Final Project

Jason Liu, Siyi.Luo

2025-06-20

Introduction

When the pandemic hit, the NBA didn't just lose games—it lost billions. I remember watching those eerie, empty arenas in 2020 and wondering: *How bad is this financially? And are the players feeling it too?* That curiosity turned into this project, where I dive into how COVID-19 affected the league, especially between 2019 and 2023.

Background Information

Based on what we found, the financial impact was brutal—at least at first. In 2019, the NBA was pulling in about $8.8 billion in revenue [1]. By the end of the next season, that number had dropped by 10%. And things didn't stop there. In 2020–21, league income fell again, this time by 19%, landing at around $6.4 billion [2]. That's nearly a one-third drop from pre-pandemic levels. Imagine running a business and suddenly losing a third of your income. It's no wonder the league had to freeze the salary cap.

So the real question is: Did average salaries actually go down? What about the superstars—were they still getting paid like kings? And how did attendance look when fans started coming back? Also I think what makes this topic so interesting isn't just the numbers—it's how resilient

the league turned out to be. One season, they were playing in a bubble. A couple years later? They hit record revenues. That's a wild ride.

Data Summary

Introduction to Datasets

The Data sources we will be using to answer our research question are the following four data sets:

Hoops Fortune (2020-2025), NBA Player salaries 2019-20,2022-2023

This dataset shows NBA player salaries from the 2022–2023 season through 2025–2026. Each row represents a player, and each column tracks their annual salary over the next few years.

ESPN NBA Attendance Report

This dataset displays NBA team attendance figures for the 2019 season, broken down by home and road games. Each row represents a team and includes total games played (always 82), the number of home and away games (41 each), total home and road attendance, and the calculated average attendance per game.

NBA 2018-2023 Overall Team Stats

This dataset presents detailed team-level performance statistics for all 30 NBA teams during the 2018–2019 regular season. Each row contains one team's season averages across a wide range of metrics: scoring (points per game), shooting efficiency (field goal %, 3P%, 2P%, FT%), rebounding (offensive, defensive, total), playmaking (assists), defense (steals, blocks), turnovers, and fouls.

Data Wrangling

Data Wrangling for NBA Player Salary

We started by loading player salary data from four NBA seasons (2019–2023). Each dataset had its own quirks—different column names, inconsistent formats, and dollar signs or commas in the salary values. So, to make things cleaner, we kept just the essentials: player names and their salaries. Then we gave each column a clear, season-specific name so everything would line up later.

Next, we deal about the formatting. Since R can't handle salary columns with symbols like $ or „ we stripped those out using regular expressions and converted the cleaned values into numbers. This step made it possible to do calculations like finding averages or comparing across seasons.

After cleaning each year's data, we stitched them together using full joins. This approach ensured that if a player appeared in one season but not another, we'd still keep their record. Any missing values? We replaced them with zeros to keep things smooth.

To dig into salary patterns, we created a new variable called IsSuperstar. It flags players in the top 5% of salaries from the 2019–20 season. That way, we can easily compare high earners with the rest of the league.

For attendance data, we calculated two new metrics: average home attendance and average road attendance. We just divided total attendance by the number of games for each category. These new variables will help us spot whether fans came back to the stands after COVID—and how quickly.

```
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
```

Data Wrangling for Attendance

Exploratory Data Analysis

Salary

Change of Salary for Average NBA Players

The chart shows that average salaries for non-superstar NBA players rose gradually between

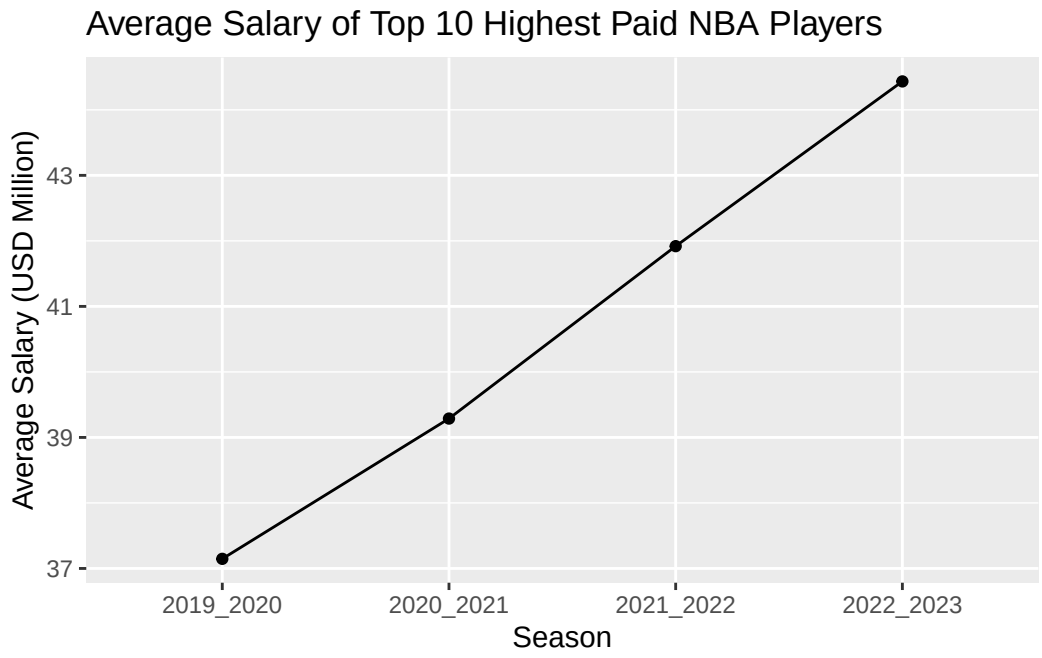
2019 and 2021, then spiked in 2022–23. In 2019–20, the average salary for these players was about $3.13 million. It climbed to roughly $3.23 million in 2020–21, then edged up slightly again to $3.26 million in 2021–22. Nothing dramatic happened during those years—just small, cautious increases.

