NBA Stats From 1999 - 2023

Research Question 1: Is a player's 3-point shooting percentage related to free throw shooting percentage?

This research examines the relationship between 3-point shooting and free-throw shooting. Understanding this correlation may provide insights into how shooting mechanics or player training affect performance in these distinct areas.

Data

We get the NBA player statistics from 1999 to 2023 from the NBA website. The data was collected by NBA analysts.

The data meets the FAIR principles:

- Find able: The data set has clear identifiers and structured attributes.
- Accessible: The data set is standard tabular data, which can be saved as CSV file.
- Inter operable: Data column names and values align with standard basketball statistical terminology.
- Reusable: The data set has vary variables from multiple years, and it's convenient to do different analyses.

We will be researching about the correlation between the 3-point goal percentage (3P%) and free throw goal percentage (FT%). These two attributes are part of data set.

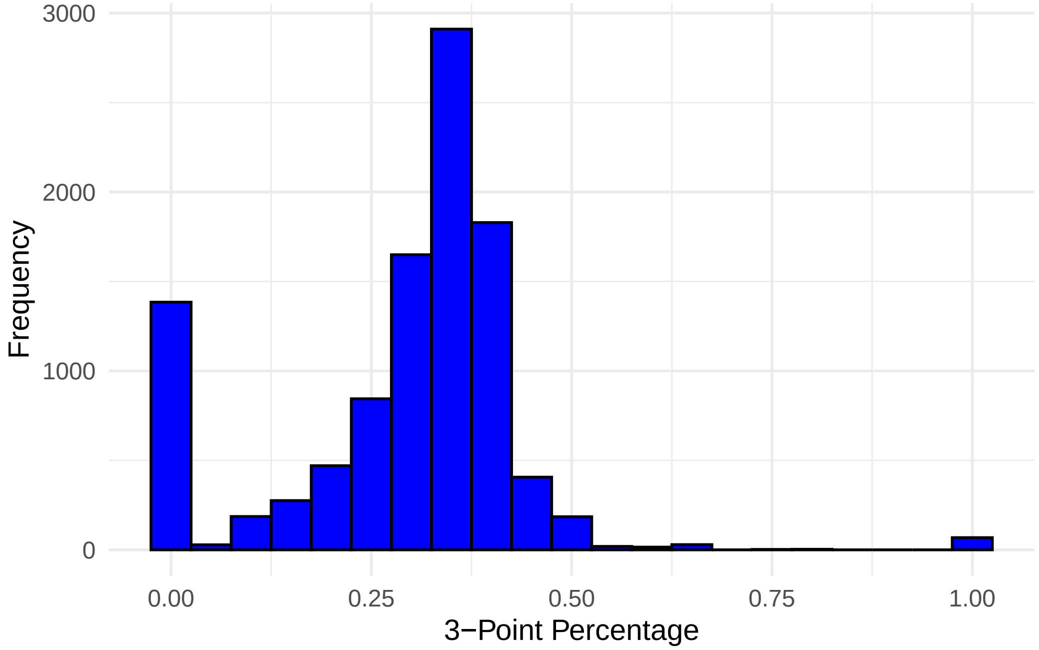


Figure 1: Distribution of 3-Point Percentage

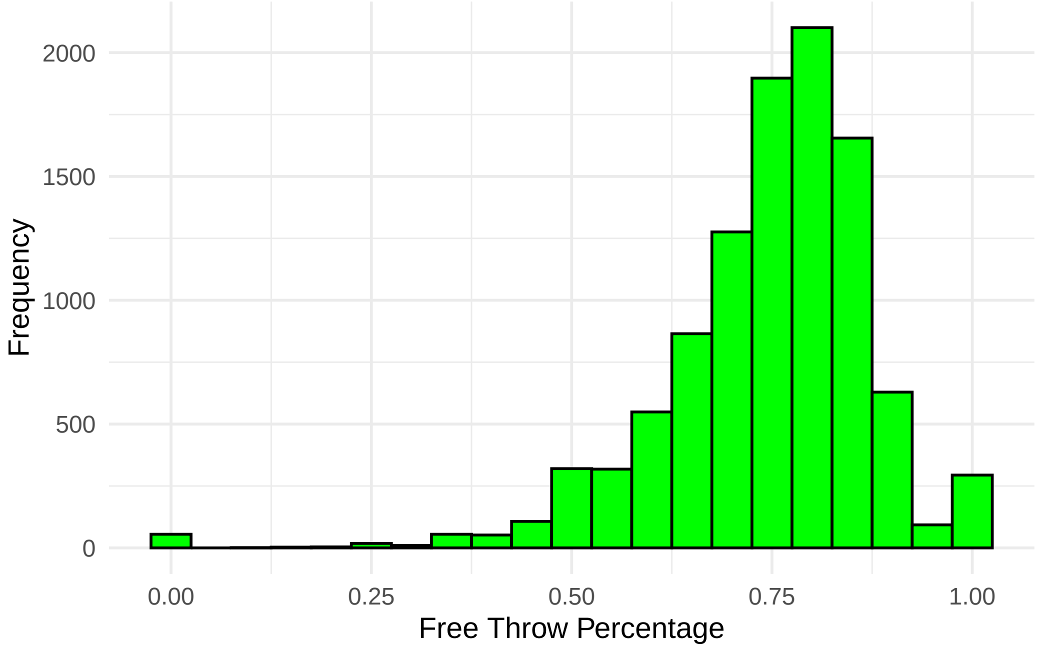


Figure 2: Distribution of Free Throw Percentage

Data Visualization

The distribution of 3-point percentages is right-skewed, with the majority of players concentrated between 0.2 and 0.35, and a notable cluster at 0.0 (players with no successful 3-point shots).

