readME

Aristoteles Canahuati

readME

```r
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   4.0.0     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- 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 becom
```

```r
library(dplyr)
gooddata <- read.csv("/Users/acanahuati/Downloads/NEWDATAvb.csv")
```

Variable Codings: - Player#: Jersey number - Height: Height in inches - Sets.Played: Sets played during the 2023-2024 season. - Aces: Aces acquired during the 2023-2024 season. - Digs: Digs acquired during the 2023-2024 season. - Year: 1: Freshman, 2: Sophomore, 3: Junior, 4: Senior - Kills: Kills acquired during the 2023-2024 season. - Blocks: Blocks acquired during the 2023-2024 season. - Position: 1/Setter, 2/Rightside, 3/Middle, 4/Outside, 5/Libero - Errors: Cumulative errors during the 2023-2024 season. - Points: Points scored over the course of the 2023-2024 season. - Hitting Percentage: (kills-errors)/attempts

Welcome to my personal dataset project for my Foundations of Data Science class at Vassar college. The goal of this project is to explore different relationships within the sport of volleyball, through the lens of our very own Vassar Men's Volleyball team (the 2023-2024 season). A few hypotheses: Certain positions will have much shorter average heights than others. Height may have a correlation with sets played or points. Older players will score more points. Let's see!

Before we begin, I would like to clarify my data obtaining / cleaning process. I obtained all of my data from one source (Vassar's Athletics Website): https://www.vassarathletics.com/sports/2009/3/5/MVB_ The original dataet was pretty large so the first thing I did was trim it down. The reason I chose the 2023-2024 season is because I feel it's roster was the most representative of an average NCAA D3 team in terms of height distribution. Creating my own dataset as a subset of the original one came with its issues: I had to turn turn the columns into rows and the rows into columns (mostly a result of my human error). I also chose the variables I found most relevant to the questions I was planning on asking (see above).

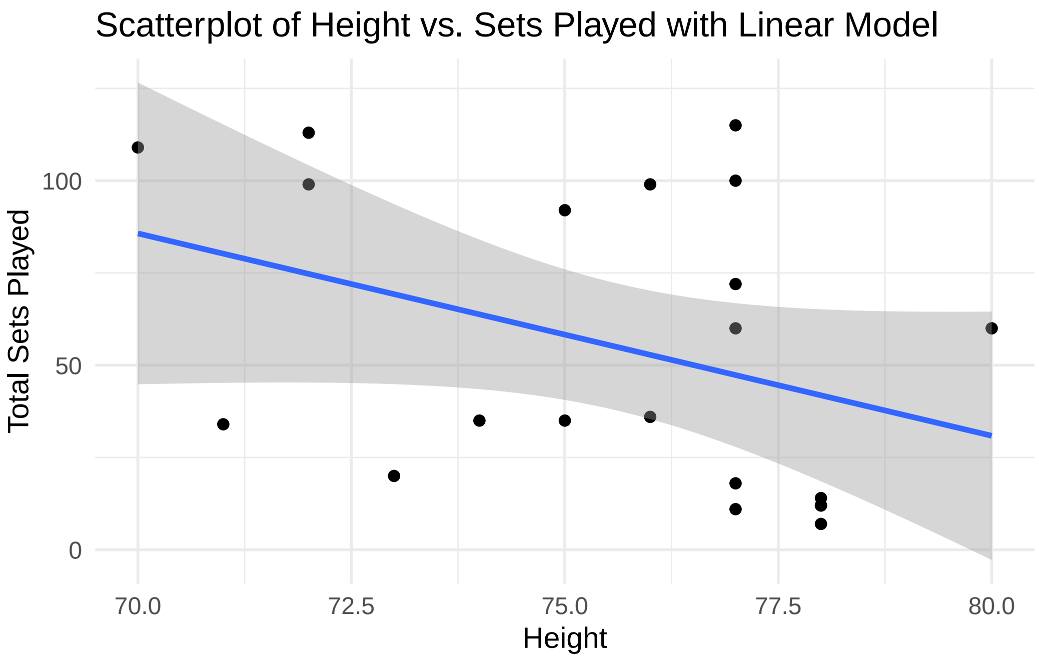
PLOTS

```r
model <- lm(Sets.Played ~ Height, data = gooddata)
summary(model)
```

```

Call:
lm(formula = Sets.Played ~ Height, data = gooddata)

Residuals:
   Min     1Q Median     3Q    Max 
-49.26 -29.85 -16.82  29.11  67.67 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  469.500    238.573   1.968   0.0638 .
Height        -5.483      3.153  -1.739   0.0982 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 37.7 on 19 degrees of freedom
Multiple R-squared:  0.1373,	Adjusted R-squared:  0.09188 
F-statistic: 3.024 on 1 and 19 DF,  p-value: 0.09824
```

