar Nielsen Aaby	M	24	*NA*	*NA*	Denmark	DEN	1920 Summer
4	4	Edgar Lindenau Aabye	M	34	*NA*	*NA*	Denmark/Sweden	DEN	1900 Summer
5	5	Christine Jacoba Aaftink	F	21	185	82	Netherlands	NED	1988 Winter

	ID	Name	S..	A..	Height	Wei...	Team	N..	Games
	<int>	<chr>	<chr>	<int>	<int>	<dbl>	<chr>	<chr>	<chr>
6	5	Christine Jacoba Aaftink	F	21	185	82	Netherlands	NED	1988 Winter

```
# 결측치 확인
missing_summary <- data %>%
  summarise_all(~ sum(is.na(.))) %>%
  gather(key = "Variable", value = "MissingCount") %>%
  arrange(desc(MissingCount))

print(missing_summary)
```

```
##    Variable MissingCount
## 1     Medal       231333
## 2    Weight        62875
## 3    Height        60171
## 4       Age         9474
## 5        ID            0
## 6      Name            0
## 7       Sex            0
## 8      Team            0
## 9       NOC            0
## 10    Games            0
## 11     Year            0
## 12   Season            0
## 13     City            0
## 14    Sport            0
## 15    Event            0
```

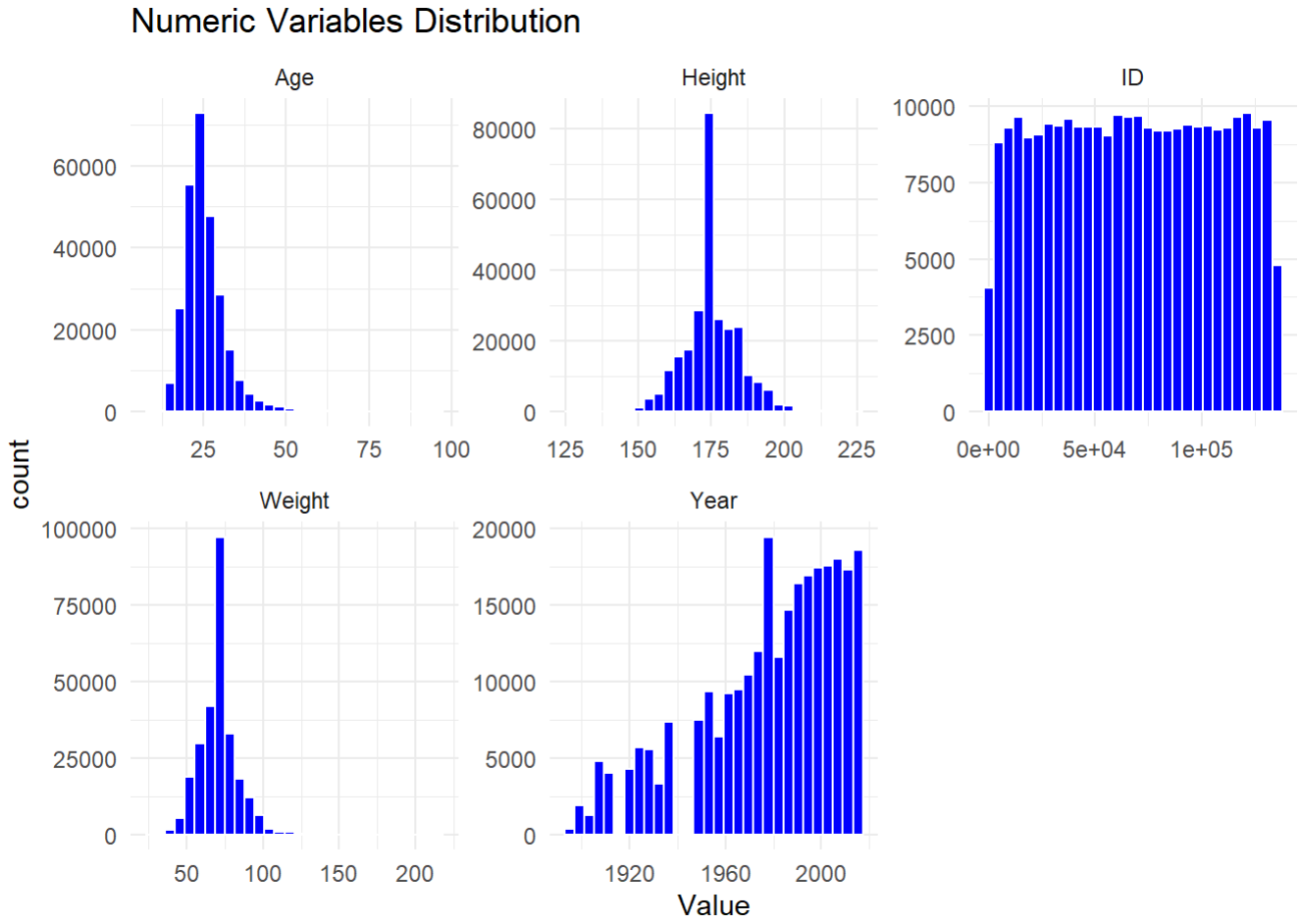
```
# 결측치 처리
# 여기서는 중앙값으로 대체
data_clean <- data %>%
  mutate(across(where(is.numeric),
                ~ ifelse(is.na(.), median(., na.rm = TRUE), .)))

# 처리 후 결측치 확인
missing_summary_clean <- data_clean %>%
  summarise_all(~ sum(is.na(.))) %>%
  gather(key = "Variable", value = "MissingCount") %>%
  arrange(desc(MissingCount))

print(missing_summary_clean)
```

```
##    Variable MissingCount
## 1     Medal       231333
## 2        ID            0
## 3      Name            0
## 4       Sex            0
## 5       Age            0
## 6    Height            0
## 7    Weight            0
## 8      Team            0
## 9       NOC            0
## 10    Games            0
## 11     Year            0
## 12   Season            0
## 13     City            0
## 14    Sport            0
## 15    Event            0
```

```
# 기초 통계 확인
summary(data_clean)
```

```
##        ID             Name               Sex                 Age
##  Min.   :     1   Length:271116      Length:271116      Min.   :10.0
##  1st Qu.: 34643   Class :character   Class :character   1st Qu.:22.0
##  Median : 68205   Mode  :character   Mode  :character   Median :24.0
##  Mean   : 68249                                        Mean   :25.5
##  3rd Qu.:102097                                        3rd Qu.:28.0
##  Max.   :135571                                        Max.   :97.0
##      Height          Weight           Team               NOC
##  Min.   :127.0   Min.   : 25.00   Length:271116      Length:271116
##  1st Qu.:170.0   1st Qu.: 63.00   Class :character   Class :character
##  Median :175.0   Median : 70.00   Mode  :character   Mode  :character
##  Mean   :175.3   Mean   : 70.54
##  3rd Qu.:180.0   3rd Qu.: 75.00
##  Max.   :226.0   Max.   :214.00
##     Games                Year         Season              City
##  Length:271116      Min.   :1896   Length:271116      Length:271116
##  Class :character   1st Qu.:1960   Class :character   Class :character
##  Mode  :character   Median :1988   Mode  :character   Mode  :character
##                     Mean   :1978
##                     3rd Qu.:2002
##                     Max.   :2016
##     Sport              Event              Medal
##  Length:271116      Length:271116      Length:271116
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```
# 주요 변수의 분포 확인
data_clean %>%
  select(where(is.numeric)) %>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Value") %>%
  ggplot(aes(x = Value)) +
  geom_histogram(bins = 30, fill = "blue", color = "white") +
  facet_wrap(~ Variable, scales = "free") +
  theme_minimal() +
  labs(title = "Numeric Variables Distribution")
```

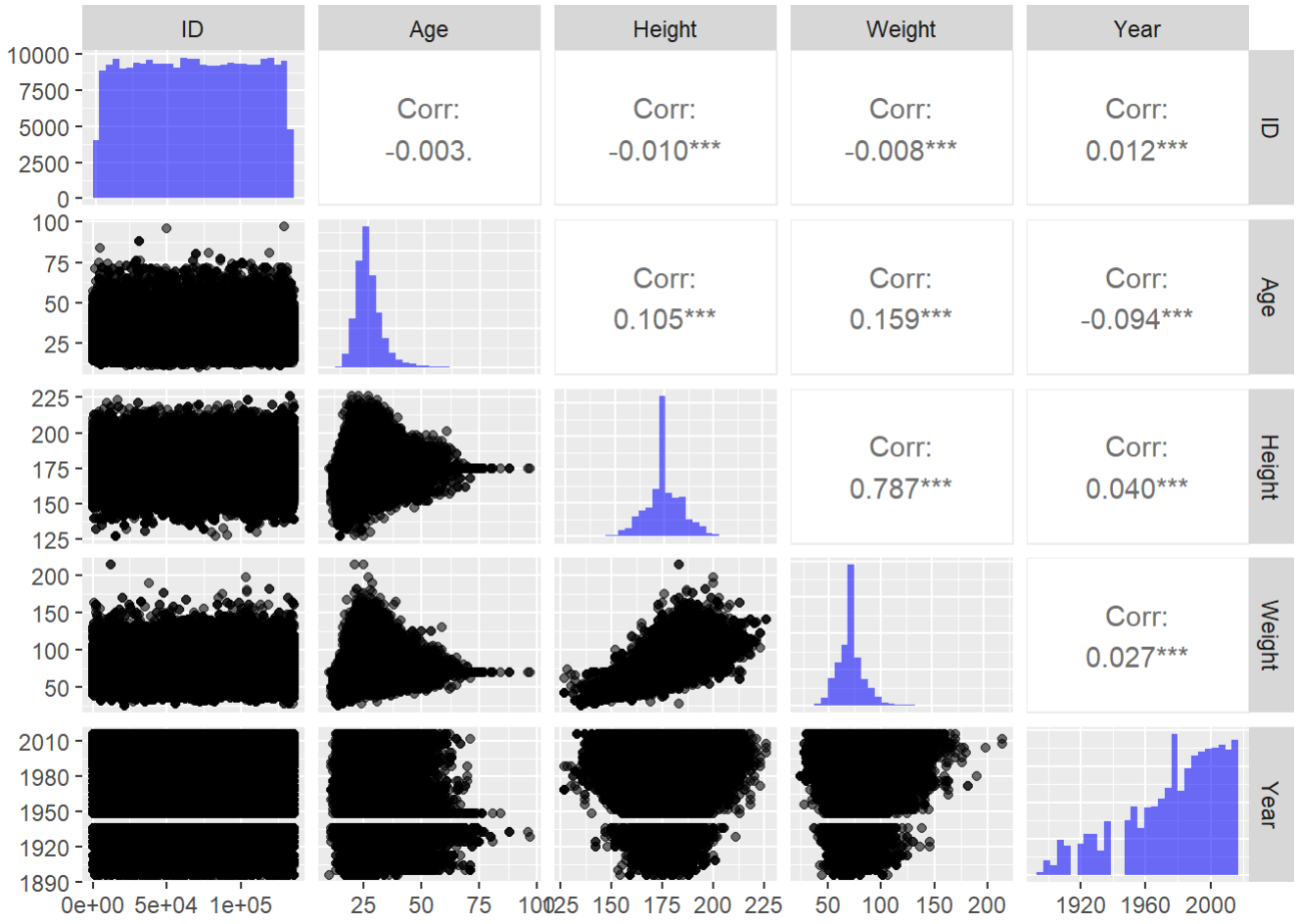


EDA

Correlation of continuous variables

```
ggpairs(data_clean,
        columns=c("ID", "Age", "Height", "Weight", "Year"),
        aes(alpha=0.001,),
        diag = list(continuous = wrap("barDiag", fill = "blue"))
)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



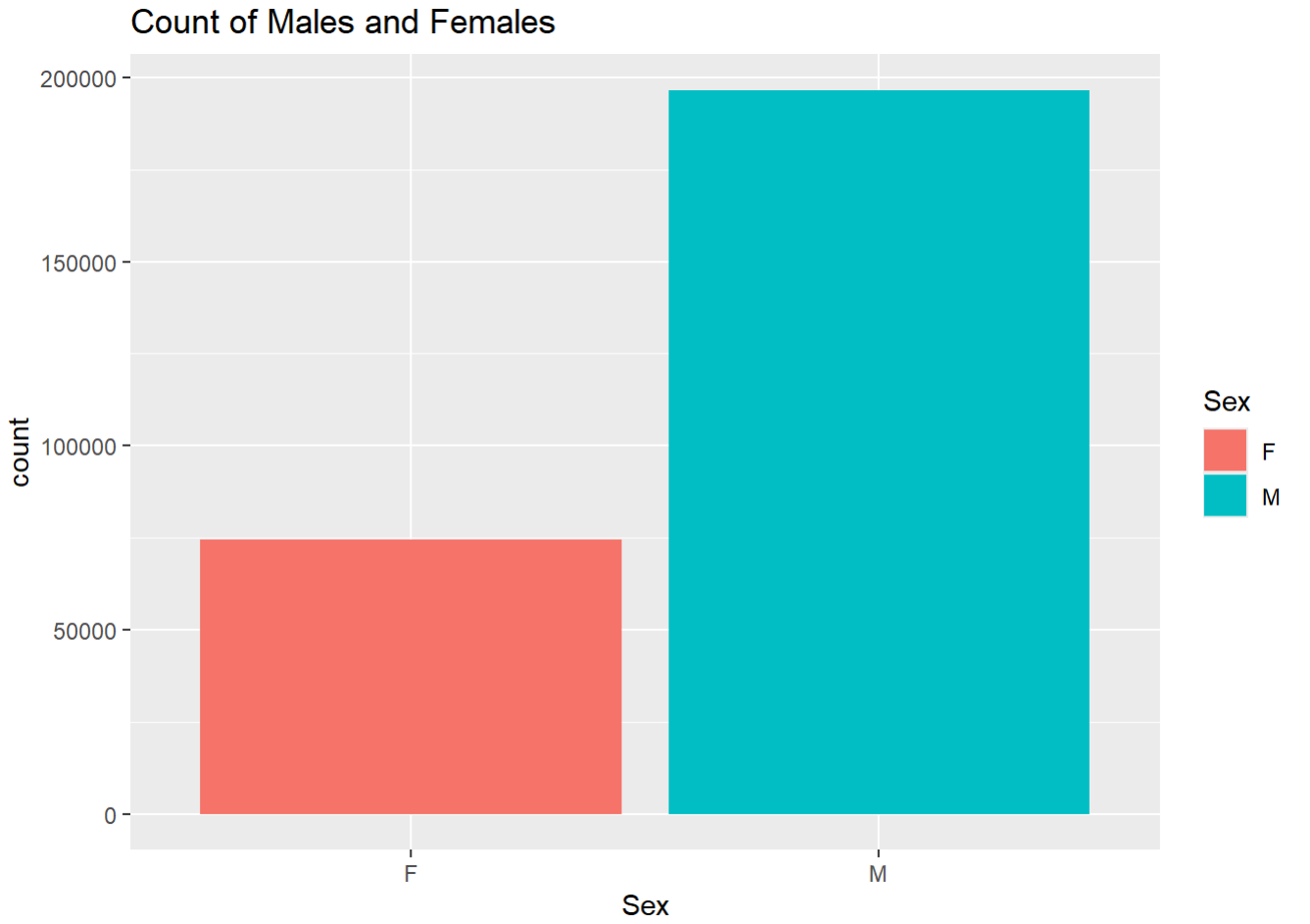
Correlation coefficent가 0.787로 양의 상관관계를 보이는 Weight와 Height를 제외한 대부분의 column 사이에는 유의미한 관계가 없는 것으로 보인다. 이는 Weight과 Height가 서로 어느 정도 영향을 미칠 수 있지만 다른 variable들은 서로 독립적인 것으로 보인다고 볼 수 있다.

The distribution of sex

```
table(data_clean['Sex'])
```

```
## Sex
##      F      M
##  74522 196594
```

```
data_clean %>%
  ggplot(aes(Sex,fill=Sex))+geom_bar()+
  ggtitle("Count of Males and Females")
```



Male이 Female보다 훨씬 많은 수를 차지하고 있음을 알 수 있다. 그래프를 통해 알 수 있는 성별 대표성의 불균형은 올림픽 대회에서 올림픽 경기에서 Male 선수들의 우세를 강조한다.

The distribution of age

```
data_clean %>%
  arrange(desc(Age)) %>%
  select(Age) %>%
  head()
```

	Age <dbl>
1	97
2	96
3	88
4	88
5	88
6	84

6 rows

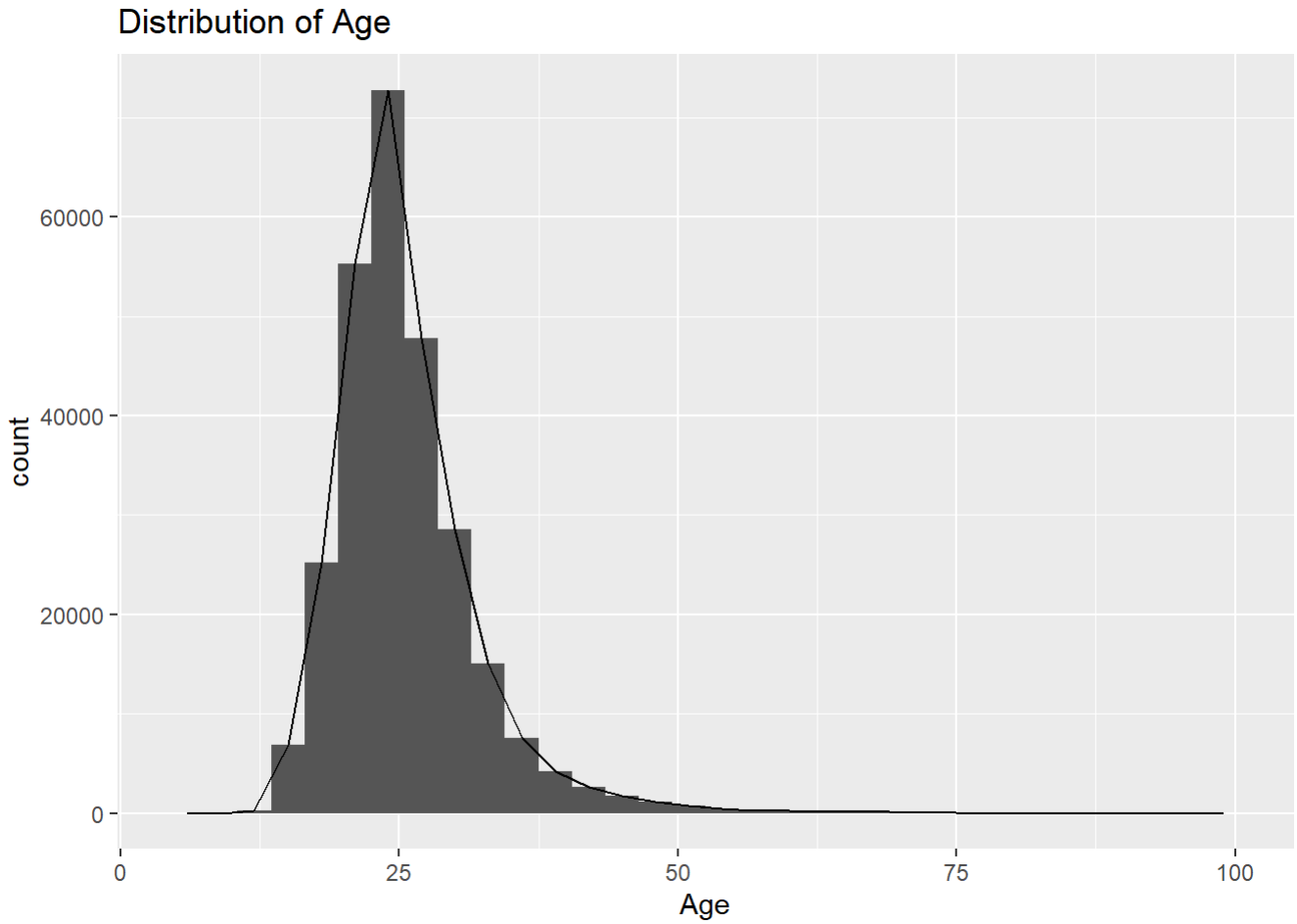
```
data_clean %>%
  group_by(Age) %>%
  summarise(n=n()) %>%
  arrange(desc(Age)) %>%
  head()
```

Age <dbl>	n <int>
97	1
96	1
88	3
84	1
81	2
80	3

6 rows

