SportsAnalytics_u3246850

2023-05-07

```
#Load required packages
# install.packages("cli")
# library(tidyverse)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.2
```

```
##
## 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)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.2
```

```
library(reshape2)
```

```
#*************2. Reading and cleaning the raw data**************************************
#*
#Load player statistics data
player_stats <- read.csv("2018-19_nba_player-statistics.csv", check.names = FALSE)

#Load player salaries data
player_salaries <- read.csv("2018-19_nba_player-salaries.csv", check.names = FALSE)

#Load team payroll data
team_payroll <- read.csv("2019-20_nba_team-payroll.csv", check.names = FALSE)

#Load team statistics data
team_stats1 <- read.csv("2018-19_nba_team-statistics_1.csv", check.names = FALSE, header = TRUE)
team_stats2 <- read.csv("2018-19_nba_team-statistics_2.csv", check.names = FALSE, header = TRUE)


# Check missing value

colSums(is.na(player_stats)) # Found missing values
```

```
## player_name         Pos         Age          Tm           G          GS
##           0           0           0           0           0           0
##          MP          FG         FGA         FG%          3P         3PA
##           0           0           0           6           0           0
##         3P%          2P         2PA         2P%        eFG%          FT
##          47           0           0          15           6           0
##         FTA         FT%         ORB         DRB         TRB         AST
##           0          43           0           0           0           0
##         STL         BLK         TOV          PF         PTS
##           0           0           0           0           0
```

```
colSums(is.na(player_salaries))
```

```
##   player_id player_name      salary
##           0           0           0
```

```
colSums(is.na(team_payroll))
```

```
## team_id    team  salary
##       0       0       0
```

```
colSums(is.na(team_stats1))
```

```
##     Rk   Team    Age      W      L     PW     PL    MOV    SOS    SRS   ORtg
##      0      0      0      0      0      0      0      0      0      0      0
##   DRtg   NRtg   Pace    FTr   3PAr    TS%   eFG%   TOV%   ORB% FT/FGA   DRB%
##      0      0      0      0      0      0      0      0      0      0      0
```

```
colSums(is.na(team_stats2))
```

```
##   Rk Team    G   MP   FG  FGA  FG%   3P  3PA  3P%   2P  2PA  2P%   FT  FTA  FT%
##    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
##  ORB  DRB  TRB  AST  STL  BLK  TOV   PF  PTS
##    0    0    0    0    0    0    0    0    0
```

```
# Check structure
str(player_stats)
```

```
## 'data.frame':    708 obs. of  29 variables:
##  $ player_name: chr  "Alex Abrines" "Quincy Acy" "Jaylen Adams" "Steven Adams" ...
##  $ Pos        : chr  "SG" "PF" "PG" "C" ...
##  $ Age        : int  25 28 22 25 21 21 25 33 21 23 ...
##  $ Tm         : chr  "OKC" "PHO" "ATL" "OKC" ...
##  $ G          : int  31 10 34 80 82 19 7 81 10 38 ...
##  $ GS         : int  2 0 1 80 28 3 0 81 1 2 ...
##  $ MP         : int  588 123 428 2669 1913 194 22 2687 120 416 ...
##  $ FG         : int  56 4 38 481 280 11 3 684 13 67 ...
##  $ FGA        : int  157 18 110 809 486 36 10 1319 39 178 ...
##  $ FG%        : num  0.357 0.222 0.345 0.595 0.576 0.306 0.3 0.519 0.333 0.376 ...
##  $ 3P         : int  41 2 25 0 3 6 0 10 3 32 ...
##  $ 3PA        : int  127 15 74 2 15 23 4 42 12 99 ...
##  $ 3P%        : num  0.323 0.133 0.338 0 0.2 0.261 0 0.238 0.25 0.323 ...
##  $ 2P         : int  15 2 13 481 277 5 3 674 10 35 ...
##  $ 2PA        : int  30 3 36 807 471 13 6 1277 27 79 ...
##  $ 2P%        : num  0.5 0.667 0.361 0.596 0.588 0.385 0.5 0.528 0.37 0.443 ...
##  $ eFG%       : num  0.487 0.278 0.459 0.595 0.579 0.389 0.3 0.522 0.372 0.466 ...
##  $ FT         : int  12 7 7 146 166 4 1 349 8 45 ...
##  $ FTA        : int  13 10 9 292 226 4 2 412 12 60 ...
##  $ FT%        : num  0.923 0.7 0.778 0.5 0.735 1 0.5 0.847 0.667 0.75 ...
##  $ ORB        : int  5 3 11 391 165 3 1 251 11 3 ...
##  $ DRB        : int  43 22 49 369 432 16 3 493 15 20 ...
##  $ TRB        : int  48 25 60 760 597 19 4 744 26 23 ...
##  $ AST        : int  20 8 65 124 184 5 6 194 13 25 ...
##  $ STL        : int  17 1 14 117 71 1 2 43 1 6 ...
##  $ BLK        : int  6 4 5 76 65 4 0 107 0 6 ...
##  $ TOV        : int  14 4 28 135 121 6 2 144 8 33 ...
##  $ PF         : int  53 24 45 204 203 13 4 179 7 47 ...
##  $ PTS        : int  165 17 108 1108 729 32 7 1727 37 211 ...
```

```
str(player_salaries)
```

```
## 'data.frame':    576 obs. of  3 variables:
##  $ player_id  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ player_name: chr  "Alex Abrines" "Quincy Acy" "Steven Adams" "Jaylen Adams" ...
##  $ salary     : int  3667645 213948 24157304 236854 2955840 77250 5285394 77250 2000000 22
347015 ...
```

```
str(team_payroll)
```

```
## 'data.frame':    30 obs. of  3 variables:
##  $ team_id: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ team   : chr  "Miami " "Golden State " "Oklahoma City " "Toronto " ...
##  $ salary : chr  "$153,171,497 " "$146,291,276 " "$144,916,427 " "$137,793,831 " ...
```

```
str(team_stats1)
```

```
## 'data.frame':    30 obs. of  22 variables:
##  $ Rk    : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Team  : chr  "Milwaukee Bucks" "Golden State Warriors" "Toronto Raptors" "Utah Jazz"
...
##  $ Age   : num  26.9 28.4 27.3 27.3 29.2 26.2 24.9 25.7 25.7 27 ...
##  $ W     : int  60 57 58 50 53 53 54 49 49 48 ...
##  $ L     : int  22 25 24 32 29 29 28 33 33 34 ...
##  $ PW    : int  61 56 56 54 53 51 51 52 50 50 ...
##  $ PL    : int  21 26 26 28 29 31 31 30 32 32 ...
##  $ MOV   : num  8.87 6.46 6.09 5.26 4.77 4.2 3.95 4.44 3.4 3.33 ...
##  $ SOS   : num  -0.82 -0.04 -0.6 0.03 0.19 0.24 0.24 -0.54 0.15 -0.57 ...
##  $ SRS   : num  8.04 6.42 5.49 5.28 4.96 4.43 4.19 3.9 3.56 2.76 ...
##  $ ORtg  : num  114 116 113 111 116 ...
##  $ DRtg  : num  105 110 107 106 111 ...
##  $ NRtg  : num  8.6 6.4 6 5.2 4.8 4.2 4.1 4.4 3.3 3.4 ...
##  $ Pace  : num  103.3 100.9 100.2 100.3 97.9 ...
##  $ FTr   : num  0.255 0.227 0.247 0.295 0.279 0.258 0.232 0.215 0.266 0.242 ...
##  $ 3PAr  : num  0.419 0.384 0.379 0.394 0.519 0.339 0.348 0.381 0.347 0.292 ...
##  $ TS%   : num  0.583 0.596 0.579 0.572 0.581 0.568 0.558 0.567 0.545 0.561 ...
##  $ eFG%  : num  0.55 0.565 0.543 0.538 0.542 0.528 0.527 0.534 0.514 0.53 ...
##  $ TOV%  : num  12 12.6 12.4 13.4 12 12.1 11.9 11.5 11.7 12.4 ...
##  $ ORB%  : num  20.8 22.5 21.9 22.9 22.8 26.6 26.6 21.6 26 21.9 ...
##  $ FT/FGA: num  0.197 0.182 0.198 0.217 0.221 0.21 0.175 0.173 0.19 0.182 ...
##  $ DRB%  : num  80.3 77.1 77.1 80.3 74.4 77.9 78 77 78.2 76.2 ...
```

```
str(team_stats2)
```

```
## 'data.frame':    30 obs. of  25 variables:
##  $ Rk  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Team: chr  "Milwaukee Bucks" "Golden State Warriors" "New Orleans Pelicans" "Philadelph
ia 76ers" ...