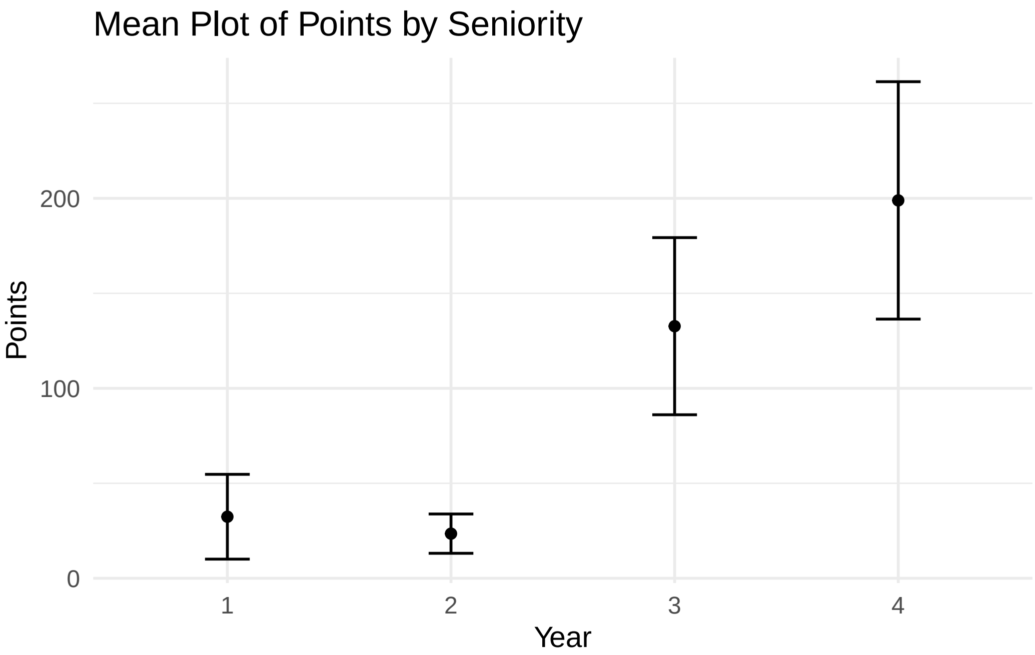
```r
# Linear Model

ggplot(gooddata, aes(x = Height, y = Sets.Played)) +
geom_point() +
geom_smooth(method = "lm") +
theme_minimal() +
   labs(
    title = "Scatterplot of Height vs. Sets Played with Linear Model",
    x = "Height",
    y = "Total Sets Played"
  )
```

```
`geom_smooth()` using formula = 'y ~ x'
```



