First Visualizations + Analyses

Aristoteles Canahuati

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!

```
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 becor
```

```
library(dplyr)
data <- read.csv("/Users/acanahuati/Downloads/updated_dataset.csv")
```

Fixing original dataset

```
data <- na.omit(data) # getting rid of missing values as they're missing because I couldn't i

gooddata <- t(data) # making rows columns and columns rows for structure

gooddata <- as.data.frame(gooddata)
```

```r
colnames(gooddata) <- gooddata[1, ]

gooddata <- gooddata[-1, ] #removing first row

str(gooddata)
```

```
'data.frame':   21 obs. of  7 variables:
 $ Height     : chr  "78.0" " 70" "74.000" "73.000" ...
 $ Sets.Played: chr  "14.0" "109" "35.000" "20.000" ...
 $ Year       : chr  " 3.0" "  1" " 1.000" " 1.000" ...
 $ Position   : chr  " 1.0" "  5" " 1.000" " 4.000" ...
 $ Errors     : chr  " 2.0" "  1" " 1.000" "13.000" ...
 $ Points     : chr  " 4.0" "  1" " 9.000" "14.000" ...
 $ HittingPct : chr  "-0.2" "  0" "-0.167" "-0.024" ...
```

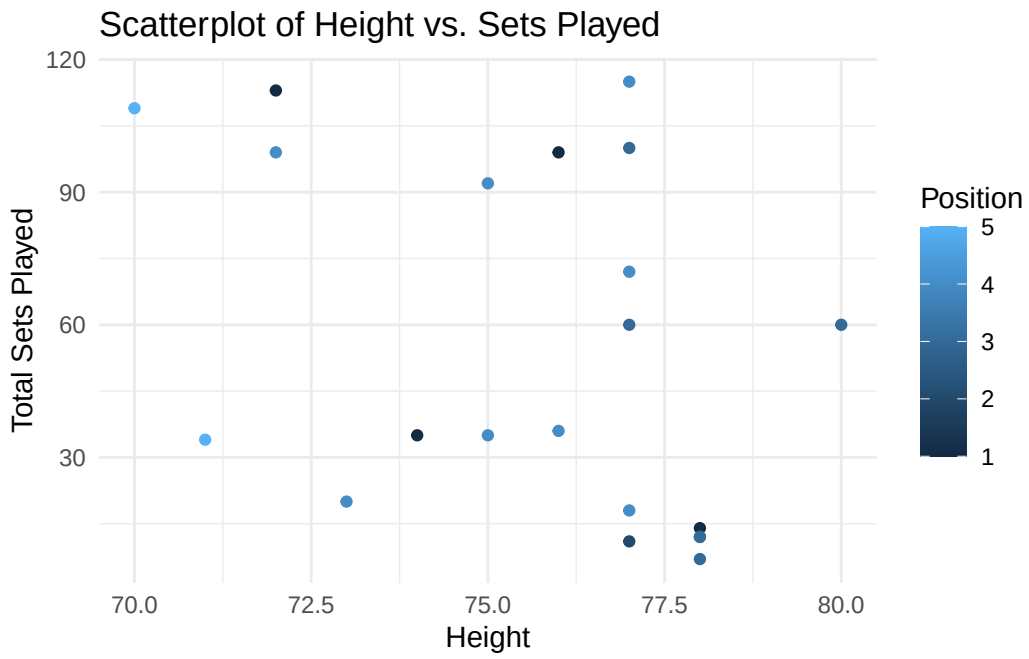
```r
head(gooddata)
```

```
    Height Sets.Played    Year Position  Errors  Points HittingPct
X1    78.0        14.0     3.0      1.0     2.0     4.0       -0.2
X3      70         109       1        5       1       1          0
X4  74.000      35.000   1.000    1.000   1.000   9.000     -0.167
X5  73.000      20.000   1.000    4.000  13.000  14.000     -0.024
X6  77.000     115.000   4.000    4.000  84.000 343.000      0.356
X7  76.000      36.000   2.000    4.000  14.000  70.500      0.361
```

```r
gooddata[] <- lapply(gooddata, function(col) {
  if (is.character(col)) {
    # Replace things like "5.0", "5.00", "12.000" → "5", "12"
    gsub("\\.0+$", "", col)
  } else {
    col
  }
})

gooddata[] <- lapply(gooddata, function(col) as.numeric(col)) # making variables numeric


# Had to adjust dataset so that variables were numeric and coherent as a whole.
```