##  $ G   : int  82 82 82 82 82 82 82 82 82 82 ...
##  $ MP  : int  19780 19805 19755 19805 19830 19855 19855 19880 19730 19930 ...
##  $ FG  : int  3555 3612 3581 3407 3384 3470 3497 3460 3541 3456 ...
##  $ FGA : int  7471 7361 7563 7233 7178 7427 7706 7305 7637 7387 ...
##  $ FG% : num  0.476 0.491 0.473 0.471 0.471 0.467 0.454 0.474 0.464 0.468 ...
##  $ 3P  : int  1105 1087 842 889 821 904 932 1015 927 930 ...
##  $ 3PA : int  3134 2824 2449 2474 2118 2520 2677 2771 2455 2731 ...
##  $ 3P% : num  0.353 0.385 0.344 0.359 0.388 0.359 0.348 0.366 0.378 0.341 ...
##  $ 2P  : int  2450 2525 2739 2518 2563 2566 2565 2445 2614 2526 ...
##  $ 2PA : int  4337 4537 5114 4759 5060 4907 5029 4534 5182 4656 ...
##  $ 2P% : num  0.565 0.557 0.536 0.529 0.507 0.523 0.51 0.539 0.504 0.543 ...
##  $ FT  : int  1471 1339 1462 1742 1853 1558 1461 1449 1354 1508 ...
##  $ FTA : int  1904 1672 1921 2258 2340 1914 2049 1803 1865 1963 ...
##  $ FT% : num  0.773 0.801 0.761 0.771 0.792 0.814 0.713 0.804 0.726 0.768 ...
##  $ ORB : int  762 797 909 892 796 967 1031 786 906 794 ...
##  $ DRB : int  3316 2990 2969 3025 2936 2968 2911 2920 2819 2679 ...
##  $ TRB : int  4078 3787 3878 3917 3732 3935 3942 3706 3725 3473 ...
##  $ AST : int  2136 2413 2216 2207 1970 1887 1917 2085 2083 2154 ...
##  $ STL : int  615 625 610 606 561 546 766 680 679 683 ...
##  $ BLK : int  486 525 441 432 385 413 425 437 363 379 ...
##  $ TOV : int  1137 1169 1215 1223 1193 1135 1145 1150 1095 1154 ...
##  $ PF  : int  1608 1757 1732 1745 1913 1669 1839 1724 1751 1701 ...
##  $ PTS : int  9686 9650 9466 9445 9442 9402 9387 9384 9363 9350 ...
```

```
# Changing the variable type for the analysis
team_payroll$salary <- as.numeric(gsub("[\\$,]", "", team_payroll$salary))

# Cleaning player stats team name mapping to abbreviation
player_stats <- player_stats%>%
  mutate(Tm = case_when(
    Tm == "ATL" ~ "Atlanta Hawks",
    Tm == "BOS" ~ "Boston Celtics",
    Tm == "BRK" ~ "Brooklyn Nets",
    Tm == "CHI" ~ "Chicago Bulls",
    Tm == "CHO" ~ "Charlotte Hornets",
    Tm == "CLE" ~ "Cleveland Cavaliers",
    Tm == "DAL" ~ "Dallas Mavericks",
    Tm == "DEN" ~ "Denver Nuggets",
    Tm == "DET" ~ "Detroit Pistons",
    Tm == "GSW" ~ "Golden State Warriors",
    Tm == "HOU" ~ "Houston Rockets",
    Tm == "IND" ~ "Indiana Pacers",
    Tm == "LAC" ~ "Los Angeles Clippers",
    Tm == "LAL" ~ "Los Angeles Lakers",
    Tm == "MEM" ~ "Memphis Grizzlies",
    Tm == "MIA" ~ "Miami Heat",
    Tm == "MIL" ~ "Milwaukee Bucks",
    Tm == "MIN" ~ "Minnesota Timberwolves",
    Tm == "NOP" ~ "New Orleans Pelicans",
    Tm == "NYK" ~ "New York Knicks",
    Tm == "OKC" ~ "Oklahoma City Thunder",
    Tm == "ORL" ~ "Orlando Magic",
    Tm == "PHI" ~ "Philadelphia 76ers",
    Tm == "PHO" ~ "Phoenix Suns",
    Tm == "POR" ~ "Portland Trail Blazers",
    Tm == "SAC" ~ "Sacramento Kings",
    Tm == "SAS" ~ "San Antonio Spurs",
    Tm == "TOR" ~ "Toronto Raptors",
    Tm == "TOT" ~ "Total",
    Tm == "UTA" ~ "Utah Jazz",
    Tm == "WAS" ~ "Washington Wizards",
    TRUE ~ NA_character_
  ))

# Cleaning team payroll data to align with the brief team name
team_payroll$team<-trimws(team_payroll$team)
team_payroll <- team_payroll %>%
  mutate(Team = case_when(
    team == "Atlanta" ~ "Atlanta Hawks",
    team == "Boston" ~ "Boston Celtics",
    team == "Brooklyn" ~ "Brooklyn Nets",
    team == "Chicago" ~ "Chicago Bulls",
    team == "Charlotte" ~ "Charlotte Hornets",
    team == "Cleveland" ~ "Cleveland Cavaliers",
    team == "Dallas" ~ "Dallas Mavericks",
    team == "Denver" ~ "Denver Nuggets",
    team == "Detroit" ~ "Detroit Pistons",
    team == "Golden State" ~ "Golden State Warriors",
    team == "Houston" ~ "Houston Rockets",
    team == "Indiana" ~ "Indiana Pacers",
    team == "LA Clippers" ~ "Los Angeles Clippers",
    team == "LA Lakers" ~ "Los Angeles Lakers",
    team == "Memphis" ~ "Memphis Grizzlies",
    team == "Miami" ~ "Miami Heat",
    team == "Milwaukee" ~ "Milwaukee Bucks",
    team == "Minnesota" ~ "Minnesota Timberwolves",
    team == "New Orleans" ~ "New Orleans Pelicans",
    team == "New York" ~ "New York Knicks",
    team == "Oklahoma City" ~ "Oklahoma City Thunder",
    team == "Orlando" ~ "Orlando Magic",
    team == "Philadelphia" ~ "Philadelphia 76ers",
    team == "Phoenix" ~ "Phoenix Suns",
    team == "Portland" ~ "Portland Trail Blazers",
    team == "Sacramento" ~ "Sacramento Kings",
    team == "San Antonio" ~ "San Antonio Spurs",
    team == "Toronto" ~ "Toronto Raptors",
    team == "Utah" ~ "Utah Jazz",
    team == "Washington" ~ "Washington Wizards",
    TRUE ~ NA_character_
  ))
```

```
# ************************ 3. Exploratory analysis ********************************

#****************** 3a checking for errors and missing values within the datasets******************

# Check structure
glimpse(player_stats)
```

```
## Rows: 708
## Columns: 29
## $ player_name <chr> "Alex Abrines", "Quincy Acy", "Jaylen Adams", "Steven Adam…
## $ Pos         <chr> "SG", "PF", "PG", "C", "C", "SF", "SG", "C", "SG", "SG", "…
## $ Age         <int> 25, 28, 22, 25, 21, 21, 25, 33, 21, 23, 20, 26, 28, 25, 25…
## $ Tm          <chr> "Oklahoma City Thunder", "Phoenix Suns", "Atlanta Hawks", …
## $ G           <int> 31, 10, 34, 80, 82, 19, 7, 81, 10, 38, 80, 19, 81, 48, 43,…
## $ GS          <int> 2, 0, 1, 80, 28, 3, 0, 81, 1, 2, 80, 1, 81, 4, 40, 0, 8, 8…
## $ MP          <int> 588, 123, 428, 2669, 1913, 194, 22, 2687, 120, 416, 2096, …
## $ FG          <int> 56, 4, 38, 481, 280, 11, 3, 684, 13, 67, 335, 65, 257, 64,…
## $ FGA         <int> 157, 18, 110, 809, 486, 36, 10, 1319, 39, 178, 568, 141, 5…
## $ `FG%`       <dbl> 0.357, 0.222, 0.345, 0.595, 0.576, 0.306, 0.300, 0.519, 0.…
## $ `3P`        <int> 41, 2, 25, 0, 3, 6, 0, 10, 3, 32, 6, 17, 96, 24, 9, 2, 7, …
## $ `3PA`       <int> 127, 15, 74, 2, 15, 23, 4, 42, 12, 99, 45, 36, 280, 77, 34…
## $ `3P%`       <dbl> 0.323, 0.133, 0.338, 0.000, 0.200, 0.261, 0.000, 0.238, 0.…
## $ `2P`        <int> 15, 2, 13, 481, 277, 5, 3, 674, 10, 35, 329, 48, 161, 40, …
## $ `2PA`       <int> 30, 3, 36, 807, 471, 13, 6, 1277, 27, 79, 523, 105, 313, 8…
## $ `2P%`       <dbl> 0.500, 0.667, 0.361, 0.596, 0.588, 0.385, 0.500, 0.528, 0.