INF1344 Final project

Group 7

r Sys.Date()

Topic

Relationship between team spending and team performance

{r include = FALSE} library(dplyr) library(ggplot2) library(readr) library(stringr)

The Data

Load the data
{r, message = F} summary_2020_21 <- read_csv("NBA Season Summary 2020_21.csv") summary_2021_22 <- read_csv("NBA Season Summary 2021_22.csv") summary_2022_23 <- read_csv("NBA Season Summary 2022_23.csv") summary_2023_24 <- read_csv("NBA Season Summary 2023_24.csv") summary_2024_25 <- read_csv("NBA Season Summary 2024_25.csv") payroll_2020_21 <- read_csv("NBA Team Payroll 2020_21.csv") payroll_2021_22 <- read_csv("NBA Team Payroll 2021_22.csv") payroll_2022_23 <- read_csv("NBA Team Payroll 2022_23.csv") payroll_2023_24 <- read_csv("NBA Team Payroll 2023_24.csv") payroll_2024_25 <- read_csv("NBA Team Payroll 2024_25.csv") cap <- read_csv("Cap.csv")

Data Assessment

Preparation steps

```
# Define mapping for team short codes
team_codes <- c(
  "Philadelphia 76ers" = "PHI",
  "Brooklyn Nets" = "BKN",
  "Milwaukee Bucks" = "MIL",
  "New York Knicks" = "NYK",
  "Atlanta Hawks" = "ATL",
  "Miami Heat" = "MIA",
  "Boston Celtics" = "BOS",
  "Washington Wizards" = "WAS",
  "Indiana Pacers" = "IND",
  "Charlotte Hornets" = "CHA",
  "Chicago Bulls" = "CHI",
```

```
  "Toronto Raptors" = "TOR",
  "Cleveland Cavaliers" = "CLE",
  "Orlando Magic" = "ORL",
  "Detroit Pistons" = "DET",
  "Utah Jazz" = "UTA",
  "Phoenix Suns" = "PHX",
  "Denver Nuggets" = "DEN",
  "Los Angeles Clippers" = "LAC",
  "Dallas Mavericks" = "DAL",
  "Portland Trail Blazers" = "POR",
  "Los Angeles Lakers" = "LAL",
  "Golden State Warriors" = "GSW",
  "Memphis Grizzlies" = "MEM",
  "San Antonio Spurs" = "SAS",
  "New Orleans Pelicans" = "NOP",
  "Sacramento Kings" = "SAC",
  "Minnesota Timberwolves" = "MIN",
  "Oklahoma City Thunder" = "OKC",
  "Houston Rockets" = "HOU"
)

# List of team summary data frames and seasons
tsummary_dfs <- list(
  "2020_21" = summary_2020_21,
  "2021_22" = summary_2021_22,
  "2022_23" = summary_2022_23,
  "2023_24" = summary_2023_24,
  "2024_25" = summary_2024_25
)

# List of team payroll data frames and seasons
tpayroll_dfs <- list(
  "2020_21" = payroll_2020_21,
  "2021_22" = payroll_2021_22,
  "2022_23" = payroll_2022_23,
  "2023_24" = payroll_2023_24,
  "2024_25" = payroll_2024_25
)

# Create list to stored processed data
processed_tsummary_list <- list()
processed_tpayroll_list <- list()
```

Data cleaning for summary data sets

```
for (season in names(tsummary_dfs)) {
  df <- tsummary_dfs[[season]]

  # Remove rows with missing values in key columns
  df <- df %>%
    filter(!is.na(W), !is.na(L), !is.na('W/L%'))
```

```
  # Remove duplicate rows, keep first occurrence
  df <- df %>%
    distinct(.keep_all = TRUE)

  # Consistency check, validate W/L% = W / (W + L)
  df <- df %>%
    mutate(WL_calc = round(W / (W + L), 3)) %>%
    filter(abs(WL_calc - `W/L%`) < 0.001)%>%  # Allow tiny rounding
difference
    select(-WL_calc)

  # Standardize team identifiers
  df <- df %>%
    mutate(Team_Code = paste0(team_codes[str_to_title(Team)], "_",
season),
           GB = as.character(GB))

  # Add column for season
  df <- df %>%
    mutate(Season = season)

  # Update processed data list
  processed_tsummary_list[[season]] <- df
}

# Aggregate Multi-Season Data
combined_tsummary_df <- bind_rows(processed_tsummary_list)

# Write aggregate data to new file
#write.csv(combined_tsummary_df,"NBA_All_Summary_Processed.csv",
row.names = FALSE)

combined_tsummary_df
```

Data cleaning for payroll data sets

```
for (season in names(tpayroll_dfs)) {
  df <- tpayroll_dfs[[season]]

  # Remove rows with missing values in key columns
  df <- df %>%
    filter(!is.na(Payroll))

  # Remove duplicate rows, keep first occurrence
  df <- df %>%
    distinct(.keep_all = TRUE)

  # Consistency check, confirm Payroll is numeric and positive
  df <- df %>%
    filter(is.numeric(Payroll),Payroll > 0)
```

```
  # Standardize team identifers
  df <- df %>%
    mutate(Team_Code = paste0(team_codes[str_to_title(Team)], "_",
season))

  # Normalize Payroll Value
  df <- df %>%
    mutate(Payroll_Millions = round(Payroll / 1e6, 3))

  # Update processed data list
  processed_tpayroll_list[[season]] <- df
}

# Aggregate Multi-Season Data
combined_tpayroll_df <- bind_rows(processed_tpayroll_list)

# Write aggregate data to new file
#write.csv(combined_tpayroll_df,"NBA_All_Payroll_Processed.csv",
row.names = FALSE)
```

Merge Season and Payroll Data by Team_Code

```
summary_payroll <- combined_tsummary_df %>%
  inner_join(combined_tpayroll_df, by="Team_Code")
```

Calculate Payroll Relative to Cap

```
relative_salary <- full_join(summary_payroll, cap, by <-
join_by(Season),
          copy=FALSE, suffix=c(".summary_payroll", ".cap"),
          keep=NULL, multiple="all") %>%
  mutate(Relative_Payroll = Payroll_Millions/Cap_Millions)
```

Select Key Columns for analysis

```
final_df <- relative_salary %>%
  select(Team_Code, W, 'W/L%',Relative_Payroll)
```

Linear Regression and Plot

```
ggplot(data = final_df, mapping = aes(x = Relative_Payroll, y = `W/L
%`)) +
  geom_point() +
  geom_smooth(method=lm)
model <- lm(`W/L%` ~ Relative_Payroll, data = final_df)
# View summary
summary(model)

final_df
```