Olympic Part5 - Season

1829008 김민영

```
library(dplyr)
```

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

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

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

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------
------------------------------------ tidyverse 1.2.1 --
```

```
## √ ggplot2 3.2.1     √ readr   1.3.1
## √ tibble  2.1.3     √ purrr   0.3.2
## √ tidyr   0.8.3     √ stringr 1.4.0
## √ ggplot2 3.2.1     √ forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------
------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

```
library(scales)
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```r
library(RColorBrewer)
olympic <- read_csv("C:/Temp/athlete_events.csv")
```

```
## Parsed with column specification:
## cols(
##   ID = col_double(),
##   Name = col_character(),
##   Sex = col_character(),
##   Age = col_double(),
##   Height = col_double(),
##   Weight = col_double(),
##   Team = col_character(),
##   NOC = col_character(),
##   Games = col_character(),
##   Year = col_double(),
##   Season = col_character(),
##   City = col_character(),
##   Sport = col_character(),
##   Event = col_character(),
##   Medal = col_character()
## )
```

```r
as_tibble(olympic)
```

```
## # A tibble: 271,116 x 15
##       ID Name  Sex     Age Height Weight Team  NOC   Games  Year Season
##    <dbl> <chr> <chr> <dbl>  <dbl>  <dbl> <chr> <chr> <chr> <dbl> <chr>
##  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 <chr>
```

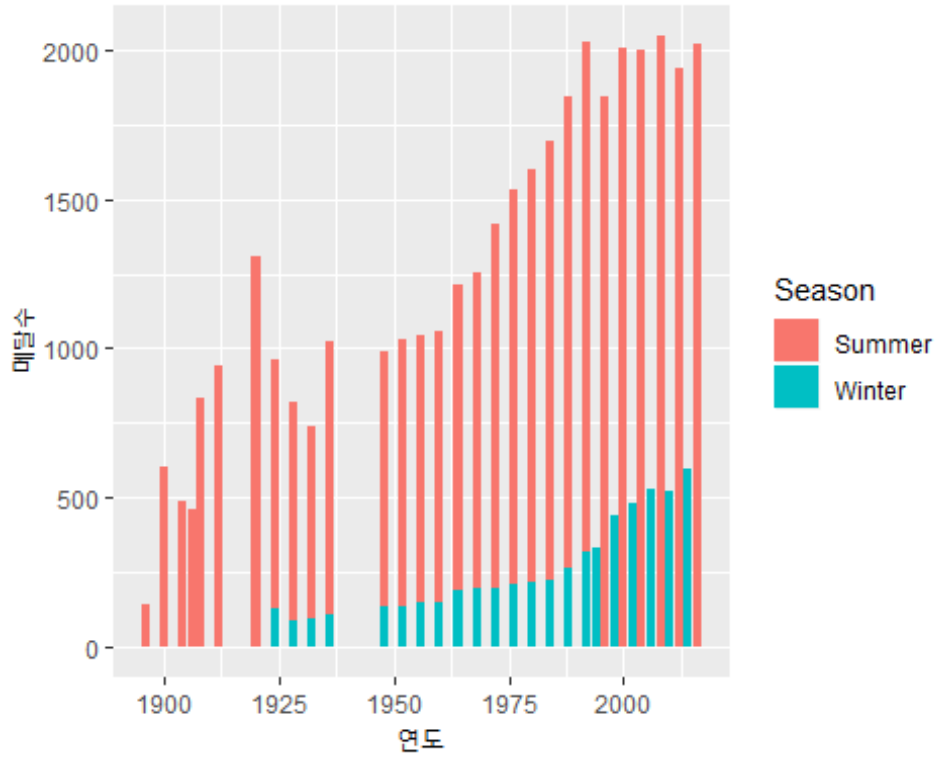
```r
olympic$Season%>%
  summary
```

```
##    Length     Class      Mode
##    271116 character character
```

