

HW5

2024-11-08

Question 1

The following is my code for the actual dataset. The total players in the 2024 roster is 192. I made the code run the website and retrieve the information from the website and gather it here on R. I then created my variables I would be using in the rest of the assignment.

```
library(rvest)
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(ggplot2)

url <- "https://www.basketball-reference.com/wnba/years/2024_totals.html"
page <- read_html(url)
player_data <- page %>%
  html_table() %>%
  .[[1]]
player_data <- player_data[, !duplicated(names(player_data))]
player_data <- player_data %>%
  select(Player, MP, `3P%`) %>%
  filter(!is.na(MP))
total_players <- nrow(player_data)
print(player_data)

## # A tibble: 192 x 3
##   Player      MP   `3P%`
##   <chr>      <chr> <chr>
## 1 Lindsay Allen  950  ".292"
## 2 Rebecca Allen  447  ".352"
## 3 Laeticia Amihere 83   ""
## 4 Ariel Atkins  1196 ".357"
## 5 Amy Atwell    59   ".231"
## 6 Shakira Austin 239   ".250"
## 7 Kierstan Bell  43   ".500"
## 8 Grace Berger   102   ".400"
## 9 Caitlin Bickle  15   ".000"
## 10 DeWanna Bonner 1271  ".294"
## # i 182 more rows
```

```
print(paste("Total players in 2024 roster:", total_players))
```

```
## [1] "Total players in 2024 roster: 192"
```

Question 2

I created a graph for the distribution of three point percentages, with the x axis labeled the percentages and the y axis labeled with the number of players.

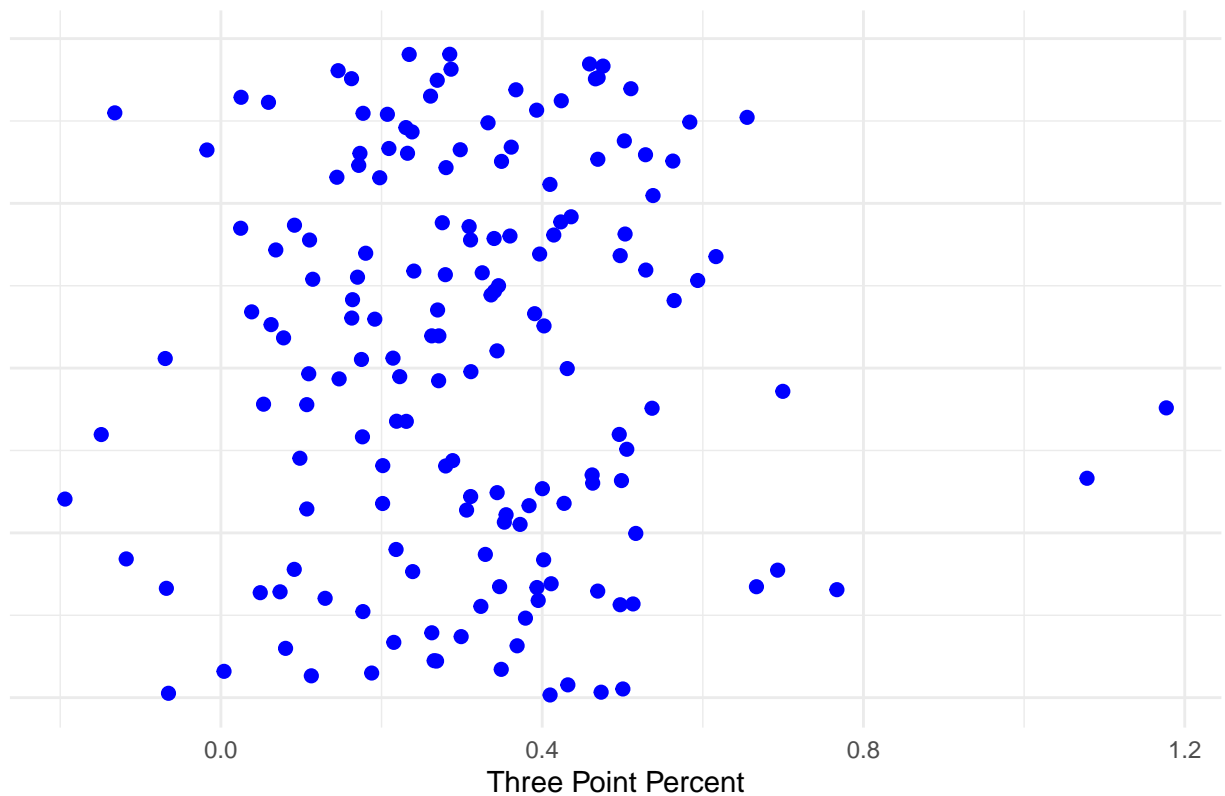
```
player_data$`3P%` <- as.numeric(player_data$`3P%`)
```

```
## Warning: NAs introduced by coercion
```

```
ggplot(player_data, aes(x = `3P%`, y = 1)) +  
  geom_jitter(color = "blue", size = 2, width = 0.2) +  
  labs(title = "Three Point Shooting Percents (2024)",  
       x = "Three Point Percent") +  
  theme_minimal() +  
  theme(axis.text.y = element_blank(),  
        axis.ticks.y = element_blank(),  
        axis.title.y = element_blank())
```

```
## Warning: Removed 30 rows containing missing values or values outside the scale range  
## (`geom_point()`).
```

Three Point Shooting Percents (2024)



Question 3

I created a graph comparing the relationship between minutes played and three point percentage using the plot function. I labeled the x axis with minutes played and y axis with the three point percentage.

```
ggplot(player_data, aes(x = MP, y = `3P%`)) +  
  geom_point(color = "blue", size = 2) +  
  labs(title = "Comparison of Three-Point Percentage and Minutes Played (2024)",  
        x = "Minutes Played",  
        y = "Three-Point Percentage") +  
  theme_minimal() +  
  theme(axis.title.x = element_text(size = 12),  
        axis.title.y = element_text(size = 12),  
        plot.title = element_text(hjust = 0.5, size = 14)) +  
  geom_smooth(method = "lm", color = "red", linetype = "dashed")
```

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

```
## Warning: Removed 30 rows containing non-finite outside the scale range
```

```
## (`stat_smooth()`).
```

```
## Warning: Removed 30 rows containing missing values or values outside the scale range
```

```
## (`geom_point()`).
```