```
data_clean %>%
  ggplot(aes(Age))+
    geom_histogram()+
    geom_freqpoly()+
    ggtitle("Distribution of Age")
```

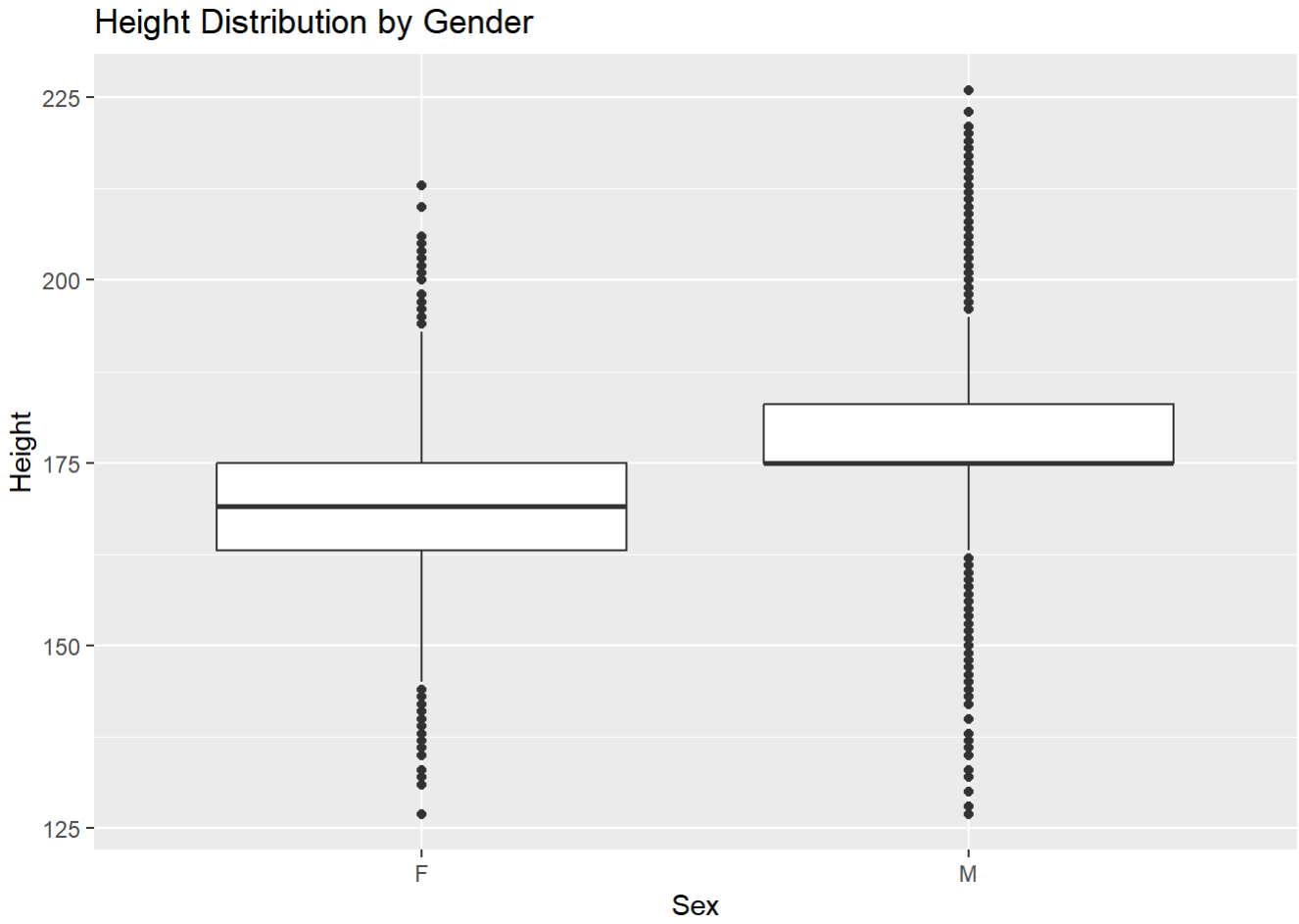
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Age의 histogram은 오른쪽으로 꼬리가 긴 분포를 나타내며, 어린 연령대에 값이 집중되어 있다. 약 25세 부근에서 피크를 보이고 있다. 약 20대 초반부터 30대 초반 사이에 대부분의 선수들의 나이가 분포하고 있음을 확인할 수 있다. 또한 100세 근처에 Outliers가 존재하는데, 이는 dataset에 나이가 많은 선수들이 존재함을 알 수 있다.

The distribution of Heights by gender

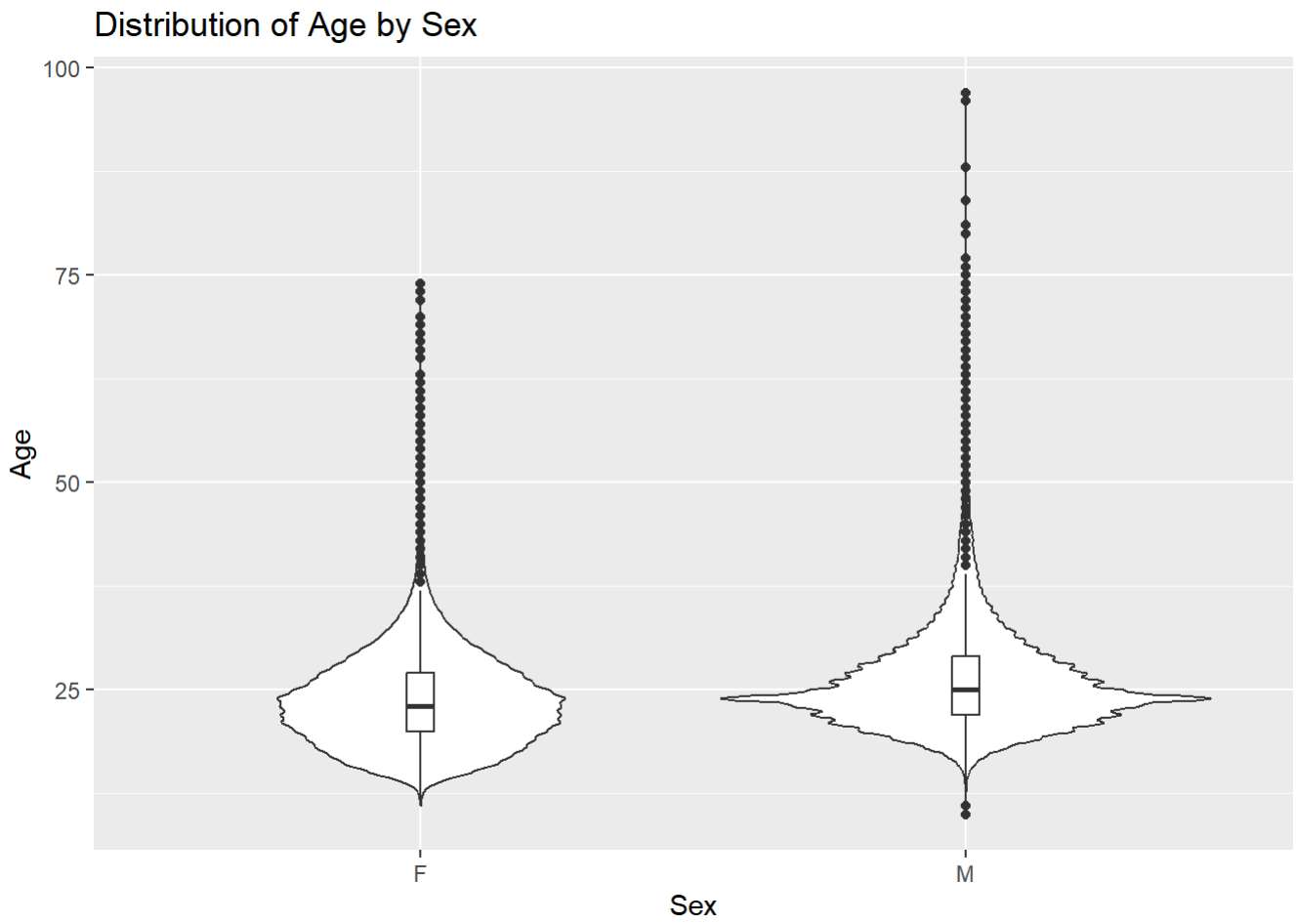
```
data_clean %>%
  ggplot(aes(Sex,Height))+
    geom_boxplot()+
    ggtitle("Height Distribution by Gender")
```



Gender(Male, Female)에 따른 Height 분포를 보여준다. 이 그래프는 Male이 평균적으로 Female보다 더 큰 Height를 가짐을 확인할 수 있으며, 이는 gender 간의 일반적인 생물학적 차이와 일치한다. 이 그래프는 Male과 Female 간의 Height 차이를 명확하게 시각적으로 보여준다.

The distribution of ages by sex

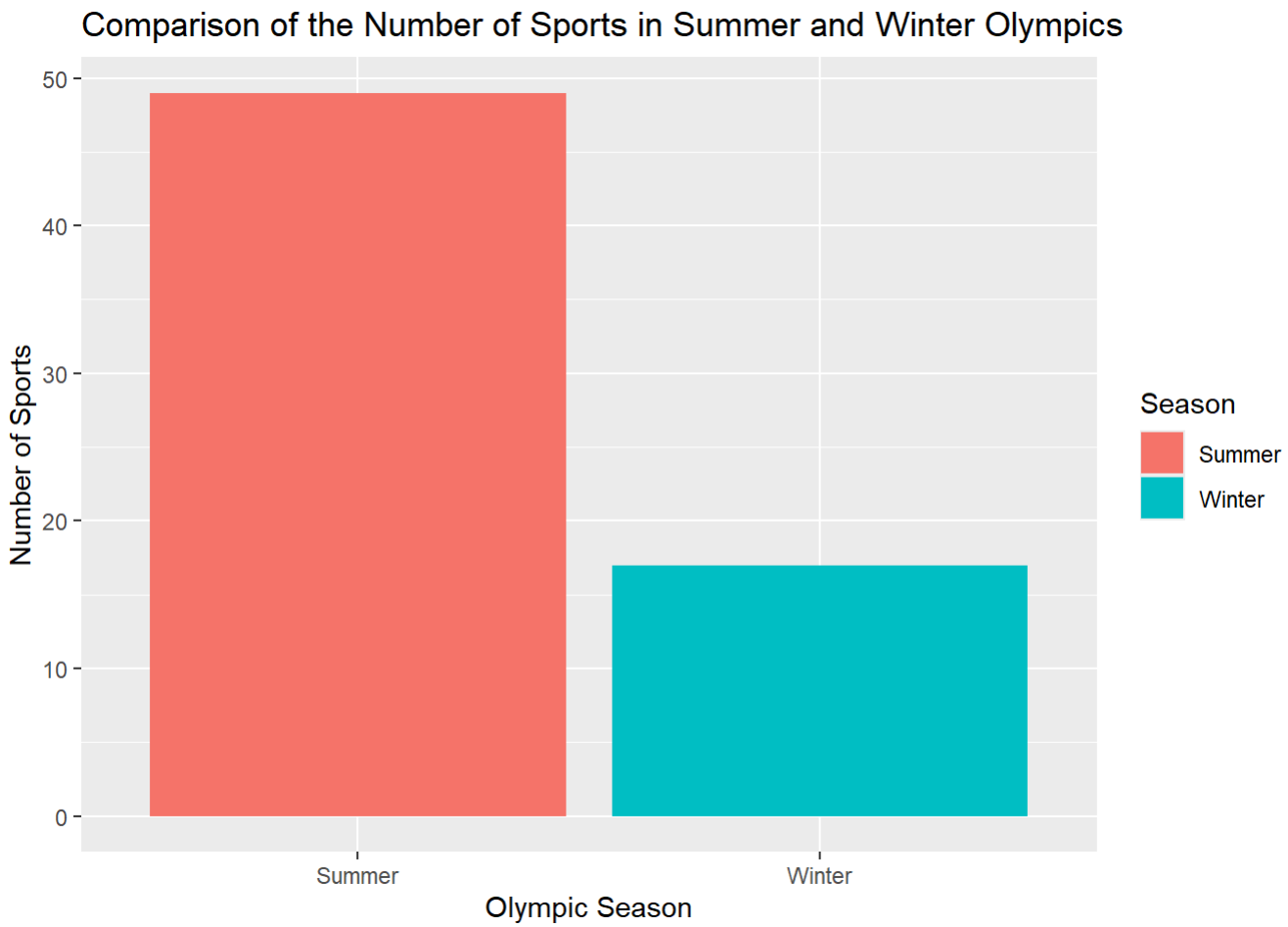
```
data_clean %>%
  ggplot(aes(Sex,Age))+
    geom_violin()+
    geom_boxplot(width=0.05)+
    ggtitle("Distribution of Age by Sex")
```



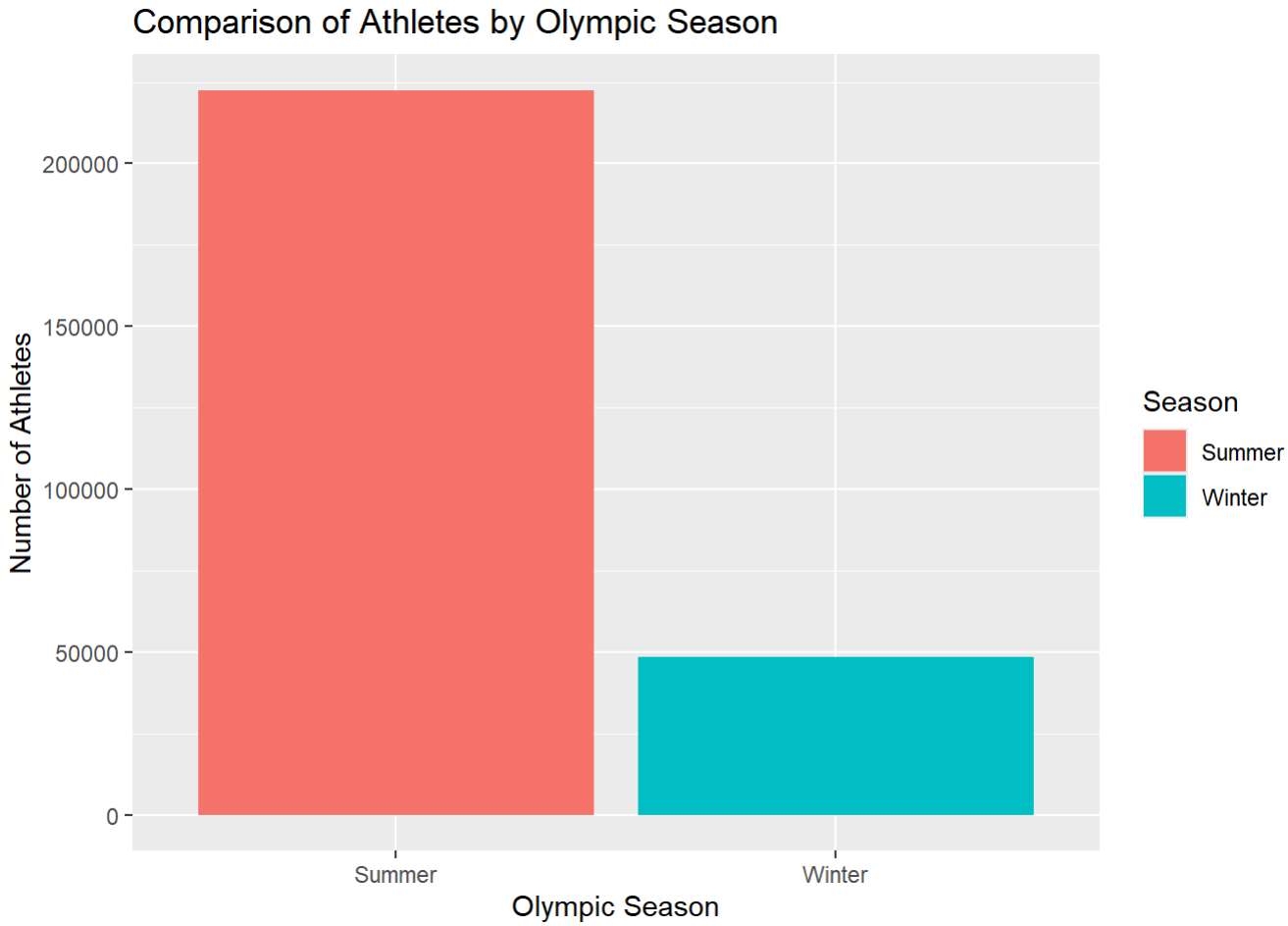
Male 선수들의 Age 분포가 Female 선수들보다 더 큰 변동성을 보임을 나타낸다. 이는 Male 선수들 사이에 Age의 다양성이 더 크다는 것을 보여주는데, 반면 Female 선수들은 상대적으로 나이 분포가 더 일관됨을 알 수 있다.

The distribution of sports and athletes by olympic season

