# Project

2024-04-17

```r
library(ggplot2)
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
```

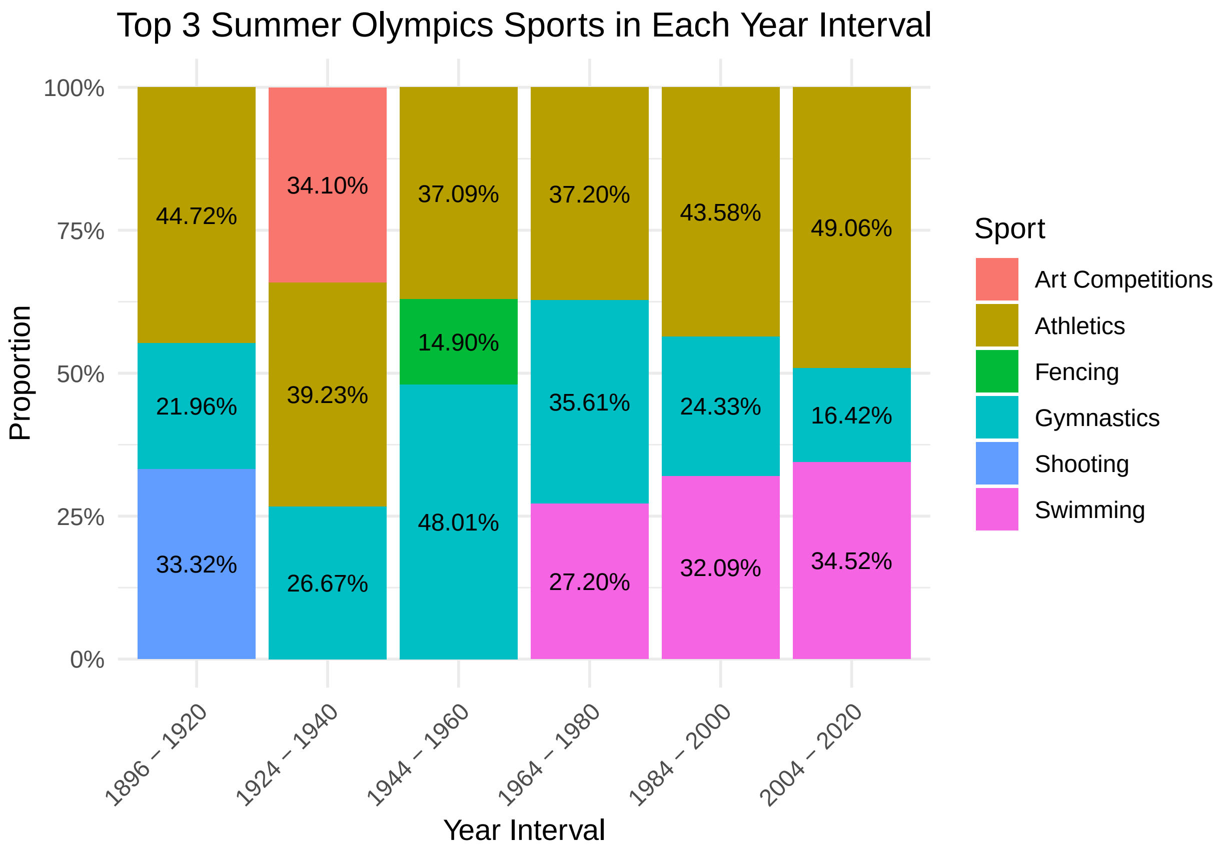
```r
summer_data <- read.csv("/Users/springkim/Downloads/archive/Athletes_summer_games.csv")
winter_data <- read.csv("/Users/springkim/Downloads/archive/Athletes_winter_games.csv")
```

```r
# 3. Trends between years and popular/dominant sports(Summer)

# 1896 - 1920
summer_data_a <- subset(summer_data, Year >= 1896 & Year <= 1920, select = c(Sport))
# 1924 - 1940
summer_data_b <- subset(summer_data, Year >= 1924 & Year <= 1940, select = c(Sport))
# 1944 - 1960
summer_data_c <- subset(summer_data, Year >= 1944 & Year <= 1960, select = c(Sport))
# 1964 - 1980
summer_data_d <- subset(summer_data, Year >= 1964 & Year <= 1980, select = c(Sport))
# 1984 - 2000
summer_data_e <- subset(summer_data, Year >= 1984 & Year <= 2000, select = c(Sport))
# 2004 - 2020
summer_data_f <- subset(summer_data, Year >= 2004 & Year <= 2020, select = c(Sport))

# Combine the data subsets into a single dataset
combined_data <- bind_rows(
  mutate(summer_data_a, YearInterval = "1896 - 1920"),
  mutate(summer_data_b, YearInterval = "1924 - 1940"),
  mutate(summer_data_c, YearInterval = "1944 - 1960"),
  mutate(summer_data_d, YearInterval = "1964 - 1980"),
  mutate(summer_data_e, YearInterval = "1984 - 2000"),
  mutate(summer_data_f, YearInterval = "2004 - 2020")
)
```

```r