…
## $ `eFG%`      <dbl> 0.487, 0.278, 0.459, 0.595, 0.579, 0.389, 0.300, 0.522, 0.…
## $ FT          <int> 12, 7, 7, 146, 166, 4, 1, 349, 8, 45, 197, 42, 150, 26, 37…
## $ FTA         <int> 13, 10, 9, 292, 226, 4, 2, 412, 12, 60, 278, 54, 173, 35, …
## $ `FT%`       <dbl> 0.923, 0.700, 0.778, 0.500, 0.735, 1.000, 0.500, 0.847, 0.…
## $ ORB         <int> 5, 3, 11, 391, 165, 3, 1, 251, 11, 3, 191, 8, 112, 24, 48,…
## $ DRB         <int> 43, 22, 49, 369, 432, 16, 3, 493, 15, 20, 481, 43, 498, 60…
## $ TRB         <int> 48, 25, 60, 760, 597, 19, 4, 744, 26, 23, 672, 51, 610, 84…
## $ AST         <int> 20, 8, 65, 124, 184, 5, 6, 194, 13, 25, 110, 76, 104, 23, …
## $ STL         <int> 17, 1, 14, 117, 71, 1, 2, 43, 1, 6, 43, 16, 68, 22, 54, 1,…
## $ BLK         <int> 6, 4, 5, 76, 65, 4, 0, 107, 0, 6, 120, 4, 33, 13, 37, 0, 1…
## $ TOV         <int> 14, 4, 28, 135, 121, 6, 2, 144, 8, 33, 103, 26, 72, 23, 58…
## $ PF          <int> 53, 24, 45, 204, 203, 13, 4, 179, 7, 47, 184, 46, 143, 48,…
## $ PTS         <int> 165, 17, 108, 1108, 729, 32, 7, 1727, 37, 211, 873, 189, 7…
```

```
glimpse(player_salaries)
```

```
## Rows: 576
## Columns: 3
## $ player_id   <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,…
## $ player_name <chr> "Alex Abrines", "Quincy Acy", "Steven Adams", "Jaylen Adam…
## $ salary      <int> 3667645, 213948, 24157304, 236854, 2955840, 77250, 5285394…
```

```
glimpse(team_payroll)
```

```
## Rows: 30
## Columns: 4
## $ team_id <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18,…
## $ team    <chr> "Miami", "Golden State", "Oklahoma City", "Toronto", "Milwauke…
## $ salary  <dbl> 153171497, 146291276, 144916427, 137793831, 130988604, 1302566…
## $ Team    <chr> "Miami Heat", "Golden State Warriors", "Oklahoma City Thunder"…
```

```
glimpse(team_stats1)
```

```
## Rows: 30
## Columns: 22
## $ Rk       <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18…
## $ Team     <chr> "Milwaukee Bucks", "Golden State Warriors", "Toronto Raptors"…
## $ Age      <dbl> 26.9, 28.4, 27.3, 27.3, 29.2, 26.2, 24.9, 25.7, 25.7, 27.0, 2…
## $ W        <int> 60, 57, 58, 50, 53, 53, 54, 49, 49, 48, 51, 48, 48, 42, 42, 3…
## $ L        <int> 22, 25, 24, 32, 29, 29, 28, 33, 33, 34, 31, 34, 34, 40, 40, 4…
## $ PW       <int> 61, 56, 56, 54, 53, 51, 51, 52, 50, 50, 48, 45, 43, 43, 41, 4…
## $ PL       <int> 21, 26, 26, 28, 29, 31, 31, 30, 32, 32, 34, 37, 39, 39, 41, 4…
## $ MOV      <dbl> 8.87, 6.46, 6.09, 5.26, 4.77, 4.20, 3.95, 4.44, 3.40, 3.33, 2…
## $ SOS      <dbl> -0.82, -0.04, -0.60, 0.03, 0.19, 0.24, 0.24, -0.54, 0.15, -0.…
## $ SRS      <dbl> 8.04, 6.42, 5.49, 5.28, 4.96, 4.43, 4.19, 3.90, 3.56, 2.76, 2…
## $ ORtg     <dbl> 113.8, 115.9, 113.1, 110.9, 115.5, 114.7, 113.0, 112.2, 110.3…
## $ DRtg     <dbl> 105.2, 109.5, 107.1, 105.7, 110.7, 110.5, 108.9, 107.8, 107.0…
## $ NRtg     <dbl> 8.6, 6.4, 6.0, 5.2, 4.8, 4.2, 4.1, 4.4, 3.3, 3.4, 2.6, 1.7, 0…
## $ Pace     <dbl> 103.3, 100.9, 100.2, 100.3, 97.9, 99.1, 97.7, 99.6, 102.8, 98…
## $ FTr      <dbl> 0.255, 0.227, 0.247, 0.295, 0.279, 0.258, 0.232, 0.215, 0.266…
## $ `3PAr`   <dbl> 0.419, 0.384, 0.379, 0.394, 0.519, 0.339, 0.348, 0.381, 0.347…
## $ `TS%`    <dbl> 0.583, 0.596, 0.579, 0.572, 0.581, 0.568, 0.558, 0.567, 0.545…
## $ `eFG%`   <dbl> 0.550, 0.565, 0.543, 0.538, 0.542, 0.528, 0.527, 0.534, 0.514…
## $ `TOV%`   <dbl> 12.0, 12.6, 12.4, 13.4, 12.0, 12.1, 11.9, 11.5, 11.7, 12.4, 1…
## $ `ORB%`   <dbl> 20.8, 22.5, 21.9, 22.9, 22.8, 26.6, 26.6, 21.6, 26.0, 21.9, 2…
## $ `FT/FGA` <dbl> 0.197, 0.182, 0.198, 0.217, 0.221, 0.210, 0.175, 0.173, 0.190…
## $ `DRB%`   <dbl> 80.3, 77.1, 77.1, 80.3, 74.4, 77.9, 78.0, 77.0, 78.2, 76.2, 7…
```

```
glimpse(team_stats2)
```

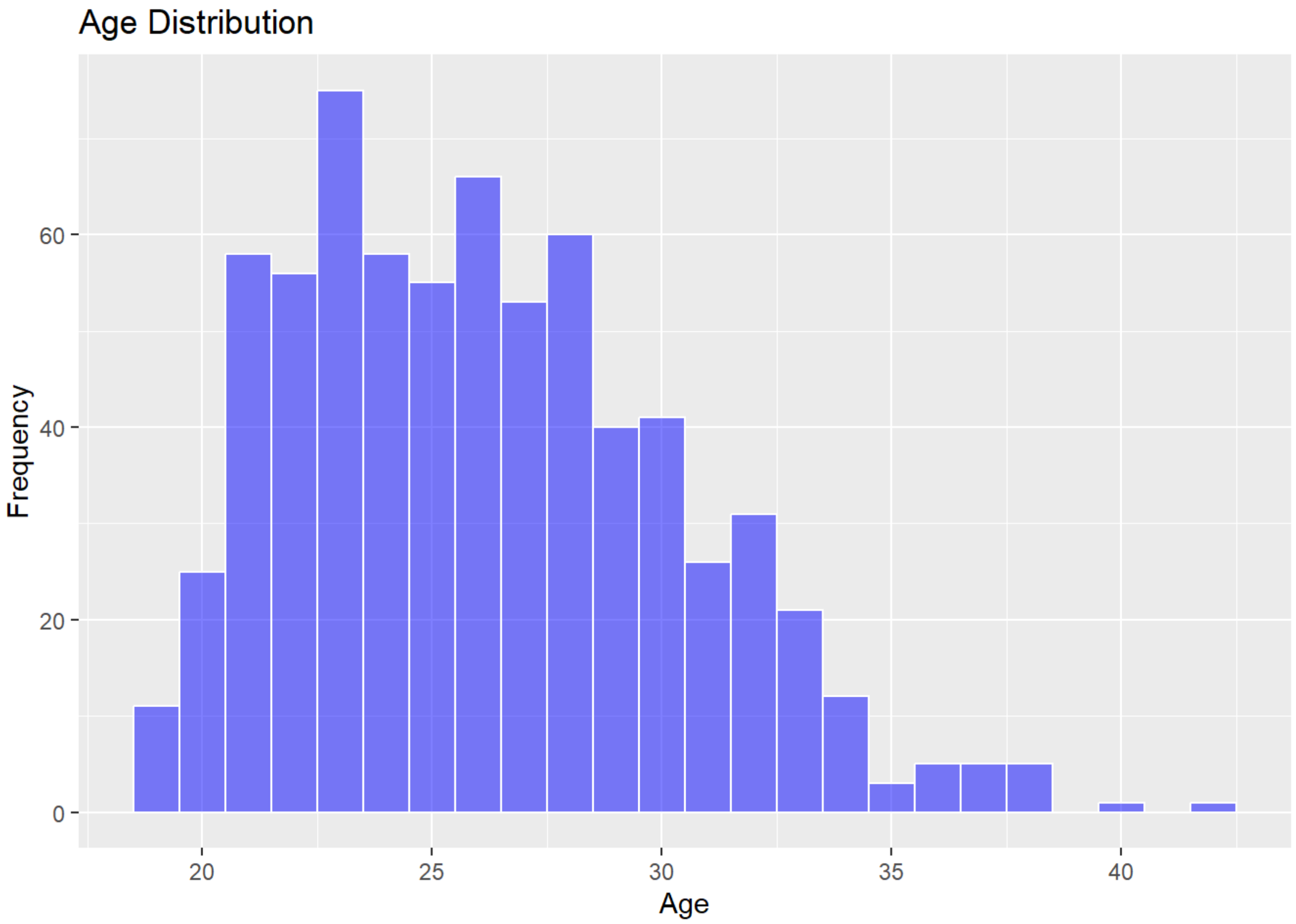
```
## Rows: 30
## Columns: 25
## $ Rk    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 1…
## $ Team  <chr> "Milwaukee Bucks", "Golden State Warriors", "New Orleans Pelican…
## $ G     <int> 82, 82, 82, 82, 82, 82, 82, 82, 82, 82, 82, 82, 82, 82, 82, 82, …
## $ MP    <int> 19780, 19805, 19755, 19805, 19830, 19855, 19855, 19880, 19730, 1…
## $ FG    <int> 3555, 3612, 3581, 3407, 3384, 3470, 3497, 3460, 3541, 3456, 3218…
## $ FGA   <int> 7471, 7361, 7563, 7233, 7178, 7427, 7706, 7305, 7637, 7387, 7163…
## $ `FG%` <dbl> 0.476, 0.491, 0.473, 0.471, 0.471, 0.467, 0.454, 0.474, 0.464, 0…
## $ `3P`  <int> 1105, 1087, 842, 889, 821, 904, 932, 1015, 927, 930, 1323, 1067,…
## $ `3PA` <int> 3134, 2824, 2449, 2474, 2118, 2520, 2677, 2771, 2455, 2731, 3721…
## $ `3P%` <dbl> 0.353, 0.385, 0.344, 0.359, 0.388, 0.359, 0.348, 0.366, 0.378, 0…
## $ `2P`  <int> 2450, 2525, 2739, 2518, 2563, 2566, 2565, 2445, 2614, 2526, 1895…
## $ `2PA` <int> 4337, 4537, 5114, 4759, 5060, 4907, 5029, 4534, 5182, 4656, 3442…
## $ `2P%` <dbl> 0.565, 0.557, 0.536, 0.529, 0.507, 0.523, 0.510, 0.539, 0.504, 0…
## $ FT    <int> 1471, 1339, 1462, 1742, 1853, 1558, 1461, 1449, 1354, 1508, 1582…
## $ FTA   <int> 1904, 1672, 1921, 2258, 2340, 1914, 2049, 1803, 1865, 1963, 2001…
## $ `FT%` <dbl> 0.773, 0.801, 0.761, 0.771, 0.792, 0.814, 0.713, 0.804, 0.726, 0…
## $ ORB   <int> 762, 797, 909, 892, 796, 967, 1031, 786, 906, 794, 836, 955, 923…
## $ DRB   <int> 3316, 2990, 2969, 3025, 2936, 2968, 2911, 2920, 2819, 2679, 2613…
## $ TRB   <int> 4078, 3787, 3878, 3917, 3732, 3935, 3942, 3706, 3725, 3473, 3449…
## $ AST   <int> 2136, 2413, 2216, 2207, 1970, 1887, 1917, 2085, 2083, 2154, 1741…
## $ STL   <int> 615, 625, 610, 606, 561, 546, 766, 680, 679, 683, 700, 675, 683,…
## $ BLK   <int> 486, 525, 441, 432, 385, 413, 425, 437, 363, 379, 405, 419, 411,…
## $ TOV   <int> 1137, 1169, 1215, 1223, 1193, 1135, 1145, 1150, 1095, 1154, 1094…
## $ PF    <int> 1608, 1757, 1732, 1745, 1913, 1669, 1839, 1724, 1751, 1701, 1803…
## $ PTS   <int> 9686, 9650, 9466, 9445, 9442, 9402, 9387, 9384, 9363, 9350, 9341…
```

```
#**********3b Checking distribution of variable****************************888
#using  summary statistics and visualizations by histogram for different data frames: player_
stats, player_salaries, team_payroll, team_stats1, and team_stats2.