The distribution of free throw percentages is left-skewed, with most players concentrated between 0.7 and 0.9, indicating a higher proficiency in free throws compared to 3-point shooting.

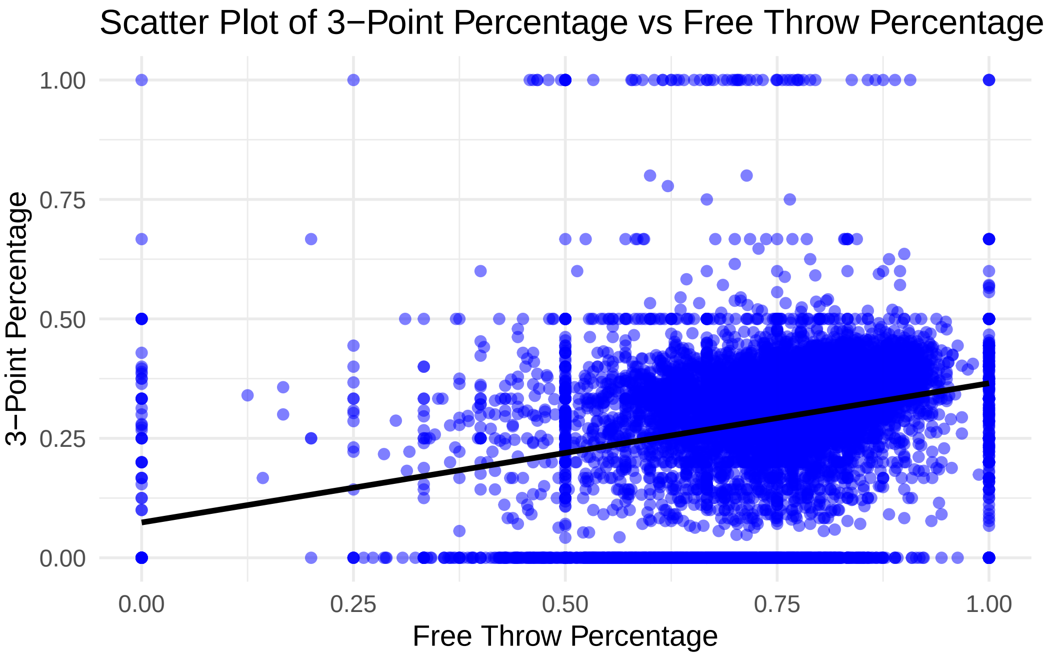


Figure 3: Scatter Plot of 3-Point Percentage vs Free Throw Percentage

While the regression line shows a positive trend, the slope is quite shallow, indicating a relatively weak correlation between these two metrics.

Result Analysis

We did Pearson correlation test between 3-point percentage and free throw percentage, and got the following result:

```
[1] "Pearson correlation coefficient: r = 0.26, p-value = 4.22e-156"
```

The Pearson correlation test between 3-point percentage and free throw percentage shows a correlation coefficient of 0.26 and a p-value of 4.22e-156. This indicates a weak positive correlation between the two variables. While the correlation is not strong, the extremely small p-value suggests that the relationship is statistically significant and not due to random chance. In practical terms, this means that players who perform better at free throws tend to have slightly better 3-point shooting percentages, but the connection between the two skills is limited. This result highlights that while both skills involve shooting mechanics, they are likely influenced by different factors such as shot distance, player roles, and situational conditions during games.

Research Question 2: How Does Age Impact Individual Player Statistics?

This research question examines the correlation between a player's age and their individual player statistics.

Data

Similarly to the first research question the data was taken from NBA player statistics from 1999 to 2023 from the NBA website. As stated above, the data meets the FAIR principles.

