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()
* Dig deeper into tidy modeling with R at https://www.tmwr.org
```

```
library(ggplot2)
library(Stat2Data)
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, 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 matches. 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.(Sackmann, Jeff 19 March 2024)

```r
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_hand: Dominant playing hand of the winner (right, left, undecided)

Winner_ht: Height of the winner in centimeters (cm)

Winner_age: Age of the winner

loser_hand: Dominant playing hand of the loser (right, left, undecided)

loser_ht: Height of the loser in centimeters (cm)

loser_age: Age of the loser

Minutes: Duration of the tennis match in minutes

```r
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
```

```r
print(loser_hand_counts)
```

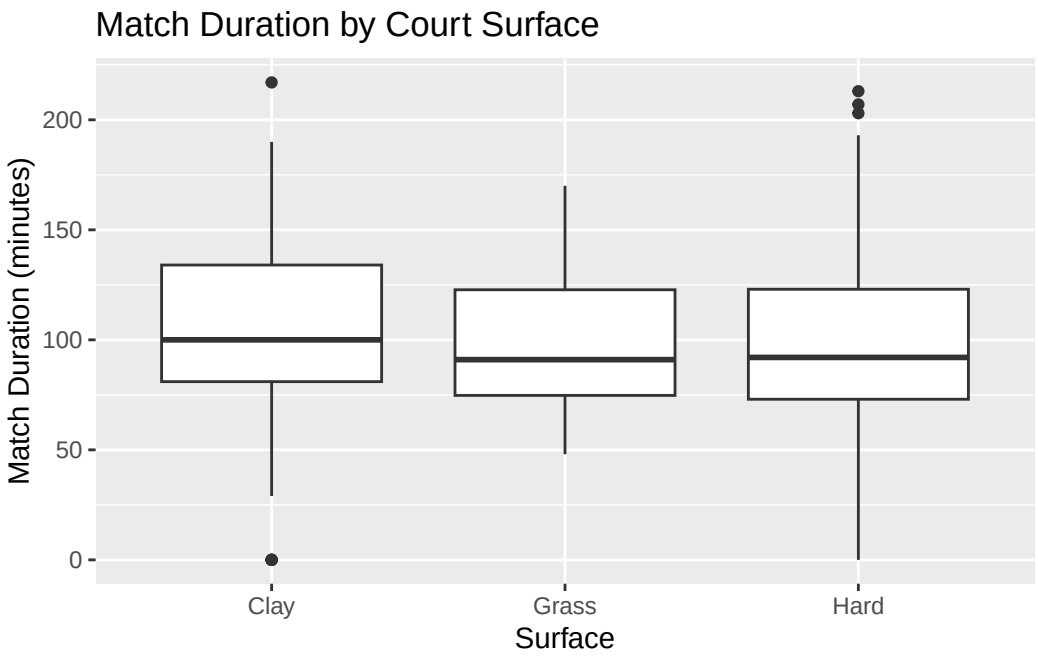
```
  loser_hand    n
1          L  107
2          R 1107
3          U   25
```

Methodology

First we wanted to visualize the relationship between the match duration and the surface that match was played on. We generated a box plot for each surface, clay, grass, and hard. The results indicated that the average match duration tends to be higher for matches played on clay.

```
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()`).
```



We are using a logistic regression model to evaluate whether the length of a tennis match is influenced by the surface the match is played on, the age of the player, the height of the player, and the playing hand of the player. This seems to be an appropriate model for our data given that it satisfies the independence and linearity assumptions in the log odds. Each observation in the dataset represents a different tennis match making this assumption hold since each tennis match is independent of each other. The outcome of one match doesn't not

influence the outcome of the other. In addition, there is independence within each player. The characteristics of one player do not directly influence the characteristics of the other. For the outcome variable, we chose to use the binary version of whether a match was above or below the average minutes of all the matches. This binary outcome generates a simpler model that makes it easier to observe and interpret.

```
library(MASS)
```

```
Attaching package: 'MASS'

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

    select
```

```
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),
         minutes = as.factor(minutes))

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

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

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -3.051243   2.321364  -1.314   0.1887
surfaceGrass          -0.246180   0.296445  -0.830   0.4063
```

```
surfaceHard            -0.198074   0.179377  -1.104   0.2695
winner_age              0.047255   0.019366   2.440   0.0147 *
as.numeric(winner_ht)   0.008849   0.012878   0.687   0.4920
winner_handR            0.121194   0.291980   0.415   0.6781
winner_handU          -12.810120 535.411309  -0.024   0.9809
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 849.49  on 618  degrees of freedom
Residual deviance: 840.17  on 612  degrees of freedom
  (620 observations deleted due to missingness)
AIC: 854.17