But in 2022–23, things changed fast. The average salary jumped to around $3.78 million, the biggest single-year increase in the period. That sharp rise matched the NBA's broader financial recovery—teams were bringing in more revenue and felt confident spending again.

League-wide, the numbers followed a similar path. The average player salary in 2019–20 was about $8.2 million [3]. In 2020–21, it dropped slightly to $7.9 million [4]. The next season saw a small recovery to $8.25 million, but by 2023–24, average salaries had soared to nearly $9.7 million [5]. Median salaries went from $3.83 million during the pandemic up to $4.6 million just a few years later [5].

Overall, non-superstar salaries didn't collapse, but they definitely slowed. And when the league bounced back, those paychecks started growing again.

Change of Salary for Top 10 Highest Paid Supstars



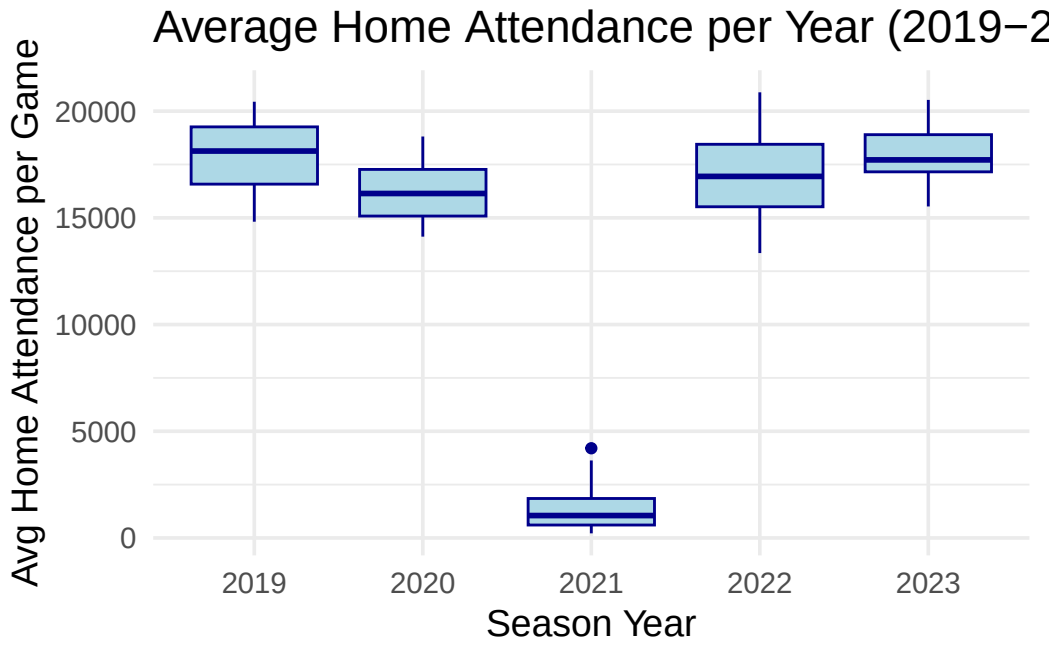
Top NBA salaries didn't flinch during the pandemic — they climbed year after year. In 2019–20, the average salary among the 10 highest-paid players sat at $37.1 million. A season later, that number rose to $39.2 million. And it didn't stop there. In 2021–22, it broke the $41 million mark, and by 2022–23, it jumped again to nearly $43.8 million. The pace is steady, and the growth is real.

What's driving this? Multi-year contracts with built-in raises played a big role. Stars like Stephen Curry didn't just hold their ground—they leveled up every season. His salary alone grew from $40.2 million in 2019–20 to over $51.9 million by 2023–24. That made him the first player to ever earn more than $50 million in a single season [6]. Meanwhile, the 10th highest-paid player also got a raise, going from around $33 million to nearly $46 million. So even at the bottom of the top 10, players earned what used to be superstar money.

The league didn't pull back on paying its biggest names—even during its most uncertain

years. These players sell jerseys, drive ratings, and fill arenas. Teams knew that. They doubled down. So while role players had to wait for the financial rebound, superstars kept cashing bigger and bigger checks. This upward curve tells you something bigger: when it comes to NBA economics, star power always wins.

