**Data Visualization in International Rugby Key**

Introduction

International Rugby Union matches have taken place since 1871 and includes people of all different backgrounds. In rugby, each team has 15 players on the field and the team that scores the most points wins. There are several tournaments in rugby, the most popular being the Six Nations and the Rugby Championship. Our focus will be to visualize both the popularity of the sport and its winners from 1871 to 2023. Doing so will require some data cleaning and wrangling.

Data Description

`winner` - winner of the match

`year` - year of the match

`competition` - name of the competition that the match was a part of

`country` - host country of the match

Exercises

```{r}

# load in the necessary packages and dataset

library(dplyr)

library(here)

rugby <- read.csv(here(".../rugby.csv"))

```

1. Use the `summarise()` and `group\_by()` functions to create a new variable that counts the number of matches per year.

```{r}

rugby\_graph <-

rugby %>%

group\_by(year) %>%

summarise(nmatches = n())

```

2. Construct a scatterplot with a smoother that displays the popularity of the sport over time. Describe the trend.

```{r}

ggplot(data = rugby\_graph, mapping = aes(x = year, y = nmatches)) +

geom\_point() +

geom\_smooth() +

theme\_minimal() +

labs(title = "Growth of International Rugby Over Time", x = "Year", y = "Number of Matches") +

theme(plot.title = element\_text(hjust = 0.5))

```

The trend has an exponential shape to it with a large growth in the number of matches in the year 2000.

3. The teams in the Rugby Championship are typically Argentina, Australia, New Zealand, and South Africa. Of those, New Zealand and South Africa are ranked the top two men's teams in the world by World Rugby. Create a line graph that displays the number of matches played by those two nations from 1871 to 2023. Use dplyr functions to filter the data.

```{r}

countries <- c("New Zealand", "South Africa")

champ\_graph <- rugby %>%

select(year, country) %>%

filter(country %in% countries) %>%

group\_by(year, country) %>%

summarise(nmatches = n())

ggplot(data = champ\_graph, mapping = aes(x = year, y = nmatches, color = country)) +

geom\_line() +

theme\_minimal() +

labs(title = "Number of Matches by the Top Nations", x = "Year", y = "Number of Matches", color = "Country") +

theme(plot.title = element\_text(hjust = 0.5))

```

4. Do you think "home field advantage" applies to the Rugby Championship? Make and interpret a scatterplot to explore this idea. The scatterplot should include average points for both home and away teams for each year of the "Rugby Championship". Hint: You'll need to use `stringr` to filter for the "Rugby Championship" competition and `dplyr` to get the average amount of points for home and away teams.

```{r}

library(stringr)

rugby\_champ <- rugby %>%

filter(competition %in% str\_subset(string = competition, pattern = "Rugby Championship"))

champ\_points <- rugby\_champ %>%

group\_by(year) %>%

summarise(avg\_home\_points = mean(home\_score, na.rm = TRUE),

avg\_away\_points = mean(away\_score, na.rm = TRUE))

ggplot(data = champ\_points, mapping = aes(x = year)) +

geom\_point(aes(y = avg\_home\_points, color = "Home")) +

geom\_point(aes(y = avg\_away\_points, color = "Away")) +

theme\_minimal() +

labs(title = "Home Team Advantage in the Rugby Championship", x = "Year", y = "Yearly Point Averages", color = "Venue") +

theme(plot.title = element\_text(hjust = 0.5))

```

There does seem to be a home team advantage when looking at the average amount of points for home and away teams in each year of the Rugby Championship. Only two years had the away teams with higher average points in the scatterplot above, while home teams had significantly higher points in the other years.

5. Let's go back to the original dataset with all of the matches. Us dplyr functions to see which nation has the most wins. Make sure to filter out any ties. Use the `head()` function to find the top three.

```{r}

top\_teams <- rugby %>%

group\_by(winner) %>%

filter(winner != "Tie") %>%

summarise(nwins = n()) %>%

arrange(desc(nwins)) %>%

head()

```

New Zealand, England, and France have the most International Rugby wins since 1871.

6. Now create a lollipop plot and a pie chart that displays the number of wins for each nation in the top 6 for total wins. Which visual is better at presenting the data. Explain.

```{r}

ggplot(data = top\_teams, aes(x = winner, y = nwins)) +

geom\_segment(aes(xend = winner, y = 0, yend = nwins),

colour = "purple") +

geom\_point(colour = "red") +

labs(title = "Total Number of Wins for the Top Six Countries") +

theme\_minimal() +

theme(plot.title = element\_text(hjust = 0.5))

```

```{r}

top\_teams\_sum <- top\_teams %>%

mutate(prop = nwins / sum(nwins))

ggplot(data = top\_teams\_sum, aes(x = "", y = prop, fill = winner)) +

geom\_bar(stat = "identity", width = 1) +

coord\_polar("y") + scale\_fill\_viridis\_d() +

theme\_void() +

labs(title = "Proportion of Rugby Wins Between Top 6 Teams")

```

The lollipop plot is better at representing the data than the pie chart because it gives numerical values that are easier to interpret than the "slices" on the pie chart. In this case, since the values are so close together, the pie chart makes it seem like all of the proportions are the same while in the lollipop plot it's clear that they are not.