1. 하계올림픽, 동계올림픽 규모 비교

a) 연도에 따른 메달 수 하계/동계 비교

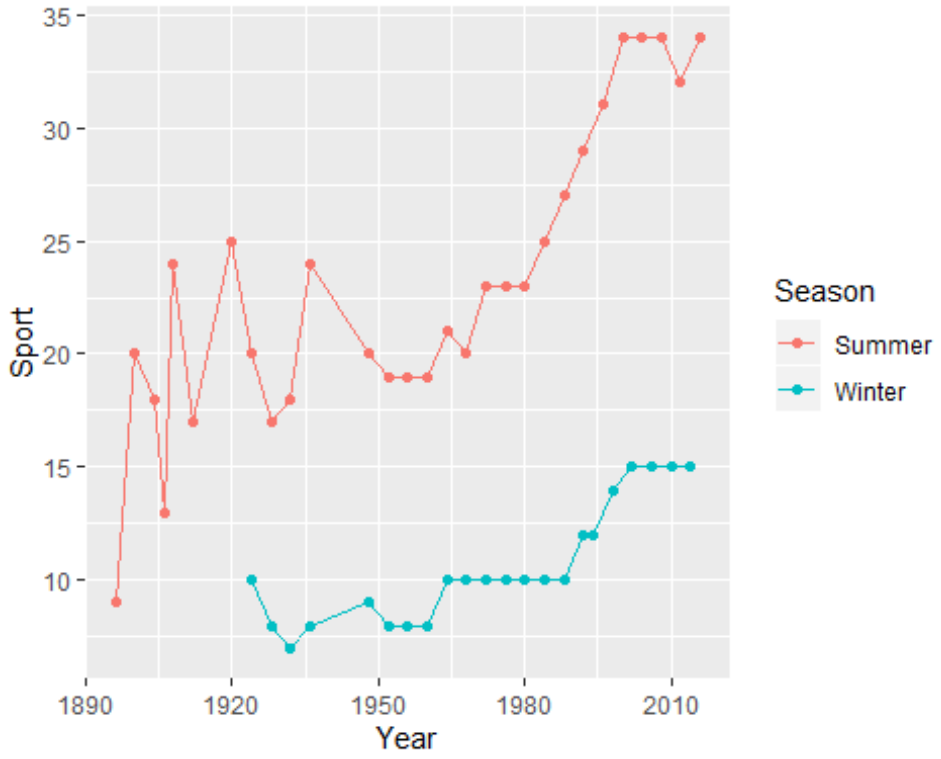
```
olympic%>%
  filter(!is.na(Medal))%>%
  group_by(Year,Season)%>%
  summarise(medal_count = n())%>%
  ggplot(aes(x=Year,y=medal_count,fill=Season))+geom_histogram(binwidth=0.5,s
tat='identity')+xlab("연도")+ylab("메달수")
```



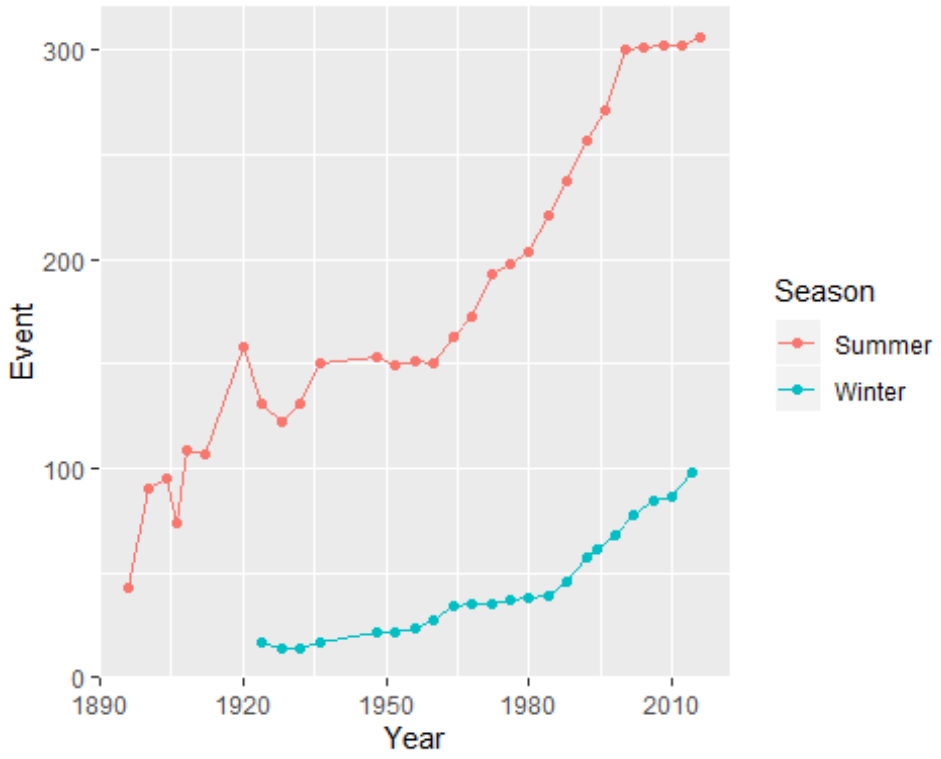
1800 년대 하계올림픽이 먼저 개최되었고, 이후 1924 년경 동계올림픽이 개최되었다는 것을 알 수 있다. 하계올림픽과 동계올림픽이 같은 연도에 개최되다가, 1994 년부터 동계올림픽, 하계올림픽이 2 년단위로 번갈아 개최되었다는 것 또한 알 수 있다. 이외에도, 하계올림픽의 메달수가 동계올림픽의 메달 수보다 월등히 많음을 알 수 있다.