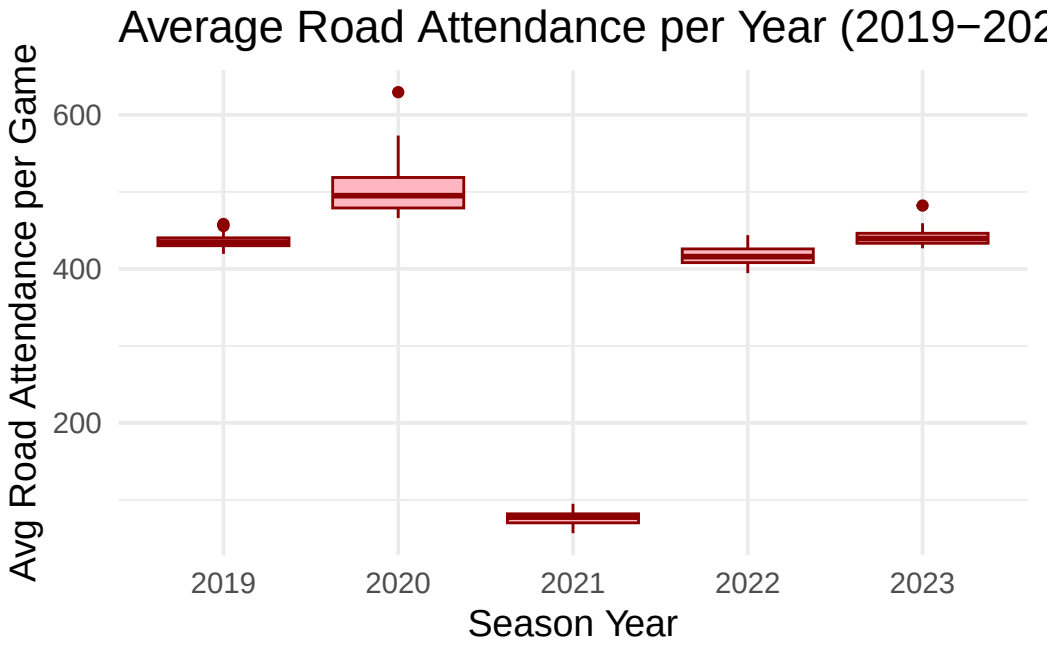
Attendance



This plot shows the average home attendance per NBA team across five seasons, from 2019 to 2023. Each box represents the spread of attendance numbers for that year. The wide boxes in 2019, 2020, 2022, and 2023 show that teams generally drew strong, consistent crowds, with most hovering between 15,000 and 20,000 fans per game. But in 2021, the box collapses. Attendance tanked league-wide, and the outlier dot near the bottom tells you just how empty some arenas were. With fans locked out or kept at limited capacity, teams played through silence. The atmosphere changed, and it showed on the court. The bounce-back in 2022 and 2023 reflects more than just people returning to seats—it signals the NBA's emotional recovery. Fans came back, arenas got

loud again, and teams once more fed off that home-court electricity. You can feel the pulse returning just by watching that box climb.

Road Attendance



This boxplot shows how average road attendance changed from 2019 to 2023 for every NBA team. Each season gets its own box, which captures how packed—or empty—the arenas were when teams hit the road. In most years—2019, 2020, 2022, and 2023—the boxes stretch wide, sitting mostly between 15,000 and 20,000. That means most arenas welcomed solid, consistent crowds. But then 2021 hits, and the whole thing collapses. The box shrinks. A single dot hovers near the bottom, showing how empty some arenas really got.

What makes this plot interesting is what it says about the culture of basketball fandom. In Europe, especially in soccer, away fans play a huge role. They travel, chant, wave flags, and turn stadiums into battlefields. In the NBA, not so much. You'll see scattered away jerseys here and there—especially for big-name teams like the Lakers—but most of the crowd still cheers for the

home team. So “road attendance” isn’t really about how many fans support the visitors. It’s more about how much draw the home team has, and how much people want to watch basketball live.

Win Rate

Table 1: Table 1 – Top 5 Teams: Performance Metrics (2018–2019)

Team	Games	FG_Percentage	ThreeP_Percentage	TwoP_Percentage	FT_Percentage	Turnovers	Assists	Points_Per_Game
Milwaukee Bucks*	82	0.48	0.35	0.56	0.77	13.9	26.0	118.1
Golden State Warriors*	82	0.49	0.38	0.56	0.80	14.3	29.4	117.7
New Orleans Pelicans	82	0.47	0.34	0.54	0.76	14.8	27.0	115.4
Philadelphia 76ers*	82	0.47	0.36	0.53	0.77	14.9	26.9	115.2
Los Angeles Clippers*	82	0.47	0.39	0.51	0.79	14.5	24.0	115.1

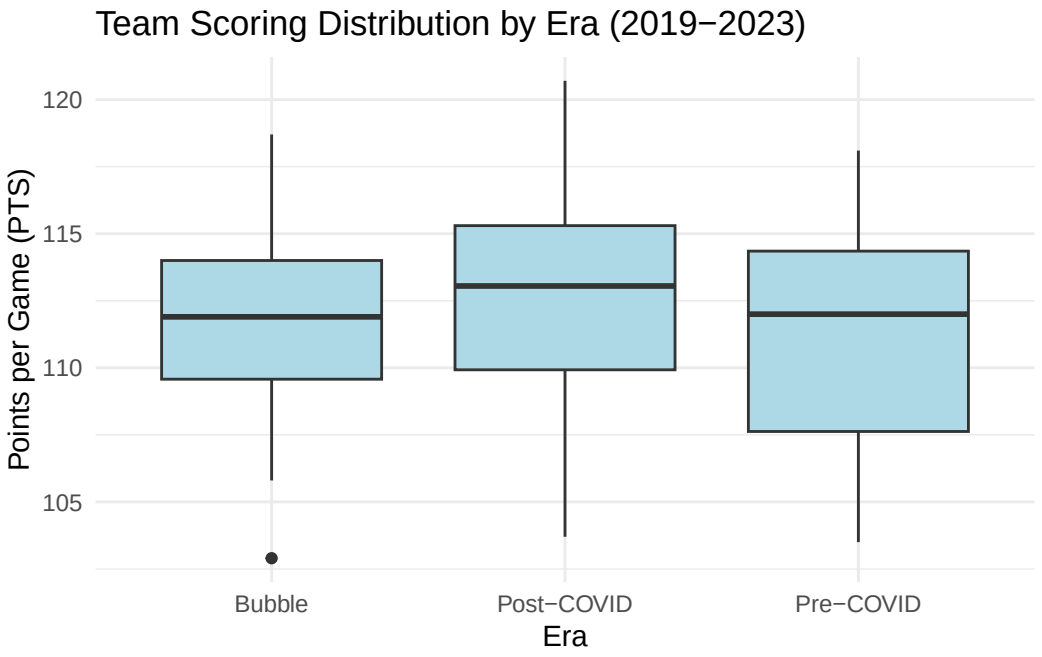
The first table shows the top five scoring teams from the 2018–2019 season and breaks down how they got buckets. These teams didn’t just outscore opponents, they truly combined solid field goal efficiency with high assist rates. For example, The Warriors led the league in ball movement and shot-making, the Bucks dominated inside with a strong two-point percentage and aggressive scoring. You can also spot their tendency to accept a few more turnovers in exchange for faster,