PLOTS

```
# Scatterplot Height vs. Sets Played
ggplot(gooddata, aes(x= Height, y= Sets.Played, color = Position)) + geom_point() + theme_mi
   labs(
    title = "Scatterplot of Height vs. Sets Played",
    x = "Height",
    y = "Total Sets Played"
  )
```



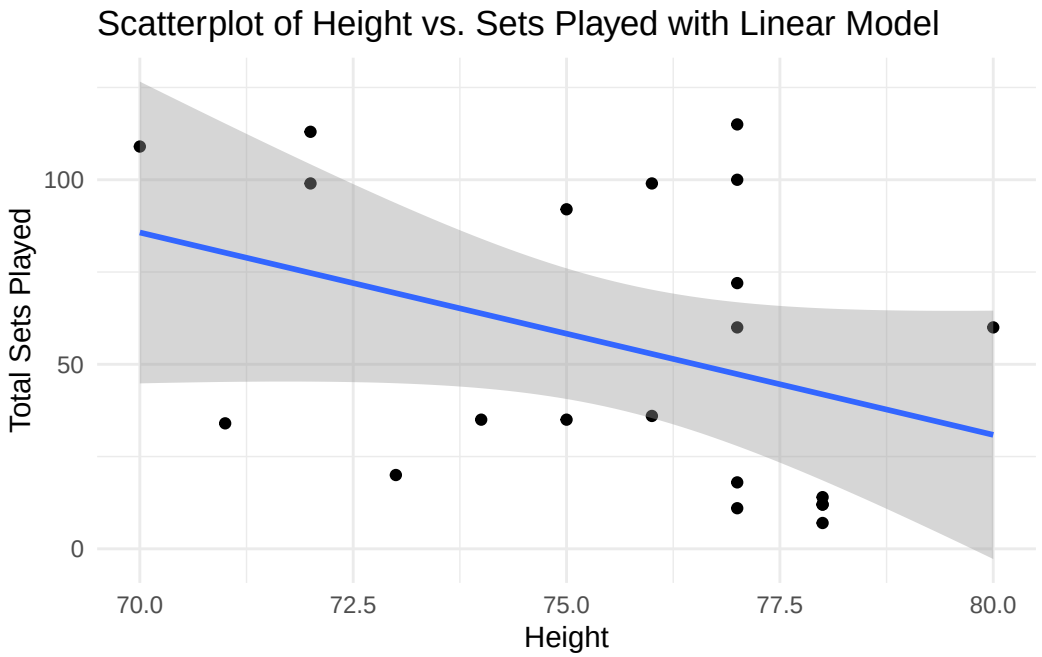

```
model <- lm(Sets.Played ~ Height, data = gooddata)
summary(model)
```

```
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
```

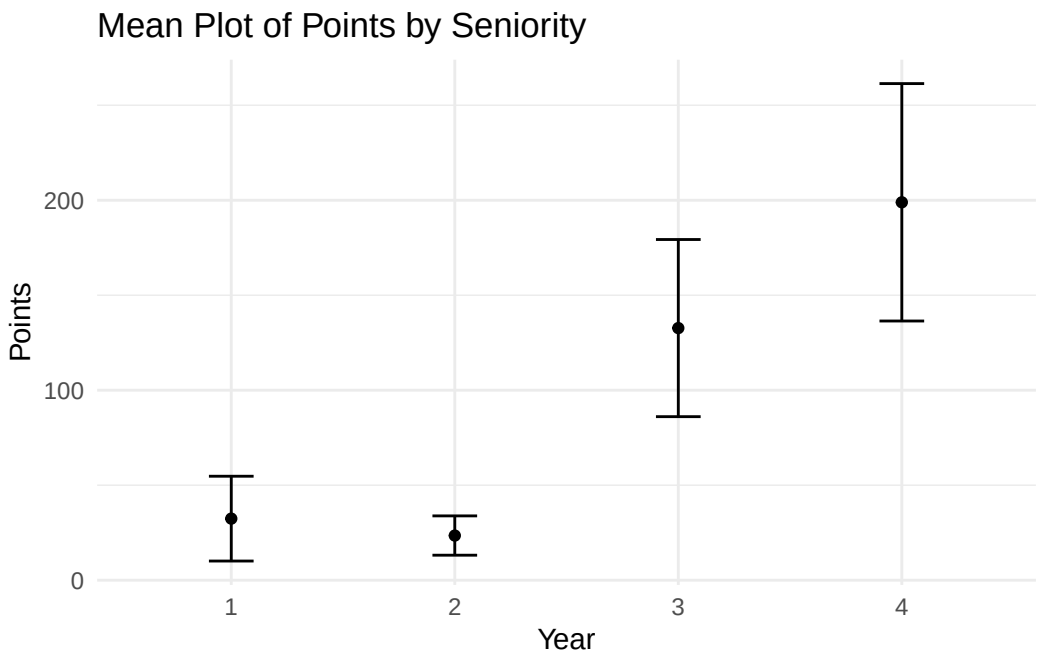
```r
# Linear Model

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'
```