```
# sports
sport_count <- data_clean %>%
  filter(!duplicated(Sport)) %>%
  group_by(Season) %>%
  summarize(Sport_Count = n())

ggplot(sport_count, aes(x = Season, y = Sport_Count, fill = Season)) +
  geom_bar(stat = "identity") +
  labs(title = "Comparison of the Number of Sports in Summer and Winter Olympics",
       x = "Olympic Season",
       y = "Number of Sports")
```



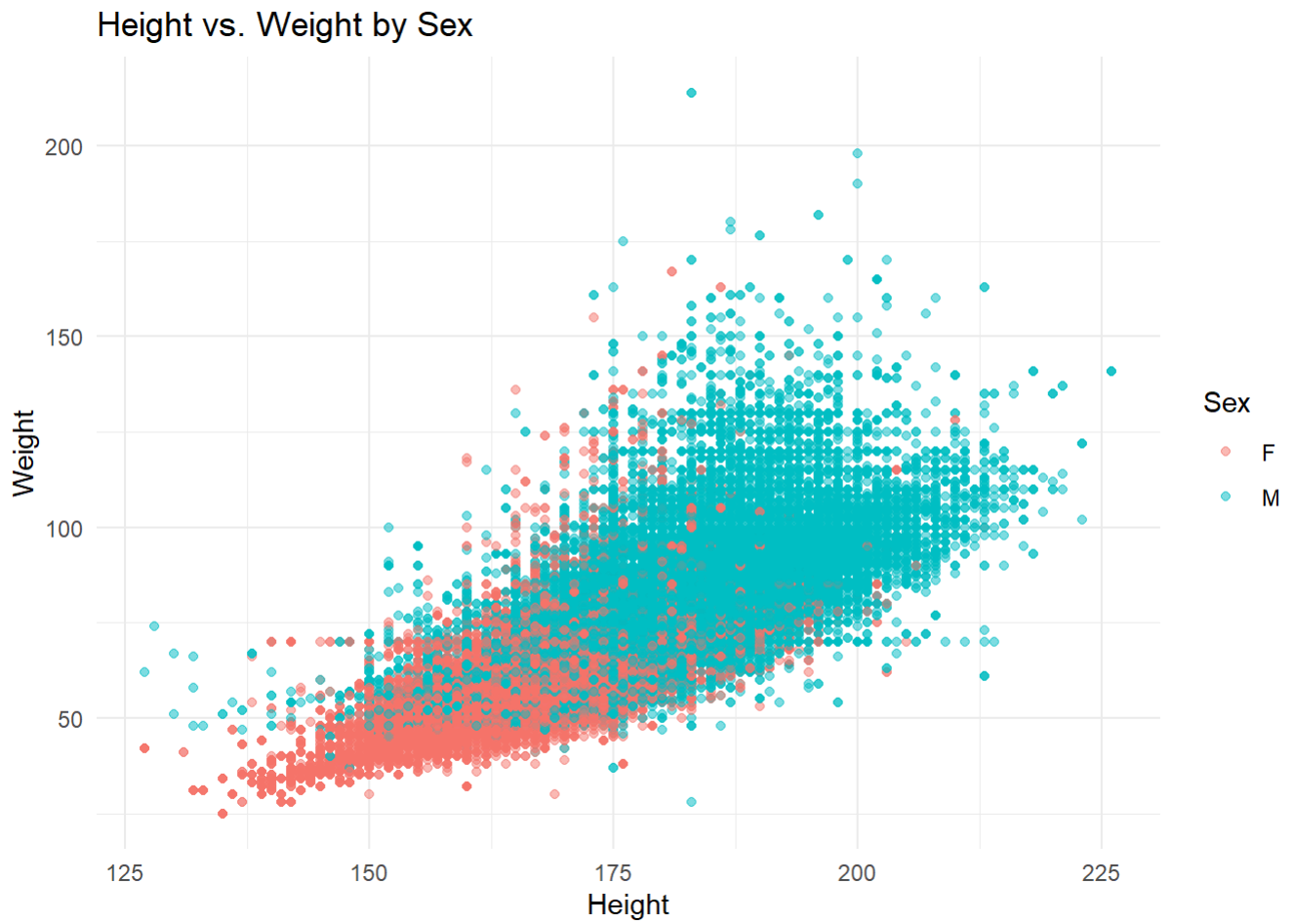

```
# season
season_count <- data_clean %>%
  group_by(Season) %>%
  summarize(Athlete_Count = n())

season_count %>%
  ggplot(aes(x = Season, y = Athlete_Count, fill = Season)) +
  geom_bar(stat = "identity") +
  labs(title = "Comparison of Athletes by Olympic Season",
       x = "Olympic Season",
       y = "Number of Athletes")
```



하계 시즌이 동계 시즌보다 종목 수가 많고, 그로 인해 선수 수 역시 전반적으로 하계가 더 많음을 확인할 수 있다.

Analyzing Height vs. Weight by Sex

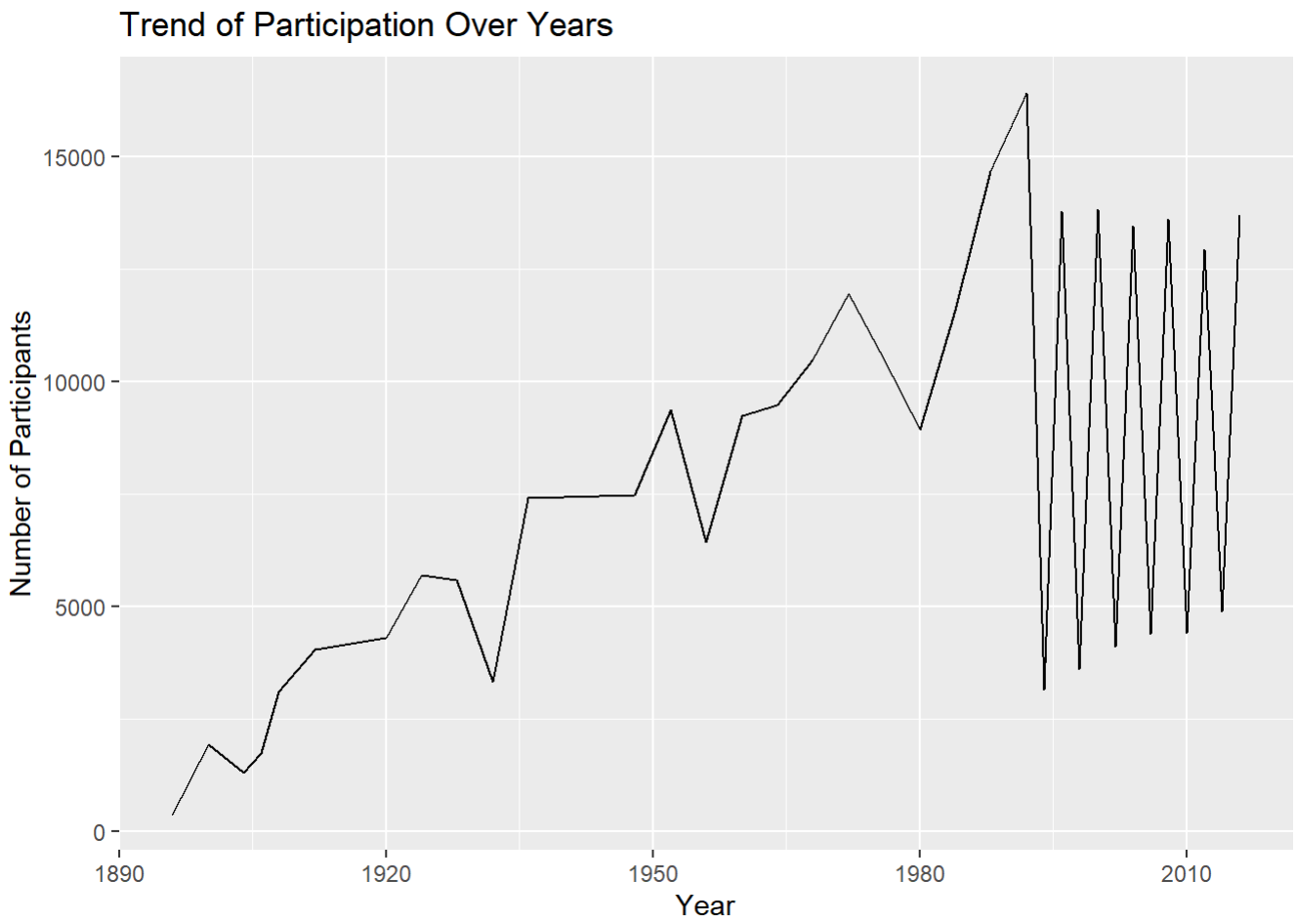
```
ggplot(data_clean, aes(x = Height, y = Weight, color = Sex)) +
  geom_point(alpha = 0.5) +
  ggtitle("Height vs. Weight by Sex") +
  theme_minimal()
```



남녀 모두 키와 몸무게 사이에 양의 상관관계를 가지며 이는 키가 큰 사람이 몸무게가 더 많이 나가는 경향이 있음을 나타낸다.

Trend of Participation Over Years

```
# 연도별로 그룹화하고 각 그룹의 크기를 계산해 연도별 참여도 구하기
participation_by_year <- data_clean %>%
  group_by(Year) %>%
  summarise(count = n())

# line plot으로 연도별 참여 추세 시각화
ggplot(participation_by_year, aes(x = Year, y = count)) +
  geom_line() +
  ggtitle('Trend of Participation Over Years') +
  xlab('Year') +
  ylab('Number of Participants')
```



1900년부터 1992년까지는 올림픽 참가자가 증가하는 추세가 나타나며 1992년에는 약 16,000명의 선수로 가장 높은 참여율을 보여준다. 그러나 1992년 이후로는 지속적으로 참여율이 변동한다는 점을 확인할 수 있다.

Analyze the number of matches held by sport

```
# 데이터셋에 존재하는 년도 확인하기
data_clean %>%
  select(Year) %>%
  distinct() %>%
  arrange(Year)
```

Year <int>
1896
1900
1904
1906
1908
1912
1920
1924
1928

Year <int>
1932

1-10 of 35 rows Previous **1** 2 3 4 Next

```
# 종목별 개최 횟수 계산
sport_counts <- data_clean %>%
  group_by(Sport) %>%
  summarise(count = n_distinct(Year)) %>%
  arrange(desc(count))

sport_counts
```

Sport <chr>	count <int>
Athletics	29
Cycling	29
Fencing	29
Gymnastics	29
Swimming	29
Rowing	28
Wrestling	28
Diving	27
Football	27
Shooting	27

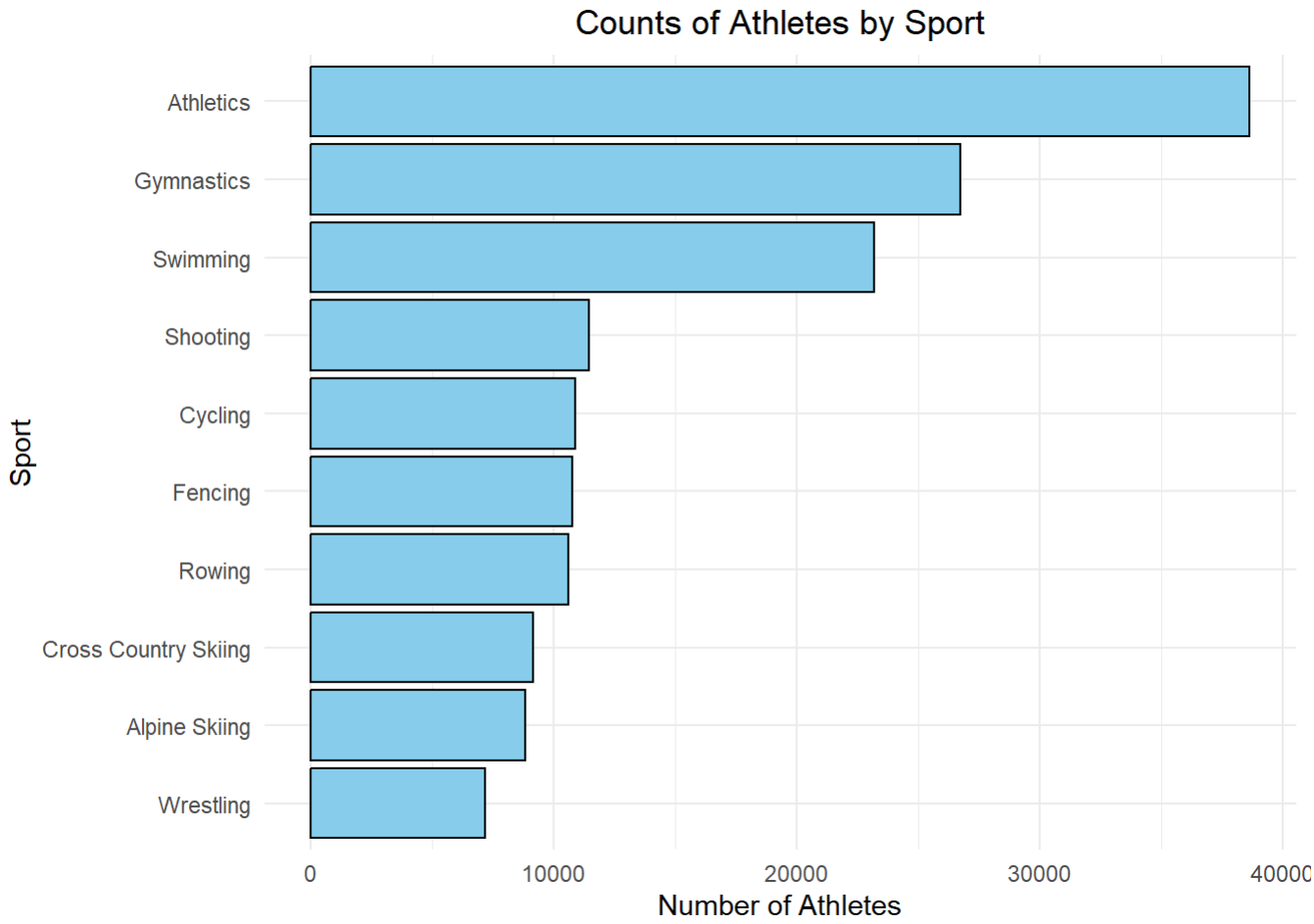
1-10 of 66 rows Previous **1** 2 3 4 5 6 7 Next

관측되는 35회 동안 29번 개최 되는 경기 종목이 있는 반면 5번 이하로 개최 되는 경기도 있음을 확인할 수 있다. 이를 통해 올림픽 경기 종목이 매년 동일하지 않다는 것을 알 수 있다.

Counts of Athletes by Sport

```
# 스포츠별 선수 수 집계
athletes_by_sport <- data_clean %>%
  group_by(Sport) %>%
  summarise(Athlete_Count = n()) %>%
  arrange(desc(Athlete_Count)) %>%
  slice_head(n = 10) # 상위 10개 스포츠 추출
```

```
# 막대그래프 시각화
ggplot(athletes_by_sport, aes(x = reorder(Sport, Athlete_Count), y = Athlete_Count)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black") +
  coord_flip() + # 스포츠 이름이 길 경우 회전
  labs(title = "Counts of Athletes by Sport",
       x = "Sport",
       y = "Number of Athletes") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```

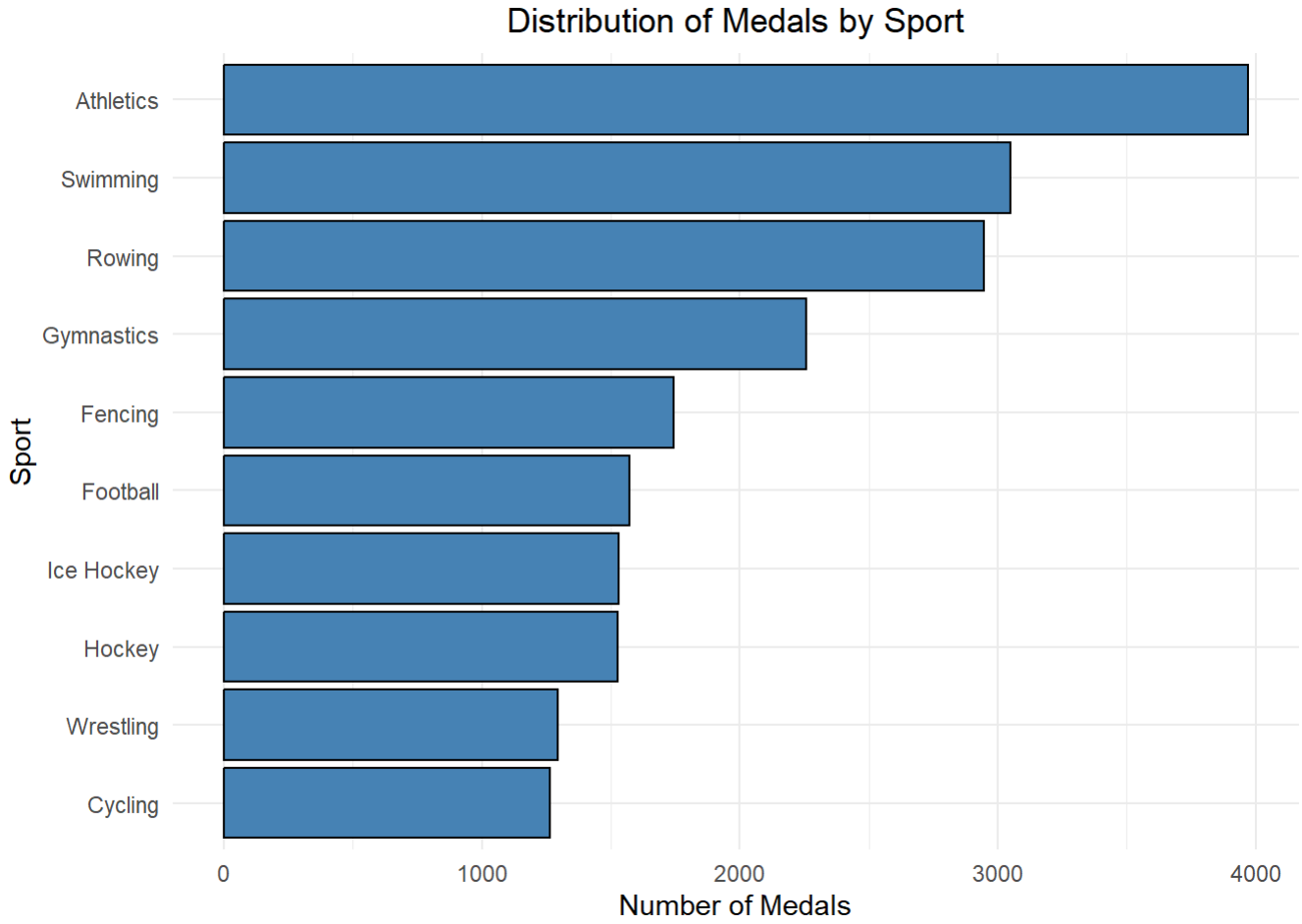


약 40,000명 이상의 선수가 육상에 참가하며, 다른 스포츠에 비해 압도적으로 높은 참여율을 보인다. 체조(Gymnastics)와 수영(Swimming)이 각각 30,000명 이상으로, 육상 다음으로 많은 선수가 참여한 스포츠로 나타난다. 사격(Shooting), 사이클링(Cycling), 펜싱(Fencing), 조정(Rowing), 크로스컨트리 스키(Cross Country Skiing), 알파인 스키(Alpine Skiing), 레슬링(Wrestling) 등이 뒤를 이으며, 대체로 10,000명 이상의 선수가 참여하였다.