freer play.

Table 2: Table 2 – Scoring Tier Summary: Shooting & Turnovers

Scoring_Tier	Avg_FG_Percentage	Avg_ThreeP_Percentage	Avg_Turnovers	Avg_Assists
Bottom 10 Teams	0.46	0.35	13.95	23.93
Middle 10 Teams	0.46	0.36	14.06	24.52
Top 5 Teams	0.48	0.37	14.48	26.66

In contrast, the second table zooms out to compare team tiers. Top scoring teams had the highest shooting percentages and assist numbers, even though their turnovers were slightly above average. Middle-tier teams looked more balanced but didn't excel in any one area. The lowest tier played cautiously, with lower turnovers and lower production.



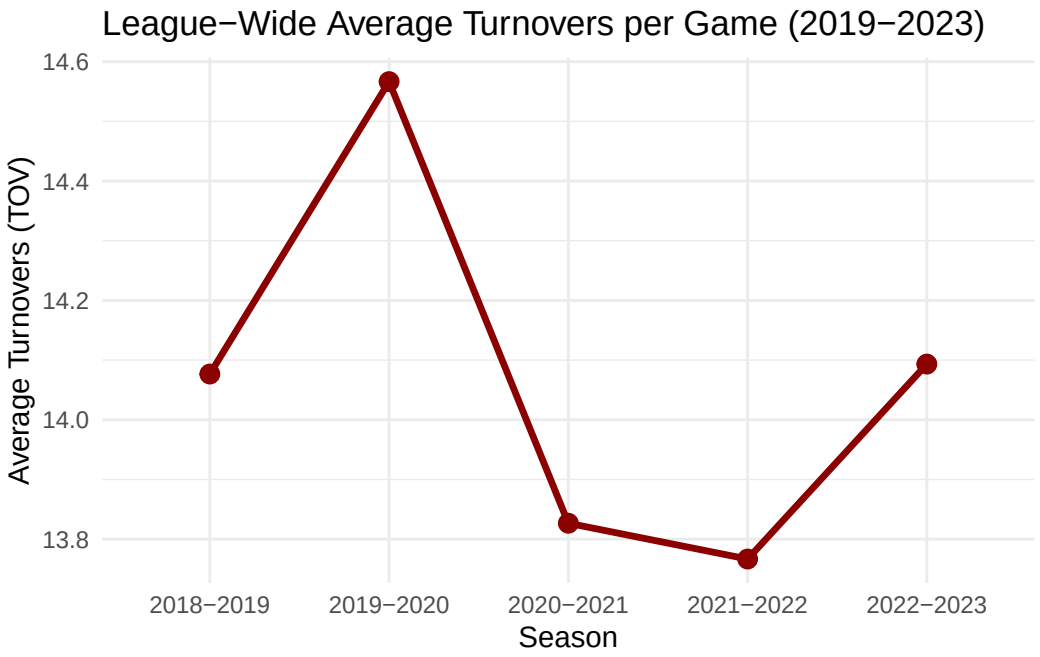
Scoring in the NBA shifted in subtle but noticeable ways across the COVID timeline. The boxplot makes it easy to see how team performance moved through three different phases—pre-COVID, the 2020 Bubble, and the seasons that followed. Even without reading the numbers, you can feel the rhythm change.

During the Bubble, games felt strange. There were no fans, no home-court energy, and no travel fatigue. Everyone played under the same roof, in the same quiet gym, every night. The box for that period looks tighter than the others. That means team scoring didn't fluctuate as much. Most teams clustered around the same point total, and only one fell way behind, which you can spot as the lone dot below the whisker. It was as if the whole league hit reset. Star power still mattered, but momentum didn't build the same way.

Then came the post-COVID seasons. Scoring ticked upward, but consistency dropped off. The box widens, stretching up and down, which shows a bigger gap between top and bottom scoring teams. Some squads bounced back stronger, while others clearly struggled to adjust. That's not surprising. Between new health rules, schedule changes, and roster shuffles, a lot of teams found themselves in unfamiliar territory.

Before all of that, the 2018–2019 season looked pretty stable. The scoring distribution wasn't too narrow or too wide. Most teams hovered around the middle, and only a few stood out. That balance makes sense. Things were normal back then—fans in the stands, regular travel, and a predictable rhythm to the year. Teams had routines they trusted, and it showed on the scoreboard.