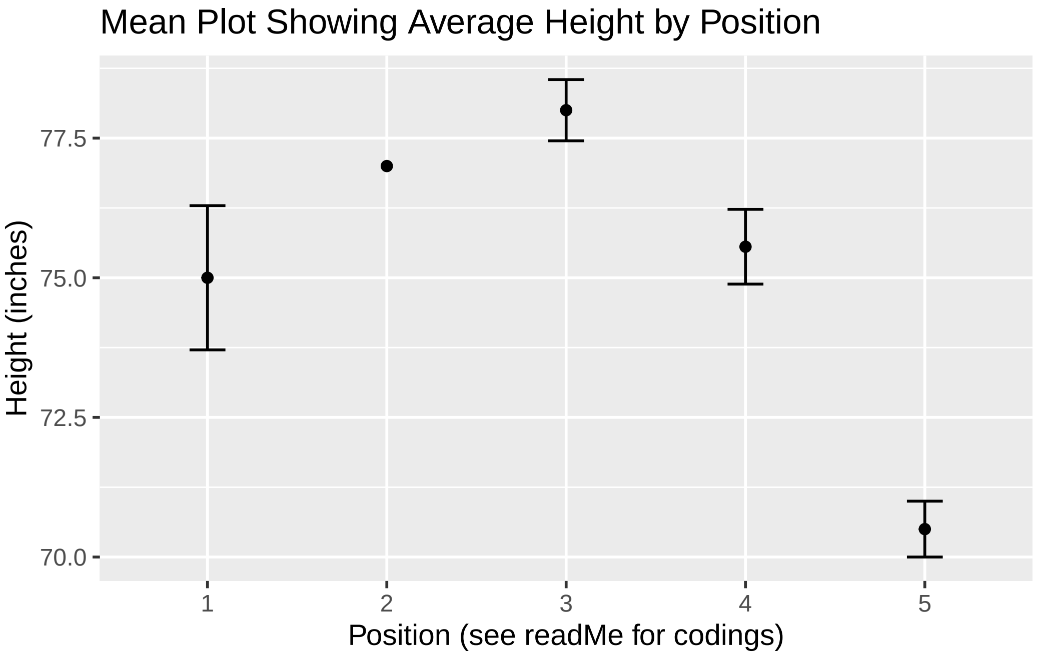
No apparent linear relationship (R-squared of 0.14 very low) between height and sets played. Overall correlation in lm is actually negative (negative slope), suggesting the opposite of what we hypothesized.

```
# Mean Plot with Error Bars Year vs. Points

ggplot(gooddata, aes(x = factor(Year), y = Points)) +
  stat_summary(fun = "mean", geom = "point") +
  stat_summary(fun.data = "mean_se", geom = "errorbar", width = 0.2) + theme_minimal() +
   labs(
    title = "Mean Plot of Points by Seniority",
    x = "Year",
    y = "Points"
  )
```



Mean value of points scored categorized by year seems to be increasing with the years (Sophomores being an outlier but not by much). Seniors have by far the most average points scored and Juniors are second by a decent margin of around 70. It may be plausible to say that, on average, older players scored more points on this team. This seems to be a trend on a lot of teams in the NCAA, as seniority seems to be associated with more responsibility.

```
# Mean Plot wwith Error Bars for Height vs Position.

ggplot(gooddata, aes(x = factor(Position), y = Height)) +
  stat_summary(fun = "mean", geom = "point") +
  stat_summary(fun.data = "mean_se", geom = "errorbar", width = 0.2) +
    labs(
    title = "Mean Plot Showing Average Height by Position",
    x = "Position (see readMe for codings)",
    y = "Height (inches)"
  )
```



This plot is very much in line with how positions are stereotyped in volleyball: middles (3) are typically the tallest on the court, as they have to block every ball; right-sides (2) are typically second tallest as they are meant to be heavy hitters that bring power and athleticism to the table; outside hitters (4) are typically shorter than the rightsides and middles, but still aren't super short as they still contribute to the offense; setters (1) don't need to be very tall to succeed (although it does help) so it makes sense that their average height isn't very tall or very short; liberos (5) are purely focused on defense and are usually the shortest players on the court, which is very much in line with the graph.