Data Loading & Processioning

```r
library(tidyverse)
library(dplyr)
library(ggplot2)
```

```r
NBAplayers = read.csv("NBA Players 1999-2023 - Sheet1.csv")
```

```r
NBAplayers3 <- NBAplayers %>%
  #Drop unnecessary columns
  select(-c(1, 4, 5, 9:21, 24, 29))

#Avg points per game
Ageppg <- NBAplayers3 %>%
  group_by(NBAplayers3[c(2)]) %>%
  summarise(Avgppg = mean(PTS, na.rm = TRUE))
```

```r
#Avg Assists Per Game
Ageast <- NBAplayers3 %>%
  group_by(NBAplayers3[c(2)]) %>%
  summarise(Avgast = mean(AST, na.rm = TRUE))

#Avg Offensive Rebounds
Ageorb <- NBAplayers3 %>%
  group_by(NBAplayers3[c(2)]) %>%
  summarise(Avgorb = mean(ORB, na.rm = TRUE))

#Avg Defensive Rebounds
Agedrb <- NBAplayers3 %>%
  group_by(NBAplayers3[c(2)]) %>%
  summarise(Avgdrb = mean(DRB, na.rm = TRUE))
```

```r
#Agestat dataframe
Agestat = data.frame(Ageppg,Ageast[-c(1)],
Ageorb[-c(1)],Agedrb[-c(1)])
```

Data Visualization

Player Age VS Average Points Per Game

```r
ggplot(
  data = Agestat,
  aes(x=Age..,y=Avgppg,fill=Age..)
) + geom_bar(stat = "identity", width = 0.5) +
labs(x = "Player Age",y = "Average Points Per Game",
title = "Average Points Scored at Different Ages in the NBA")
```

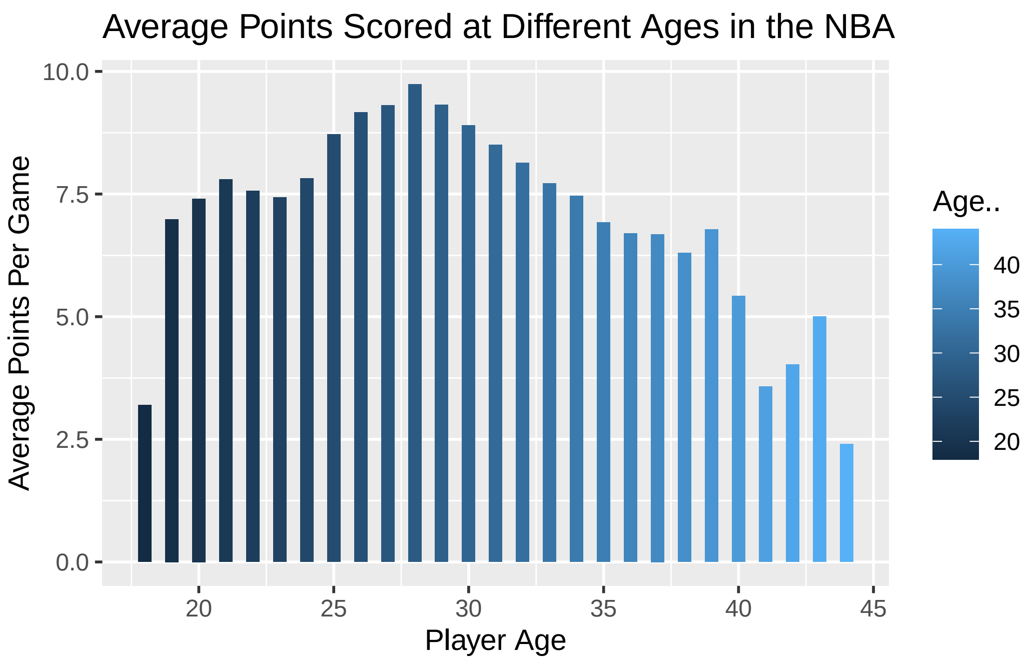


Figure 4

This figure shows negative concavity. It also shows that even with generally negative concavity, there was a spike in average points per game for the ages: 18, 39, 40, 42, and 44.

Player Age VS Average Assists Per Game

```
ggplot(data = Agestat,
       aes(x=Age..,y=Avgast,fill=Age..)
) + geom_bar(stat = "identity", width = 0.5) +
  labs(x = "Player Age", y = "Average Assists Per Game",
       title = "Average Assists Made at Different Ages in the NBA")
```