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



In the 2019–2020 season, teams coughed up the ball more than at any other point in this five-year stretch. The timing's no coincidence. That year threw the entire league into a storm. COVID hit, the season got suspended, and players had to adjust to life in the bubble. Routines broke down. Chemistry suffered. You can't expect crisp passing when half the roster's in quarantine or trying to mentally reset in a hotel room. Before that, in 2018–2019, things looked pretty steady. Average turnovers hovered just above 14. It wasn't perfect basketball, but it wasn't messy either. Players knew their systems. Coaches stuck to their rotations. Then, boom—pandemic season. Turnovers shot up. It makes sense. New lineups. Empty arenas. Shortened prep. Everything felt off.

But what's wild is what came next. After that chaotic peak, turnovers dipped in 2020–2021 and hit the lowest point in 2021–2022. That's not what you'd expect. You'd think post-COVID adjustments would keep mistakes high, but teams cleaned up. Maybe the league emphasized ball control. Maybe younger players got more reps and grew up fast. Whatever the reason, teams tightened their grip. Then in 2022–2023, the number jumps again. Not as high as the bubble year,

but enough to raise eyebrows. Was it fatigue? Roster turnover? Riskier offense? Possibly all of the above. What matters is the trend doesn't stay flat. The league moves through phases—disruption, correction, and recalibration.

I suppose stats like turnovers tell truth in a more hidden way, we don't really notice them until they pile up, it's not just about the mistakes, it's also about how teams handle pressure, adapted to the unknown and fought to regain control, even though in some scenario things do get out of hand.

Conclusion

In conclusion, the NBA's financial and salary swings around COVID proved sharp but short-lived. After suffering an unprecedented revenue hit, teams bounced back quickly—thanks to the league's strong market pull and nimble operations—which confirms that the pandemic's impact remained temporary. Looking ahead, new broadcast deals and expanding markets should keep revenues and the salary cap climbing. At the same time, the NBA will rely on revenue sharing, luxury-tax rules, and similar measures to preserve competitive balance and fiscal health, ensuring stability under any external shock. Fan turnout, scoring rhythms, and turnover trends also reflect the league's cycle of disruption and renewal. The bubble season pushed average turnovers to a five-year high. Then in 2021–22, teams tightened ball control and drove turnovers down to multi-year lows. As schedules and travel normalized in 2022–23, turnovers ticked up slightly again. Attendance, too, rebounded far faster in the post-COVID era than we expected. By weathering chaos, making corrections, and emerging stronger, the NBA and its teams have shown their real power when facing forces beyond their control.

Reference

1. https://frontofficesports.com/nba-tops-10b-in-revenue-for-first-time-ever/#:~:text=%2A%20For%20the%2
2. https://runrepeat.com/nba-revenue-statistics
3. https://www.thehoopsgeek.com/average-nba-salary/
4. https://runrepeat.com/salary-analysis-in-the-nba-1991-2019
5. https://fansided.com/posts/what-is-the-average-and-median-nba-salary-for-2023-24-by-position-01hbrqp8mv6f
6. https://en.wikipedia.org/wiki/List_of_highest-paid_NBA_players_by_season