Distribution of Medals by Sport

```
# 메달 데이터 필터링 (NA 제외)
medal_data <- data_clean %>%
  filter(!is.na(Medal))

# 스포츠별 메달 수 집계
medals_by_sport <- medal_data %>%
  group_by(Sport) %>%
  summarise(Medal_Count = n()) %>%
  arrange(desc(Medal_Count)) %>%
  slice_head(n = 10)
```

```
# 막대그래프 시각화
ggplot(medals_by_sport, aes(x = reorder(Sport, Medal_Count), y = Medal_Count)) +
  geom_bar(stat = "identity", fill = "steelblue", color = "black") +
  coord_flip() + # 스포츠 이름이 길 경우 회전
  labs(title = "Distribution of Medals by Sport",
       x = "Sport",
       y = "Number of Medals") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```

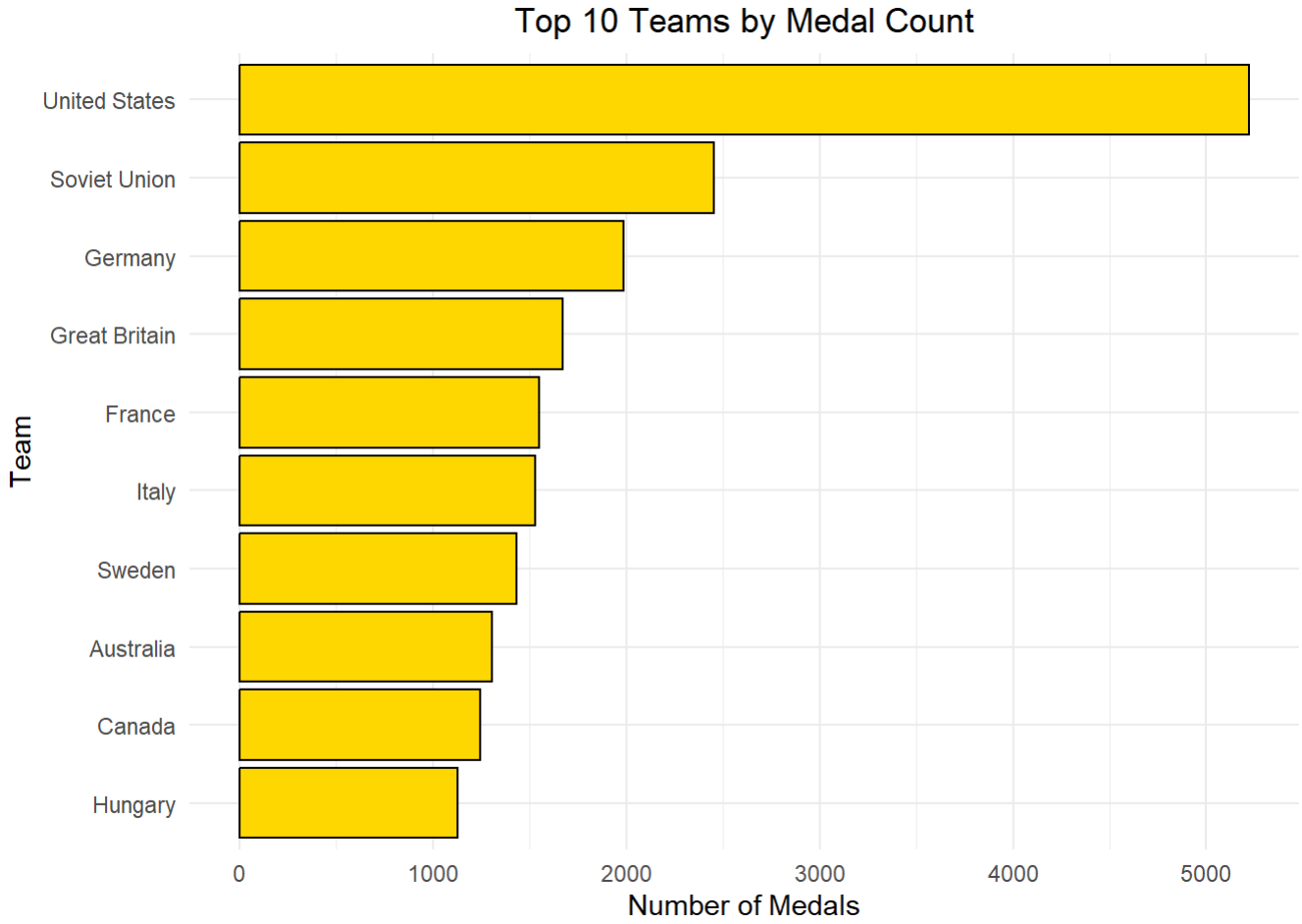


육상이 모든 스포츠 중 가장 많은 메달을 배출했으며, 약 4000개의 메달로 다른 스포츠들보다 월등히 높은 수치를 기록한다. 수영 (Swimming)과 조정 (Rowing)이 각각 2위와 3위를 차지하며, 수영은 3000개 이상의 메달을 기록한다. 체조(Gymnastics), 펜싱(Fencing), 축구(Football), 아이스하키(Ice Hockey), 하키(Hockey), 레슬링(Wrestling), 사이클링(Cycling)이 상위권에 포함되었으며, 각각 약 1000~2000개의 메달 범위에 위치한다.

Distribution of Medals by Team

```
# Team(국가 또는 팀)별 메달 수 계산
medals_by_team <- data_clean %>%
  filter(!is.na(Medal)) %>% # 결측치 제외
  group_by(Team) %>%        # Team 열 기준 그룹화
  summarise(Medal_Count = n()) %>%
  arrange(desc(Medal_Count)) %>%
  slice_head(n = 10)        # 상위 10개 팀 추출
```

```
# 막대그래프 시각화
ggplot(medals_by_team, aes(x = reorder(Team, Medal_Count), y = Medal_Count)) +
  geom_bar(stat = "identity", fill = "gold", color = "black") +
  coord_flip() + # 그래프 회전
  labs(title = "Top 10 Teams by Medal Count",
       x = "Team",
       y = "Number of Medals") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



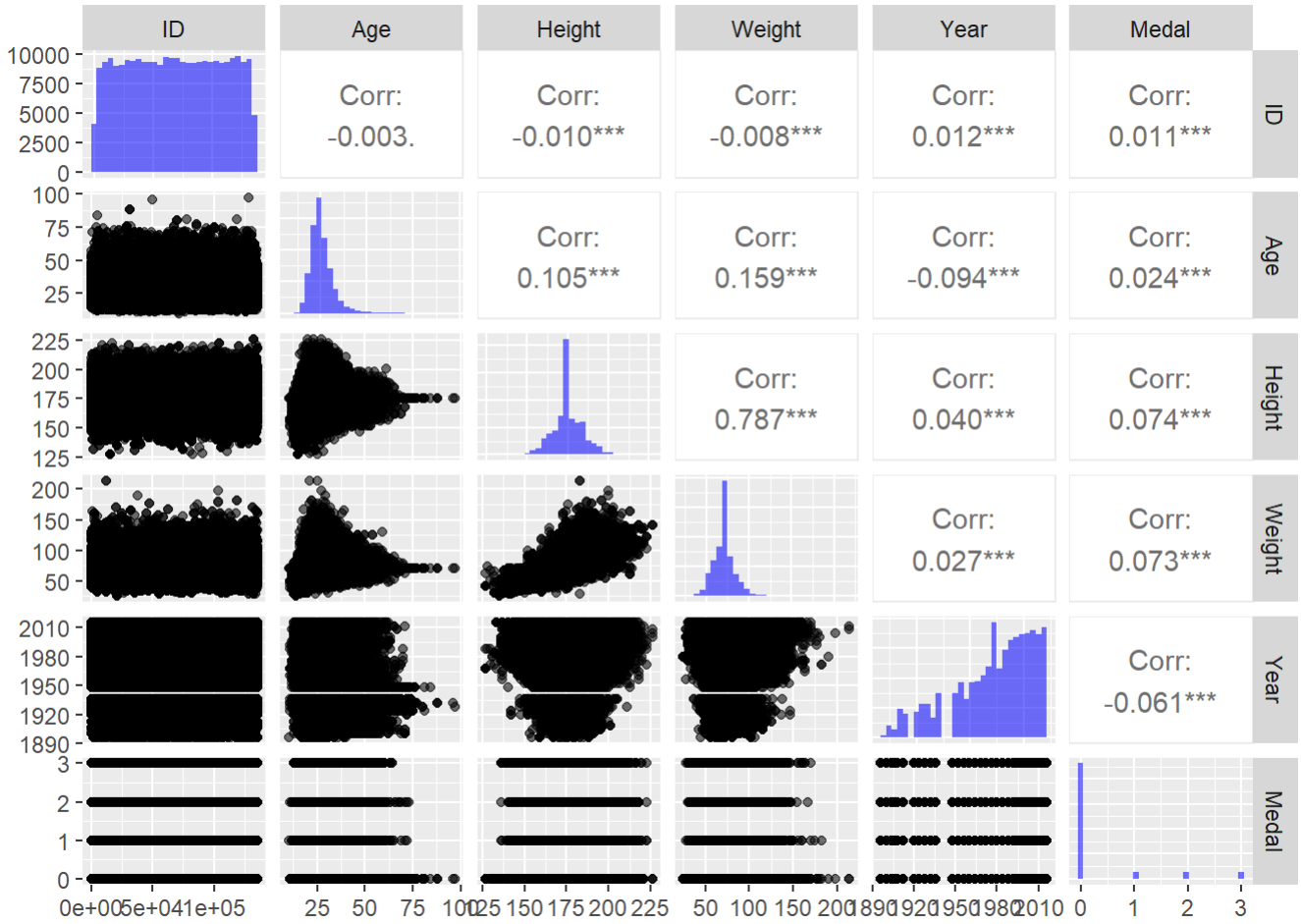
미국은 5000개 이상의 메달을 획득하며 1위를 차지하고 있다. 소련은 약 2500개 이상의 메달로 2위를 기록하며, 미국의 절반 수준의 메달 수를 보유하고 있다. 상위 10개 팀 중 대부분이 유럽 국가들로 구성되어 있다. 미국과 소련을 제외하면, 유럽 국가들이 스포츠 강국임을 알 수 있다.

Correlation Analysis

Exploring Correlation with Medal Winning

```
# categorical 변수인 Medal을 numerical 변수로 인코딩
data_encoded <- data_clean
medal_mapping <- c('Gold' = 3, 'Silver' = 2, 'Bronze' = 1)
data_encoded$Medal <- medal_mapping[data_encoded$Medal]

# Medla을 수상하지 못해 NA로 입력된 값 0으로 대체
data_encoded$Medal[is.na(data_encoded$Medal)] <- 0

# 인코딩 된 데이터프레임을 이용해 상관관계 행렬 계산
correlation_matrix <- data_encoded %>%
  select(where(is.numeric)) %>%
  cor()

# 상관관계 그래프 시각화
ggpairs(data_encoded,
        columns=c("ID", "Age", "Height", "Weight", "Year","Medal"),
        aes(alpha=0.001,),
        diag = list(continuous = wrap("barDiag", fill = "blue"))
)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

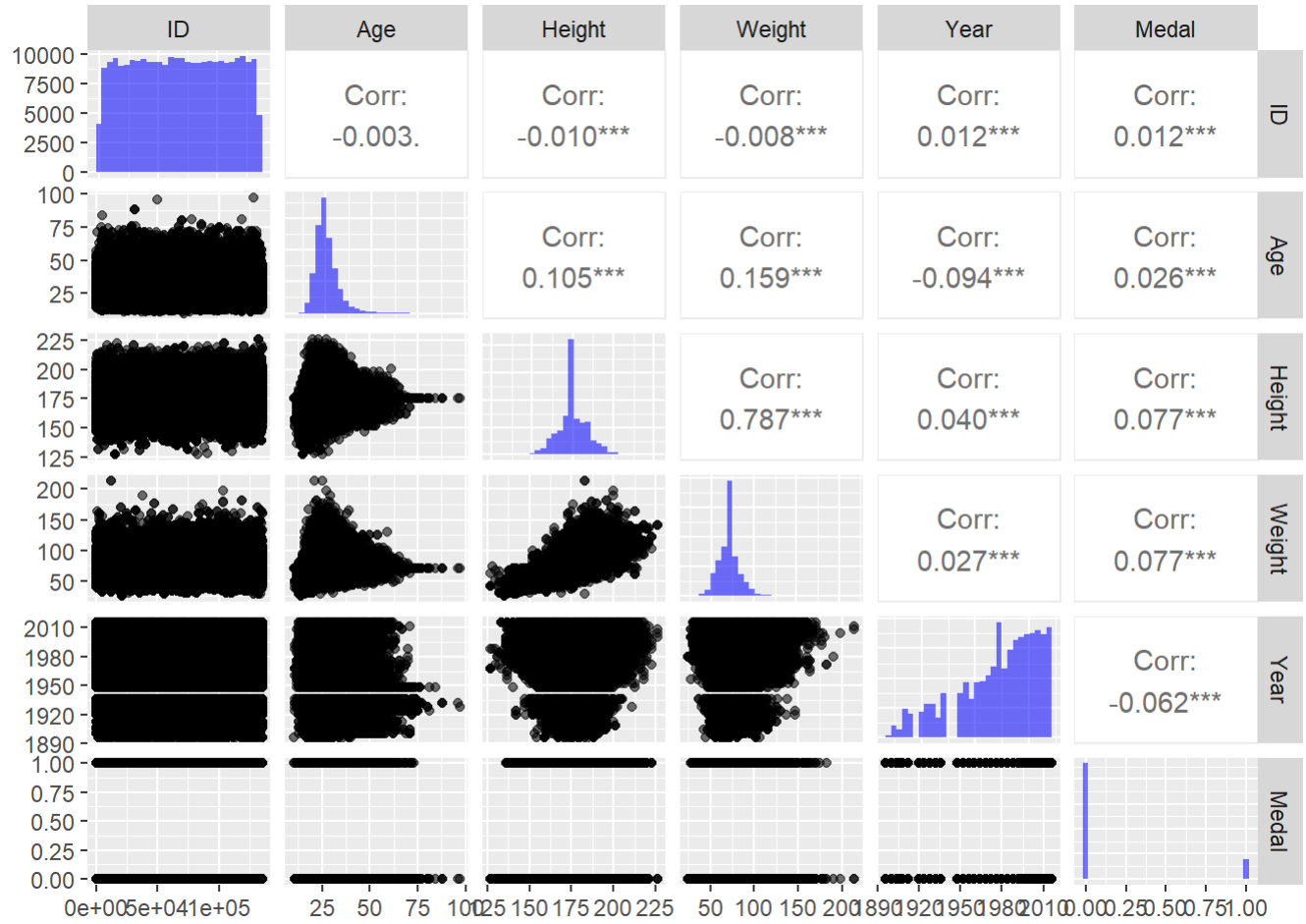


Categorical 변수인 Medal을 Numerical 변수로 인코딩하여 획득한 메달 종류와 다른 변수 사이에 상관관계가 있는지 나타내는 그래프이다. 가장 높은 값이 약 0.08인 것을 통해 획득한 메달 종류와 변수 간의 상관관계는 없는 것으로 보이고 즉, 획득한 메달 종류와 데이터 세트의 다른 변수 사이에 유의미한 관계가 관찰되지 않음을 시사한다.

Analyzing Correlation with Medal Winning

```
# 'Medal' 값을 이진 인코딩: 메달을 받은 경우 1, 메달을 받지 않은 경우 0
data_encoded$Medal <- ifelse(data_encoded$Medal > 0, 1, 0)

# 인코딩 된 데이터프레임을 이용해 상관관계 행렬 계산
correlation_matrix <- data_encoded %>%
  select(where(is.numeric)) %>%
  cor()

# 상관관계 그래프 시각화
ggpairs(data_encoded,
        columns=c("ID", "Age", "Height", "Weight", "Year","Medal"),
        aes(alpha=0.001,),
        diag = list(continuous = wrap("barDiag", fill = "blue"))
)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Medal을 획득했는지 안했는지 이진 분류로 인코딩을 다시 진행하여 메달 획득과 다른 변수 사이에 상관관계가 있는지 나타내는 그래프이다. 가장 높은 값이 약 0.08인 것을 통해 메달 획득과 변수 간의 상관관계는 없는 것으로 보이고 즉, 메달 획득과 데이터 세트의 다른 변수 사이에 유의미한 관계가 관찰되지 않음을 시사한다.

각 변수와 메달 획득 여부 간의 개별적인 선형 상관관계는 크지 않은 것으로 나타났다. 이는 메달 획득이 단일 변수가 아니라 여러 변수의 조합과 상호작용에 의해 결정되기 때문일 수 있다. 예를 들어, 키, 몸무게, 나이 등 여러 요인이 결합될 때 메달 획득에 더 큰 영향을 미칠 수 있다.

따라서 이후 분석에서는 다변량 탐색을 통해 메달 획득과 다른 변수들 간의 관계를 심층적으로 조사하고, 변수 간 상호작용이 메달 획득에 어떤 영향을 미치는지 탐구할 예정이다.

주제 심화 탐구

계절별 종목의 게임 수 계산

가설: 하계 종목 내에서는 Athletics 종목이 다른 종목에 비해 대중적 인기가 높으며, 경기 진행이 용이하기 때문에 게임 수가 많을 것이다. 반면에 생소한 종목은 하계 내에서 게임 진행 수가 적을 것이다. 또한 동계 종목 내에서는 스키, 아이스하키 등의 익숙한 종목이 게임 수가 많을 것이다.

```r
# 필요한 패키지 로드
library(dplyr)
library(ggplot2)

# 계절별 종목별 게임 횟수 계산
season_sport_game_count <- data_encoded %>%
  group_by(Season, Sport) %>%
  summarize(Game_Count = n(), .groups = "drop") %>%
  arrange(Season, desc(Game_Count))

# 상위 5개 종목 확인 (계절별)
top_season_sports <- season_sport_game_count %>%
  group_by(Season) %>%
  slice_head(n = 5) # 각 계절별 상위 5개 종목

# 결과 출력
print("계절별 상위 종목 (게임 수 기준):")
```

```
## [1] "계절별 상위 종목 (게임 수 기준):"
```

```r
print(top_season_sports)
```

```
## # A tibble: 10 × 3
## # Groups:   Season [2]
##    Season Sport                Game_Count
##    <chr>  <chr>                     <int>
##  1 Summer Athletics                 38624
##  2 Summer Gymnastics                26707
##  3 Summer Swimming                  23195
##  4 Summer Shooting                  11448
##  5 Summer Cycling                   10859
##  6 Winter Cross Country Skiing       9133
##  7 Winter Alpine Skiing              8829
##  8 Winter Speed Skating              5613
##  9 Winter Ice Hockey                 5456
## 10 Winter Biathlon                   4893
```

