Predicting the Winner of the 2026 Men's FIFA World Cup

Diana Batista Capellan, Kate Miller, Isabel Sumy

Introduction

A brief description of the motivations behind your project, the goal of your project, why it is interesting or important, and why machine learning is a reasonable approach.

Exploratory Data Analysis

```r
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
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v forcats   1.0.0     v readr     2.1.5
v ggplot2   3.4.4     v stringr   1.5.1
v lubridate 1.9.3     v tibble    3.2.1
v purrr     1.0.2     v tidyr     1.3.1
```

```
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```
library(readr)

data1 <- read.csv("fifa_ranking-2023-07-20.csv")

data2 <- read.csv("worldcups.csv")

data3 <- read.csv("wcmatches.csv")

# Use columns from data 3: Home score and away score
# Use columns from data 1: Rank
# Add these columns to data 2

names(data1)[names(data1) == "country_full"] <- "Country"
names(data2)[names(data2) == "winner"] <- "Country"
names(data3)[names(data3) == "winning_team"] <- "Country"

first_join <- left_join(data2, data1, by = "Country")

second_join <- left_join(first_join, data3, by = "Country")
```

Illustration / Figure

A figure or a diagram that illustrates the overall model or idea of your project. The idea is to make your report more accessible, especially to readers who are starting by skimming your work. For the project, taking a picture of a hand-drawn diagram is fine, as long as it's legible. PowerPoint is another option. You will not be penalized for hand-drawn illustrations – you are graded on the design and illustrative power

```
library(dplyr)
library(ggplot2)

# Filter matches where the home team won
home_wins <- filter(second_join, outcome == "H")

# Filter matches where the away team won
```

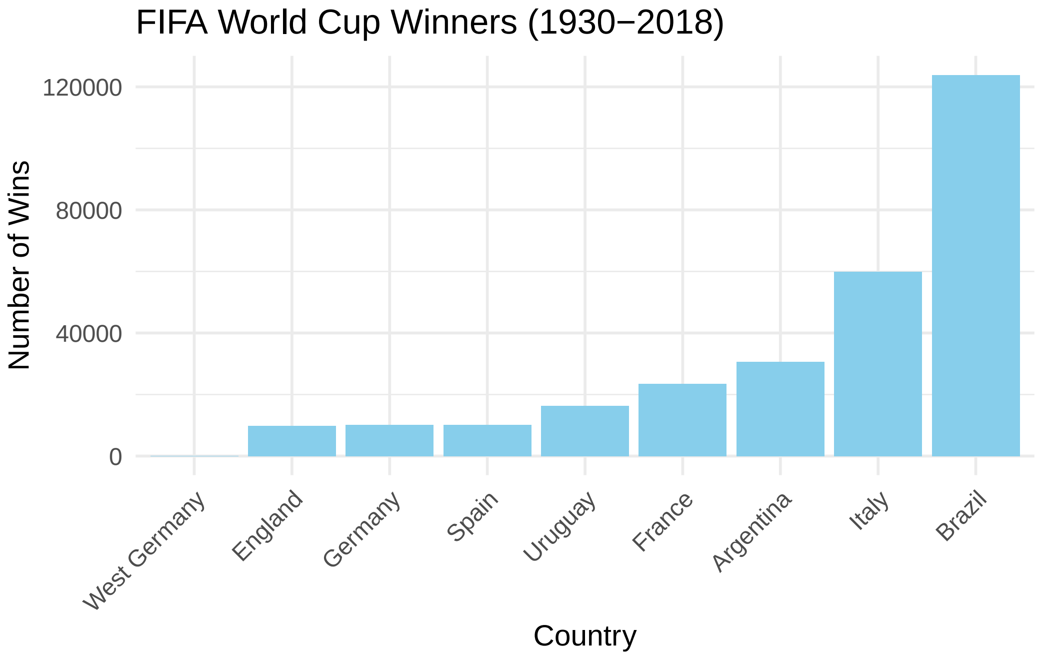
```r
away_wins <- filter(second_join, outcome == "A")

# Combine the home and away wins
all_wins <- bind_rows(home_wins, away_wins)

# Aggregate the data to count the number of wins for each country
world_cup_wins <- all_wins %>%
  group_by(Country) %>%
  summarise(Wins = n())

# Sort the data by the number of wins in descending order
world_cup_wins <- world_cup_wins[order(-world_cup_wins$Wins),]

# Create the bar plot
ggplot(world_cup_wins, aes(x = reorder(Country, Wins), y = Wins)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "FIFA World Cup Winners (1930-2018)",
       x = "Country",
       y = "Number of Wins") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



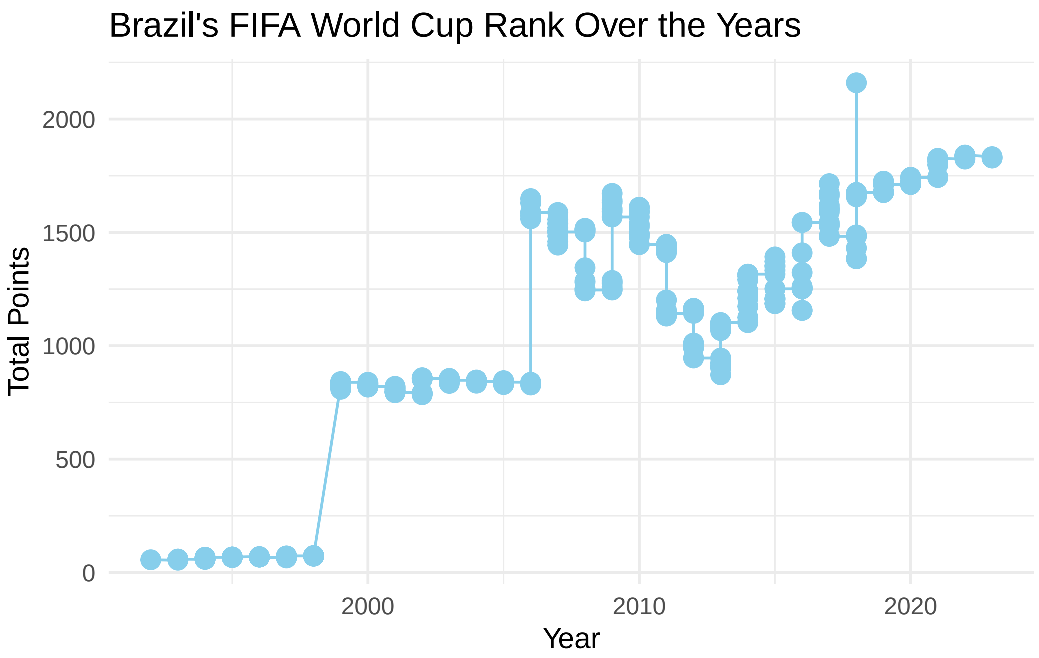
```r
library(dplyr)
library(ggplot2)