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))
```

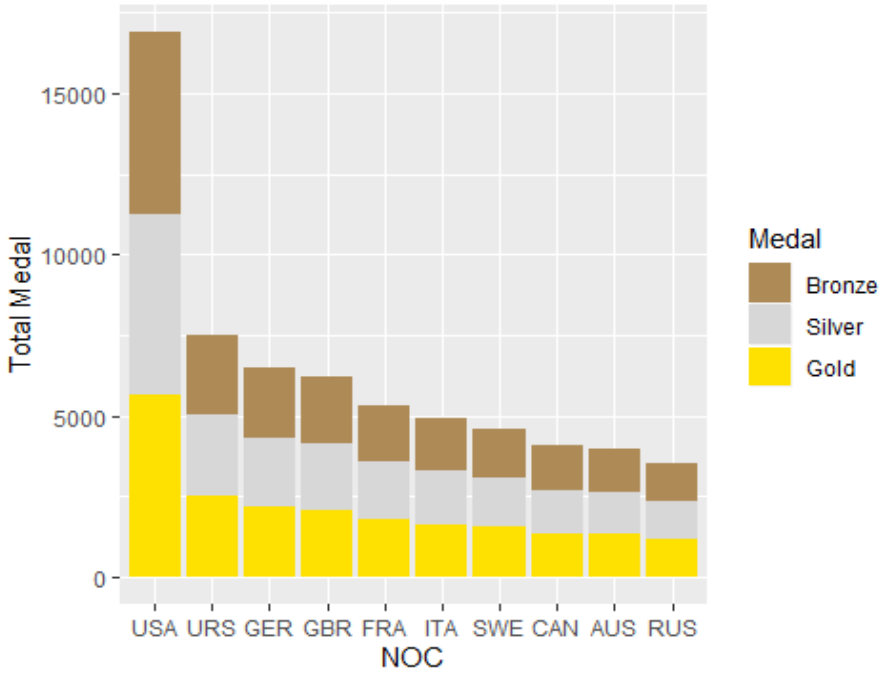


하계, 동계 올림픽의 Sport(종목)와 Event(세부 종목) 수를 연도별로 살펴본 결과, 하계올림픽이 동계올림픽에 비해 종목 수/세부 종목 수가 월등히 많다. 앞서 메달 수가 동계 대비 하계가 더 많았던 이유도 이러한 종목 수와 세부 종목 수의 차이에서 기인함을 알 수 있다. 위 두 결과를 종합하면, 하계올림픽이 동계올림픽보다 메달 수, 종목 수, 세부 종목수가 많은 것을 볼 수 있었다. 따라서 하계올림픽이 동계올림픽보다 더 규모가 크고 인기있다고 볼 수 있다.

2. 하계올림픽/동계올림픽 메달수와 나라의 기후간의 연관성

먼저 메달수의 특성을 살펴보기 위해, 상위 10 개국의 메달의 비율에 대한 그림을 그려보았다.

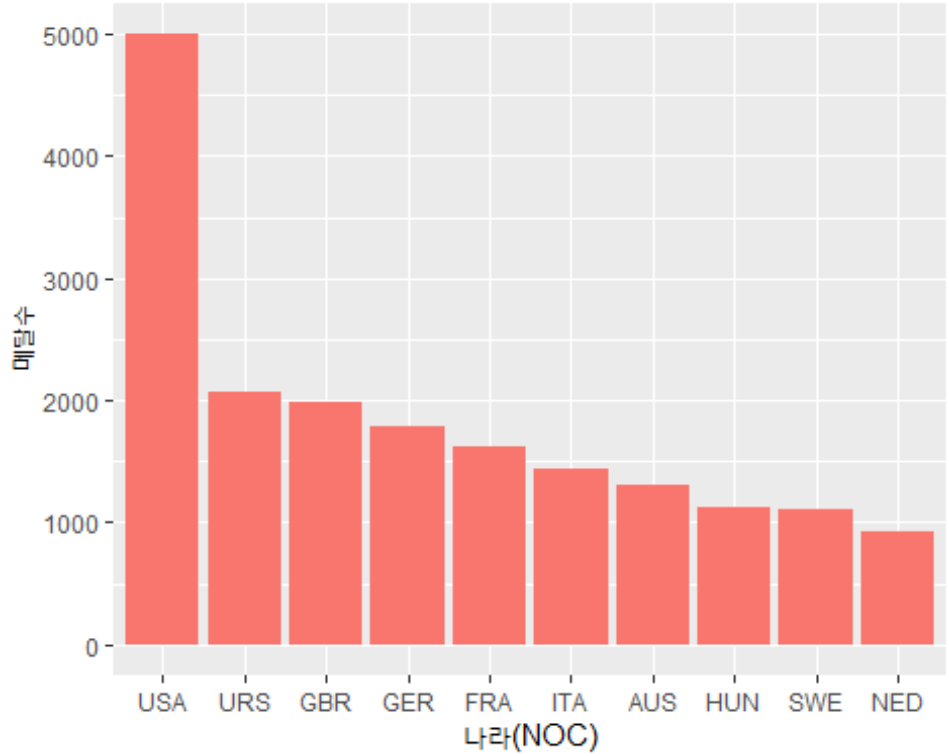
```
olympicTot<-olympic%>%
  filter(Medal!=0)%>%
  group_by(NOC)%>%
  summarise(medal_tot=n(),na.rm=TRUE)%>%
  select(-na.rm)%>%
  filter(rank(desc(medal_tot))<=10)

