Final Project

Alexia Crisologo and Olivia Encarnacion

```
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 core tidyverse packages ------------------------ tidyverse 2.0.0 --
v forcats   1.0.0     v readr     2.1.4
v ggplot2   3.4.3     v stringr   1.5.0
v lubridate 1.9.2     v tibble    3.2.1
v purrr     1.0.2     v tidyr     1.3.0

-- 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 becor
```

```
library(tidymodels)
```

```
-- Attaching packages ------------------------------------ tidymodels 1.1.1 --
v broom        1.0.5     v rsample      1.2.0
v dials        1.2.0     v tune         1.1.2
v infer        1.0.5     v workflows    1.1.3
v modeldata    1.2.0     v workflowsets 1.0.1
v parsnip      1.1.1     v yardstick    1.2.0
v recipes      1.0.8
-- Conflicts --------------------------------------- tidymodels_conflicts() --
x scales::discard() masks purrr::discard()
x dplyr::filter()   masks stats::filter()
x recipes::fixed()  masks stringr::fixed()
x dplyr::lag()      masks stats::lag()
x yardstick::spec() masks readr::spec()
x recipes::step()   masks stats::step()
* Use suppressPackageStartupMessages() to eliminate package startup messages
```

```
library(ggplot2)
tennis <- read.csv("data/wta_matches_qual_itf_2023.csv")
```

Introduction and Data

Research Question

Does the court type impact the duration of a women's single tennis match, considering the winner's height, age, and playing hand?

The data we chose was the Women's Tennis Association data set from Awesomedata's GitHub repository (https://github.com/JeffSackmann/tennis_wta/blob/master/wta_matches_qual_itf_2023.csv) The data was created in 2023 and was collected from the International Tennis Federation. The data contains 34,323 observations and 49 variables. The variables of interest in our research include surface, the winner and loser id, winner and loser IOC (country), winner and loser height, winner and loser age, and winner and loser hand. These variables will help us answer the question of the court types impact on the duration of tennis matchs. There are some NA variables corresponding to the players height. Since this is our variable of interest, we will be dropping all NA values corresponding to height. This will leave us with 1,256 obeservations. Additionally, since grand slams are played on either clay, grass, or hard courts, we will be dropping the matches that were played on carpet. This will then leave us with 1,239 observations. The motivation behind this project is to help Women's Tennis players better prepare for the length of their match based on the surface they will be playing on. With the Olympics coming up, this data will help the tennis player better prepare for a match.

```
tennis <- tennis %>%
  filter(!is.na(winner_ht) & !is.na(loser_ht) & surface != "Carpet")
```

Variables of Interest

Surface: Surface the match was played on (clay, grass, or hard)

Winner_id: Identification of the winner

Winner_hand: Dominant playing hand of the winner (right, left, undecided)

Winner_ht: Height of the winner in centimeters (cm)

Winner_ioc: Country of the winner as assigned by the International Olympic Committee

Winner_age: Age of the winner

Minutes: Duration of the tennis match in minutes

```
winner_hand_counts <- tennis |>
  count(winner_hand)

loser_hand_counts <- tennis |>
  count(loser_hand)

print(winner_hand_counts)
```

```
  winner_hand    n
1           L  116
2           R 1110
3           U   13
```