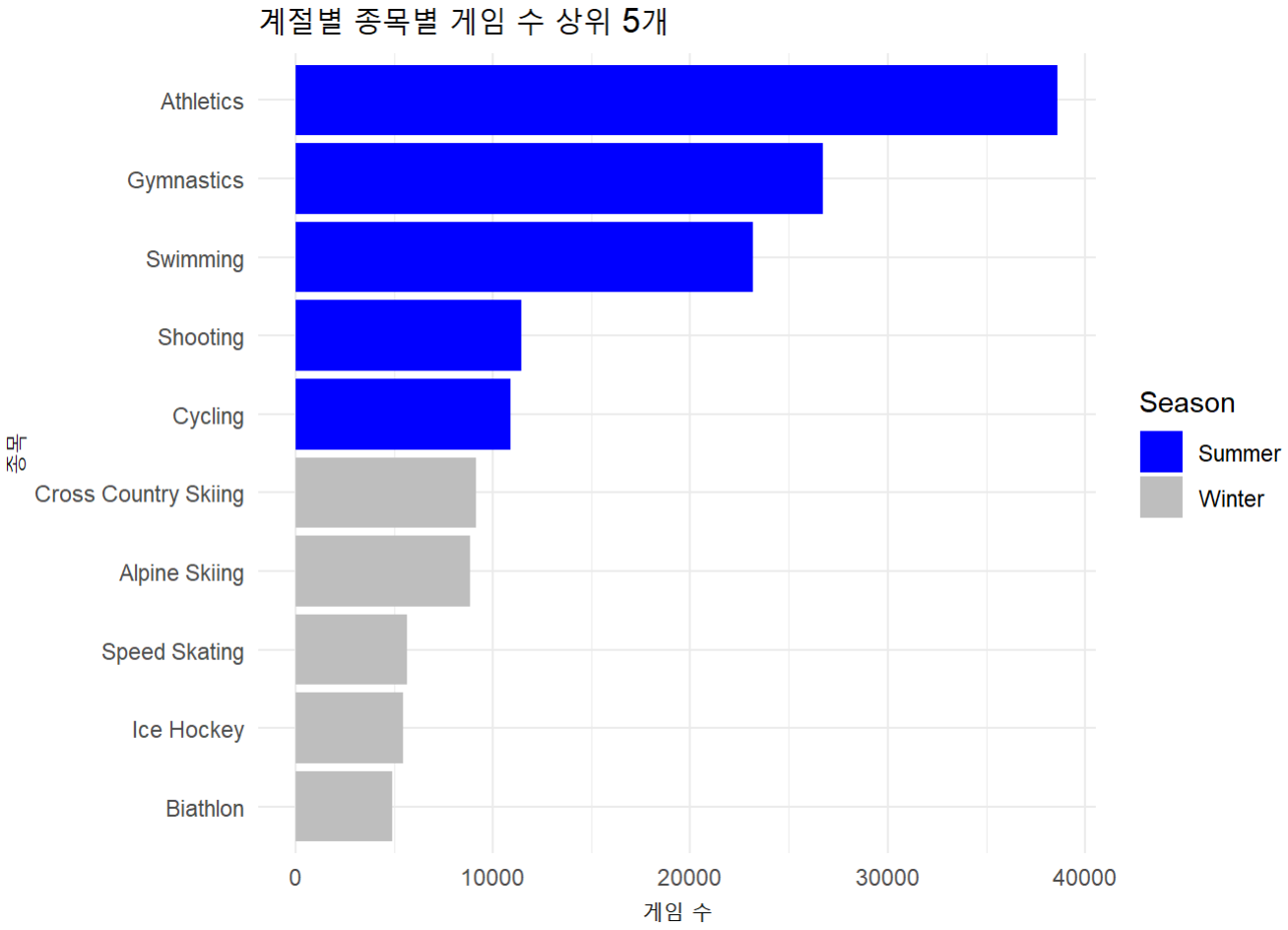
```r
# 하위 5개 종목 확인 (계절별)
low_season_sports <- season_sport_game_count %>%
  group_by(Season) %>%
  slice_tail(n = 5) # 각 계절별 하위 5개 종목

# 결과 출력
print("계절별 하위 종목 (게임 수 기준):")
```

```
## [1] "계절별 하위 종목 (게임 수 기준):"
```

```r
print(low_season_sports)
```

```
## # A tibble: 10 × 3
## # Groups:   Season [2]
##    Season Sport               Game_Count
##    <chr>  <chr>                    <int>
##  1 Summer Jeu De Paume                11
##  2 Summer Alpinism                     4
##  3 Summer Roque                        4
##  4 Summer Basque Pelota                2
##  5 Summer Aeronautics                  1
##  6 Winter Snowboarding               936
##  7 Winter Curling                    463
##  8 Winter Skeleton                   199
##  9 Winter Military Ski Patrol         24
## 10 Winter Alpinism                    21
```

```r
# 시각화: 계절별 상위 종목
ggplot(top_season_sports, aes(x = reorder(Sport, Game_Count), y = Game_Count, fill = Season)) +
  geom_bar(stat = "identity", position = "dodge") +
  coord_flip() +
  labs(
    title = "계절별 종목별 게임 수 상위 5개",
    x = "종목",
    y = "게임 수"
  ) +
  theme_minimal() +
  scale_fill_manual(values = c("Summer" = "blue", "Winter" = "gray"))
```



해석: 예측한 대로 여름에 진행하는 종목 내에서는 Athletics 종목에서 게임 수가 가장 높게 나왔으며, 유일하게 게임 수가 30,000이 넘는다. Basque Pelota와 Aeronautics 등의 상대적으로 생소한 종목이 게임 수가 적게 나왔다. 겨울에 진행한 종목 내에서는 예측한 대로 두 종류의 스키 종목인 Cross Country Skiing와 Alpine Skiing가 상위 두 종목으로 나타났고, Alpinism과 같은 생소한 종목이 적은 게임 수를 기록했다.

키와 몸무게를 합친 값으로 국가별 분석

가설: 추운 지역, 즉 북유럽(노르웨이, 스웨덴, 핀란드)이나 러시아처럼 추운 기후와 넓은 지형을 가진 국가의 선수들은 대체로 키가 크고 체격이 클 것이다. 반면에 더운 지역, 즉 열대 기후 지역(예: 아프리카, 동남아시아) 선수들은 체구가 비교적 작을 것이다.

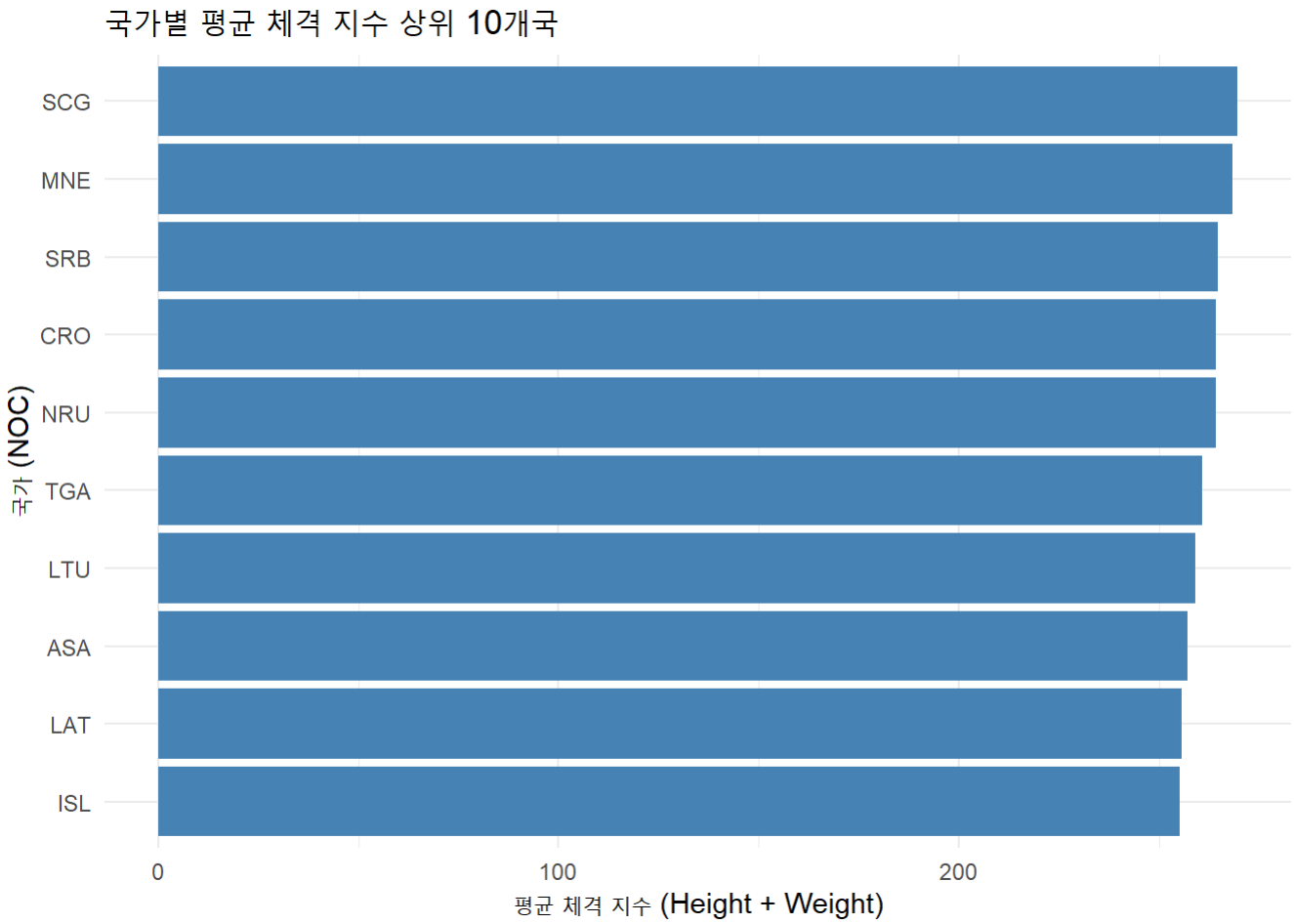
```
# 필요한 패키지 로드
library(dplyr)
library(ggplot2)

# 국가별 키와 몸무게 합산 계산
country_body_size <- data_encoded %>%
  filter(!is.na(Height), !is.na(Weight)) %>% # 결측값 제외
  mutate(
    Body_Size = Height + Weight # 키와 몸무게를 합산한 값
  ) %>%
  group_by(NOC) %>%
  summarize(
    Avg_Body_Size = mean(Body_Size, na.rm = TRUE), # 국가별 평균 체격 지수
    Athlete_Count = n(), # 국가별 선수 수
    .groups = "drop"
  ) %>%
  arrange(desc(Avg_Body_Size)) # 체격 지수 기준 정렬

# 결과 출력
print("국가별 평균 체격 지수:")
```

```
## [1] "국가별 평균 체격 지수:"
```

```
print(country_body_size)
```

```
## # A tibble: 230 × 3
##    NOC   Avg_Body_Size Athlete_Count
##    <chr>         <dbl>         <int>
##  1 SCG            269.           321
##  2 MNE            268.            94
##  3 SRB            265.           392
##  4 CRO            264.           876
##  5 NRU            264.            13
##  6 TGA            261.            46
##  7 LTU            259.           654
##  8 ASA            257.            37
##  9 LAT            256.           951
## 10 ISL            255.           627
## # ℹ 220 more rows
```

```r
# 체격 지수 상위 국가 시각화
top_body_size_countries <- country_body_size %>% slice_max(Avg_Body_Size, n = 10)

ggplot(top_body_size_countries, aes(x = reorder(NOC, Avg_Body_Size), y = Avg_Body_Size)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +
  labs(
    title = "국가별 평균 체격 지수 상위 10개국",
    x = "국가 (NOC)",
    y = "평균 체격 지수 (Height + Weight)"
  ) +
  theme_minimal()
```



해석: 상위 10개국은 SCG, MNE, SRB, CRO, NRU, TGA, LTU, ASA, LAT, ISL로 나타나며, 각국 선수들의 평균 키와 몸무게의 합이 다른 국가들에 비해 높다. 이 중 아이슬란드, 리투아니아, 라트비아는 북유럽과 동유럽에 위치하며 추운 기후의 영향을 받는다. 따라서 가설에서 예측한 이유와 매우 유사하며, 추운 지역에서 체온 유지를 위해 키가 크고 체격이 큰 신체적 특성이 발달했음을 생리학적으로 설명할 수 있다. 반면에 예측과 반대의 결과도 몇 나라에서 보이는데, NRU(나우루)와 TGA(통가)는 남태평양의 섬나라로, 따뜻한 지역이다. 이들의 체격 지수가 높은 이유는 역도, 럭비 등 특정 종목에서 활약하는 체격이 큰 선수들의 영향을 받은 것으로 예측할 수 있다.

BMI 값(몸무게/키*키)으로 국가별 분석

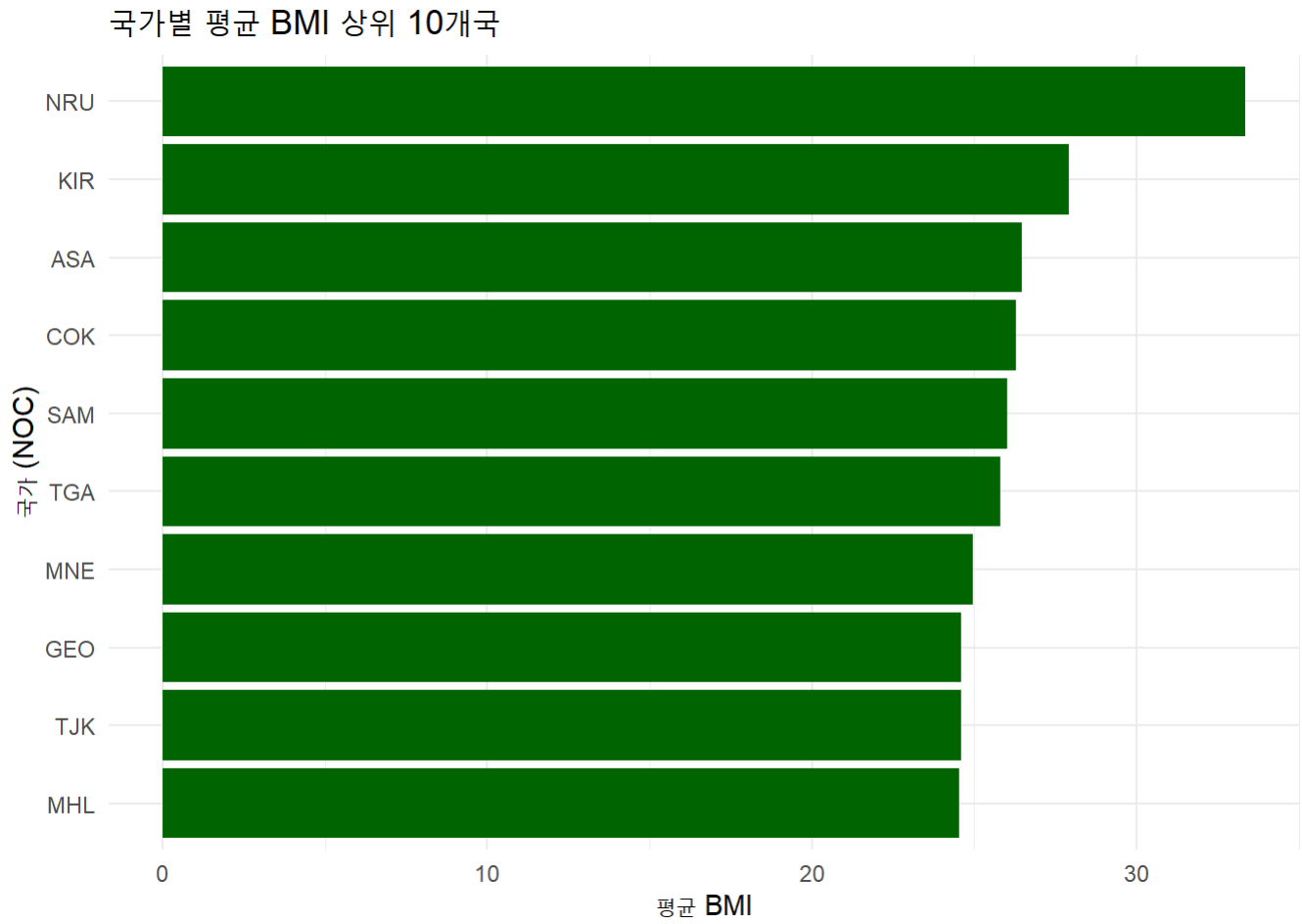
```
# 같은 변수를 사용했으나, 키와 몸무게의 합산으로는 키와 몸무게의 비율, 즉 마른 사람과 체격이 큰 사람을 비교하기는 어려우므로 BMI 수치를 이용해 그래프를 그려보았다.

# 필요한 패키지 로드
library(dplyr)
library(ggplot2)

# 국가별 BMI와 키 대비 몸무게 비율 계산
country_body_stats_extended <- data_encoded %>%
  filter(!is.na(Height), !is.na(Weight)) %>% # 결측값 제외
  mutate(
    Height_m = Height / 100, # 키를 미터로 변환
    BMI = Weight / (Height_m^2), # BMI 계산
  ) %>%
  group_by(NOC) %>%
  summarize(
    Avg_BMI = mean(BMI, na.rm = TRUE), # 평균 BMI
    Athlete_Count = n(), # 국가별 선수 수
    .groups = "drop"
  ) %>%
  arrange(desc(Avg_BMI)) # 평균 BMI 기준 정렬

# 결과 출력
print("국가별 BMI 및 키 대비 몸무게 비율:")
```

```
## [1] "국가별 BMI 및 키 대비 몸무게 비율:"
```

```
print(country_body_stats_extended)
```

```
## # A tibble: 230 × 3
##    NOC   Avg_BMI Athlete_Count
##    <chr>   <dbl>         <int>
##  1 NRU      33.3            13
##  2 KIR      27.9            11
##  3 ASA      26.5            37
##  4 COK      26.3            40
##  5 SAM      26.0            63
##  6 TGA      25.8            46
##  7 MNE      25.0            94
##  8 GEO      24.6           286
##  9 TJK      24.6            70
## 10 MHL      24.5            14
## # ℹ 220 more rows
```

```
# BMI 상위 국가 시각화
top_bmi_countries <- country_body_stats_extended %>% slice_max(Avg_BMI, n = 10)

ggplot(top_bmi_countries, aes(x = reorder(NOC, Avg_BMI), y = Avg_BMI)) +
  geom_bar(stat = "identity", fill = "darkgreen") +
  coord_flip() +
  labs(
    title = "국가별 평균 BMI 상위 10개국",
    x = "국가 (NOC)",
    y = "평균 BMI"
  ) +
  theme_minimal()
```

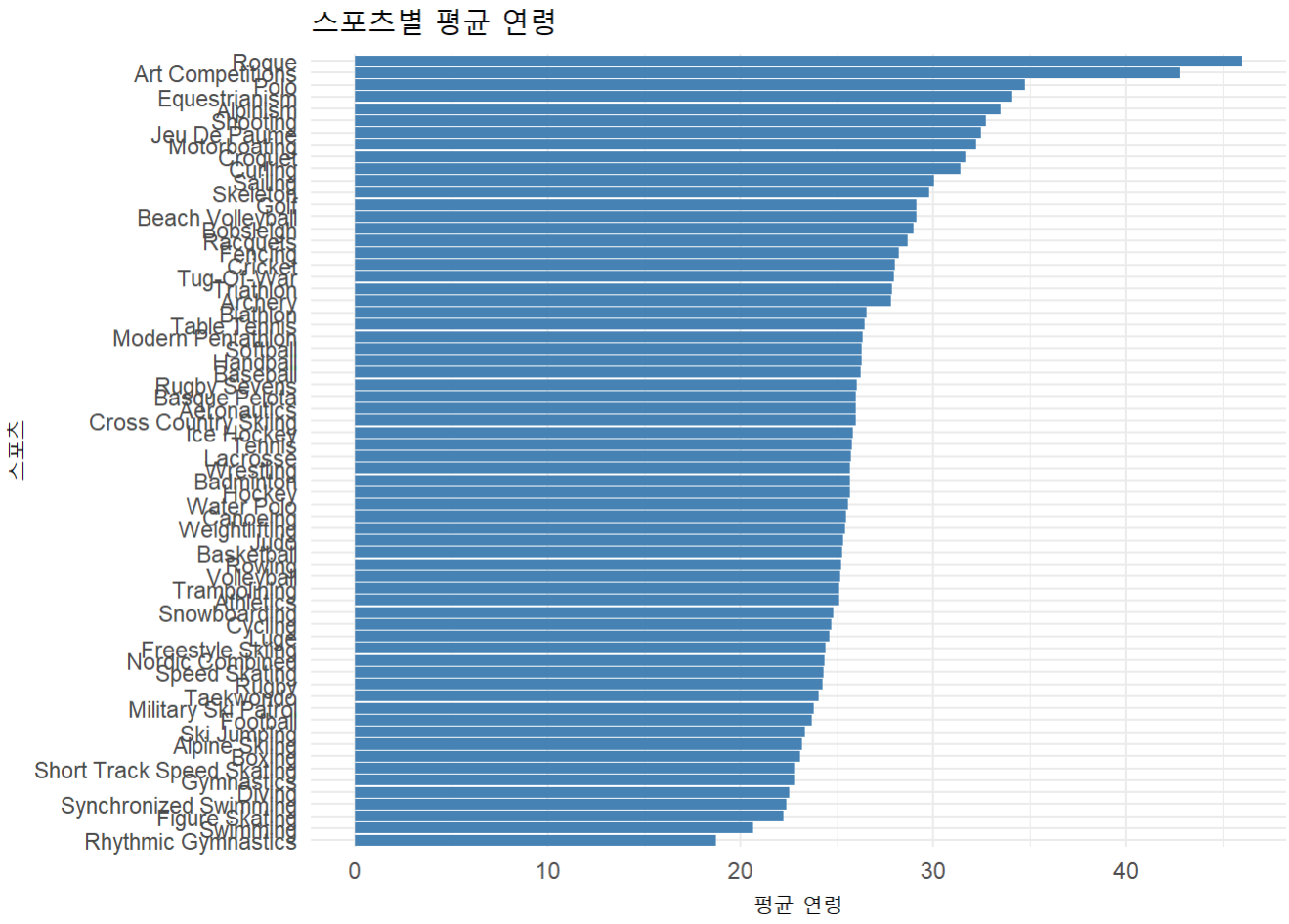


해석: 상위 10개국으로 NRU (나우루), KIR (키리바시), ASA (아메리칸 사모아), COK (쿡 제도), SAM (사모아), TGA (통가), MNE (몬테네그로), GEO (조지아), TJK (타지키스탄), MHL (마셜 제도)가 선정되었다. 상위 10개국 중 NRU(나우루), KIR(키리바시), ASA(아메리칸 사모아), COK(쿡 제도), SAM(사모아), TGA(통가)가 태평양 섬 국가이므로, 가설과 일치하는 결과가 나왔음을 알 수 있다. 예외적으로 MHL (마셜 제도), COK (쿡 제도) 등은 태평양 섬 국가가 아니며 추운 지역 국가가 아니기 때문에, 활동하는 선수 수가 적어서 특정 선수의 BMI가 국가 평균에 큰 영향을 미쳤을 것이라고 예상해볼 수 있다.