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"))
```



전체 메달수가 많은 국가가 각 금, 은, 동 메달의 수도 많다는 것을 알 수 있다. 따라서 하계/동계올림픽 메달수 비교에 전체 메달수만을 사용하기로 하였다.

하계올림픽 메달 획득 수 상위 10 개국

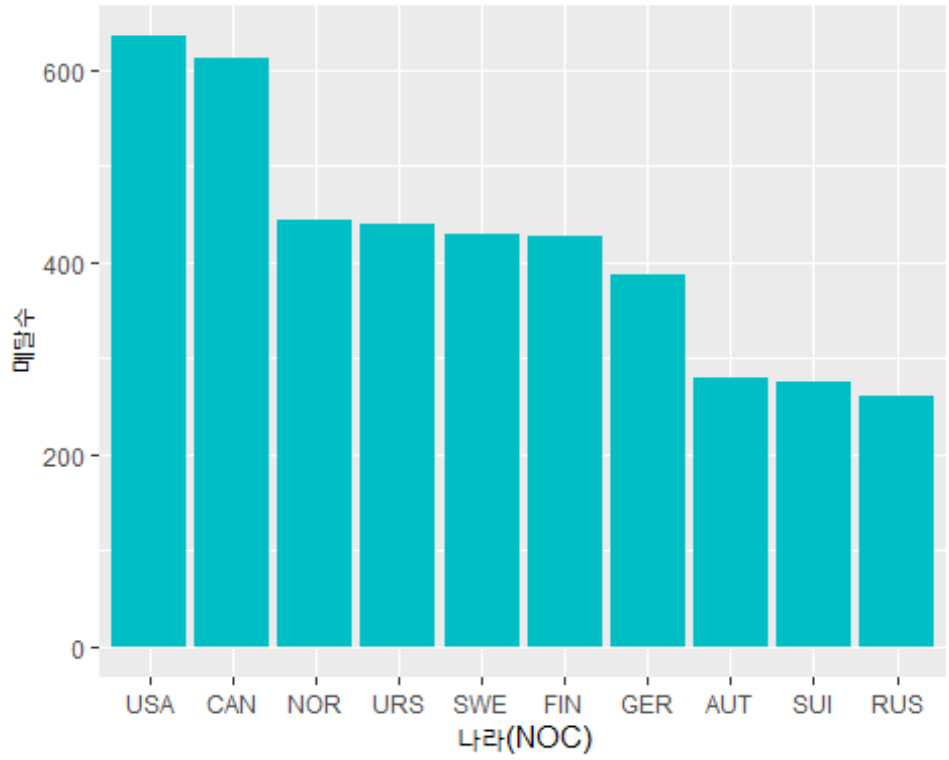
```
summer <- olympic %>%
  filter(Season=='Summer',!is.na(Medal))
summer %>%
  group_by(NOC)%>%
  summarise(medal_count=n())%>%
  filter(rank(desc(medal_count))<=10)%>%
  arrange(desc(medal_count))%>%
  ggplot(aes(x=reorder(NOC,-medal_count,sum),y=medal_count))+geom_bar(stat="i
dentity",fill="#F8766D",show.legend=FALSE)+xlab("나라(NOC)")+ylab("메달수")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