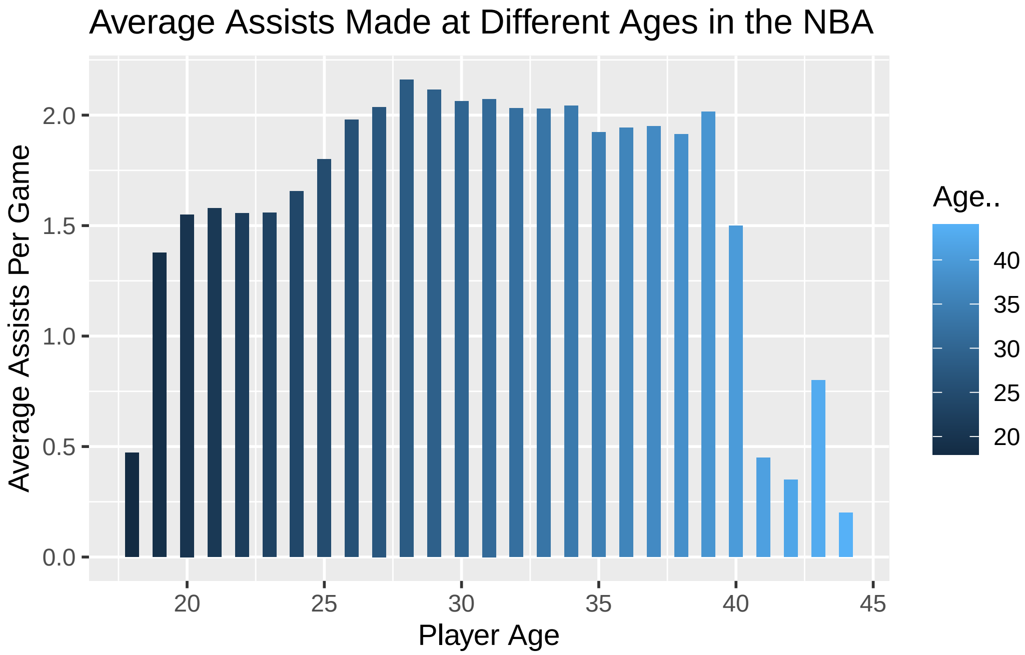


Figure 5

This figure has a more aggressive negative concave to it. Like previously we see that at age 39 and 44 there is a large spike in average assists per game.

Player Age VS Average Rebounds Per Game

```
ggplot(data = Agestat,
       aes(x=Age..,y=(Avgorb+Avgdrb),fill=Age..)
) + geom_bar(stat = "identity", width = 0.5) +
  labs(x = "Player Age", y = "Average Rebounds Per Game",
       title = "Average Rebounds at Different Ages in the NBA")
```

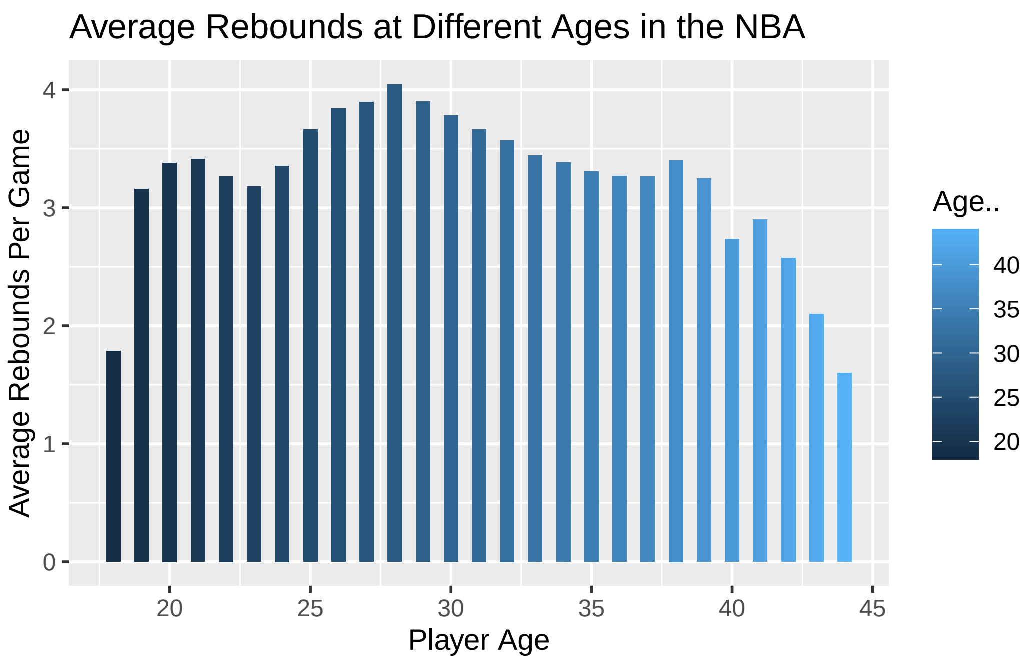


Figure 6

This figure like the others has negative concavity as well. Also like the previous figures there is a spike in average rebounds per game for the ages: 42 and 44.

Result Analysis

From looking at the above visuals we can conclude that generally speaking players in their mid 20's to mid 30's perform the best in all 3 aspects. We can see there are some outliers among all 3 visuals occurring at the age 39 and higher. Due to the negative concavity, we can conclude that the general body of players start out with lower stats, progress to peak throughout the middle of their career, then decrease in performance as they are about to retire.

Research Question 3: Are offensive or defensive statistics more important to a team's success?

This research question is intended to search for correlation between a team's win rate and their offensive or defensive statistics to find out whether offense or defense is more important to a team's success.

Data

The data for this research question was found on basketball-reference.com, which is a site that records and maintains data from the NBA. The subset of data used is advanced team statistics from 1999-2023. A team's win rate in a particular season is calculated by their wins divided by total games played. The offense rating of a team is recorded on the website as the average number of points a team scores per 100 possessions while the defense rating of a team is recorded as the average number of points the team concedes per 100 possessions.