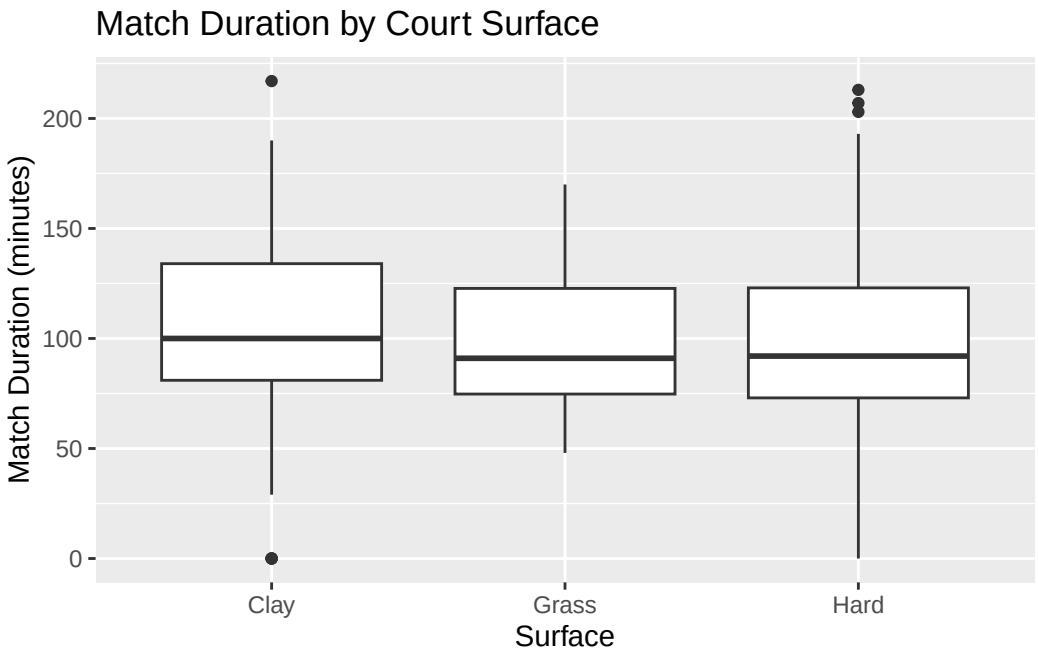
```
print(loser_hand_counts)
```

```
  loser_hand    n
1          L  107
2          R 1107
3          U   25
```

Methodology

```
ggplot(data = tennis, aes(x = surface, y = minutes)) +
  geom_boxplot() +
  labs(title = "Match Duration by Court Surface",
       x = "Surface",
       y = "Match Duration (minutes)")
```

```
Warning: Removed 620 rows containing non-finite values (`stat_boxplot()`).
```




```
library(MASS)
```

```
Attaching package: 'MASS'

The following object is masked from 'package:dplyr':

    select
```

```r
tennis_binary <- tennis |>
  mutate(minutes = case_when(short = minutes <= 60 ~ "short",
         medium = minutes >= 60 & minutes <= 100 ~ "medium",
         long = minutes >= 100 ~ "long"))

winner_mins <- polr(as.factor(minutes) ~ surface + winner_age +
                    as.numeric(winner_ht) + winner_hand, data = tennis_binary)
summary(winner_mins)
```

```
Re-fitting to get Hessian

Call:
polr(formula = as.factor(minutes) ~ surface + winner_age + as.numeric(winner_ht) +
    winner_hand, data = tennis_binary)

Coefficients:
                        Value Std. Error t value
surfaceGrass          0.33532    0.28470  1.1778
surfaceHard           0.24840    0.17325  1.4338
winner_age           -0.03828    0.01845 -2.0743
as.numeric(winner_ht) -0.01016    0.01250 -0.8123
winner_handR          -0.11421    0.27374 -0.4172
winner_handU           0.64323    1.79462  0.3584

Intercepts:
             Value   Std. Error t value
long|medium  -2.9470  2.2467    -1.3117
medium|short -0.2195  2.2479    -0.0976

Residual Deviance: 1117.097
AIC: 1133.097
(620 observations deleted due to missingness)
```

```r
loser_mins <- polr(as.factor(minutes) ~ surface + loser_age +
                   as.numeric(loser_ht) + loser_hand, data = tennis_binary)
summary(loser_mins)
```

```
Re-fitting to get Hessian
```

```
Call:
polr(formula = as.factor(minutes) ~ surface + loser_age + as.numeric(loser_ht) +
    loser_hand, data = tennis_binary)

Coefficients:
                         Value Std. Error t value
surfaceGrass          0.292832    0.28684  1.0209
surfaceHard           0.242758    0.17399  1.3953
loser_age             0.004285    0.01884  0.2274
as.numeric(loser_ht) -0.029149    0.01228 -2.3736
loser_handR          -0.517624    0.28281 -1.8303
loser_handU           0.357646    0.74934  0.4773

Intercepts:
             Value   Std. Error t value
long|medium  -5.4166  2.2128    -2.4478
medium|short -2.6713  2.2086    -1.2095

Residual Deviance: 1111.578
AIC: 1127.578
(620 observations deleted due to missingness)
```