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)
```

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)
```

```
## ------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, th
en dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------
```

```
##
## Attaching package: 'plyr'
```

```
## The following object is masked from 'package:purrr':
##
##     compact
```

```
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
library(rworldmap)
```

```
## Loading required package: sp
```

```
## ### Welcome to rworldmap ###
```

```
## For a short introduction type :   vignette('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')
```

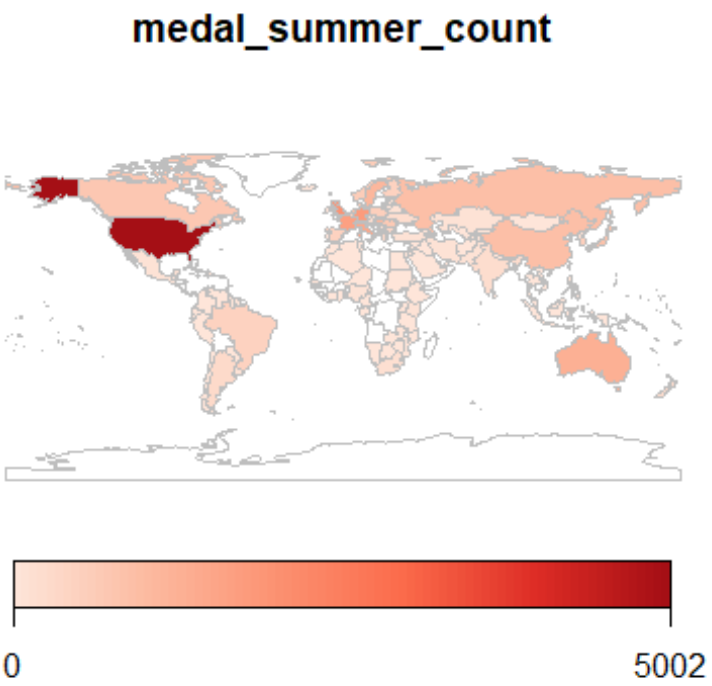
```
## Parsed with column specification:
## cols(
##   Flag = col_character(),
##   Country = col_character(),
##   IOC = col_character(),
##   FIFA = col_character(),
##   ISO = col_character()
## )
```

```
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

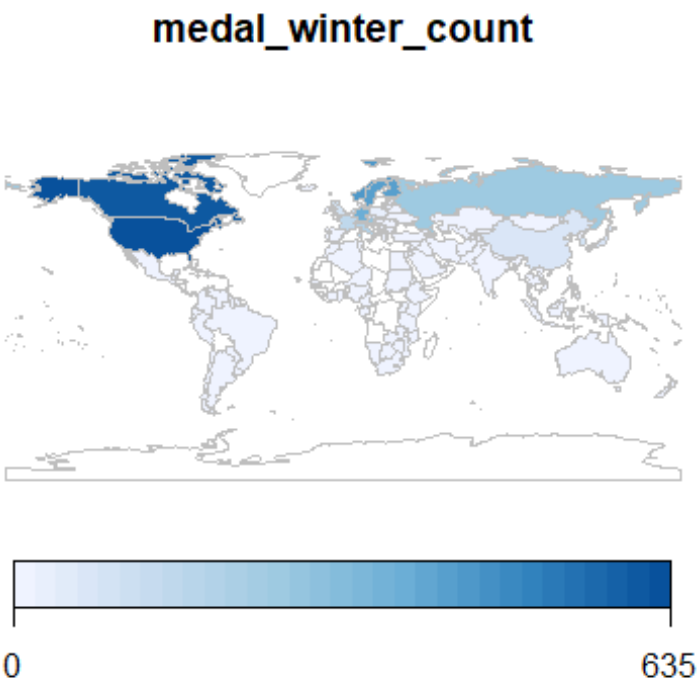
```
options(repr.plot.width=25, repr.plot.height=25)