연령과 스포츠 관계

가설: 골프와 같은 비교적 활동 강도가 낮은 스포츠에서는 선수들이 경력을 쌓으며 기술적인 완성도를 높일 수 있어 연령대가 높은 경향을 보일 것이다. 반면, 활동적이고 유연성이 중요한 스포츠는 신체적 전성기에 있는 젊은 연령대의 선수들이 두각을 나타낼 가능성이 크다. 따라서 종목별 특성에 따라 선수들의 연령 분포는 다르게 나타날 것이다.

```r
# 필요한 패키지 로드
library(dplyr)
library(ggplot2)

# 스포츠별 평균 연령 계산
age_summary <- data_encoded %>%
  filter(!is.na(Age)) %>% # 연령 데이터가 존재하는 경우만 선택
  group_by(Sport) %>%
  summarize(
    Mean_Age = mean(Age, na.rm = TRUE),
    Min_Age = min(Age, na.rm = TRUE),
    Max_Age = max(Age, na.rm = TRUE),
    .groups = "drop"
  ) %>%
  arrange(Mean_Age) # 평균 연령 기준으로 정렬

# 평균 연령 시각화
ggplot(age_summary, aes(x = reorder(Sport, Mean_Age), y = Mean_Age)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +
  labs(
    title = "스포츠별 평균 연령",
    x = "스포츠",
    y = "평균 연령"
  ) +
  theme_minimal()
```



```
# 가장 어린 선수와 나이 많은 선수가 많은 종목 확인
youngest_sports <- age_summary %>% slice_head(n = 5)
oldest_sports <- age_summary %>% slice_tail(n = 5)

# 결과 출력
print("평균 연령이 낮은 종목 (Top 5):")
```

```
## [1] "평균 연령이 낮은 종목 (Top 5):"
```

```
print(youngest_sports)
```

```
## # A tibble: 5 × 4
##   Sport                Mean_Age Min_Age Max_Age
##   <chr>                   <dbl>   <dbl>   <dbl>
## 1 Rhythmic Gymnastics      18.7      13      30
## 2 Swimming                 20.6      11      46
## 3 Figure Skating           22.3      11      52
## 4 Synchronized Swimming    22.4      15      40
## 5 Diving                   22.5      12      51
```

```
print("평균 연령이 높은 종목 (Top 5):")
```

```
## [1] "평균 연령이 높은 종목 (Top 5):"
```

```
print(oldest_sports)
```

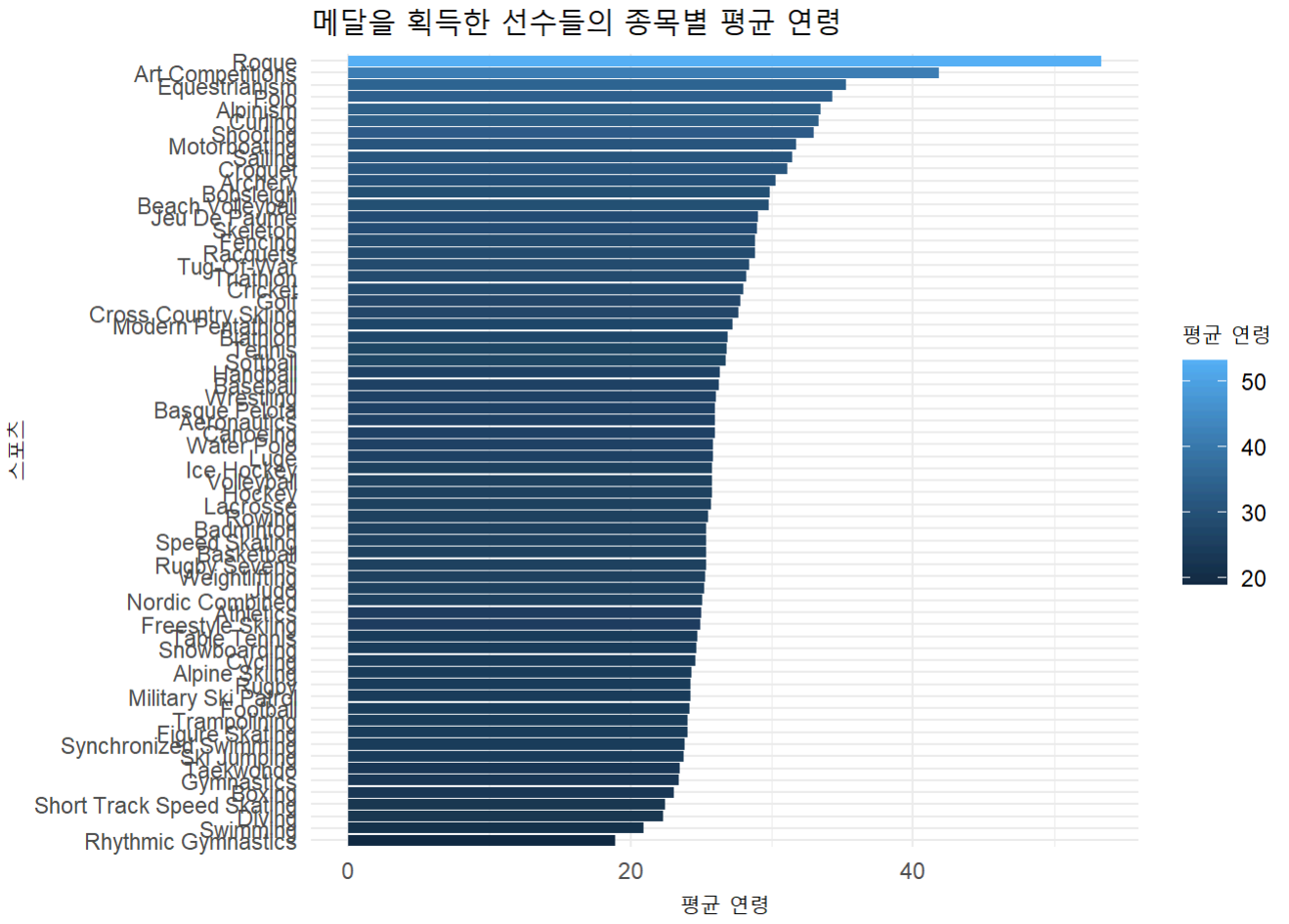
```
## # A tibble: 5 × 4
##   Sport            Mean_Age Min_Age Max_Age
##   <chr>               <dbl>   <dbl>   <dbl>
## 1 Alpinism             33.5      22      57
## 2 Equestrianism        34.1      16      72
## 3 Polo                 34.7      21      53
## 4 Art Competitions     42.8      14      97
## 5 Roque                46        24      64
```

해석: 스포츠 별 평균 연령이 가장 높은 종목은 Roque이며, 평균 연령이 45세 이상인 것을 알 수 있다. Roque는 골프와 비슷한 형태의 종목이기 때문에, 처음 가설을 세웠던 것과 같이 평균 연령이 높다는 것을 알 수 있다. 반면에 Rhythmic Gymnastics, 즉 리듬체조 종목에서 평균 연령이 20살 미만으로 가장 낮다고 나타났다. 이는 리본이나 곤봉, 공 등을 이용해 온 몸을 활용하는 활동적인 운동이며, 유연성을 요구하기 때문에 평균 연령이 매우 낮음을 짐작할 수 있다.

스포츠별 메달 획득 평균 연령 분석

가설: 메달을 획득한 선수의 평균 연령은 종목에 따라 다를 것이다.

```
# 메달을 획득한 선수들만 필터링 (NA 제외)
df_medals <- data_clean %>% filter(Medal != "NA")

# 종목별로 메달을 획득한 선수들의 평균 연령 계산
avg_age_by_sport <- df_medals %>%
  group_by(Sport) %>%
  summarize(Average_Age = mean(Age, na.rm = TRUE)) %>%
  arrange(desc(Average_Age))  # 평균 연령이 높은 순으로 정렬

# 시각화: 종목별 평균 연령
ggplot(avg_age_by_sport, aes(y = reorder(Sport, Average_Age), x = Average_Age, fill = Average_A
ge)) +
  geom_bar(stat = "identity") +  # 막대그래프
  labs(title = "메달을 획득한 선수들의 종목별 평균 연령",
       x = "평균 연령",
       y = "스포츠",
       fill = "평균 연령") +
  theme_minimal() +
  theme(axis.text.y = element_text(angle = 0, hjust = 1))
```



해석 : 설정한 가설대로, 메달을 획득한 선수의 평균 연령은 종목에 따라 차이가 나는 것을 확인할 수 있다. 이 그래프는 앞서 살펴본 연령과 스포츠 간 관계를 보여주는 그래프와 유사하게 나타나며, 이를 통해 종목별 참가 연령과 메달 획득 연령이 비슷한 분포를 보인다고 해석할 수 있다.

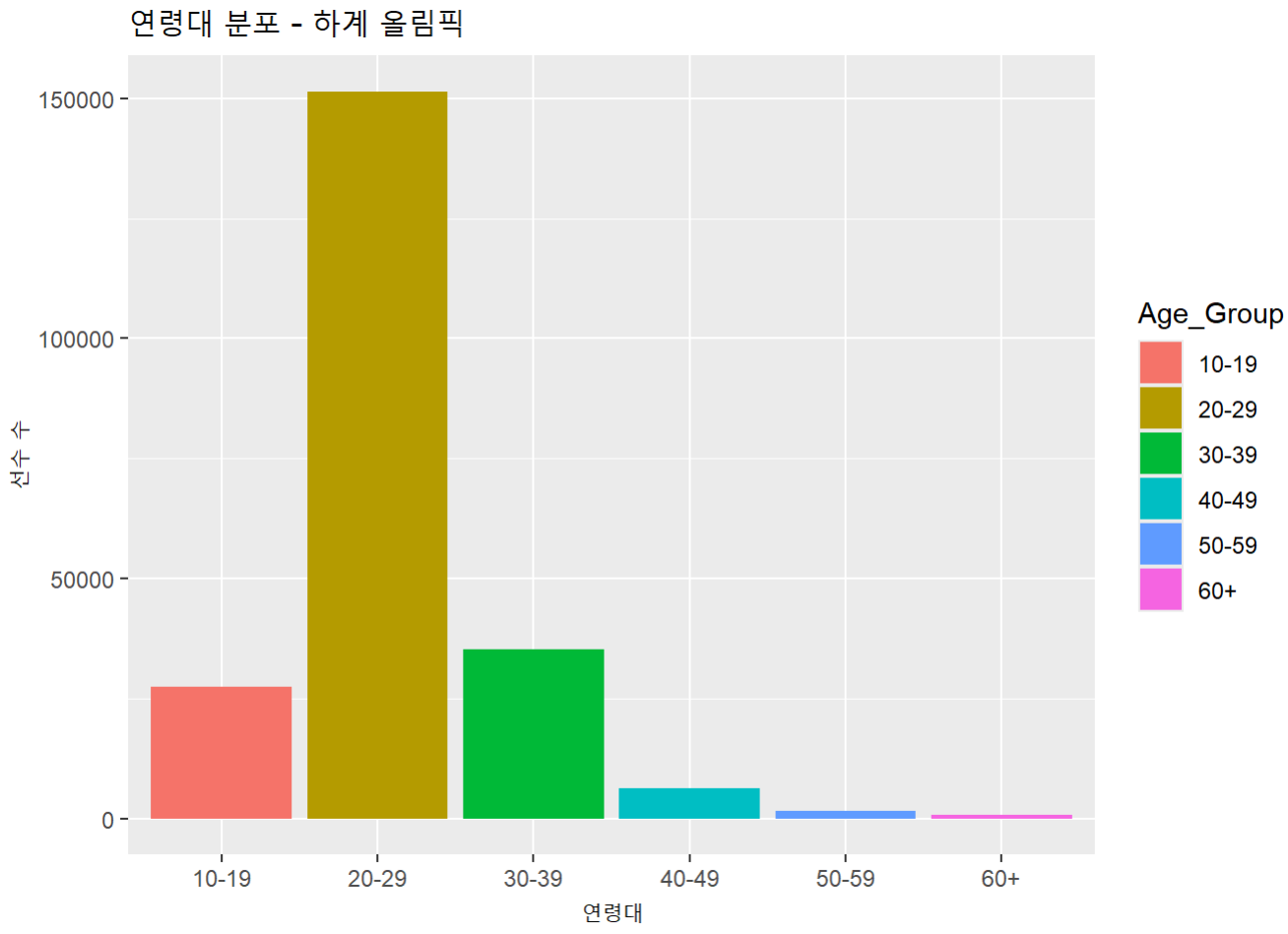
올림픽 시즌에 따른 선수 연령대 분포

가설 : 올림픽 계절에 상관없이 20대 연령층의 선수들의 분포가 가장 많을 것이다.

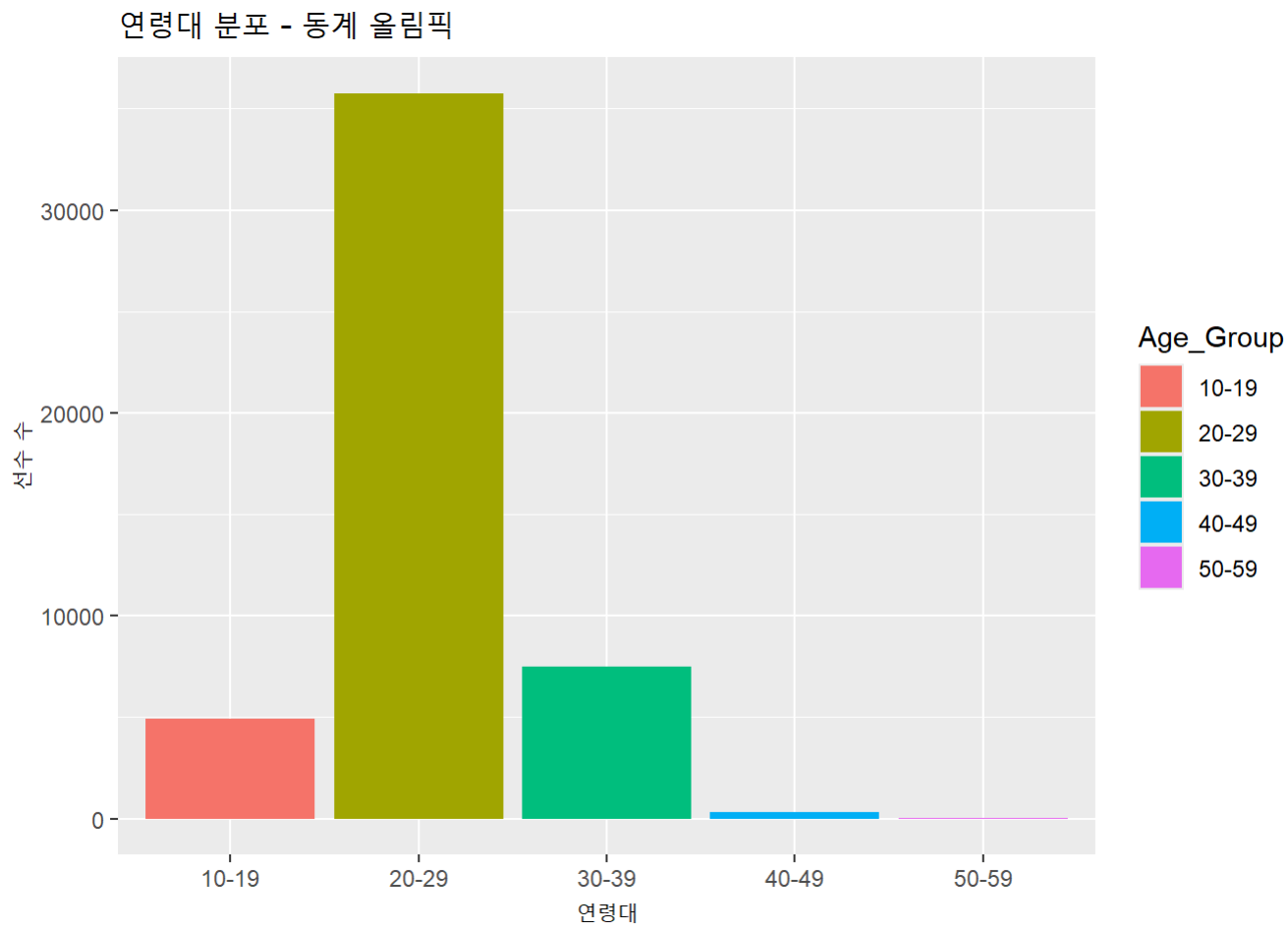
```r
# 연령대 생성
df <- data_clean %>%
  mutate(Age_Group = cut(Age,
                         breaks = c(0, 10, 20, 30, 40, 50, 60, Inf),
                         labels = c("0-9", "10-19", "20-29", "30-39", "40-49", "50-59", "60+"),
                         right = FALSE))

# 데이터 집계
age_dist <- df %>%
  group_by(Season, Age_Group) %>%
  summarise(Count = n())
```

```
## `summarise()` has grouped output by 'Season'. You can override using the
## `.groups` argument.
```

```r
# 하계 올림픽 시각화
summer_data <- age_dist %>% filter(Season == "Summer")
ggplot(summer_data, aes(x = Age_Group, y = Count, fill = Age_Group)) +
  geom_bar(stat = "identity") +
  labs(title = "연령대 분포 - 하계 올림픽",
       x = "연령대",
       y = "선수 수")
```



```
# 동계 올림픽 시각화
winter_data <- age_dist %>% filter(Season == "Winter")
ggplot(winter_data, aes(x = Age_Group, y = Count, fill = Age_Group)) +
  geom_bar(stat = "identity") +
  labs(title = "연령대 분포 - 동계 올림픽",
       x = "연령대",
       y = "선수 수")
```



해석 : 올림픽 참가 연령 분포를 살펴보면 예상대로 20대 선수들이 가장 큰 비중을 차지하는 것을 확인할 수 있다. 흥미롭게도 만 16세 이상이라는 참가 연령 제한이 있음에도 불구하고 10대 선수들의 참가율이 적지 않게 나타났다. 반면, 50~60대 이상의 고령 선수들의 참가율은 매우 낮으며, 특히 동계 올림픽에서는 60대 이상의 참가자가 없는 것을 확인할 수 있는데 이를 통해 동계 스포츠의 선수 수명이 하계 스포츠보다 상대적으로 짧을 수 있다는 점을 추측할 수 있다.