Data Visualizations

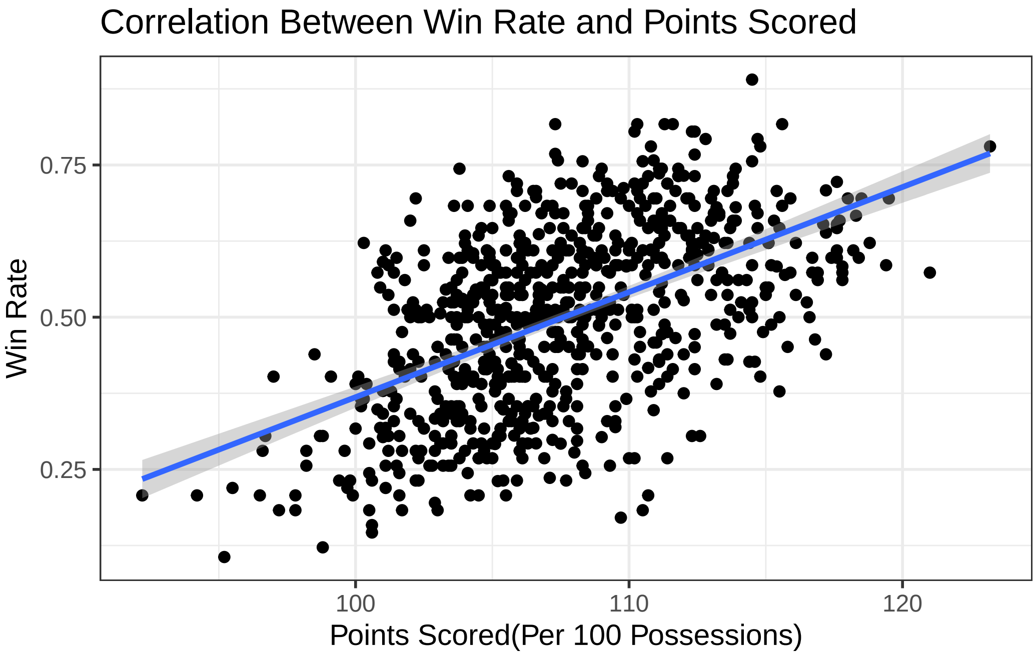
```
#Load libraries
library(tidyverse)
library(dplyr)
library(ggplot2)

#Read csv
NBAteamsA = read.csv("NBA Teams ADvanced 1999-2023 - Sheet1.csv")
View(NBAteamsA)

#Create WinRate comparison data frame
WinRate = (NBAteamsA[4]/(NBAteamsA[5]+NBAteamsA[4]))
WinRateComp = data.frame(NBAteamsA[2],WinRate,NBAteamsA[6],NBAteamsA[7],NBAteamsA[16])
View(WinRateComp)
```

Win Rate Vs Offense(Points Scored)

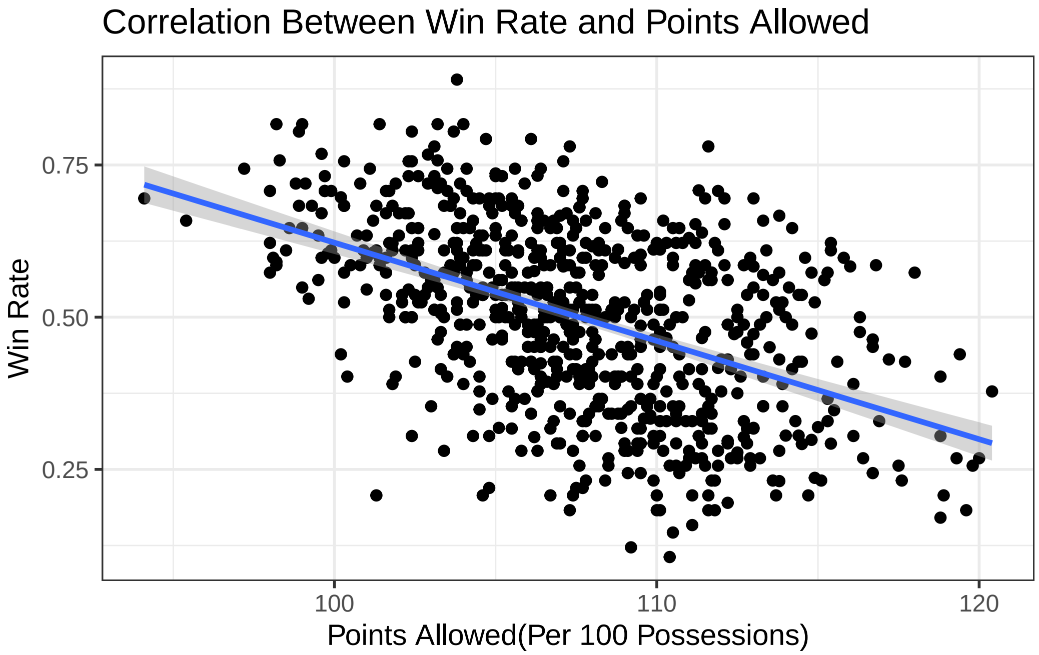
```
ggplot(
  data = WinRateComp,
  aes(x=as.numeric(unlist(WinRateComp[3])),y=as.numeric(unlist(WinRateComp[2])))
) +
  geom_point() +
  labs(x = "Points Scored(Per 100 Possessions)",
       y = "Win Rate",
       title = "Correlation Between Win Rate and Points Scored")+
 theme_bw() + stat_smooth(method = "lm",
                          formula = y ~ x)
```



Most of the data in the plot is clustered in the middle indicating some correlation exists. There are few outliers. The regression line shows a positive slope. Teams with higher average points scored tend to win more.