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)))

## Warning in rwmGetColours(colourPalette, numColours): 6 colours specified
## and 80 required, using interpolation to calculate colours
```



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

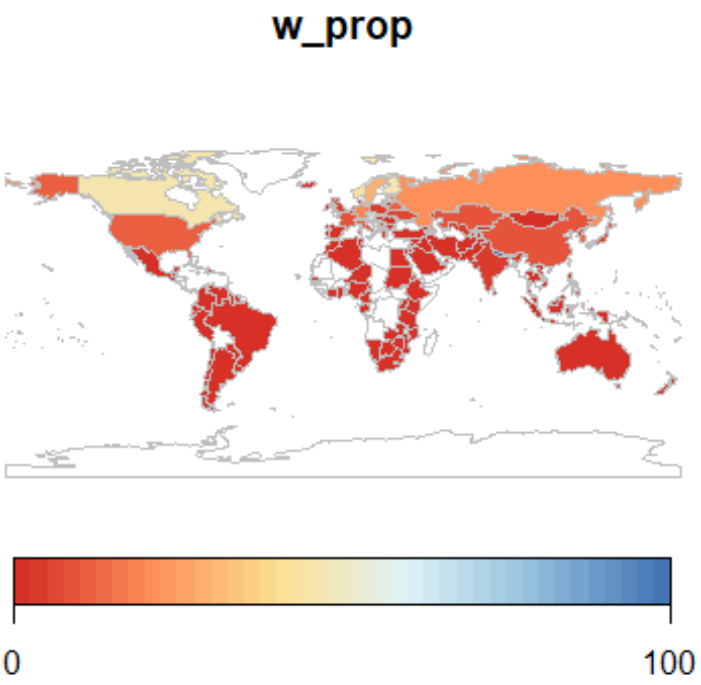
```
options(repr.plot.width=25, repr.plot.height=25)

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)))