# Summary of player
summary(player_stats)
```

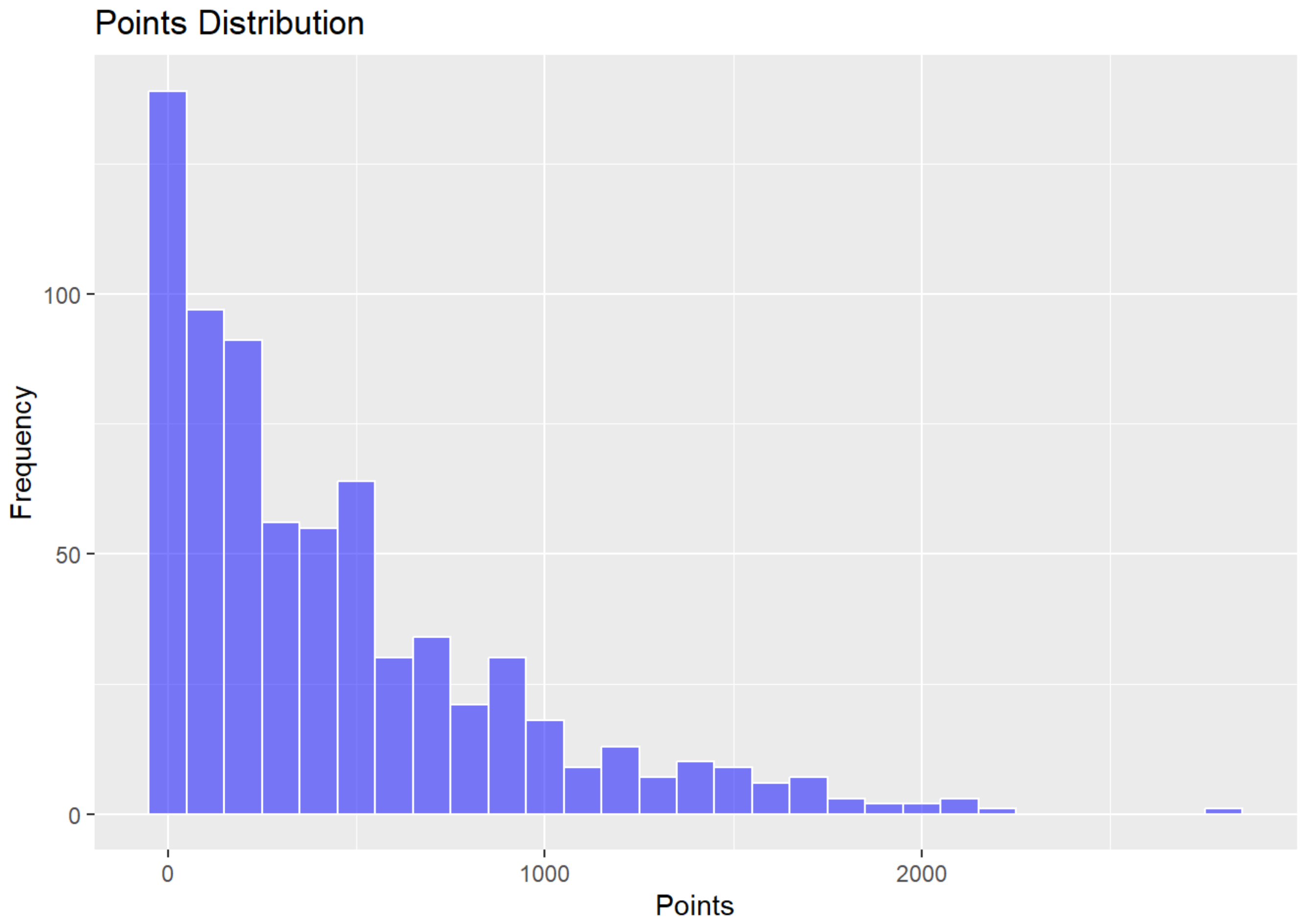
```
##  player_name            Pos                 Age              Tm
##  Length:708         Length:708         Min.   :19.00   Length:708
##  Class :character   Class :character   1st Qu.:23.00   Class :character
##  Mode  :character   Mode  :character   Median :26.00   Mode  :character
##                                        Mean   :26.14
##                                        3rd Qu.:29.00
##                                        Max.   :42.00
##
##        G               GS              MP                FG
##  Min.   : 1.00   Min.   : 0.00   Min.   :   1.0   Min.   :  0.0
##  1st Qu.:19.00   1st Qu.: 0.00   1st Qu.: 245.2   1st Qu.: 32.0
##  Median :44.00   Median : 6.00   Median : 788.0   Median :108.5
##  Mean   :42.88   Mean   :19.85   Mean   : 972.3   Mean   :162.6
##  3rd Qu.:68.00   3rd Qu.:32.00   3rd Qu.:1579.5   3rd Qu.:236.2
##  Max.   :82.00   Max.   :82.00   Max.   :3028.0   Max.   :843.0
##
##       FGA               FG%               3P              3PA
##  Min.   :   0.00   Min.   :0.0000   Min.   :  0.00   Min.   :   0.0
##  1st Qu.:  72.75   1st Qu.:0.4000   1st Qu.:  4.00   1st Qu.:  13.0
##  Median : 256.00   Median :0.4340   Median : 26.00   Median :  79.0
##  Mean   : 355.42   Mean   :0.4373   Mean   : 46.12   Mean   : 130.1
##  3rd Qu.: 526.00   3rd Qu.:0.4850   3rd Qu.: 69.25   3rd Qu.: 200.0
##  Max.   :1909.00   Max.   :1.0000   Max.   :378.00   Max.   :1028.0
##                    NA's   :6
##       3P%              2P             2PA              2P%
##  Min.   :0.000   Min.   :  0.0   Min.   :   0.0   Min.   :0.0000
##  1st Qu.:0.286   1st Qu.: 18.0   1st Qu.:  40.0   1st Qu.:0.4500
##  Median :0.335   Median : 71.0   Median : 138.0   Median :0.5000
##  Mean   :0.315   Mean   :116.5   Mean   : 225.3   Mean   :0.4923
##  3rd Qu.:0.372   3rd Qu.:164.2   3rd Qu.: 314.5   3rd Qu.:0.5480
##  Max.   :1.000   Max.   :674.0   Max.   :1277.0   Max.   :1.0000
##  NA's   :47                                       NA's   :15
##       eFG%              FT              FTA              FT%
##  Min.   :0.0000   Min.   :  0.00   Min.   :  0.00   Min.   :0.0000
##  1st Qu.:0.4700   1st Qu.: 11.00   1st Qu.: 15.00   1st Qu.:0.6840
##  Median :0.5080   Median : 39.00   Median : 51.00   Median :0.7630
##  Mean   :0.5002   Mean   : 69.88   Mean   : 91.01   Mean   :0.7396
##  3rd Qu.:0.5517   3rd Qu.: 94.00   3rd Qu.:123.00   3rd Qu.:0.8250
##  Max.   :1.5000   Max.   :754.00   Max.   :858.00   Max.   :1.0000
##  NA's   :6                                          NA's   :43
##       ORB              DRB             TRB              AST
##  Min.   :  0.00   Min.   :  0.0   Min.   :   0.00   Min.   :  0.00
##  1st Qu.:  7.00   1st Qu.: 32.0   1st Qu.:  41.75   1st Qu.: 16.00
##  Median : 23.00   Median :102.5   Median : 128.50   Median : 56.00
##  Mean   : 41.01   Mean   :139.1   Mean   : 180.12   Mean   : 96.32
##  3rd Qu.: 54.00   3rd Qu.:199.0   3rd Qu.: 258.00   3rd Qu.:124.25
##  Max.   :423.00   Max.   :809.0   Max.   :1232.00   Max.   :784.00
##
##       STL              BLK              TOV               PF
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.0
##  1st Qu.:  7.00   1st Qu.:  3.00   1st Qu.: 11.00   1st Qu.: 24.0
##  Median : 21.00   Median : 10.00   Median : 36.00   Median : 73.5
##  Mean   : 30.58   Mean   : 19.29   Mean   : 53.52   Mean   : 84.1
##  3rd Qu.: 46.00   3rd Qu.: 25.00   3rd Qu.: 75.00   3rd Qu.:131.2
##  Max.   :170.00   Max.   :199.00   Max.   :387.00   Max.   :292.0
```

```
##
##       PTS
##  Min.   :   0.00
##  1st Qu.:  82.75
##  Median : 294.00
##  Mean   : 441.29
##  3rd Qu.: 634.00
##  Max.   :2818.00
##
```

```
# Distributon of player age
ggplot(player_stats, aes(x = Age)) +
  geom_histogram(binwidth = 1, color = "white", fill = "blue", alpha = 0.5) +
  labs(title = "Age Distribution", x = "Age", y = "Frequency")
```



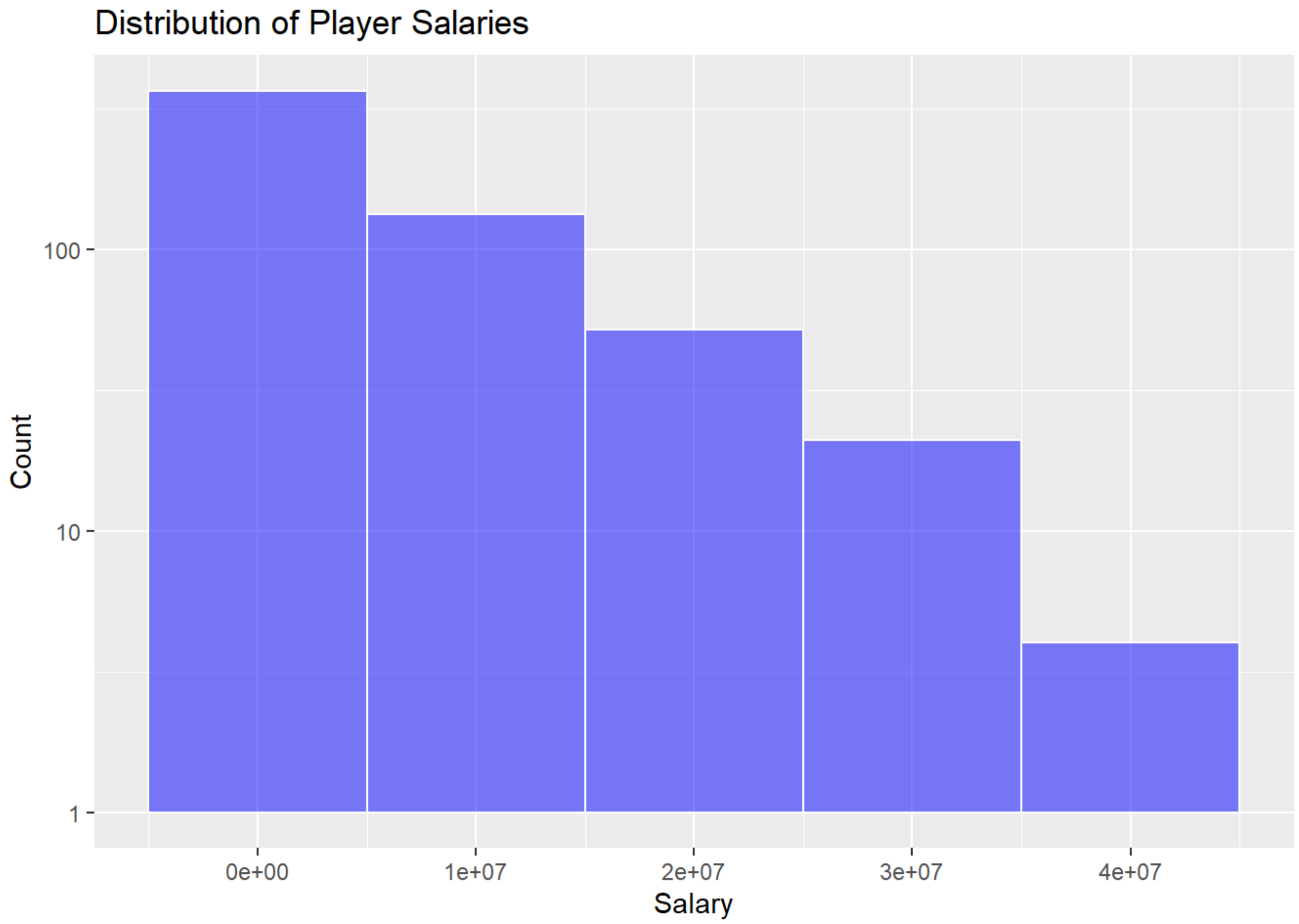
```
# Distribution of player points
ggplot(player_stats, aes(x = PTS)) +
  geom_histogram(binwidth = 100, color = "white", fill = "blue", alpha = 0.5) +
  labs(title = "Points Distribution", x = "Points", y = "Frequency")
```



```
# Summary of player salary
summary(player_salaries)
```

```
##    player_id      player_name            salary
##  Min.   :  1.0   Length:576         Min.   :   47370
##  1st Qu.:144.8   Class :character   1st Qu.: 1349383
##  Median :288.5   Mode  :character   Median : 2530560
##  Mean   :288.5                      Mean   : 6258149
##  3rd Qu.:432.2                      3rd Qu.: 9000000
##  Max.   :576.0                      Max.   :37457154
```

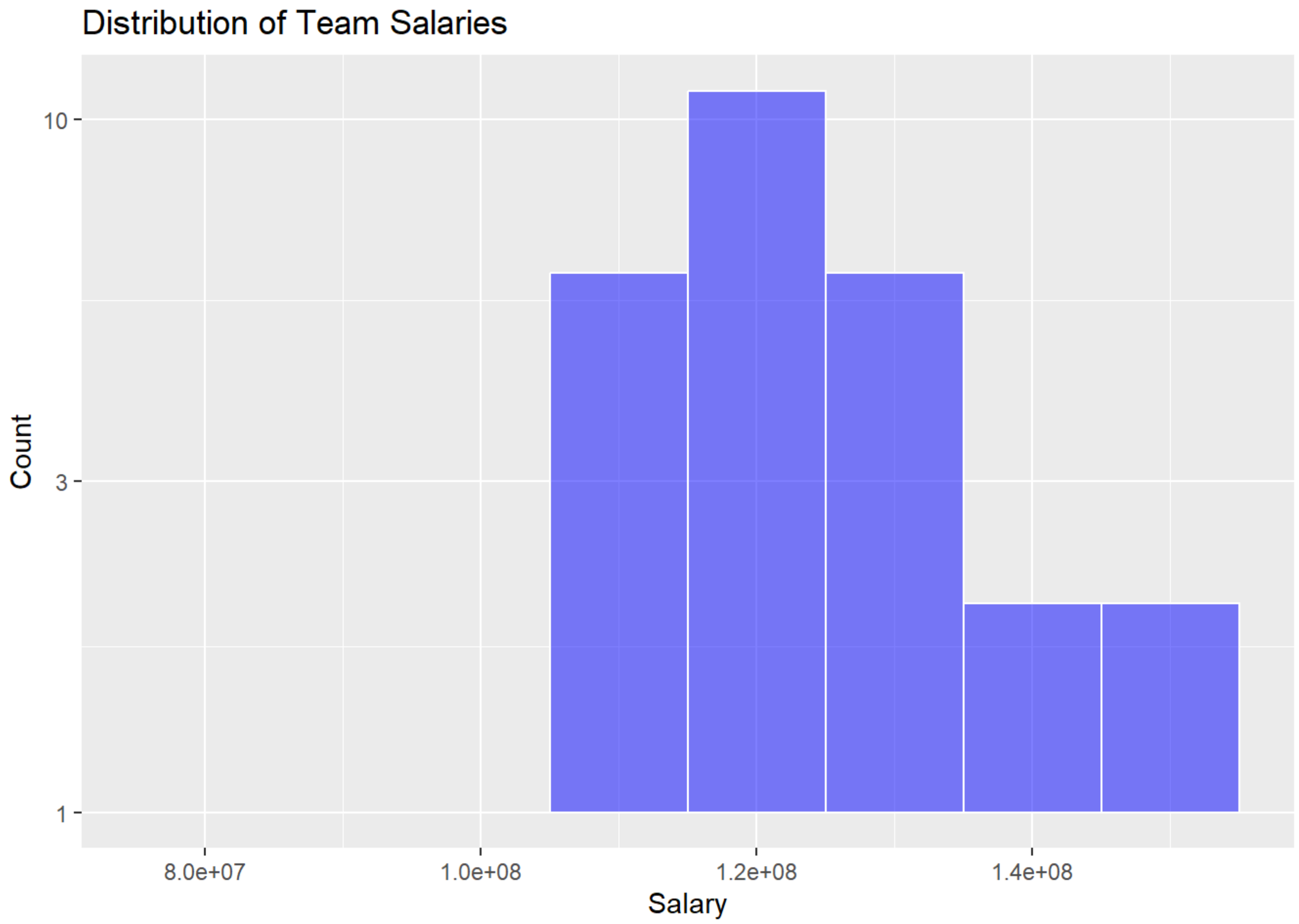
```
# Distribution of player Salary
ggplot(player_salaries, aes(x = salary)) +
  geom_histogram(binwidth = 10000000, color = "white", fill = "blue", alpha = 0.5) +
  scale_y_log10() +
  labs(title = "Distribution of Player Salaries",
       x = "Salary",
       y = "Count")
```



```
# Summary of team salary
summary(team_payroll)
```

```
##     team_id          team               salary              Team          
##  Min.   : 1.00   Length:30          Min.   : 79180081   Length:30         
##  1st Qu.: 8.25   Class :character   1st Qu.:113968170   Class :character  
##  Median :15.50   Mode  :character   Median :121508324   Mode  :character  
##  Mean   :15.50                      Mean   :120157121                     
##  3rd Qu.:22.75                      3rd Qu.:126382440                     
##  Max.   :30.00                      Max.   :153171497
```