Number of Fisher Scoring iterations: 12
```

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

```
Call:
glm(formula = minutes ~ surface + loser_age + as.numeric(loser_ht) +
    loser_hand, family = "binomial", data = tennis_binary)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -6.106356   2.286333  -2.671  0.00757 **
surfaceGrass         -0.211784   0.298861  -0.709  0.47855
surfaceHard          -0.201920   0.180160  -1.121  0.26238
loser_age             0.009084   0.019564   0.464  0.64240
as.numeric(loser_ht)  0.030503   0.012685   2.405  0.01618 *
loser_handR           0.537760   0.308438   1.743  0.08125 .
loser_handU          -0.434522   0.873335  -0.498  0.61881
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 849.49  on 618  degrees of freedom
Residual deviance: 837.20  on 612  degrees of freedom
  (620 observations deleted due to missingness)
AIC: 851.2

Number of Fisher Scoring iterations: 4
```

```
exp(coef(loser_mins))
```

```
       (Intercept)         surfaceGrass          surfaceHard
       0.002228658          0.809139613          0.817160408
         loser_age as.numeric(loser_ht)          loser_handR
       1.009125753          1.030972719          1.712166551
       loser_handU
       0.647574123
```

```
exp(coef(winner_mins))
```

```
        (Intercept)          surfaceGrass           surfaceHard
       4.730011e-02          7.817814e-01          8.203089e-01
         winner_age as.numeric(winner_ht)          winner_handR
       1.048390e+00          1.008888e+00          1.128843e+00
       winner_handU
       2.732973e-06
```

We found the average duration of a match from all of the data provided, which happened to be 101.3 minutes. We will use this to determine if the court type causes a match to last longer or shorter than the average duration of a match throughout our project.

Loser For the loser of a match on a court with a surface of grass is predicted to have 0.810 times the odds of the match running over 101.3 minutes compared to a surface of clay while adjusting for age, height, and playing hand.

For the loser of a match on a court with a surface of hard is predicted to have 0.817 times the odds of the match running over 101.3 minutes compared to a surface of clay while adjusting for age, height, and playing hand.

For the loser of a match on a court with an age one year older than another is predicted to have 1.00913 times the odds of the match running over 101.3 minutes while adjusting for playing surface, height, and playing hand.

For the loser of a match on a court with a height 1 cm taller than another is predicted to have 1.031 times the odds of the match running over 101.3 minutes while adjusting for playing surface, age, and playing hand.

For the loser of a match on a court who plays with their right hand is predicted to have 1.712 times the odds of the match running over 101.3 minutes, compared to that of a left handed player while adjusting for playing surface, age, and height.

For the loser of a match on a court who plays with either hand is predicted to have 0.648 times the odds of the match running over 101.3 minutes, compared to that of a left handed player while adjusting for playing surface, age, and height.

Winner For the winner of a match on a court with a surface of grass is predicted to have 0.0782 times the odds of the match running over 101.3 minutes compared to a surface of clay while adjusting for age, height, and playing hand.

For the winner of a match on a court with a surface of hard is predicted to have 0.0820 times the odds of the match running over 101.3 minutes compared to a surface of clay while adjusting for age, height, and playing hand.

For the winner of a match on a court with an age one year older than another is predicted to have 1.0484 times the odds of the match running over 101.3 minutes while adjusting for playing surface, height, and playing hand.

For the winner of a match on a court with a height 1 cm taller than another is predicted to have 1.00889 times the odds of the match running over 101.3 minutes while adjusting for playing surface, age, and playing hand.

For the winner of a match on a court who plays with their right hand is predicted to have 1.129 times the odds of the match running over 101.3 minutes, compared to that of a left handed player while adjusting for playing surface, age, and height.

For the winner of a match on a court who plays with either hand is predicted to have 0.00000273 times the odds of the match running over 101.3 minutes, compared to that of a left handed player while adjusting for playing surface, age, and height.

We then checked to see if the linearity condition was met for the continuous predictors that we used on our variables. Our plots indicated that the winner age and loser age passed the linearity assumption. The linearity condition was not met for the predictors winner_ht and loser_ht. To address this violation of the linearity assumption, we used a log transformation on these variables.

We produced two models as seen below:

p(1-p) = the odds of the match duration being above 101.3 minutes.

$$log(p/(1-p)) = \beta_0 + \beta_1(surface) + \beta_2(winner\_age) + \beta_3(log(winner\_ht)) + \beta_4(winner\_hand)$$

$$log(p/(1-p)) = \beta_0 + \beta_1(surface) + \beta_2(log(loser\_age)) + \beta_3(log(loser\_ht)) + \beta_4(loser\_hand)$$

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

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

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -9.07179   11.46850  -0.791   0.4289
surfaceGrass                -0.24619    0.29645  -0.830   0.4063
surfaceHard                 -0.19786    0.17937  -1.103   0.2700
winner_age                   0.04722    0.01937   2.438   0.0148 *
log(as.numeric(winner_ht))   1.46557    2.22021   0.660   0.5092
winner_handR                 0.12186    0.29193   0.417   0.6764
winner_handU               -12.80678  535.41132  -0.024   0.9809
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 849.49  on 618  degrees of freedom
Residual deviance: 840.20  on 612  degrees of freedom
  (620 observations deleted due to missingness)
AIC: 854.2

Number of Fisher Scoring iterations: 12
```

```
loser_mins <- glm(minutes ~ surface + loser_age +
                    log(as.numeric(loser_ht)) + loser_hand, data = tennis_binary,
                    famil = "binomial")
summary(loser_mins)
```

```
Call:
glm(formula = minutes ~ surface + loser_age + log(as.numeric(loser_ht)) +
    loser_hand, family = "binomial", data = tennis_binary)

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)               -28.530883  11.347591  -2.514   0.0119 *
surfaceGrass               -0.210907   0.298914  -0.706   0.4805
surfaceHard                -0.201553   0.180189  -1.119   0.2633
loser_age                   0.009225   0.019566   0.471   0.6373
log(as.numeric(loser_ht))   5.375154   2.197977   2.446   0.0145 *
loser_handR                 0.539225   0.308468   1.748   0.0805 .
loser_handU                -0.434222   0.873307  -0.497   0.6190
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 849.49  on 618  degrees of freedom
Residual deviance: 836.99  on 612  degrees of freedom
  (620 observations deleted due to missingness)
AIC: 850.99

Number of Fisher Scoring iterations: 4
```