## Warning in rwmGetColours(colourPalette, numColours): 6 colours specified
## and 31 required, using interpolation to calculate colours
```



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")

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



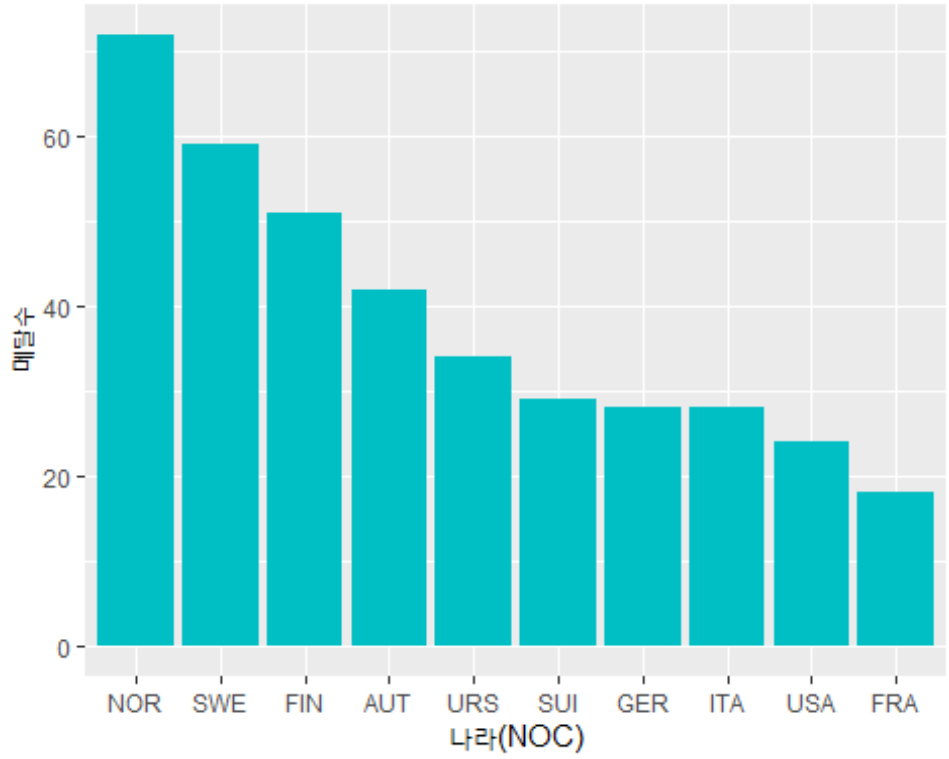
출전 인원이 많은 나라일수록 medal count 가 높게 잡힐 가능성이 크므로, 출전 인원의 영향을 배제하기 위해 나라별 하계올림픽, 동계올림픽 메달 비율에 대해 살펴보았다. 저위도지역에서는 동계올림픽보다 하계올림픽에 메달을 훨씬 많이 따는 것을 볼 수 있었고 노르웨이, 캐나다 등 고위도 지역, 히말라야 산맥 근처의 산지인 네팔 등 에서는 동계올림픽 메달 비율이 높아지는 것을 볼 수 있다. 다만, 대부분의 나라가 하계올림픽 메달 수가 많기 때문에 0~50 사이의 비중이 높다.

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

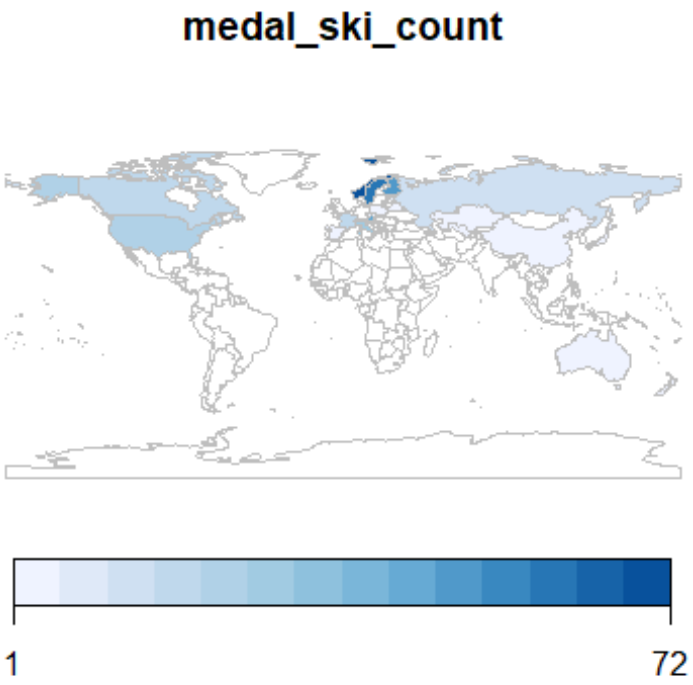
스키 종목(설상 종목)