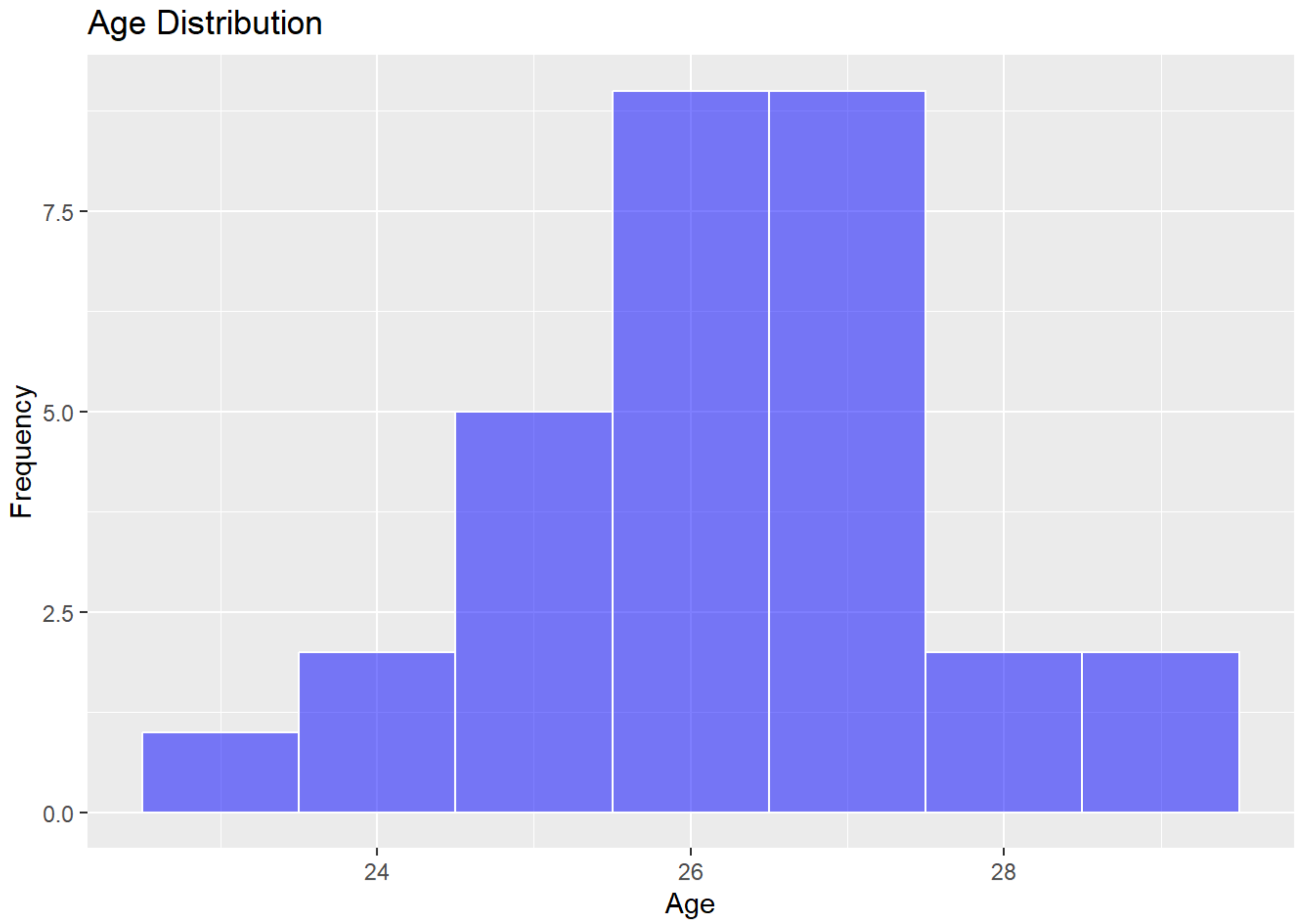
```
# Distribution of team Salary
ggplot(team_payroll, aes(x = salary)) +
  geom_histogram(binwidth = 10000000, color = "white", fill = "blue", alpha = 0.5) +
  scale_y_log10() +
  labs(title = "Distribution of Team Salaries",
       x = "Salary",
       y = "Count")
```



```
summary(team_stats1)
```

```
##        Rk            Team                Age              W
##  Min.   : 1.00   Length:30          Min.   :23.40   Min.   :17.00
##  1st Qu.: 8.25   Class :character   1st Qu.:25.48   1st Qu.:33.00
##  Median :15.50   Mode  :character   Median :26.30   Median :41.50
##  Mean   :15.50                      Mean   :26.28   Mean   :41.00
##  3rd Qu.:22.75                      3rd Qu.:27.00   3rd Qu.:49.75
##  Max.   :30.00                      Max.   :29.20   Max.   :60.00
##        L               PW              PL             MOV
##  Min.   :22.00   Min.   :19.00   Min.   :21.00   Min.   :-9.610
##  1st Qu.:32.25   1st Qu.:37.00   1st Qu.:31.25   1st Qu.:-1.665
##  Median :40.50   Median :40.50   Median :41.50   Median :-0.150
##  Mean   :41.00   Mean   :41.10   Mean   :40.90   Mean   : 0.001
##  3rd Qu.:49.00   3rd Qu.:50.75   3rd Qu.:45.00   3rd Qu.: 3.812
##  Max.   :65.00   Max.   :61.00   Max.   :63.00   Max.   : 8.870
##       SOS              SRS                 ORtg            DRtg
##  Min.   :-0.820   Min.   :-9.390000   Min.   :104.5   Min.   :105.2
##  1st Qu.:-0.325   1st Qu.:-1.327500   1st Qu.:108.3   1st Qu.:108.3
##  Median : 0.110   Median :-0.425000   Median :110.7   Median :110.2
##  Mean   :-0.003   Mean   :-0.003333   Mean   :110.4   Mean   :110.4
##  3rd Qu.: 0.240   3rd Qu.: 3.815000   3rd Qu.:112.5   3rd Qu.:112.6
##  Max.   : 0.730   Max.   : 8.040000   Max.   :115.9   Max.   :117.6
##       NRtg                Pace             FTr              3PAr
##  Min.   :-9.900000   Min.   : 96.60   Min.   :0.2150   Min.   :0.2860
##  1st Qu.:-1.650000   1st Qu.: 98.22   1st Qu.:0.2425   1st Qu.:0.3325
##  Median :-0.150000   Median : 99.90   Median :0.2570   Median :0.3475
##  Mean   :-0.003333   Mean   :100.04   Mean   :0.2588   Mean   :0.3588
##  3rd Qu.: 3.925000   3rd Qu.:101.55   3rd Qu.:0.2692   3rd Qu.:0.3832
##  Max.   : 8.600000   Max.   :103.90   Max.   :0.3260   Max.   :0.5190
##       TS%              eFG%             TOV%            ORB%
##  Min.   :0.5290   Min.   :0.4900   Min.   :10.90   Min.   :19.40
##  1st Qu.:0.5505   1st Qu.:0.5140   1st Qu.:11.93   1st Qu.:21.75
##  Median :0.5555   Median :0.5255   Median :12.40   Median :22.60
##  Mean   :0.5596   Mean   :0.5242   Mean   :12.40   Mean   :22.89
##  3rd Qu.:0.5710   3rd Qu.:0.5317   3rd Qu.:12.85   3rd Qu.:24.40
##  Max.   :0.5960   Max.   :0.5650   Max.   :14.30   Max.   :26.60
##      FT/FGA            DRB%
##  Min.   :0.1680   Min.   :72.50
##  1st Qu.:0.1825   1st Qu.:76.25
##  Median :0.1960   Median :77.10
##  Mean   :0.1983   Mean   :77.07
##  3rd Qu.:0.2100   3rd Qu.:77.97
##  Max.   :0.2580   Max.   :80.30
```

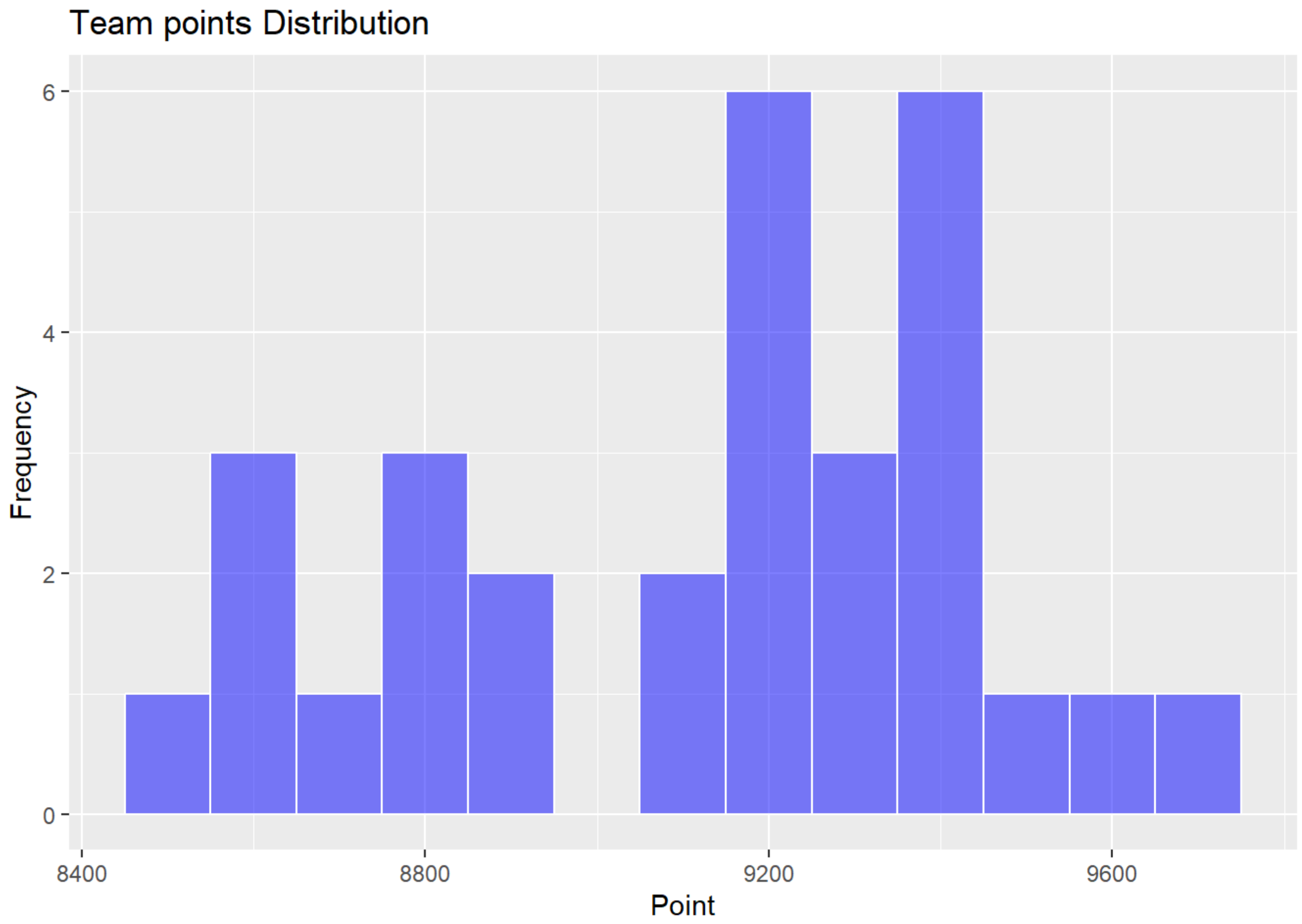
```
# Distribution of team age
ggplot(team_stats1, aes(x = Age)) +
  geom_histogram(binwidth = 1, color = "white", fill = "blue", alpha = 0.5) +
  labs(title = "Age Distribution", x = "Age", y = "Frequency")
```



```
summary(team_stats2)
```

```
##        Rk            Team                 G            MP              FG
##  Min.   : 1.00   Length:30          Min.   :82   Min.   :19705   Min.   :3113
##  1st Qu.: 8.25   Class :character   1st Qu.:82   1st Qu.:19780   1st Qu.:3272
##  Median :15.50   Mode  :character   Median :82   Median :19805   Median :3391
##  Mean   :15.50                      Mean   :82   Mean   :19815   Mean   :3369
##  3rd Qu.:22.75                      3rd Qu.:82   3rd Qu.:19855   3rd Qu.:3466
##  Max.   :30.00                      Max.   :82   Max.   :19980   Max.   :3612
##       FGA            FG%               3P              3PA
##  Min.   :6924   Min.   :0.4330   Min.   : 745.0   Min.   :2071
##  1st Qu.:7189   1st Qu.:0.4500   1st Qu.: 830.8   1st Qu.:2405
##  Median :7306   Median :0.4615   Median : 927.5   Median :2602
##  Mean   :7315   Mean   :0.4605   Mean   : 931.8   Mean   :2625
##  3rd Qu.:7424   3rd Qu.:0.4708   3rd Qu.:1009.5   3rd Qu.:2815
##  Max.   :7706   Max.   :0.4910   Max.   :1323.0   Max.   :3721
##       3P%               2P            2PA            2P%               FT
##  Min.   :0.3290   Min.   :1895   Min.   :3442   Min.   :0.4790   Min.   :1231
##  1st Qu.:0.3480   1st Qu.:2322   1st Qu.:4535   1st Qu.:0.5070   1st Qu.:1340
##  Median :0.3525   Median :2474   Median :4716   Median :0.5175   Median :1451
##  Mean   :0.3555   Mean   :2437   Mean   :4691   Mean   :0.5202   Mean   :1450
##  3rd Qu.:0.3590   3rd Qu.:2564   3rd Qu.:4998   3rd Qu.:0.5343   3rd Qu.:1532
##  Max.   :0.3920   Max.   :2739   Max.   :5182   Max.   :0.5650   Max.   :1853
##       FTA            FT%              ORB              DRB            TRB
##  Min.   :1575   Min.   :0.6950   Min.   : 718.0   Min.   :2563   Min.   :3311
##  1st Qu.:1741   1st Qu.:0.7482   1st Qu.: 794.5   1st Qu.:2769   1st Qu.:3607
##  Median :1900   Median :0.7715   Median : 833.5   Median :2864   Median :3720
##  Mean   :1892   Mean   :0.7670   Mean   : 848.5   Mean   :2855   Mean   :3704
##  3rd Qu.:1987   3rd Qu.:0.7917   3rd Qu.: 908.2   3rd Qu.:2932   3rd Qu.:3803
##  Max.   :2340   Max.   :0.8190   Max.   :1031.0   Max.   :3316   Max.   :4078
##       AST            STL             BLK             TOV             PF
##  Min.   :1646   Min.   :501.0   Min.   :195.0   Min.   : 992   Min.   :1487
##  1st Qu.:1917   1st Qu.:563.0   1st Qu.:380.5   1st Qu.:1103   1st Qu.:1653
##  Median :2016   Median :621.5   Median :415.5   Median :1148   Median :1712
##  Mean   :2016   Mean   :626.0   Mean   :406.2   Mean   :1155   Mean   :1714
##  3rd Qu.:2132   3rd Qu.:682.2   3rd Qu.:439.2   3rd Qu.:1204   3rd Qu.:1762
##  Max.   :2413   Max.   :766.0   Max.   :525.0   Max.   :1397   Max.   :1932
##       PTS
##  Min.   :8490
##  1st Qu.:8826
##  Median :9184
##  Mean   :9119
##  3rd Qu.:9379
##  Max.   :9686
```

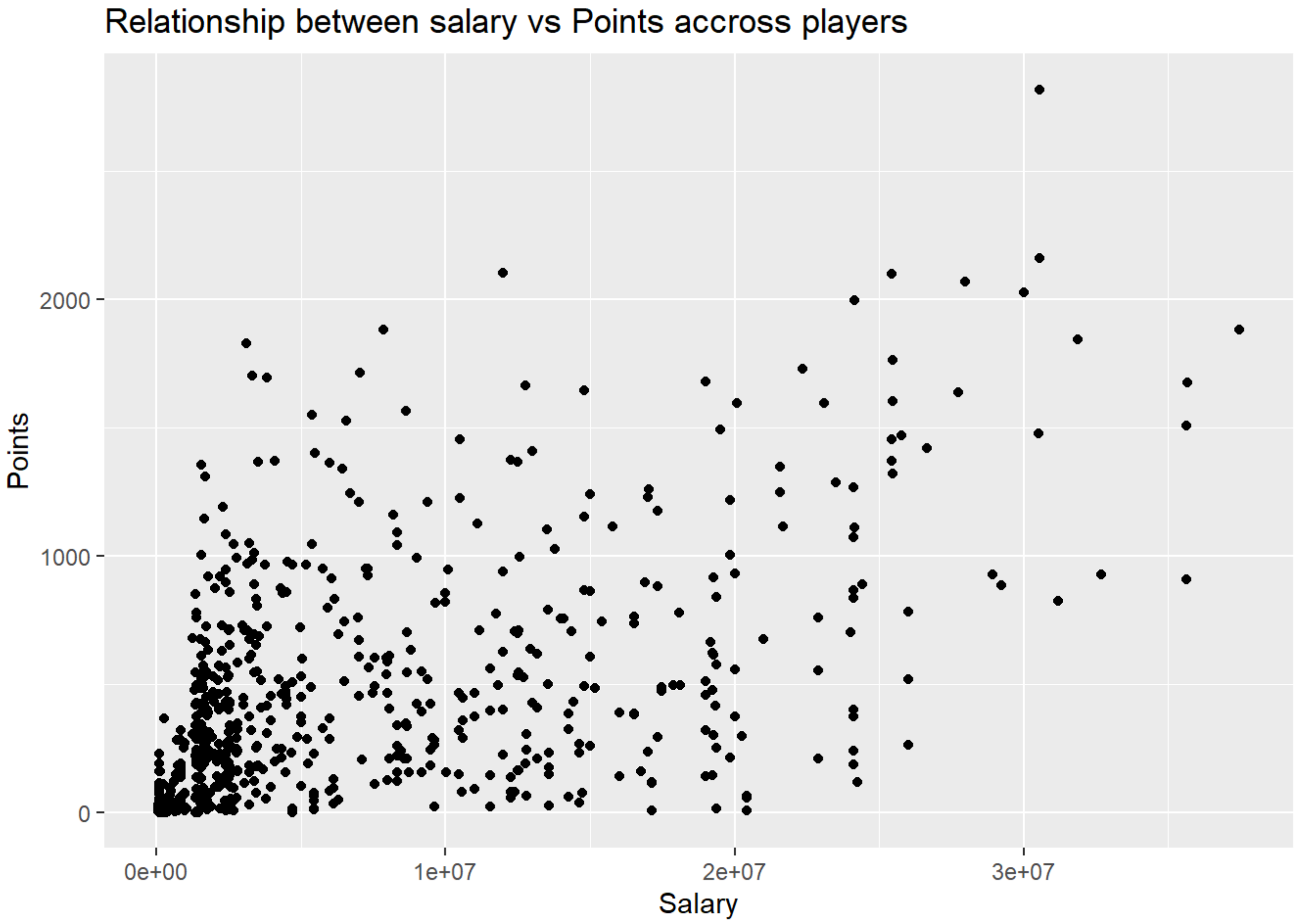
```
ggplot(team_stats2, aes(x = PTS)) +
  geom_histogram(binwidth = 100, color = "white", fill = "blue", alpha = 0.5) +
  labs(title = "Team points Distribution", x = "Point", y = "Frequency")
```



```
#********** 3c checking for relationships between variables, or differences between groups***
**********

# Merge player_stats and player_salaries datasets
player_stats_salaries <- left_join( player_stats,player_salaries, by = "player_name")
colSums(is.na(player_stats_salaries))
```

```
## player_name         Pos         Age          Tm           G          GS 
##           0           0           0           0           0           0 
##          MP          FG         FGA         FG%          3P         3PA 
##           0           0           0           6           0           0 
##         3P%          2P         2PA         2P%        eFG%          FT 
##          47           0           0          15           6           0 
##         FTA         FT%         ORB         DRB         TRB         AST 
##           0          43           0           0           0           0 
##         STL         BLK         TOV          PF         PTS   player_id 
##           0           0           0           0           0          22 
##      salary 
##          22
```

```
# Removing null values rows in salaries
player_stats_salaries <- player_stats_salaries[complete.cases(player_stats_salaries$player_i
d), ]

# Scatter plot to understand the relationship between salary and PTS of a player
ggplot(player_stats_salaries, aes(x = salary, y = PTS)) +
  geom_point() +
  labs(title = "Relationship between salary vs Points accross players",x = "Salary", y = "Poi
nts")
```

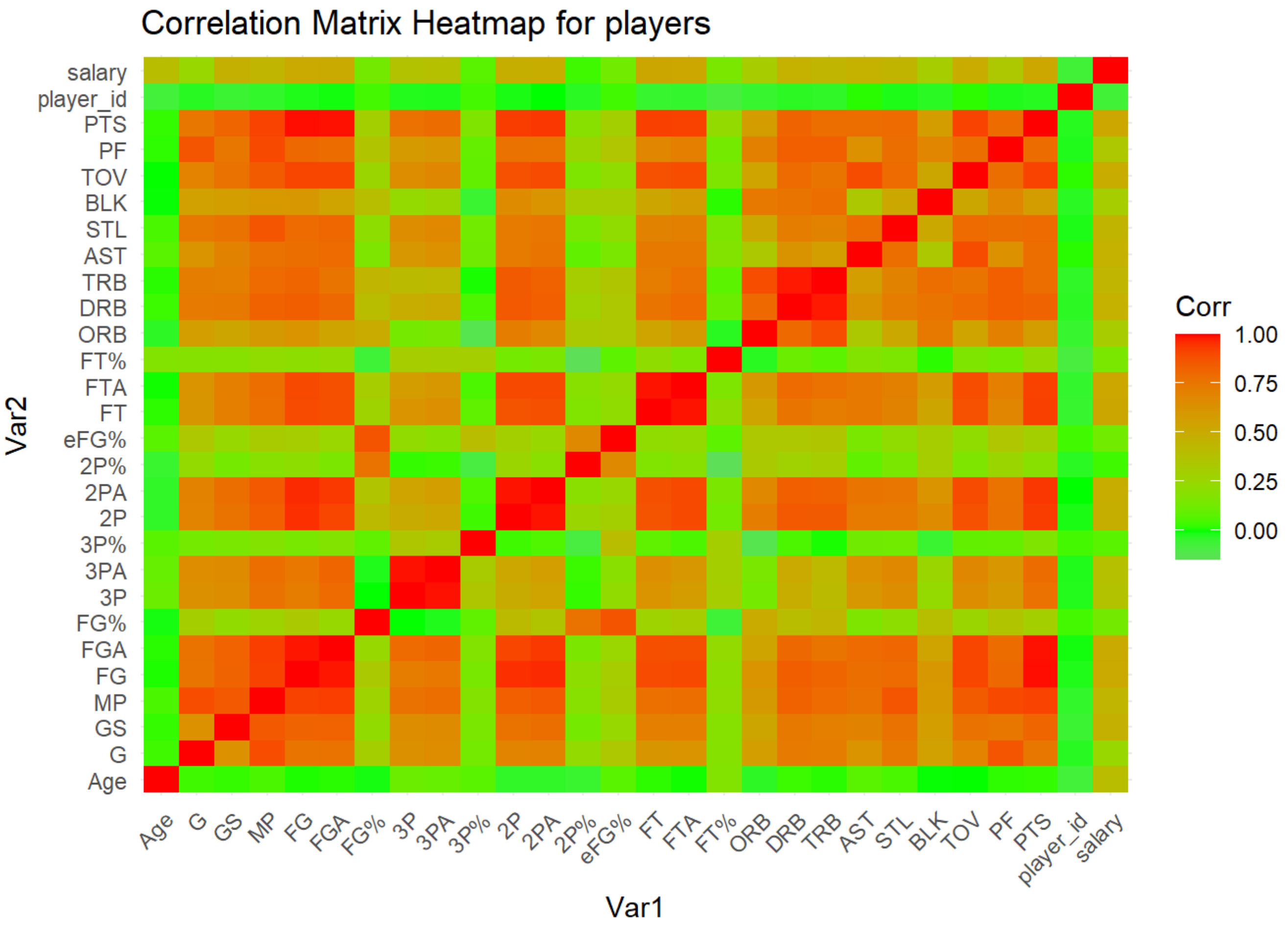


```
# There seems to be an increased positive correlation between salary of a player and points
```

```
# Correlation accross different player statistics

# find numeric variables
player_stats_salaries_omit<-na.omit(player_stats_salaries)
num_vars <- sapply(player_stats_salaries_omit, is.numeric)

# subset dataframe to include only numeric variables
players_corr <- player_stats_salaries_omit[,num_vars]

# calculate correlation matrix
#cor(players_corr) calculates the correlation matrix for the players_corr dataset, which cont
ains the relevant numeric variables from the players dataset.
cor_matrix <- cor(players_corr)
cor_matrix
```

```
##                     Age          G          GS          MP           FG
## Age         1.0000000000  0.0359558  0.02538208  0.05150767  0.009139349
## G           0.0359557956  1.0000000  0.63213869  0.89255064  0.759286248
## GS          0.0253820776  0.6321387  1.00000000  0.85379068  0.821968231
## MP          0.0515076673  0.8925506  0.85379068  1.00000000  0.918061332
## FG          0.0091393489  0.7592862  0.82196823  0.91806133  1.000000000
## FGA         0.0169983954  0.7658724  0.82248085  0.92886393  0.987968162
## FG%        -0.0095865074  0.3037297  0.22583815  0.27461953  0.330788312
## 3P          0.1131953212  0.6372458  0.65235210  0.76892838  0.728557516
## 3PA         0.1017385350  0.6558404  0.66018795  0.78779832  0.744927343
## 3P%         0.0761883889  0.1334317  0.15122653  0.18161596  0.145824277
## 2P         -0.0363798902  0.6910183  0.76394881  0.83618387  0.956983916
## 2PA        -0.0367620072  0.7008662  0.77915080  0.85459404  0.964780681
## 2P%        -0.0479275798  0.2277981  0.14082115  0.19505687  0.215490135
## eFG%        0.0762496378  0.3408862  0.24168325  0.31727432  0.310455200
## FT          0.0195584825  0.6054723  0.71221647  0.77529616  0.897165111
## FTA         0.0039865865  0.6197502  0.71319743  0.78252251  0.901472061
## FT%         0.1762853515  0.1837874  0.18870067  0.22264285  0.208723489
## ORB        -0.0308311730  0.5690956  0.53234858  0.59120968  0.617571136
## DRB         0.0338183370  0.7346048  0.74326540  0.82839962  0.840372721
## TRB         0.0168786144  0.7180038  0.71414310  0.79534188  0.811890498
## AST         0.0764047780  0.6196993  0.69733943  0.76888682  0.778694948
## STL         0.0508178682  0.7396360  0.76743348  0.86785135  0.794999817
## BLK        -0.0026024958  0.5503855  0.56786901  0.59183193  0.601568347
## TOV         0.0009743469  0.6979516  0.76991458  0.85009949  0.911698085
## PF          0.0207331284  0.8691229  0.74704226  0.89939050  0.804680076
## PTS         0.0242972333  0.7496931  0.81824208  0.91481614  0.993611330
## player_id -0.0696552960 -0.0265067 -0.05485056 -0.03872867 -0.016203768
## salary      0.3991686107  0.2499993  0.47301761  0.44380454  0.508959812
##                    FGA          FG%          3P         3PA         3P%
## Age         0.016998395 -0.009586507  0.11319532  0.10173854  0.07618839
## G           0.765872421  0.303729678  0.63724584  0.65584044  0.13343171
## GS          0.822480855  0.225838152  0.65235210  0.66018795  0.15122653
## MP          0.928863931  0.274619531  0.76892838  0.78779832  0.18161596
## FG          0.987968162  0.330788312  0.72855752  0.74492734  0.14582428
## FGA         1.000000000  0.245063475  0.79673618  0.81755041  0.17618564
## FG%         0.245063475  1.000000000 -0.00151524 -0.01800268  0.08951609
## 3P          0.796736185 -0.001515240  1.00000000  0.99140014  0.34737439
## 3PA         0.817550413 -0.018002683  0.99140014  1.00000000  0.31766627
## 3P%         0.176185643  0.089516090  0.34737439  0.31766627  1.00000000
## 2P          0.912878007  0.419281279  0.49847572  0.52283573  0.03741406
## 2PA         0.938393486  0.360999250  0.54360305  0.56818587  0.06113968
## 2P%         0.153306224  0.762901437  0.02602411  0.03301764 -0.08421263
## eFG%        0.257666167  0.872661299  0.22733035  0.19622051  0.39944725
## FT          0.887405547  0.275339233  0.61527949  0.64193402  0.08764911
## FTA         0.883536556  0.308329421  0.57047499  0.60138273  0.05498292
## FT%         0.233563592 -0.061717617  0.30817876  0.30018534  0.30175619
## ORB         0.536589836  0.501928457  0.13628123  0.15134646 -0.12224031
## DRB         0.800596415  0.401967688  0.49095061  0.51185031  0.05506158
## TRB         0.758784265  0.446922637  0.41063400  0.43074212  0.00704765
## AST         0.796084851  0.163075697  0.60298230  0.63259614  0.12451615
## STL         0.809712094  0.210021789  0.64559119  0.67308692  0.12678048
## BLK         0.540607933  0.395536699  0.23645114  0.25799982 -0.05187657
## TOV         0.911729112  0.262332484  0.64514454  0.67741375  0.09235809
## PF          0.791465389  0.356154560  0.58833754  0.60909884  0.09879845
## PTS        0.991228434  0.294912577  0.77343988  0.78957291  0.16571001
## player_id -0.007586202  0.041972235 -0.01892704 -0.01783505  0.04262209
## salary     0.505311664  0.135655205  0.37466576  0.38572227  0.07638686
##                    2P           2PA         2P%       eFG%          FT
## Age       -0.03637989 -0.0367620072 -0.04792758 0.07624964  0.01955848
## G          0.69101828  0.7008662104  0.22779814 0.34088622  0.60547227
## GS         0.76394881  0.7791508005  0.14082115 0.24168325  0.71221647
## MP         0.83618387  0.8545940435  0.19505687 0.31727432  0.77529616
## FG         0.95698392  0.9647806805  0.21549014 0.31045520  0.89716511
## FGA        0.91287801  0.9383934863  0.15330622 0.25766617  0.88740555
## FG%        0.41928128  0.3609992500  0.76290144 0.87266130  0.27533923
## 3P         0.49847572  0.5436030462  0.02602411 0.22733035  0.61527949
## 3PA        0.52283573  0.5681858678  0.03301764 0.19622051  0.64193402
## 3P%        0.03741406  0.0611396822 -0.08421263 0.39944725  0.08764911
## 2P         1.00000000  0.9907527216  0.26169715 0.29661527  0.87481953
## 2PA        0.99075272  1.0000000000  0.19926104 0.25045436  0.88288233
## 2P%        0.26169715  0.1992610392  1.00000000 0.67180261  0.17095059
## eFG%       0.29661527  0.2504543596  0.67180261 1.00000000  0.22124252
## FT         0.87481953  0.8828823264  0.17095059 0.22124252  1.00000000
## FTA        0.89924829  0.9016881439  0.19267542 0.23262610  0.98974303
## FT%        0.13362033  0.1536230404 -0.14758454 0.08335902  0.21884995
## ORB        0.72386136  0.6759652157  0.33392954 0.34243475  0.53501155
## DRB        0.85560650  0.8368945074  0.27844053 0.34117889  0.76148536
## TRB        0.85357988  0.8258174846  0.30567972 0.35562728  0.72870319
## AST        0.73009467  0.7579885764  0.09090060 0.14980320  0.74003684
## STL        0.73268183  0.7531636296  0.14978198 0.22230326  0.69741186
## BLK        0.66117927  0.6177049889  0.30812401 0.30651631  0.53192974
## TOV        0.88056216  0.8963495371  0.16043056 0.21966923  0.87954807
## PF         0.76918408  0.7654879626  0.26111250 0.35145152  0.68336503
## PTS        0.92988758  0.9426480567  0.19374292 0.29728277  0.92534382
## player_id -0.01249020 -0.0001379758 -0.02765477 0.03932673 -0.04534284
## salary     0.48543168  0.4906210569  0.03765851 0.12792197  0.52545505
##                    FTA         FT%         ORB         DRB         TRB
## Age        0.003986587  0.17628535 -0.03083117  0.03381834  0.01687861
## G          0.619750213  0.18378735  0.56909557  0.73460482  0.71800377
## GS         0.713197431  0.18870067  0.53234858  0.74326540  0.71414310
## MP         0.782522514  0.22264285  0.59120968  0.82839962  0.79534188
## FG         0.901472061  0.20872349  0.61757114  0.84037272  0.81189050
## FGA        0.883536556  0.23356359  0.53658984  0.80059641  0.75878427
## FG%        0.308329421 -0.06171762  0.50192846  0.40196769  0.44692264
## 3P         0.570474994  0.30817876  0.13628123  0.49095061  0.41063400
## 3PA        0.601382726  0.30018534  0.15134646  0.51185031  0.43074212
## 3P%        0.054982915  0.30175619 -0.12224031  0.05506158  0.00704765
## 2P         0.899248293  0.13362033  0.72386136  0.85560650  0.85357988
## 2PA        0.901688144  0.15362304  0.67596522  