aggregated_data <- combined_data %>%
  group_by(YearInterval, Sport) %>%
  summarise(Count = n()) %>%
  arrange(YearInterval, desc(Count)) %>%
  group_by(YearInterval) %>%
  mutate(Rank = row_number()) %>%
  filter(Rank <= 3) %>%
  ungroup() %>%
  group_by(YearInterval) %>%
  mutate(Total = sum(Count)) %>%
  ungroup() %>%
  mutate(Proportion = Count / Total)
```

```
## ‘summarise()‘ has grouped output by ’YearInterval’. You can override using the
## ‘.groups‘ argument.
```

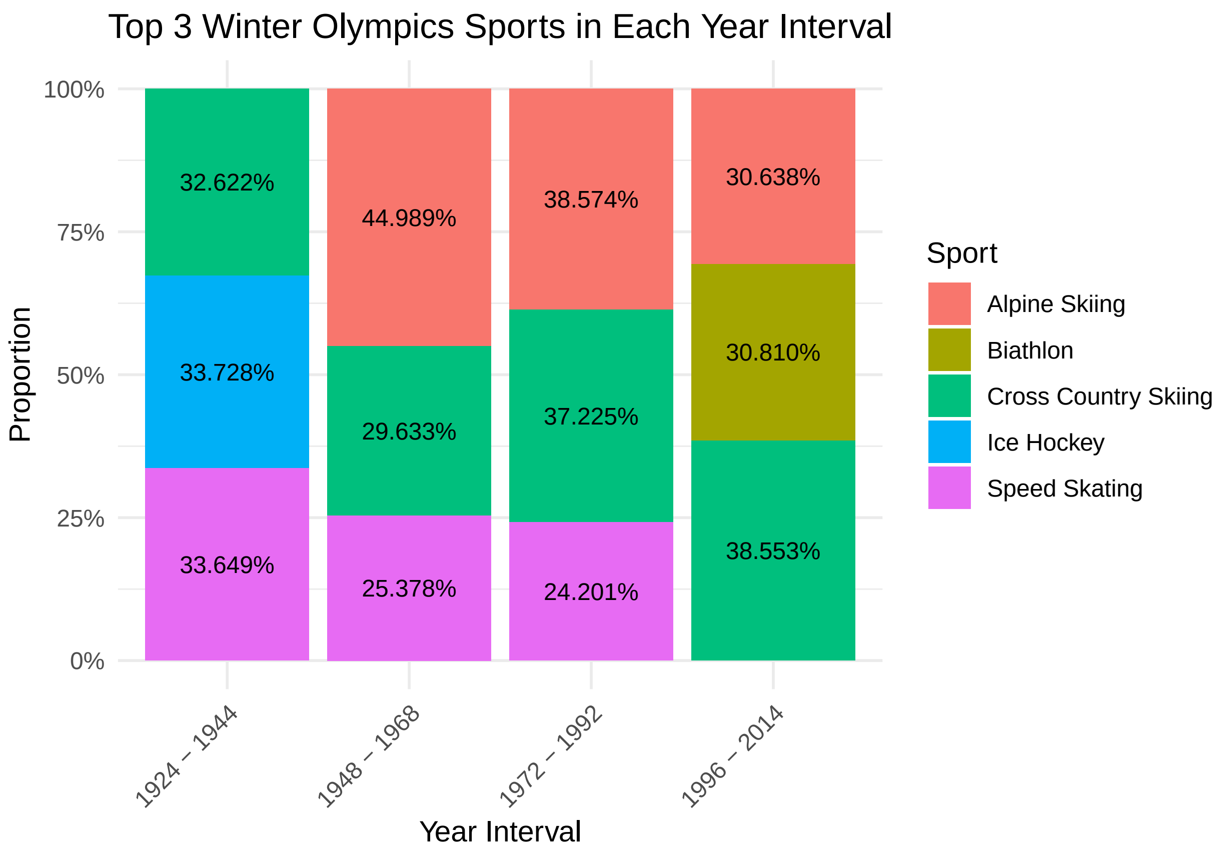
```r
ggplot(aggregated_data, aes(x = YearInterval, y = Proportion, fill = Sport)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = scales::percent(Proportion)), position = position_stack(vjust = 0.5), size = 3,
  labs(x = "Year Interval", y = "Proportion", fill = "Sport") +
  ggtitle("Top 3 Summer Olympics Sports in Each Year Interval") +
  scale_y_continuous(labels = scales::percent) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust = 0.5))
```



```r
# 4. Trends between years and popular/dominant sports(Winter)

# 1924 - 1944
winter_data_a <- subset(winter_data, Year >= 1924 & Year <= 1944, select = c(Sport))
# 1948 - 1968
winter_data_b <- subset(winter_data, Year >= 1948 & Year <= 1968, select = c(Sport))
# 1972 - 1992
winter_data_c <- subset(winter_data, Year >= 1972 & Year <= 1992, select = c(Sport))

# 1996 - 2014
winter_data_d <- subset(winter_data, Year >= 1996 & Year <= 2014, select = c(Sport))

