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 tidymodels_prefer() to resolve common conflicts.
```

```
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
```

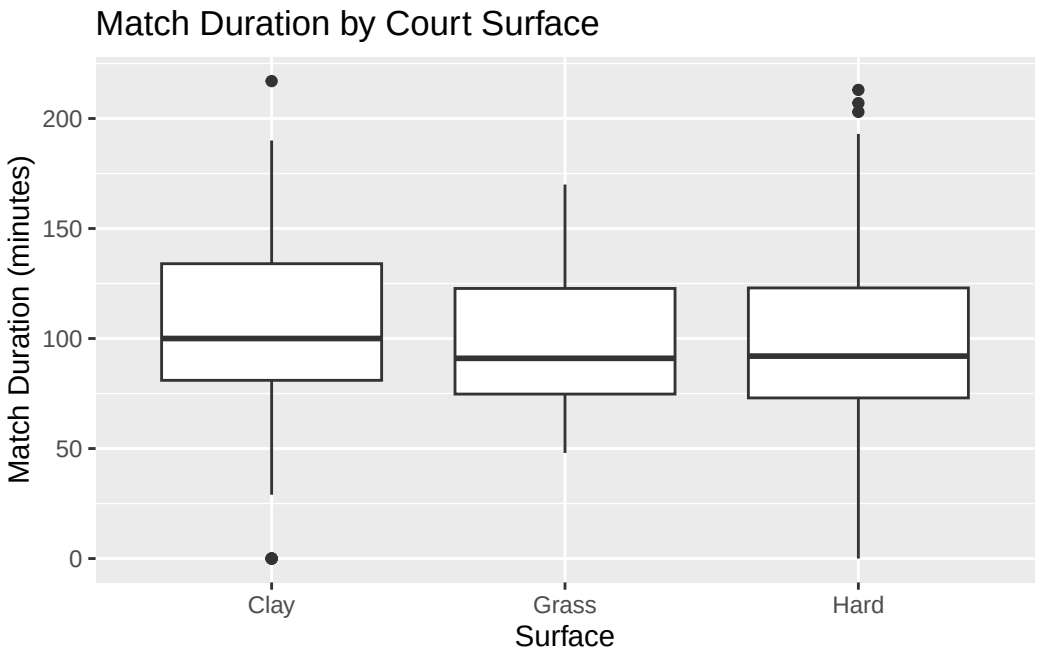
```
print(loser_hand_counts)
```

```
  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()`).
```




```
mean_game_duration <- mean(tennis$minutes, na.rm = TRUE)
print(mean_game_duration)
```

```
[1] 101.2973
```

```
tennis_binary <- tennis |>
  mutate(minutes = case_when(minutes >= "101.3"~ 1,
                             minutes <= "101.3" ~ 0))
```

```r
log_winner <- glm(as.factor(minutes) ~ surface  +
             winner_age + as.numeric(winner_ht) + winner_hand,
    data = tennis_binary,
    family = "binomial")
summary(log_winner)
```

```
Call:
glm(formula = as.factor(minutes) ~ surface + winner_age + as.numeric(winner_ht) +
    winner_hand, family = "binomial", data = tennis_binary)

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            1.418e+01  8.743e+02   0.016    0.987
surfaceGrass           4.842e-01  7.905e-01   0.612    0.540
surfaceHard            6.803e-01  4.695e-01   1.449    0.147
winner_age             4.342e-03  5.467e-02   0.079    0.937
as.numeric(winner_ht)  2.212e-02  3.506e-02   0.631    0.528
winner_handR          -1.525e+01  8.743e+02  -0.017    0.986
winner_handU           2.396e-01  6.581e+03   0.000    1.000

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 183.39  on 618  degrees of freedom
Residual deviance: 176.78  on 612  degrees of freedom
  (620 observations deleted due to missingness)
AIC: 190.78

Number of Fisher Scoring iterations: 17
```

```r
log_loser <- glm(as.factor(minutes) ~ surface  +
             loser_age + as.numeric(loser_ht) + loser_hand,
    data = tennis_binary,
    family = "binomial")
summary(log_loser)
```

```
Call:
glm(formula = as.factor(minutes) ~ surface + loser_age + as.numeric(loser_ht) +
```

```
    loser_hand, family = "binomial", data = tennis_binary)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -7.826e+00  6.580e+00  -1.189   0.2343
surfaceGrass          6.143e-01  7.960e-01   0.772   0.4402
surfaceHard           8.074e-01  4.740e-01   1.704   0.0885 .
loser_age            -9.559e-03  5.465e-02  -0.175   0.8612
as.numeric(loser_ht)  6.018e-02  3.675e-02   1.638   0.1015
loser_handR           7.055e-01  6.506e-01   1.085   0.2781
loser_handU           1.499e+01  1.374e+03   0.011   0.9913
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 183.39  on 618  degrees of freedom
Residual deviance: 176.51  on 612  degrees of freedom
  (620 observations deleted due to missingness)
AIC: 190.51

Number of Fisher Scoring iterations: 16
```