Win Rate Vs Defense(Points Allowed)

```
ggplot(
  data = WinRateComp,
  aes(x=as.numeric(unlist(WinRateComp[4])),y=as.numeric(unlist(WinRateComp[2])))
) +
  geom_point() +
  labs(x = "Points Allowed(Per 100 Possessions)",
       y = "Win Rate",
       title = "Correlation Between Win Rate and Points Allowed")+
  theme_bw() + stat_smooth(method = "lm",
                           formula = y ~ x)
```



Again most of the data is clustered in the center of the graph indicating a correlation exists. There are more outliers in this graph. The regression line shows a negative slope. Teams with higher average points allowed tend to win less.

Result Analysis

The graphics and correlation tests indicate that there is a moderately strong positive correlation between the number of points that a NBA team and a moderately strong negative correlation between the number of points a NBA team allows. The absolute values of the correlations are relatively close(0.54 for points scored and -0.48 for points allowed), but it appears scoring points is more important to a team winning than preventing the other team from scoring. Thus, offense has a stronger correlation with winning that defense in the NBA.

```r
#Signficance test for win rate and offense
correlation2 <- cor.test(WinRateComp$W..,WinRateComp$ORtg , method = "pearson")

correlation2_result <- paste0(
  "Pearson correlation coefficient: r = ", round(correlation2$estimate, 2),
  ", p-value = ", signif(correlation2$p.value, 3)
)
```

```
#Signficance test for win rate and defense
correlation3 <- cor.test(WinRateComp$W..,WinRateComp$DRtg , method = "pearson")

correlation3_result <- paste0(
  "Pearson correlation coefficient: r = ", round(correlation3$estimate, 2),
  ", p-value = ", signif(correlation3$p.value, 3)
)
```