Code Appendix

```
# Load Packages

library(dplyr)

library(readr)

# Load the datasets

PlayerSalary2019_2020 <- read.csv("~/Desktop/Stat 184/STAT184_FinalProject/PlayerSalary/

PlayerSalary2020_2021 <- read.csv("~/Desktop/Stat 184/STAT184_FinalProject/PlayerSalary/

PlayerSalary2021_2022 <- read.csv("~/Desktop/Stat 184/STAT184_FinalProject/PlayerSalary/

PlayerSalary2022_2023 <- read.csv("~/Desktop/Stat 184/STAT184_FinalProject/PlayerSalary/



# The way we will tidy the data will be this way:

# we will try to keep the only following columns: team,player,salary, (General Data Wran
```

```r
PlayerSalary2019_2020 <- PlayerSalary2019_2020 %>%
  rename(player = player, team = team, Season2019_2020 = salary) %>%   # Adjust if neede
  select(player, Season2019_2020)

PlayerSalary2020_2021 <- PlayerSalary2020_2021 %>%
  rename(player = name, team = team, Season2020_2021 = salary) %>%     # Adjust if neede
  select(player, Season2020_2021)

PlayerSalary2021_2022 <- PlayerSalary2021_2022 %>%
  rename(player = name, team = team, Season2021_2022 = salary) %>%
  select(player, Season2021_2022)

# Joins the data sets together (use of joins)

PlayerSalary2019_2022 <- PlayerSalary2019_2020 %>%
  full_join(PlayerSalary2020_2021, by = "player") %>%
  full_join(PlayerSalary2021_2022, by = "player")

# Fix the NAs, Change all NAs to 0

PlayerSalary2019_2022[is.na(PlayerSalary2019_2022)] <- 0
```

```r
# we need to fix the column name for some of the data sets. such as 2022.23, 2023.24 in

colnames(PlayerSalary2022_2023)[3:6] <- c("Season2022_2023", "Season2023_2024", "Season2

PlayerSalary2022_2023 <- PlayerSalary2022_2023 %>%
  rename(player = Player.Name) %>%
  select(player, Season2022_2023)
# Need to also do the wrangling to get ride of the $ sign and comma signs (use of regula

PlayerSalary2022_2023$Season2022_2023 <- gsub("\\$","",PlayerSalary2022_2023$Season2022_
PlayerSalary2022_2023$Season2022_2023 <- gsub("\\,","",PlayerSalary2022_2023$Season2022_

# Now fully join everything all together.

PlayerSalary2019_2023 <- PlayerSalary2019_2022 %>%
  full_join(PlayerSalary2022_2023, by = "player")

PlayerSalary2019_2023[is.na(PlayerSalary2019_2023)] <- 0
# Create 1 variable, IsSuperStar, to determine if player is a SuperStar, the ones that h

PlayerSalary2019_2023$IsSuperstar <- PlayerSalary2019_2023$Season2019_2020 >= quantile(P
PlayerSalary2019_2023 <- PlayerSalary2019_2023 %>%
  mutate(across(starts_with("Season"), ~ as.numeric(gsub("[$,]", "", as.character(.)))))
```

```r
Attendance <- read.csv("~/Desktop/Stat 184/STAT184_FinalProject/Attendance/2019_2023_NBA

# Create 2 Variables
Attendance <- Attendance %>%
  mutate(AvgHomeAttendance = HomeAttendance / HomeGame) %>%
  mutate(AvgRoadAttendance = RoadAttendance / RoadGame)



# load packages
library(dplyr)
library(tidyr)
library(ggplot2)



# Create a new dataset for Non superstars, AKA Average Players
non_superstars <- PlayerSalary2019_2023 %>%
  filter(IsSuperstar == FALSE)



non_superstars_long <- non_superstars %>%
  pivot_longer(cols = starts_with("Season"), names_to = "Season", values_to = "Salary")
```

```r
# Clean season names for plotting
non_superstars_long$Season <- gsub("Season", "", non_superstars_long$Season)


# Calculate average salary per year
avg_salary_by_year <- non_superstars_long %>%
  group_by(Season) %>%
  summarise(AverageSalary = mean(Salary, na.rm = TRUE))


ggplot(avg_salary_by_year, aes(x = Season, y = AverageSalary / 1e6)) +
  geom_line(group = 1) +
  geom_point() +
  labs(
    title = "Average NBA Player Salary (Non-Superstars)",
    x = "Season",
    y = "Average Salary (USD Million)"
  )




top10_long <- PlayerSalary2019_2023 %>%
  select(player, starts_with("Season")) %>%
  pivot_longer(
```

```
    cols = starts_with("Season"),
    names_to = "Season",
    values_to = "Salary"
  ) %>%
  mutate(
    Season = gsub("Season", "", Season),
    Salary = as.numeric(gsub("[$,]", "", as.character(Salary)))
  )

# Step 2: For each season, filter top 10 earners
top10_highest_paid <- top10_long %>%
  group_by(Season) %>%
  arrange(desc(Salary), .by_group = TRUE) %>%
  slice_head(n = 10)

# Step 3: Calculate average salary per season
avg_salary_top10 <- top10_highest_paid %>%
  group_by(Season) %>%
  summarise(AverageSalary = mean(Salary))

# Step 4: Plot
ggplot(avg_salary_top10, aes(x = Season, y = AverageSalary / 1e6)) +
```

```
  geom_line(group = 1) +
  geom_point() +
  labs(
    title = "Average Salary of Top 10 Highest Paid NBA Players",
    x = "Season",
    y = "Average Salary (USD Million)"
  )


ggplot(Attendance, aes(x = as.factor(Year), y = AvgHomeAttendance)) +
  geom_boxplot(fill = "lightblue", color = "darkblue") +
  labs(
    title = "Average Home Attendance per Year (2019-2023)",
    x = "Season Year",
    y = "Avg Home Attendance per Game"
  ) +
  theme_minimal(base_size = 14)
ggplot(Attendance, aes(x = as.factor(Year), y = AvgRoadAttendance)) +
  geom_boxplot(fill = "lightpink", color = "darkred") +
  labs(
    title = "Average Road Attendance per Year (2019-2023)",
    x = "Season Year",
    y = "Avg Road Attendance per Game"
```

```
  ) +