올림픽 시즌에 따른 성별 메달 획득 비율 분석

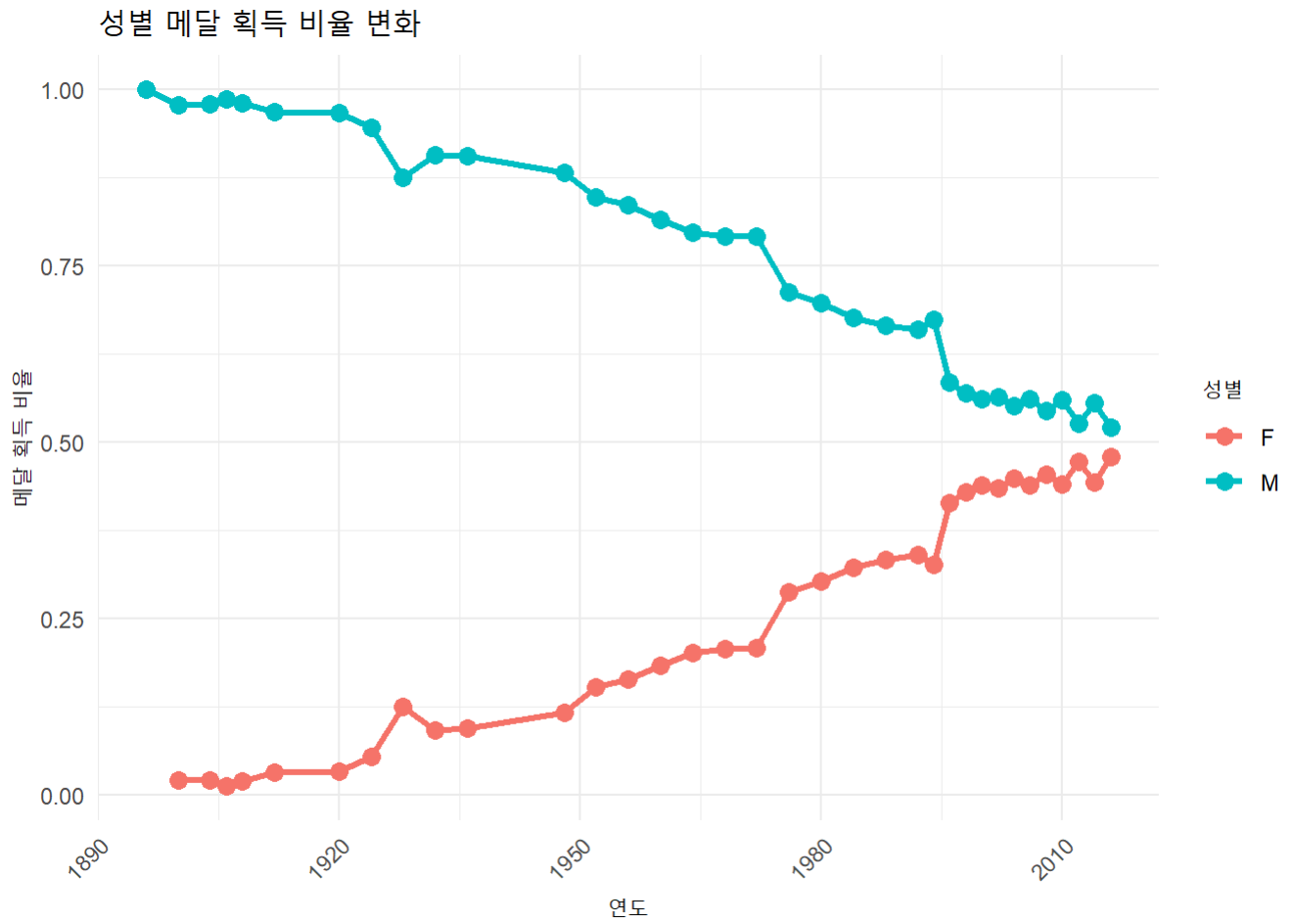
가설 : 남성의 메달 획득 비율이 여성보다 높을 것이다.

```
# 성별 메달 획득 비율 계산
medal_ratio_data <- data %>%
  filter(!is.na(Medal)) %>%
  group_by(Year, Sex) %>%
  summarise(Medal_Count = n()) %>%
  group_by(Year) %>%
  mutate(Total_Medals = sum(Medal_Count),
         Medal_Ratio = Medal_Count / Total_Medals) %>%
  arrange(Year, Sex)
```

```
## `summarise()` has grouped output by 'Year'. You can override using the
## `.groups` argument.
```

```
# 성별 메달 획득 비율 변화 시각화
ggplot(medal_ratio_data, aes(x=Year, y=Medal_Ratio, color=Sex, group=Sex)) +
  geom_line(size=1.2) +
  geom_point(size=3) +
  labs(title="성별 메달 획득 비율 변화",
       x="연도",
       y="메달 획득 비율",
       color="성별") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle=45, hjust=1))
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



해석 : 초기인 1890년대부터 1920년대까지는 남성의 메달 획득 비율이 거의 100%로, 초창기 올림픽에서는 여성의 참여가 사실상 불가능했음을 보여준다. 이후 1920년대부터 1980년대까지 여성의 메달 획득 비율이 점진적으로 증가했으며, 특히 1980년대에 들어서 급격히 상승하는 경향이 나타난다. 이는 여성 스포츠가 활성화되고, 올림픽에 더 많은 여성 종목이 추가된 결과로 해석할 수 있다. 1980년대 이후 현대에는 남성과 여성의 메달 획득 비율이 점차 비슷해지면서 최근에는 거의 50:50에 가까워졌다. 이러한 변화는 시간이 지남에 따라 스포츠 대회의 성평등이 크게 개선되었음을 보여준다.

국가별 메달 획득 경향을 하계, 동계 올림픽별로 분석

가설 : 특정 기후를 가진 국가별로 하계 올림픽과 동계 올림픽 중 특정 계절에서 메달을 더 많이 획득하는 경향이 있다.

```
# 필요한 열만 선택
filtered_data <- data_encoded %>%
  select(Team, Season, Sport, Medal) %>%
  filter(Medal!=0) # 메달 데이터가 있는 행만 사용
```

```
# 국가별 하계 및 동계 메달 획득 수 요약
medal_summary <- filtered_data %>%
  group_by(Team, Season) %>%
  summarise(Total_Medals = n()) %>%
  spread(key = Season, value = Total_Medals, fill = 0)
```

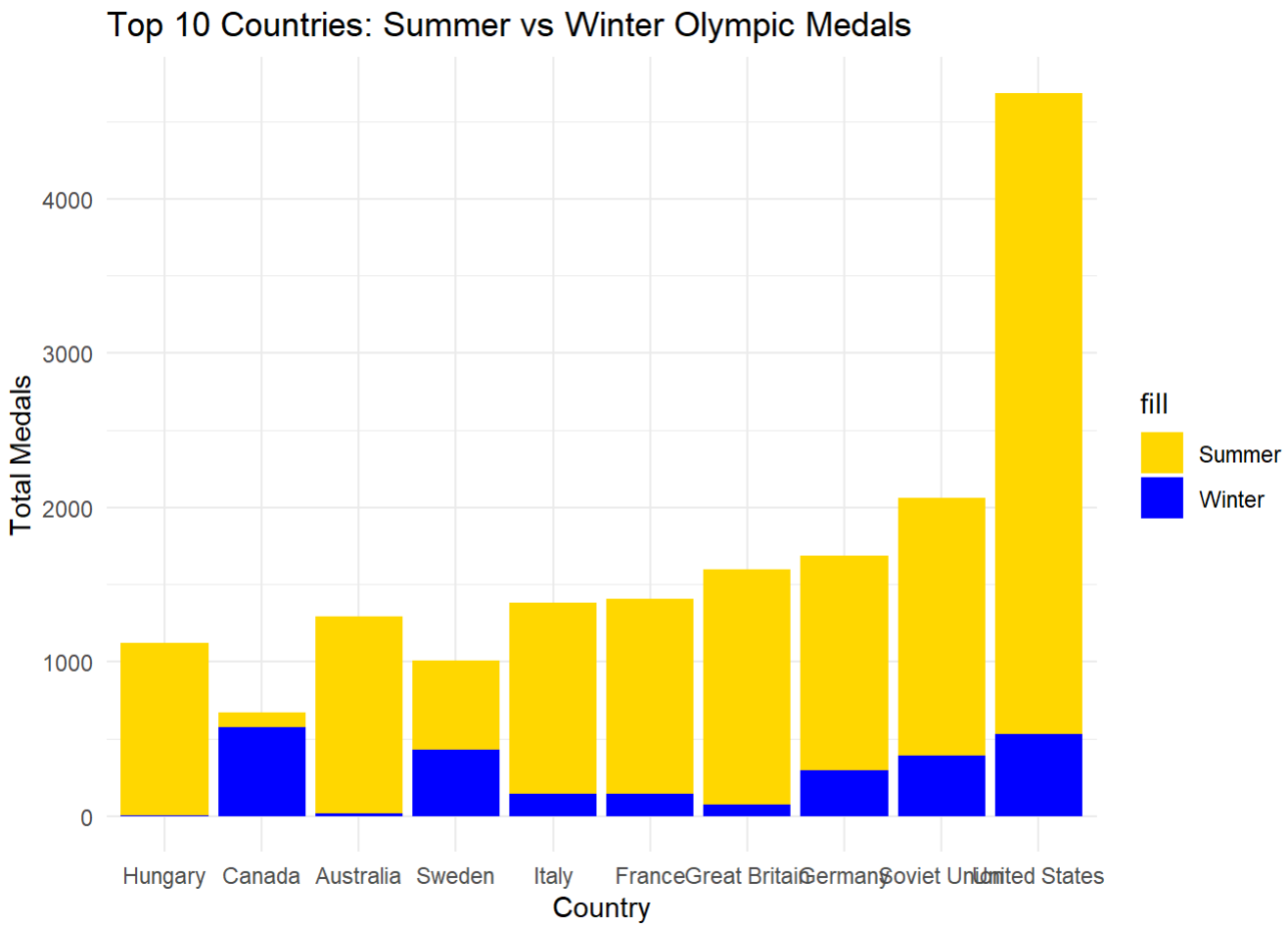
```
## `summarise()` has grouped output by 'Team'. You can override using the
## `.groups` argument.
```

```
# 결과 확인
head(medal_summary)
```

Team <chr>	**Summer** <dbl>	**Winter** <dbl>
A North American Team	4	0
Afghanistan	2	0
Algeria	17	0
Ali-Baba II	5	0
Amateur Athletic Association	5	0
Amstel Amsterdam	4	0

6 rows

```
# 하계와 동계 메달 비교 (상위 10개 국가)
top10_countries <- medal_summary %>%
  arrange(desc(Summer + Winter)) %>%
  head(10)

ggplot(top10_countries, aes(x = reorder(Team, Summer + Winter))) +
  geom_bar(aes(y = Summer, fill = "Summer"), stat = "identity", position = "dodge") +
  geom_bar(aes(y = Winter, fill = "Winter"), stat = "identity", position = "dodge") +
  labs(title = "Top 10 Countries: Summer vs Winter Olympic Medals",
       x = "Country", y = "Total Medals") +
  theme_minimal() +
  scale_fill_manual(values = c("Summer" = "gold", "Winter" = "blue"))
```



Dataset 내에서 하계 스포츠가 동계 스포츠보다 상대적으로 많이 존재하고 있어 하계와 동계를 구분해서 알아보기가 힘들다고 판단된다. 따라서 하계 스포츠와 동계 스포츠를 각각 구분해서 알아보겠다.

```
# 하계와 동계 데이터 분리
summer_data <- filtered_data %>% filter(Season == "Summer")
winter_data <- filtered_data %>% filter(Season == "Winter")
```

하계 분석: 국가별 메달 요약

```
# 하계 국가별 메달 수 요약
summer_medals <- summer_data %>%
  group_by(Team) %>%
  summarise(Total_Medals = n()) %>%
  arrange(desc(Total_Medals))

# 상위 10개 국가 확인
head(summer_medals, 10)
```

Team <chr>	Total_Medals <int>
United States	4686
Soviet Union	2061
Germany	1687
Great Britain	1598
France	1408

Team <chr>	Total_Medals <int>
Italy	1384
Australia	1290
Hungary	1123
Sweden	1006
Russia	894

1-10 of 10 rows

동계 분석: 국가별 메달 요약

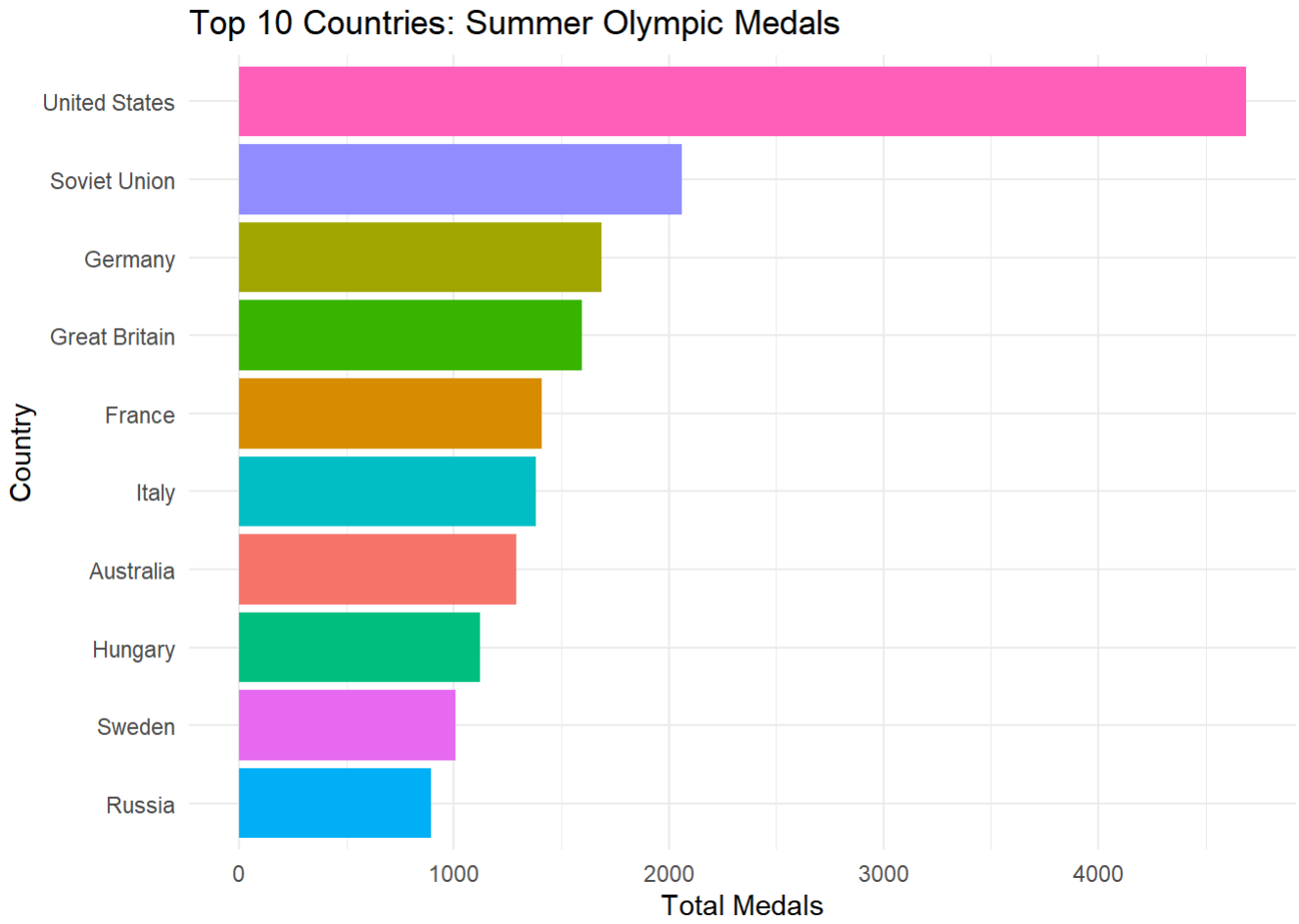
```
# 동계 국가별 메달 수 요약
winter_medals <- winter_data %>%
  group_by(Team) %>%
  summarise(Total_Medals = n()) %>%
  arrange(desc(Total_Medals))