Appendix

```
library(tidyverse)
library(ggplot2)

player_data <- read.csv("NBA Players 1999-2023 - Sheet1.csv")
cleaned_data <- player_data %>%
  select(`Player`,`X3P.`, `FT.`) %>%
  rename(player_name = `Player`,three_point_pct = `X3P.`, free_throw_pct = `FT.`) %>%
  drop_na()
ggplot(cleaned_data, aes(x = three_point_pct)) +
  geom_histogram(binwidth = 0.05, fill = "blue", color = "black") +
  labs(x = "3-Point Percentage", y = "Frequency") +
  theme_minimal()
ggplot(cleaned_data, aes(x = free_throw_pct)) +
  geom_histogram(binwidth = 0.05, fill = "green", color = "black") +
  labs(x = "Free Throw Percentage", y = "Frequency") +
  theme_minimal()

ggplot(cleaned_data, aes(x = free_throw_pct, y = three_point_pct)) +
  geom_point(color = "blue",alpha=0.5) +
  geom_smooth(method = "lm", se = FALSE, color = "black") +
  labs(
    title = "Scatter Plot of 3-Point Percentage vs Free Throw Percentage",
    x = "Free Throw Percentage",
    y = "3-Point Percentage"
  ) +
  theme_minimal()

correlation <- cor.test(cleaned_data$three_point_pct, cleaned_data$free_throw_pct, method = "

correlation_result <- paste0(
  "Pearson correlation coefficient: r = ", round(correlation$estimate, 2),
  ", p-value = ", signif(correlation$p.value, 3)
)

correlation_result

library(tidyverse)
library(dplyr)
library(ggplot2)


NBAplayers = read.csv("NBA Players 1999-2023 - Sheet1.csv")


NBAplayers3 <- NBAplayers %>%
  #Drop unnecessary columns
  select(-c(1, 4, 5, 9:21, 24, 29))

#Avg points per game
Ageppg <- NBAplayers3 %>%
  group_by(NBAplayers3[c(2)]) %>%
  summarise(Avgppg = mean(PTS, na.rm = TRUE))

#Avg Assists Per Game
Ageast <- NBAplayers3 %>%
  group_by(NBAplayers3[c(2)]) %>%
  summarise(Avgast = mean(AST, na.rm = TRUE))

#Avg Offensive Rebounds
Ageorb <- NBAplayers3 %>%
  group_by(NBAplayers3[c(2)]) %>%
  summarise(Avgorb = mean(ORB, na.rm = TRUE))

#Avg Defensive Rebounds
Agedrb <- NBAplayers3 %>%
  group_by(NBAplayers3[c(2)]) %>%
  summarise(Avgdrb = mean(DRB, na.rm = TRUE))


#Agestat dataframe
Agestat = data.frame(Ageppg,Ageast[-c(1)],
Ageorb[-c(1)],Agedrb[-c(1)])
```

```r
ggplot(
  data = Agestat,
  aes(x=Age..,y=Avgppg,fill=Age..)
) + geom_bar(stat = "identity", width = 0.5) +
labs(x = "Player Age",y = "Average Points Per Game",
title = "Average Points Scored at Different Ages in the NBA")

ggplot(data = Agestat,
       aes(x=Age..,y=Avgast,fill=Age..)
) + geom_bar(stat = "identity", width = 0.5) +
  labs(x = "Player Age", y = "Average Assists Per Game",
       title = "Average Assists Made at Different Ages in the NBA")

ggplot(data = Agestat,
       aes(x=Age..,y=(Avgorb+Avgdrb),fill=Age..)
) + geom_bar(stat = "identity", width = 0.5) +
  labs(x = "Player Age", y = "Average Rebounds Per Game",
       title = "Average Rebounds at Different Ages in the NBA")


#Load libraries
library(tidyverse)
library(dplyr)
library(ggplot2)

#Read csv
NBAteamsA = read.csv("NBA Teams ADvanced 1999-2023 - Sheet1.csv")
View(NBAteamsA)

#Create WinRate comparison data frame
WinRate = (NBAteamsA[4]/(NBAteamsA[5]+NBAteamsA[4]))
WinRateComp = data.frame(NBAteamsA[2],WinRate,NBAteamsA[6],NBAteamsA[7],NBAteamsA[16])
View(WinRateComp)


ggplot(
  data = WinRateComp,
  aes(x=as.numeric(unlist(WinRateComp[3])),y=as.numeric(unlist(WinRateComp[2])))
) +
  geom_point() +
  labs(x = "Points Scored(Per 100 Possessions)",
       y = "Win Rate",
       title = "Correlation Between Win Rate and Points Scored")+
 theme_bw() + stat_smooth(method = "lm",
                          formula = y ~ x)
ggplot(
  data = WinRateComp,
  aes(x=as.numeric(unlist(WinRateComp[4])),y=as.numeric(unlist(WinRateComp[2])))
) +
  geom_point() +
  labs(x = "Points Allowed(Per 100 Possessions)",
       y = "Win Rate",
       title = "Correlation Between Win Rate and Points Allowed")+
  theme_bw() + stat_smooth(method = "lm",
                           formula = y ~ x)
#Signficance test for win rate and offense
correlation2 <- cor.test(WinRateComp$W..,WinRateComp$ORtg , method = "pearson")

correlation2_result <- paste0(
  "Pearson correlation coefficient: r = ", round(correlation2$estimate, 2),
  ", p-value = ", signif(correlation2$p.value, 3)
)
#Signficance test for win rate and defense
correlation3 <- cor.test(WinRateComp$W..,WinRateComp$DRtg , method = "pearson")

correlation3_result <- paste0(
  "Pearson correlation coefficient: r = ", round(correlation3$estimate, 2),
  ", p-value = ", signif(correlation3$p.value, 3)
)

#Load libraries
library(tidyverse)
library(dplyr)
library(ggplot2)

#Read csv files
NBAplayers = read.csv("NBA Players 1999-2023 - Sheet1.csv")
NBAteams = read.csv("NBA Teams 1999-2023 - Sheet1.csv")
NBAteamsA = read.csv("NBA Teams ADvanced 1999-2023 - Sheet1.csv")

View(NBAteamsA)
NBAplayers3 <- NBAplayers %>%
  #Drop unnecessary columns
  select(-c(1, 4, 5, 9:21, 24, 29))
```

```
#Avg points per game