# Filter Brazil's rank data
brazil_rank <- filter(data1, Country == "Brazil")

# Extract the year from the rank_date column
brazil_rank$year <- substr(brazil_rank$rank_date, 1, 4)

# Convert year to numeric
brazil_rank$year <- as.integer(brazil_rank$year)

# Create the line plot
ggplot(brazil_rank, aes(x = year, y = total_points)) +
  geom_line(color = "skyblue") +
  geom_point(color = "skyblue", size = 3) +
  labs(title = "Brazil's FIFA World Cup Rank Over the Years",
       x = "Year",
       y = "Total Points") +
  theme_minimal()
```



Background & Related Work (2 points)

A description of 1-2 related work in the field, to provide reader a sense of what has already been done in this area, e.g. papers or existing products/software that do a related thing.

- 2/2 Briefly describes 1-2 prior work related to your project to put your project into context. Your descriptions need not be complete, but should contain important work
- 1/2 Background that has omissions or factual incorrectness, but otherwise places your project into context.
- 0/1 Background contains too much information not related to your project, or has major omissions of content provided to you by your instructor or TA., or does not sufficiently put your project into context.

Data Processing

Describe the data that you have collected and cleaned. Be clear and specific when describing what you've done, so that a classmate can reproduce your work. Show some statistics and examples of your data.

- 4/4 Clearly describes sources of data, and the steps you took to clean and format your data. Statistics and data example are well-chosen, and gives readers a "feel" for your data.
- 3/4 Mostly clear description, but some aspects of the data processing steps are vague. Statistics and data example are somewhat illustrative/helpful.
- 2/4 Vague description or missing key information about where your data comes from or what you did. No example data shown, or the ones shown are not illustrative.
- 1/4 Incomplete information.

Architecture

A description of the final model. Do not describe all the intermediate models that you have tried. Instead, present the model (or models) whose quantitative results you will show. These should be your most interesting models. Be as specific as you can while being concise. Readers should be able to reproduce a model similar enough to yours and obtain a similar performance.

Baseline Model

Describe a simple, baseline model that you will compare your neural network against. This can be a simple model that you build.

Quantitative Results

A description of the quantitative measures of your result. What measurements can you use to illustrate how your model performs?

Qualitative Results

Include some sample outputs of your model, to help your readers better understand what your model can do. The qualitative results should also put your quantitative results into context (e.g. Why did your model perform well? Is there a type of input that the model does not do well on?)

Discussion

Discuss your results. Do you think your model is performing well? Why or why not? What is unusual, surprising, or interesting about your results? What did you learn?

Ethical Considerations

Description of a use of the system that could give rise to ethical issues. Are there limitations of your model? Your training data?

(Note that the expectations are higher here than in the project proposal.)

Conclusion(Optional)

Summarize the whole report.