# 상위 10개 국가 확인
head(winter_medals, 10)
```

Team <chr>	Total_Medals <int>
Canada	575
United States	533
Norway	443
Sweden	428
Finland	426
Soviet Union	390
Germany	297
Austria	244
Russia	216
Switzerland	183

1-10 of 10 rows

하계 올림픽

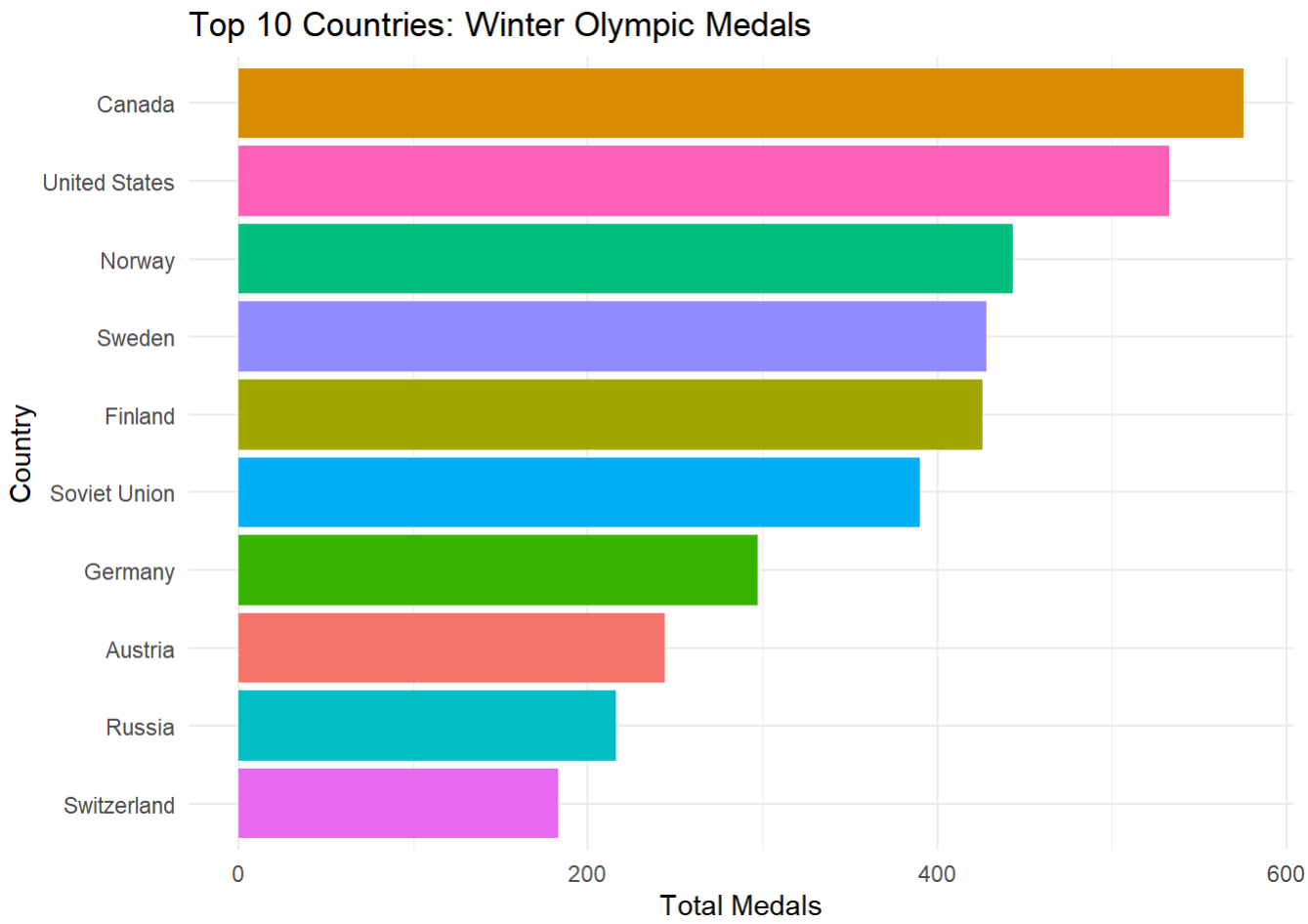
```
# 상위 10개 국가 시각화 (하계)
top_summer <- summer_medals %>% head(10)

ggplot(top_summer, aes(x = reorder(Team, Total_Medals), y = Total_Medals, fill = Team)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Top 10 Countries: Summer Olympic Medals",
       x = "Country", y = "Total Medals") +
  theme_minimal() +
  theme(legend.position = "none")
```



United States의 하계 올림픽에서 Medal 획득 수가 다른 상위국가에 비해서 현저히 높게 나타나는 것을 확인할 수 있다.

동계 올림픽

```
# 상위 10개 국가 시각화 (동계)
top_winter <- winter_medals %>% head(10)

ggplot(top_winter, aes(x = reorder(Team, Total_Medals), y = Total_Medals, fill = Team)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Top 10 Countries: Winter Olympic Medals",
       x = "Country", y = "Total Medals") +
  theme_minimal() +
  theme(legend.position = "none")
```



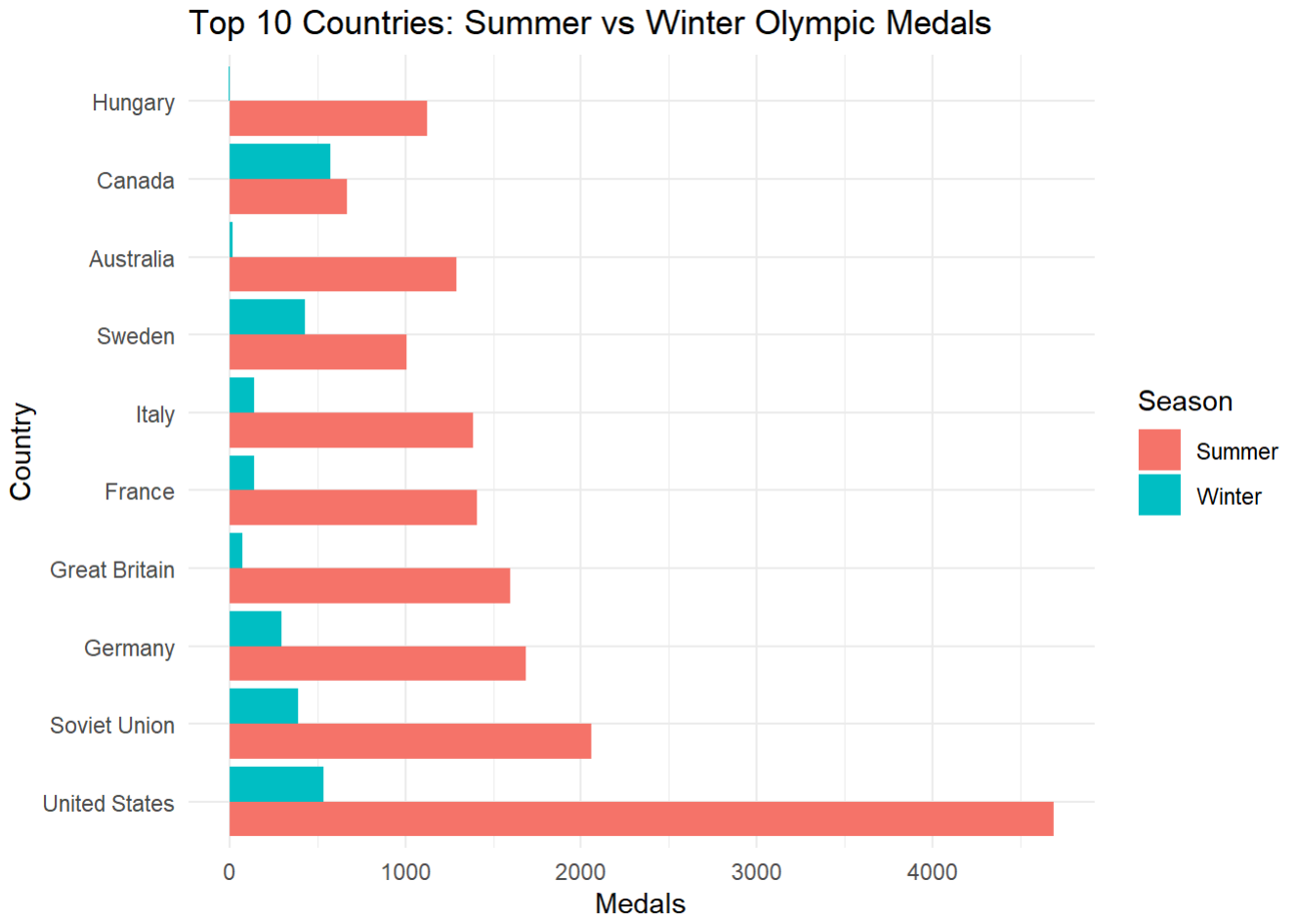
동계 올림픽에서는 United States와 Canada의 Medal 획득 수가 엄청나게 높게 나타나는데, 위의 하계 올림픽 Medal 획득 수에 비해서 Canada가 동계에서 더 많은 성과를 보였음을 확인할 수 있다. 그리고 상위 국가도 하계 올림픽 그래프와 다르게 나타나는데, 상위 국가 중 Norway, Sweden, Finland 등이 상위권에 자리 잡고 있음을 알 수 있다.

위의 하계 올림픽 그래프와 동계 올림픽 그래프를 한 눈에 알아보기 위해 각 계절별로 비율로 나타내서 다시 하계, 동계 올림픽의 데이터를 시각화해보겠다.

```
# 하계와 동계 데이터를 결합
summer_medals$Season <- "Summer"
winter_medals$Season <- "Winter"

combined_medals <- bind_rows(
  summer_medals %>% rename(Medals = Total_Medals),
  winter_medals %>% rename(Medals = Total_Medals)
)

# 상위 국가들만 필터링 (하계와 동계를 포함한 상위 10개 국가)
top_countries <- combined_medals %>%
  group_by(Team) %>%
  summarise(Total_Medals = sum(Medals)) %>%
  arrange(desc(Total_Medals)) %>%
  head(10) %>%
  pull(Team)

filtered_combined <- combined_medals %>%
  filter(Team %in% top_countries)
```

하계와 동계를 하나의 그래프로 나타내기

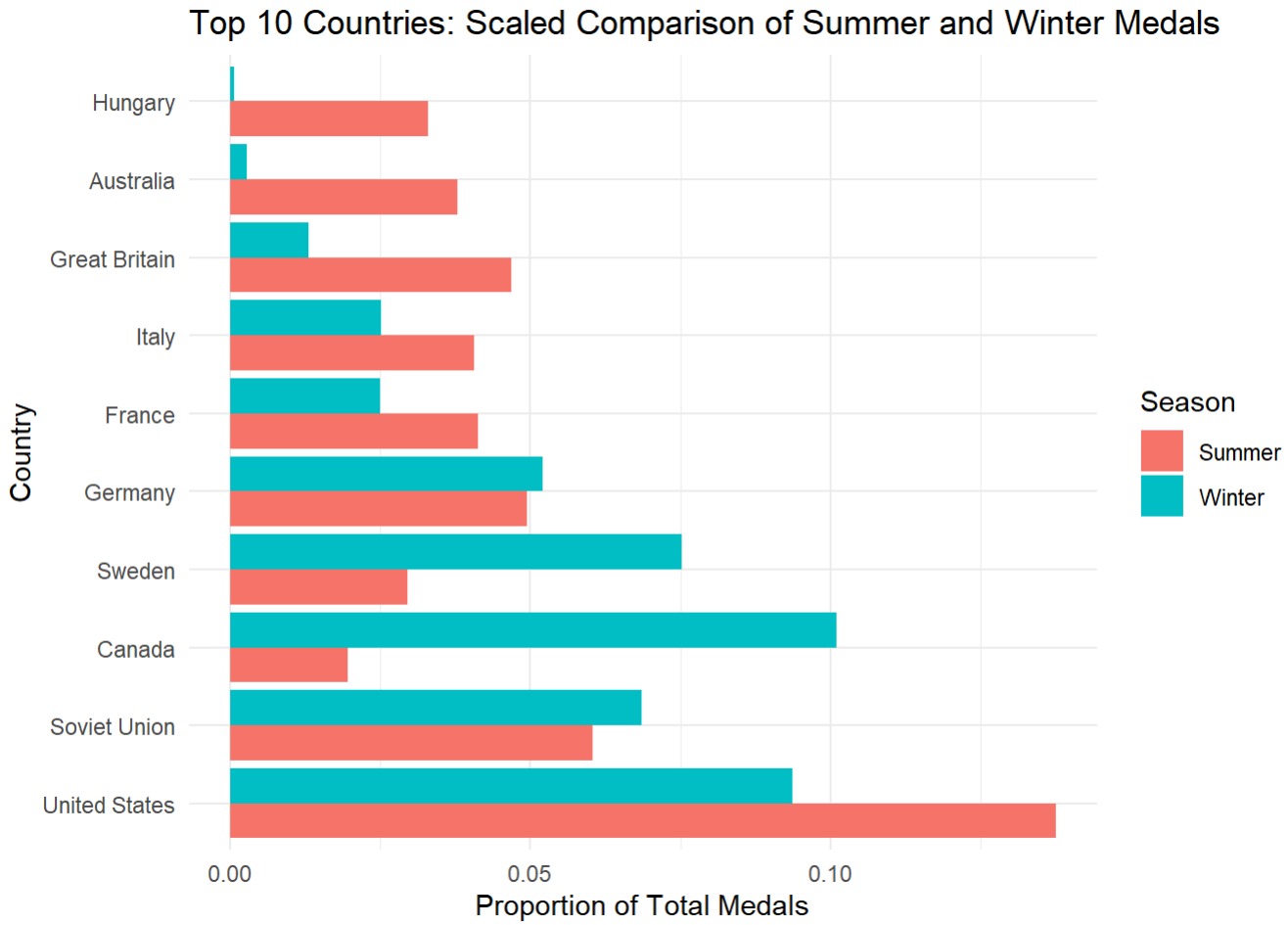
```
# 시각화
ggplot(filtered_combined, aes(x = reorder(Team, -Medals), y = Medals, fill = Season)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Top 10 Countries: Summer vs Winter Olympic Medals",
       x = "Country", y = "Medals",
       fill = "Season") +
  theme_minimal() +
  coord_flip()
```



전체적으로 하계 스포츠의 Medal 수가 훨씬 많이 존재하는 것을 확인할 수 있는데, 이는 위 하계 스포츠와 동계 스포츠 관련 EDA에서 이유를 확인할 수 있다. 더 세부적으로 알아보기 위해 x 범위를 각 Season별로 비율로 스케일링하여 비율로 나타내어 분석하겠다.

하계와 동계 메달 획득 비율로 표준화(x축 스케일링)

```
# 하계와 동계 각각 메달 수를 전체 비율로 변환
scaled_combined <- combined_medals %>%
  group_by(Season) %>%
  mutate(Scaled_Medals = Medals / sum(Medals))

# 시각화
ggplot(scaled_combined %>% filter(Team %in% top_countries),
       aes(x = reorder(Team, -Scaled_Medals), y = Scaled_Medals, fill = Season)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Top 10 Countries: Scaled Comparison of Summer and Winter Medals",
       x = "Country", y = "Proportion of Total Medals",
       fill = "Season") +
  theme_minimal() +
  coord_flip()
```



United States는 하계, 동계 모두 Medal을 많이 획득하였음을 확인할 수 있는데, 그 중에서도 동계보다는 하계 올림픽에서 획득한 Medal의 비율이 더 높음을 확인할 수 있다. 이는 기후의 영향을 받을 수 있는데, 미국은 다양한 기후를 가진 국가이지만, 동계 스포츠를 위한 환경(특히 눈과 얼음)은 북부 지역에 한정된다. 반면에, 하계 스포츠는 전역에서 쉽게 접근 가능하다는 이유로 접근할 수 있다. 또한 미국 전역에는 하계 스포츠를 위한 훈련 시설이 더 많이 분포하고 있기에 하계 스포츠의 Medal 획득 비율이 동계 스포츠보다 더 유리함을 알아낼 수 있다.

Canada도 United States의 북부 지역과 맞닿아있는 지역이다. 이에 따라 겨울에 눈과 얼음이 풍부하여 스키, 스노보드, 아이스하키, 스피드 스케이팅 등 동계 스포츠에 적합한 환경을 제공하기에, 이런 자연적 특성 덕분에 쉽게 동계 스포츠에 접할 수 있다. 또한 전국적으로 동계 스포츠 훈련 시설이 잘 갖춰져 있기에 동계 스포츠의 Medal 획득 비율이 하계 올림픽에 비해서 높은 이유를 알아볼 수 있다.

Sweden도 Canada와 유사한 이유로 동계 스포츠의 Medal 획득 비율이 하계 스포츠의 3배정도로 훨씬 높은 것을 알아볼 수 있다.

하계 올림픽의 Medal 획득 비율이 상대적으로 엄청 높게 나타나는 국가들은 Hungary, Australia이다.

Hungary에서 하계 올림픽의 Medal 획득 비율이 동계에 비해서 굉장히 크게 나타나고 있다. 여기서는 위의 United States와 Canada에서 본 기후와 같은 큰 이유는 보이지 않고, Hungary의 하계 스포츠에 집중적인 자원 투입과 지원이 이루어진 결과로 알아볼 수 있다.

Australia는 대부분의 지역에서 온화하고 더운 날씨가 지속된된다. 여름에는 평균 기온이 25도에서 35도 사이로 올라가며, 이는 하계 스포츠에 매우 적합한 환경을 제공한다. 수영, 육상, 사이클링, 테니스 등 하계 올림픽의 많은 종목들이 고온의 날씨에서 훈련할 수 있기 때문에, 오스트레일리아 선수들은 자연스럽게 하계 종목에서 경쟁력이 높기에, 하계 올림픽의 Medal 획득 비율이 동계에 비해 높게 나타나는 이유를 알아볼 수 있다.

<최종 해석>

1. 하계 올림픽 vs 동계 올림픽

- 하계 스포츠는 대체로 동계 스포츠에 비해 Sport(참가 종목)의 수와 Medal 획득 수가 더 많다.(이는 위 EDA에서 확인할 수 있다.) 데이터에서 하계 스포츠 Medal 획득 비율이 높은 국가는 미국(United States), 헝가리(Hungary), 호주(Australia) 등이며, 이는 기후와 훈련 환경의 영향을 받는다.

2. 국가별 하계 올림픽, 동계 올림픽 Medal 획득 비율

[하계 스포츠 강국]

2-1. 미국(United States) - 동계 스포츠보다 하계 스포츠에서 훨씬 많은 메달 획득했다. 이는 동계 스포츠를 위한 적합한 환경을 가지는 북부지역보다 미국의 전역에서 쉽게 접근 가능한 하계 스포츠 인프라 덕분인 것으로 예상할 수 있다.

2-2. 호주(Australia) - 온화하고 더운 기후의 영향으로 하계 스포츠에서 강세를 보인다. 따라서 동계 스포츠는 상대적으로 참여와 인프라가 적음을 예상해볼 수 있다.

2-4. 헝가리(Hungary) - 하계 스포츠에서 압도적으로 Medal 획득 비율이 높다. 하지만 헝가리에서는 기후가 큰 영향을 끼치지는 않으며, 헝가리의 하계 스포츠에 대한 국가적 지원과 스포츠 문화 때문인 것을 추측할 수 있다.

[동계 스포츠 강국]

2-5. 캐나다(Canada) - 동계 스포츠에 적합한 자연 환경(눈과 얼음)과 훈련 시설로 인해 동계 스포츠에서 강세하다.

2-6. 스웨덴(Sweden) - 동계 올림픽의 메달 획득 비율이 높으며, 하계 스포츠보다 약 3배 높은 비율로 나타났다.

3. 시각화를 통한 결과

- 상위 10개 국가를 하계와 동계로 비교하면 Medal의 수는 하계 스포츠에서의 메달 획득이 훨씬 많으나, 동계 스포츠 강국의 Medal 획득 비율은 해당 국가들에서 기후별로 특화된 스포츠에 따라 높게 나타나는 것을 확인할 수 있다.

가설은 데이터 분석을 통해 유효함을 확인할 수 있다.

국가별 기후가 '특정 국가'에서는 하계와 동계 올림픽에서의 성과 차이를 크게 좌우한다고 볼 수 있다. 동계 스포츠는 특정 자연적 환경 조건을 필요로 하기 때문에 환경을 만족하는 소수의 국가에서는 강세를 보이는 반면, 하계 스포츠는 상대적으로 접근성이 높아 더 많은 국가가 성과를 거두는 경향이 있음을 확인할 수 있다.