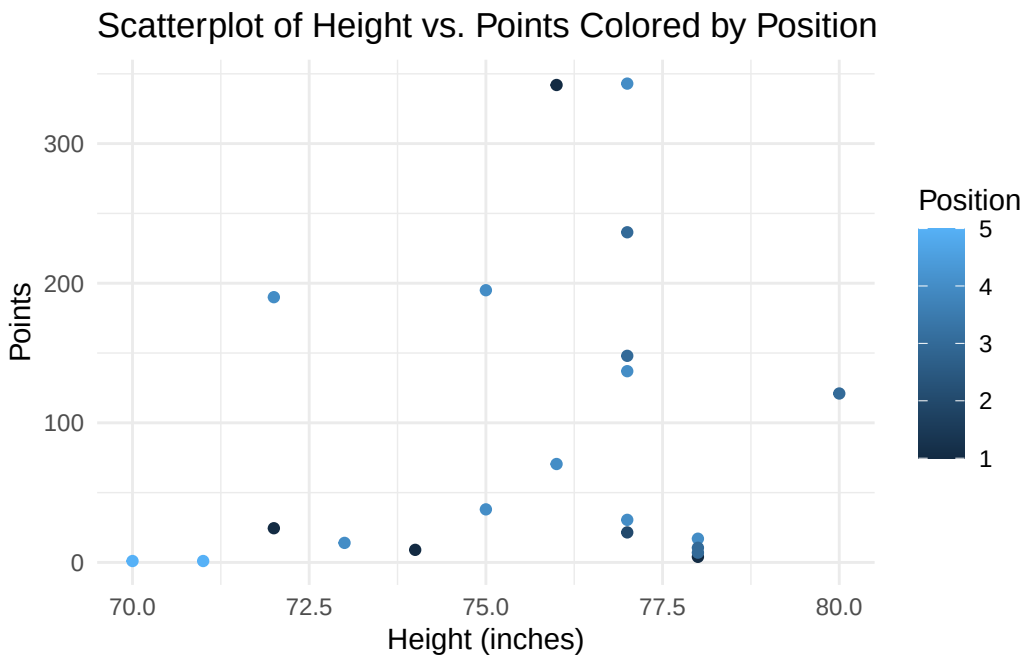
No apparent linear relationship between height and sets played, even by position. Overall correlation in lm is actually negative, suggesting the opposite of what we hypothesized. However, R-squared of 0.14 suggests model not really applicable.

```
# 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.

```
# Scatterplot Height vs. Points
ggplot(gooddata, aes(x= Height, y= Points, color = Position)) + geom_point() + theme_minimal
   labs(
    title = "Scatterplot of Height vs. Points Colored by Position",
    x = "Height (inches)",
    y = "Points"
  )
```

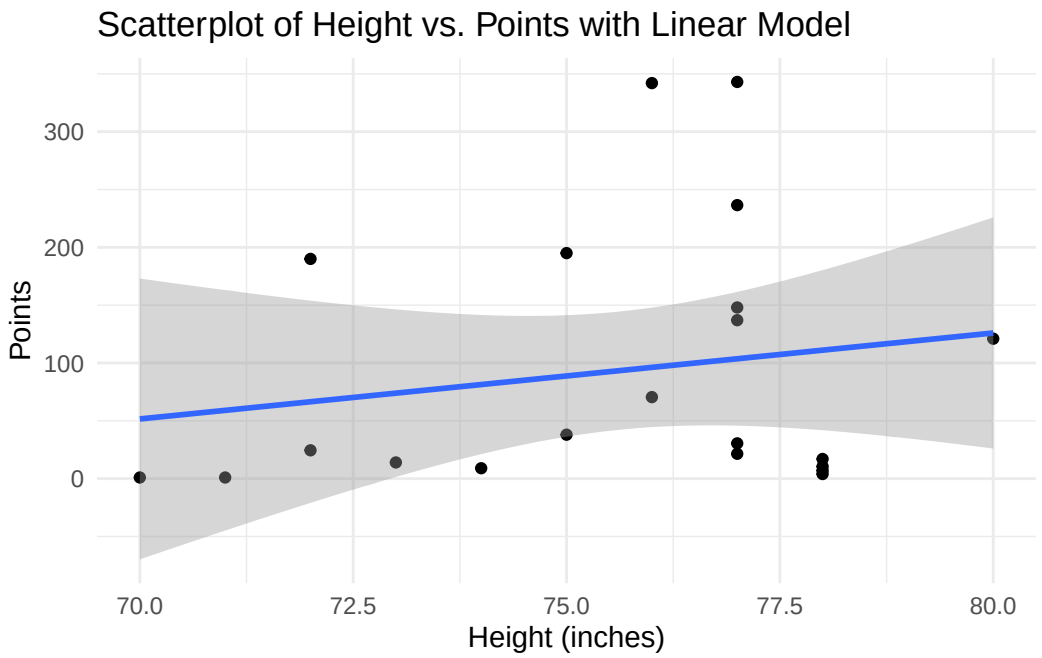