Ageppg <- NBAplayers3 %>%
  group_by(NBAplayers3[c(2)]) %>%
  summarise(Avgppg = mean(PTS, na.rm = TRUE))

#Avg Assists Per Game
Ageast <- NBAplayers3 %>%
  group_by(NBAplayers3[c(2)]) %>%
  summarise(Avgast = mean(AST, na.rm = TRUE))

#Avg Minutes Per Game
Agemp <- NBAplayers3 %>%
  group_by(NBAplayers3[c(2)]) %>%
  summarise(Avgmp = mean(MP, na.rm = TRUE))

#Avg Offensive Rebounds
Ageorb <- NBAplayers3 %>%
  group_by(NBAplayers3[c(2)]) %>%
  summarise(Avgorb = mean(ORB, na.rm = TRUE))

#Avg Defensive Rebounds
Agedrb <- NBAplayers3 %>%
  group_by(NBAplayers3[c(2)]) %>%
  summarise(Avgdrb = mean(DRB, na.rm = TRUE))

#Avg Steals
Agestl <- NBAplayers3 %>%
  group_by(NBAplayers3[c(2)]) %>%
  summarise(Avgstl = mean(STL, na.rm = TRUE))

#Avg Blocks
Ageblk <- NBAplayers3 %>%
  group_by(NBAplayers3[c(2)]) %>%
  summarise(Avgblk = mean(BLK, na.rm = TRUE))

#Avg Turnovers
Agetov <- NBAplayers3 %>%
  group_by(NBAplayers3[c(2)]) %>%
  summarise(Avgtov = mean(TOV, na.rm = TRUE))

#Agestat dataframe
Agestat = data.frame(Ageppg,Ageast[-c(1)],
```

```r
Agemp[-c(1)],Ageorb[-c(1)],Agedrb[-c(1)],Agestl[-c(1)],
Ageblk[-c(1)],Agetov[-c(1)])
View(Agestat)

#Visual 1
ggplot(
  data = Agestat,
  aes(x=Age..,y=Avgppg,fill=Age..)
) + geom_bar(stat = "identity", width = 0.5) +
labs(x = "Player Age",y = "Average Points Per Game",
title = "Average Points Scored at Different Ages in the NBA")

#Visual 2
ggplot(data = Agestat,
       aes(x=Age..,y=Avgast,fill=Age..)
) + geom_bar(stat = "identity", width = 0.5) +
  labs(x = "Player Age", y = "Average Assists Per Game",
       title = "Average Assists Made at Different Ages in the NBA")

#Visual 3
ggplot(data = Agestat,
       aes(x=Age..,y=(Avgorb+Avgdrb),fill=Age..)
) + geom_bar(stat = "identity", width = 0.5) +
  labs(x = "Player Age", y = "Average Rebounds Per Game",
       title = "Average Rebounds at Different Ages in the NBA")


NBAplayers2 <- NBAplayers %>%
  #Drop unnecessary columns
  select(-c(1, 3:13, 15:20, 22:31))
NBAplayers2 = na.omit(NBAplayers2)
NBAplayers22 = as.numeric(unlist(NBAplayers2[2]))
NBAplayers23 = as.numeric(unlist(NBAplayers2[3]))
NBAplayers21 = data.frame(NBAplayers2[1],NBAplayers22,NBAplayers23)
Thrperc = NBAplayers21 %>% group_by(Player) %>%
  summarise(Thrperc = mean(NBAplayers22))
FTperc = NBAplayers21 %>% group_by(Player) %>%
  summarise(FTperc = mean(NBAplayers23))

Shotcomp = data.frame(Thrperc,FTperc[-c(1)])
View(Shotcomp)
```

```
#Visual 4
ggplot(
  data = Shotcomp,
  aes(x=as.numeric(unlist(Shotcomp[3])),y=as.numeric(unlist(Shotcomp[2])))
  ) +
    geom_point() +
    labs(x = "% of Free Throws Made",
         y = "% of 3s Made",
         title = "Free Throw% Compared to 3 Point% in NBA Players") +
    theme_bw() + stat_smooth(method = "lm",
                             formula = y ~ x)


WinRate = (NBAteamsA[4]/(NBAteamsA[5]+NBAteamsA[4]))
WinRateComp = data.frame(NBAteamsA[2],WinRate,NBAteamsA[6],NBAteamsA[7],NBAteamsA[16])
View(WinRateComp)

#Visual 5
ggplot(
  data = WinRateComp,
  aes(x=as.numeric(unlist(WinRateComp[3])),y=as.numeric(unlist(WinRateComp[2])))
) +
  geom_point() +
  labs(x = "Points Scored(Per 100 Possessions)",
       y = "Win Rate",
       title = "Correlation Between Win Rate and Points Scored")+
 theme_bw() + stat_smooth(method = "lm",
                          formula = y ~ x)

#Visual 6
ggplot(
  data = WinRateComp,
  aes(x=as.numeric(unlist(WinRateComp[4])),y=as.numeric(unlist(WinRateComp[2])))
) +
  geom_point() +
  labs(x = "Points Allowed(Per 100 Possessions)",
       y = "Win Rate",
       title = "Correlation Between Win Rate and Points Allowed")+
  theme_bw() + stat_smooth(method = "lm",
                           formula = y ~ x)

## Data Loading & Preprocessing
```

```
player_data <- read.csv("NBA Players 1999-2023 - Sheet1.csv")
player_data

cleaned_data <- player_data %>%
  select(`Player`,`X3P.`, `FT.`) %>%
  rename(player_name = `Player`,three_point_pct = `X3P.`, free_throw_pct = `FT.`) %>%
  drop_na()

cleaned_data

## Data Exploration
#Visual 7
ggplot(cleaned_data, aes(x = free_throw_pct, y = three_point_pct)) +
  geom_point(alpha = 0.7) +
  labs(
    title = "Scatter Plot of 3-Point Percentage vs Free Throw Percentage",
    x = "Free Throw Percentage",
    y = "3-Point Percentage"
  ) +
  theme_minimal()


## Correlation Test
correlation <- cor.test(cleaned_data$three_point_pct, cleaned_data$free_throw_pct, method = "

correlation

if (correlation$p.value < 0.05) {
  cat("Found significant correlation between 3-point shooting percentage and free throw shoo
} else {
  cat("Found no significant correlation between 3-point shooting percentage and free throw s
}
```

NBA and ABA league index. Basketball. (n.d.). https://www.basketball-reference.com/leagues/NBA_2000_per

NBA. (n.d.). *All Time Leaders | Stats | NBA.com*. Www.nba.com. https://www.nba.com/stats/alltime