  theme_minimal(base_size = 14)



# Load the necessary libraries
library(ggplot2)   # For plotting
library(dplyr)     # For data manipulation
library(readr)     # For reading CSV files


# ---- READ EACH SEASON FILE ----


# Read the CSV files you uploaded (make sure the paths are correct)
data_2019 <- read.csv("~/Desktop/Stat 184/STAT184_FinalProject/NBA_Teams_Stat/NBA 2018-2
data_2020 <- read.csv("~/Desktop/Stat 184/STAT184_FinalProject/NBA_Teams_Stat/NBA 2019-2
data_2021 <- read.csv("~/Desktop/Stat 184/STAT184_FinalProject/NBA_Teams_Stat/NBA 2020-2
data_2022 <- read.csv("~/Desktop/Stat 184/STAT184_FinalProject/NBA_Teams_Stat/NBA 2021-2
# If you have the 2023 data file, load it like this
data_2023 <- read.csv("~/Desktop/Stat 184/STAT184_FinalProject/NBA_Teams_Stat/NBA 2022-2


# ---- ADD SEASON COLUMN ----
data_2019$Season <- "2018-2019"
data_2020$Season <- "2019-2020"
```

```r
data_2021$Season <- "2020-2021"

data_2022$Season <- "2021-2022"

data_2023$Season <- "2022-2023"  # Only if you have the 2023 file


# ---- COMBINE ALL YEARS INTO ONE DATAFRAME ----

all_data <- bind_rows(data_2019, data_2020, data_2021, data_2022, data_2023)


# Power Houses

# Golden State Warriors, Houston Rockets, San Antonio Spurs, Los Angeles Lakers, Boston
```

```r
# see if they are power houses

traditional_teams <- c("Golden State Warriors", "Houston Rockets", "San Antonio Spurs",

                       "Los Angeles Lakers", "Boston Celtics", "Toronto Raptors",

                       "Milwaukee Bucks", "Philadelphia 76ers")


all_data <- all_data %>%

  mutate(IsTraditional = ifelse(Team %in% traditional_teams, "Traditional", "Other"))
```

```r
# determine Pre-Covid, Covid, Post-Covid


all_data <- all_data %>%
  mutate(Era = case_when(
    Season == "2018-2019" ~ "Pre-COVID",
    Season == "2019-2020" ~ "Bubble",
    Season %in% c("2020-2021", "2021-2022", "2022-2023") ~ "Post-COVID"
  ))


library(knitr)


# Select and clean columns, display only top 5 teams
table1 <- data_2019 %>%
  select(
    Team,
    Games = G,
    FG_Percentage = FG.,
    ThreeP_Percentage = X3P.,
    TwoP_Percentage = X2P.,
    FT_Percentage = FT.,
    Turnovers = TOV,
    Assists = AST,
```

```
    Points_Per_Game = PTS
  ) %>%
  head(5)


# Print table
kable(table1, caption = "Table 1 - Top 5 Teams: Performance Metrics (2018-2019)", digits




table2 <- data_2019 %>%
  mutate(
    Scoring_Tier = case_when(
      rank(-PTS) <= 5 ~ "Top 5 Teams",
      rank(-PTS) <= 15 ~ "Middle 10 Teams",
      TRUE ~ "Bottom 10 Teams"
    )
  ) %>%
  group_by(Scoring_Tier) %>%
  summarise(
    Avg_FG_Percentage = mean(FG., na.rm = TRUE),
    Avg_ThreeP_Percentage = mean(X3P., na.rm = TRUE),
    Avg_Turnovers = mean(TOV, na.rm = TRUE),
```

```
    Avg_Assists = mean(AST, na.rm = TRUE)
  ) %>%
  head(5)

# Print table
kable(table2, caption = "Table 2 - Scoring Tier Summary: Shooting & Turnovers", digits =




ggplot(all_data, aes(x = Era, y = PTS)) +
  geom_boxplot(fill = "lightblue") +
  labs(
    title = "Team Scoring Distribution by Era (2019-2023)",
    x = "Era",
    y = "Points per Game (PTS)"
  ) +
  theme_minimal()




avg_turnovers <- all_data %>%
  group_by(Season) %>%
```

```
  summarise(AverageTOV = mean(TOV, na.rm = TRUE))


ggplot(avg_turnovers, aes(x = Season, y = AverageTOV, group = 1)) +
  geom_line(color = "darkred", size = 1.2) +
  geom_point(size = 3, color = "darkred") +
  labs(
    title = "League-Wide Average Turnovers per Game (2019-2023)",
    x = "Season",
    y = "Average Turnovers (TOV)"
  ) +
  theme_minimal()


# Load Packages
library(dplyr)
library(readr)
# Load the datasets
PlayerSalary2019_2020 <- read.csv("~/Desktop/Stat 184/STAT184_FinalProject/PlayerSalary/
PlayerSalary2020_2021 <- read.csv("~/Desktop/Stat 184/STAT184_FinalProject/PlayerSalary/
PlayerSalary2021_2022 <- read.csv("~/Desktop/Stat 184/STAT184_FinalProject/PlayerSalary/
PlayerSalary2022_2023 <- read.csv("~/Desktop/Stat 184/STAT184_FinalProject/PlayerSalary/


# The way we will tidy the data will be this way:
```

```r
# we will try to keep the only following columns: team,player,salary, (General Data Wran

PlayerSalary2019_2020 <- PlayerSalary2019_2020 %>%
  rename(player = player, team = team, Season2019_2020 = salary) %>%   # Adjust if neede
  select(player, Season2019_2020)


PlayerSalary2020_2021 <- PlayerSalary2020_2021 %>%
  rename(player = name, team = team, Season2020_2021 = salary) %>%     # Adjust if neede
  select(player, Season2020_2021)


PlayerSalary2021_2022 <- PlayerSalary2021_2022 %>%
  rename(player = name, team = team, Season2021_2022 = salary) %>%
  select(player, Season2021_2022)


# Joins the data sets together (use of joins)


PlayerSalary2019_2022 <- PlayerSalary2019_2020 %>%
  full_join(PlayerSalary2020_2021, by = "player") %>%
  full_join(PlayerSalary2021_2022, by = "player")


# Fix the NAs, Change all NAs to 0


PlayerSalary2019_2022[is.na(PlayerSalary2019_2022)] <- 0
```

```r
# we need to fix the column name for some of the data sets. such as 2022.23, 2023.24 in

colnames(PlayerSalary2022_2023)[3:6] <- c("Season2022_2023", "Season2023_2024", "Season2