```
model <- lm(Points ~ Height, data = gooddata)
summary(model)
```

```
Call:
lm(formula = Points ~ Height, data = gooddata)

Residuals:
    Min      1Q  Median      3Q     Max
-107.08  -73.14  -50.62   44.35  245.79

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -468.686    708.454  -0.662    0.516
Height         7.433      9.363   0.794    0.437

Residual standard error: 111.9 on 19 degrees of freedom
Multiple R-squared:  0.0321,    Adjusted R-squared:  -0.01884
F-statistic: 0.6302 on 1 and 19 DF,  p-value: 0.4371
```

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

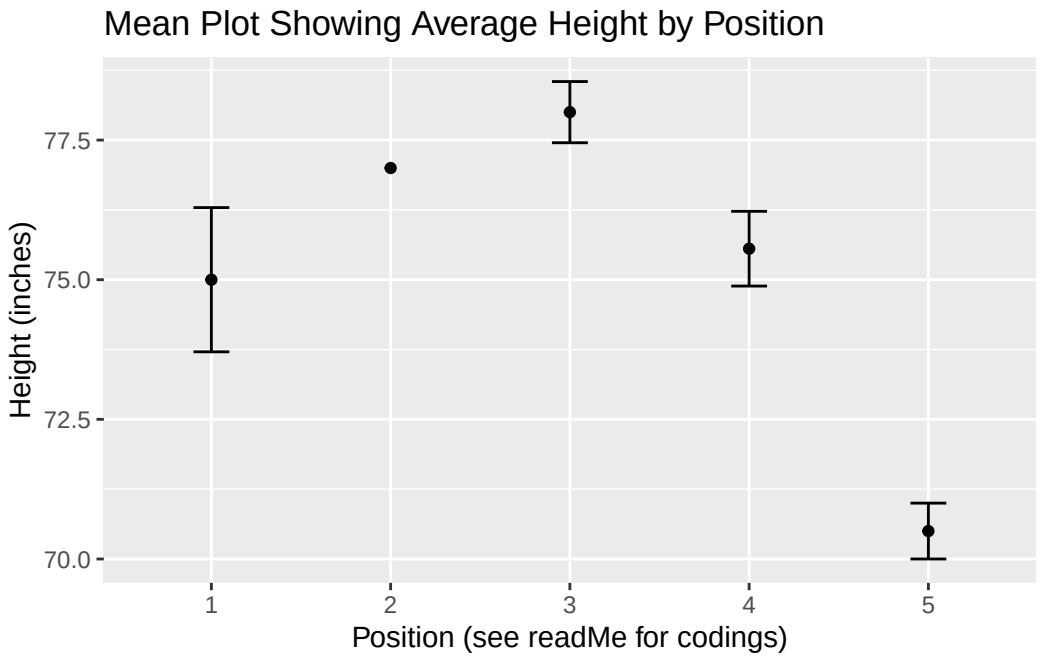
```
`geom_smooth()` using formula = 'y ~ x'
```



No apparent linear relationship between height and points, even by position. Overall correlation in lm is positive, which would be in line with what we hypothesized. However, R-squared of 0.03 suggests model not really applicable.

```
# 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.