combined_data <- bind_rows(
  mutate(winter_data_a, YearInterval = "1924 - 1944"),
  mutate(winter_data_b, YearInterval = "1948 - 1968"),
  mutate(winter_data_c, YearInterval = "1972 - 1992"),
  mutate(winter_data_d, YearInterval = "1996 - 2014")
)



aggregated_data <- combined_data %>%
  group_by(YearInterval, Sport) %>%
  summarise(Count = n()) %>%
  arrange(YearInterval, desc(Count)) %>%
  group_by(YearInterval) %>%
  mutate(Rank = row_number()) %>%
  filter(Rank <= 3) %>%
  ungroup() %>%
  group_by(YearInterval) %>%
  mutate(Total = sum(Count)) %>%
  ungroup() %>%
  mutate(Proportion = Count / Total)
```

```
## ‘summarise()‘ has grouped output by ’YearInterval’. You can override using the
## ‘.groups‘ argument.
```

```r
ggplot(aggregated_data, aes(x = YearInterval, y = Proportion, fill = Sport)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = scales::percent(Proportion)), position = position_stack(vjust = 0.5), size = 3,
  labs(x = "Year Interval", y = "Proportion", fill = "Sport") +
  ggtitle("Top 3 Winter Olympics Sports in Each Year Interval") +
  scale_y_continuous(labels = scales::percent) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust = 0.5))
```



```
# Participation (Original/Given Source)
# Do sports with higher global participation tend to dominate the Olympic medal standings? (majority nu
# Scatter plot / Summer
```