Results

After transforming the height variable to address the violation of linearity, we analyzed each estimate at the p=0.05 level.

For losers:

- Surface doesn't have a significant effect on the match duration being above or below the threshold of 101.3 minutes
- The age of the loser has no significant effect on the match duration
- The height of the loser has a significant effect on the match duration
- The dominant playing hand of the loser doesn't have a significant effect on the match duration

For winners:

- Surface doesn't have a significant effect on the match duration
- The age of the winner has a significant effect on the match duration
- The height of the winner has no significant effect on the match duration
- The dominant playing had of the winner doesn't have a significant effect on the match duration

Overall, the analysis suggest that the age of the winner and height of the loser significantly influence the duration of tennis matches, while other factors seem to show no significant effects.

Discussion

Overall, we found that the duration of matches is not actually predicted by the surface of the court, but rather age for winners and height for losers. Originally, we did some research and found that clay courts typically host shorter matches, so we expected to see this in the results. We decided to research further and found that in earlier years the court type mattered more and clay had shorter match durations, but in recent years this is not as significant. With a p value of 0.148, we found age to be significant at the alpha = 0.05 level and in our results of match duration. This proves that if the winner is older than the loser, the odds the match lasts longer than 101.3 minutes is 1.0484 times the odds that of players with equal age. This shows that those who are younger typically have shorter games, so if an individual is older than their opponent, they should be prepared to have a match longer than average to beat their opponent. In a similar manner with a significant p value of 0.0145 at the alpha = 0.05 value , if the loser is taller than their opponent they have 1.031 times the odds of their shorter opponent at the match lasting longer than 101.3 minutes which is the average. In the broader context of tennis matches, if the opponent is taller and loses it is expected that their match will last longer than average, and if the opponent is older and wins the match is expected to last longer than average. There may be limitations in our data since we do not mention bmi of the players to compare their weights, and height does not necessarily correspond to wingspan which may also be an advantage of tennis players. Additionally, we do not mention the country they play in, which may be seen as a home field advantage, so we don't talk about their home country and the impact of playing inside versus out has on their game durations. In the future, if we were to continue our research, we think it would be beneficial to see if longer game durations typically led to more wins or if shorter games did. The data is reliable, and the creator, Jeff Sackmann gets his data from Tennis Abstract, and he includes some of the work he created with his own set as well.He has updated the stats for years, and just recently made his finishing touches on the 2023 dataset on March 19 2024, so he has put a lot of effort into his data.

Sackmann, Jeff. 19 March 2024. *WTA Tennis Rankings, Results, and Stats.* creative commons. https://github.com/JeffSackmann/tennis_wta/blob/master/README.md.