PlayerSalary2022_2023 <- PlayerSalary2022_2023 %>%
  rename(player = Player.Name) %>%
  select(player, Season2022_2023)
# Need to also do the wrangling to get ride of the $ sign and comma signs (use of regula

PlayerSalary2022_2023$Season2022_2023 <- gsub("\\$","",PlayerSalary2022_2023$Season2022_
PlayerSalary2022_2023$Season2022_2023 <- gsub("\\,","",PlayerSalary2022_2023$Season2022_



# Now fully join everything all together.

PlayerSalary2019_2023 <- PlayerSalary2019_2022 %>%
  full_join(PlayerSalary2022_2023, by = "player")

PlayerSalary2019_2023[is.na(PlayerSalary2019_2023)] <- 0
# Create 1 variable, IsSuperStar, to determine if player is a SuperStar, the ones that h

PlayerSalary2019_2023$IsSuperstar <- PlayerSalary2019_2023$Season2019_2020 >= quantile(P
PlayerSalary2019_2023 <- PlayerSalary2019_2023 %>%
  mutate(across(starts_with("Season"), ~ as.numeric(gsub("[$,]", "", as.character(.)))))
```

```r
Attendance <- read.csv("~/Desktop/Stat 184/STAT184_FinalProject/Attendance/2019_2023_NBA

# Create 2 Variables
Attendance <- Attendance %>%
  mutate(AvgHomeAttendance = HomeAttendance / HomeGame) %>%
  mutate(AvgRoadAttendance = RoadAttendance / RoadGame)

# load packages
library(dplyr)
library(tidyr)
library(ggplot2)



# Create a new dataset for Non superstars, AKA Average Players
non_superstars <- PlayerSalary2019_2023 %>%
  filter(IsSuperstar == FALSE)



non_superstars_long <- non_superstars %>%
  pivot_longer(cols = starts_with("Season"), names_to = "Season", values_to = "Salary")

# Clean season names for plotting
```

```r
non_superstars_long$Season <- gsub("Season", "", non_superstars_long$Season)


# Calculate average salary per year
avg_salary_by_year <- non_superstars_long %>%
  group_by(Season) %>%
  summarise(AverageSalary = mean(Salary, na.rm = TRUE))



ggplot(avg_salary_by_year, aes(x = Season, y = AverageSalary / 1e6)) +
  geom_line(group = 1) +
  geom_point() +
  labs(
    title = "Average NBA Player Salary (Non-Superstars)",
    x = "Season",
    y = "Average Salary (USD Million)"
  )


ggplot(Attendance, aes(x = as.factor(Year), y = AvgHomeAttendance)) +
  geom_boxplot(fill = "lightblue", color = "darkblue") +
  labs(
    title = "Average Home Attendance per Year (2019-2023)",
    x = "Season Year",
```

```
    y = "Avg Home Attendance per Game"
  ) +
  theme_minimal(base_size = 14)


ggplot(Attendance, aes(x = as.factor(Year), y = AvgRoadAttendance)) +
  geom_boxplot(fill = "lightpink", color = "darkred") +
  labs(
    title = "Average Road Attendance per Year (2019-2023)",
    x = "Season Year",
    y = "Avg Road Attendance per Game"
  ) +
  theme_minimal(base_size = 14)


# Load the necessary libraries
library(ggplot2)   # For plotting
library(dplyr)     # For data manipulation
library(readr)     # For reading CSV files


# ---- READ EACH SEASON FILE ----


# Read the CSV files you uploaded (make sure the paths are correct)
data_2019 <- read.csv("~/Desktop/Stat 184/STAT184_FinalProject/NBA_Teams_Stat/NBA 2018-2
```

```
data_2020 <- read.csv("~/Desktop/Stat 184/STAT184_FinalProject/NBA_Teams_Stat/NBA 2019-2
data_2021 <- read.csv("~/Desktop/Stat 184/STAT184_FinalProject/NBA_Teams_Stat/NBA 2020-2
data_2022 <- read.csv("~/Desktop/Stat 184/STAT184_FinalProject/NBA_Teams_Stat/NBA 2021-2
# If you have the 2023 data file, load it like this
data_2023 <- read.csv("~/Desktop/Stat 184/STAT184_FinalProject/NBA_Teams_Stat/NBA 2022-2


# ---- ADD SEASON COLUMN ----
data_2019$Season <- "2018-2019"
data_2020$Season <- "2019-2020"
data_2021$Season <- "2020-2021"
data_2022$Season <- "2021-2022"
data_2023$Season <- "2022-2023"  # Only if you have the 2023 file


# ---- COMBINE ALL YEARS INTO ONE DATAFRAME ----
all_data <- bind_rows(data_2019, data_2020, data_2021, data_2022, data_2023)


# Power Houses
# Golden State Warriors, Houston Rockets, San Antonio Spurs, Los Angeles Lakers, Boston


# see if they are power houses
traditional_teams <- c("Golden State Warriors", "Houston Rockets", "San Antonio Spurs",
                       "Los Angeles Lakers", "Boston Celtics", "Toronto Raptors",
```

```
                      "Milwaukee Bucks", "Philadelphia 76ers")


all_data <- all_data %>%
  mutate(IsTraditional = ifelse(Team %in% traditional_teams, "Traditional", "Other"))



# determine Pre-Covid, Covid, Post-Covid


all_data <- all_data %>%
  mutate(Era = case_when(
    Season == "2018-2019" ~ "Pre-COVID",
    Season == "2019-2020" ~ "Bubble",
    Season %in% c("2020-2021", "2021-2022", "2022-2023") ~ "Post-COVID"
  ))


ggplot(all_data, aes(x = Era, y = PTS)) +
  geom_boxplot(fill = "lightblue") +
  labs(
    title = "Team Scoring Distribution by Era (2019-2023)",
    x = "Era",
    y = "Points per Game (PTS)"
  ) +
```

```
  theme_minimal()


avg_turnovers <- all_data %>%
  group_by(Season) %>%
  summarise(AverageTOV = mean(TOV, na.rm = TRUE))


ggplot(avg_turnovers, aes(x = Season, y = AverageTOV, group = 1)) +
  geom_line(color = "darkred", size = 1.2) +
  geom_point(size = 3, color = "darkred") +
  labs(
    title = "League-Wide Average Turnovers per Game (2019-2023)",
    x = "Season",
    y = "Average Turnovers (TOV)"
  ) +
  theme_minimal()
```