```
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 = "NOC")
```

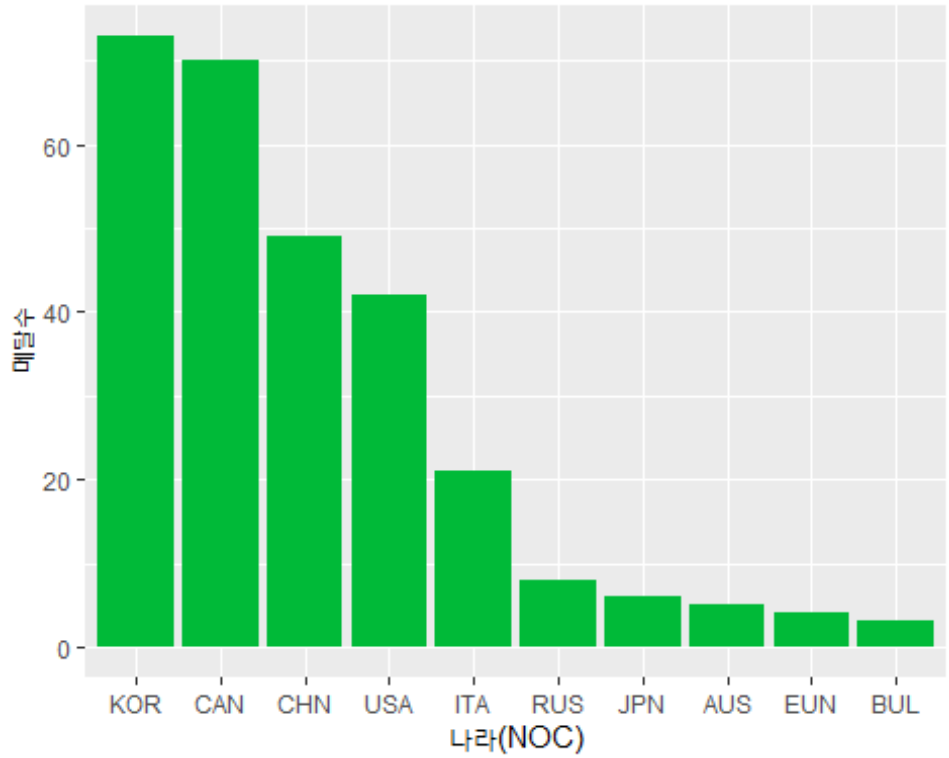
```
mapCountryData(sPDF4,nameColumnToPlot='medal_ski_count',colourPalette=colourPalette, catMethod ='fixedWidth',numCats = length(table(sPDF4$medal_ski_count)))
```



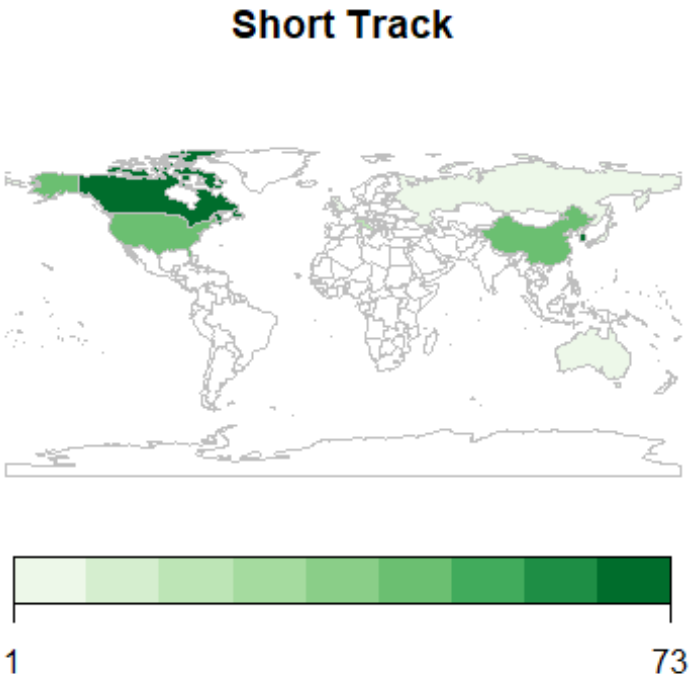
먼저 자연환경의 영향을 많이 받는 스키 종목에 대해 살펴보았다. 노르웨이, 스웨덴, 핀란드, 호주, 러시아, 독일, 이탈리아, 미국, 프랑스 순으로 높았다. 이를 통해 스키종목에서 추운나라가 우세하다는 것을 알 수 있었다. 스키경기를 하기 위해서는 특이한 지형, 즉 눈덮인 산지가 필요하기 때문에 위와 같은 결과가 나왔음을 추측할 수 있다.

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
")

## 10 codes from your data successfully matched countries in the map
## 3 codes from your data failed to match with a country code in the map
## 233 codes from the map weren't represented in your data

mapCountryData(sPDF5,nameColumnToPlot='medal_st_count',mapTitle="Short Track",
colourPalette=colourPalette, catMethod ='fixedWidth',numCats = length(table(s
PDF5$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.