0.83689451  0.82581748
## 2P%        0.192675424 -0.14758454  0.33392954  0.27844053  0.30567972
## eFG%       0.232626104  0.08335902  0.34243475  0.34117889  0.35562728
## FT         0.989743034  0.21884995  0.53501155  0.76148536  0.72870319
## FTA        1.000000000  0.16122661  0.59648345  0.79562372  0.77200424
## FT%        0.161226607  1.00000000 -0.02557424  0.11353687  0.07877030
## ORB        0.596483447 -0.02557424  1.00000000  0.80382717  0.89266935
## DRB        0.795623723  0.11353687  0.80382717  1.00000000  0.98566381
## TRB        0.772004237  0.07877030  0.89266935  0.98566381  1.00000000
## AST        0.739871420  0.17396838  0.33662352  0.62053850  0.56564254
## STL        0.711327578  0.15718449  0.51599126  0.72422087  0.69507412
## BLK        0.576483410  0.01866424  0.73763207  0.75796914  0.78350839
## TOV        0.892682902  0.16201830  0.54346830  0.79350258  0.75536039
## PF         0.712484831  0.13377752  0.70800873  0.83896943  0.83647811
## PTS        0.920994204  0.23264245  0.57167313  0.82172490  0.78474346
## player_id -0.040692587 -0.08265517 -0.04602645 -0.02973384 -0.03558305
## salary     0.523546852  0.14898805  0.31309167  0.46852862  0.44379427
##                  AST         STL          BLK          TOV          PF
## Age       0.07640478  0.05081787 -0.002602496 0.0009743469  0.02073313
## G         0.61969931  0.73963602  0.550385548 0.6979516367  0.86912295
## GS        0.69733943  0.76743348  0.567869013 0.7699145811  0.74704226
## MP        0.76888682  0.86785135  0.591831931 0.8500994853  0.89939050
## FG        0.77869495  0.79499982  0.601568347 0.9116980853  0.80468008
## FGA       0.79608485  0.80971209  0.540607933 0.9117291119  0.79146539
## FG%       0.16307570  0.21002179  0.395536699 0.2623324841  0.35615456
## 3P        0.60298230  0.64559119  0.236451136 0.6451445389  0.58833754
## 3PA       0.63259614  0.67308692  0.257999825 0.6774137502  0.60909884
## 3P%       0.12451615  0.12678048 -0.051876572 0.0923580902  0.09879845
## 2P        0.73009467  0.73268183  0.661179272 0.8805621563  0.76918408
## 2PA       0.75798858  0.75316363  0.617704989 0.8963495371  0.76548796
## 2P%       0.09090060  0.14978198  0.308124009 0.1604305559  0.26111250
## eFG%      0.14980320  0.22230326  0.306516311 0.2196692250  0.35145152
## FT        0.74003684  0.69741186  0.531929736 0.8795480659  0.68336503
## FTA       0.73987142  0.71132758  0.576483410 0.8926829021  0.71248483
## FT%       0.17396838  0.15718449  0.018664245 0.1620182979  0.13377752
## ORB       0.33662352  0.51599126  0.737632074 0.5434683032  0.70800873
## DRB       0.62053850  0.72422087  0.757969139 0.7935025769  0.83896943
## TRB       0.56564254  0.69507412  0.783508386 0.7553603917  0.83647811
## AST       1.00000000  0.78888428  0.337664017 0.8935094745  0.63175366
## STL       0.78888428  1.00000000  0.513793150 0.7938385320  0.77888334
## BLK       0.33766402  0.51379315  1.000000000 0.5250976690  0.68149740
## TOV       0.89350947  0.79383853  0.525097669 1.0000000000  0.77978013
## PF        0.63175366  0.77888334  0.681497402 0.7797801265  1.00000000
## PTS       0.78625184  0.79487184  0.571616581 0.9157329910  0.79216628
## player_id 0.01675645 -0.01363147 -0.028744863 0.0193153848 -0.01813525
## salary    0.46848054  0.45041132  0.309187498 0.4940764436  0.34157979
##                   PTS     player_id      salary
## Age        0.02429723 -0.0696552960  0.39916861
## G          0.74969314 -0.0265067030  0.24999926
## GS         0.81824208 -0.0548505598  0.47301761
## MP         0.91481614 -0.0387286716  0.44380454
## FG         0.99361133 -0.0162037681  0.50895981
## FGA        0.99122843 -0.0075862023  0.50531166
## FG%        0.29491258  0.0419722349  0.13565520
## 3P         0.77343988 -0.0189270394  0.37466576
## 3PA        0.78957291 -0.0178350501  0.38572227
## 3P%        0.16571001  0.0426220935  0.07638686
## 2P         0.92988758 -0.0124902018  0.48543168
## 2PA        0.94264806 -0.0001379758  0.49062106
## 2P%        0.19374292 -0.0276547677  0.03765851
## eFG%       0.29728277  0.0393267322  0.12792197
## FT         0.92534382 -0.0453428386  0.52545505
## FTA        0.92099420 -0.0406925872  0.52354685
## FT%        0.23264245 -0.0826551652  0.14898805
## ORB        0.57167313 -0.0460264468  0.31309167
## DRB        0.82172490 -0.0297338354  0.46852862
## TRB        0.78474346 -0.0355830540  0.44379427
## AST        0.78625184  0.0167564475  0.46848054
## STL        0.79487184 -0.0136314730  0.45041132
## BLK        0.57161658 -0.0287448628  0.30918750
## TOV        0.91573299  0.0193153848  0.49407644
## PF         0.79216628 -0.0181352471  0.34157979
## PTS        1.00000000 -0.0230590296  0.51979003
## player_id -0.02305903  1.0000000000 -0.06351243
## salary     0.51979003 -0.0635124327  1.00000000
```

```
#melt() function from the reshape2 package is used to convert the matrix into a long-format data frame that can be used to create a heatmap with ggplot2

cor_df <- melt(cor_matrix)
colnames(cor_df) <- c("Var1", "Var2", "Corr")

# create heatmap by ggplot()and theme_minimal() function is used to apply a minimal theme to the plot
ggplot(data = cor_df, aes(x = Var1, y = Var2, fill = Corr)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "green", midpoint = 0) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Correlation Matrix Heatmap for players")
```



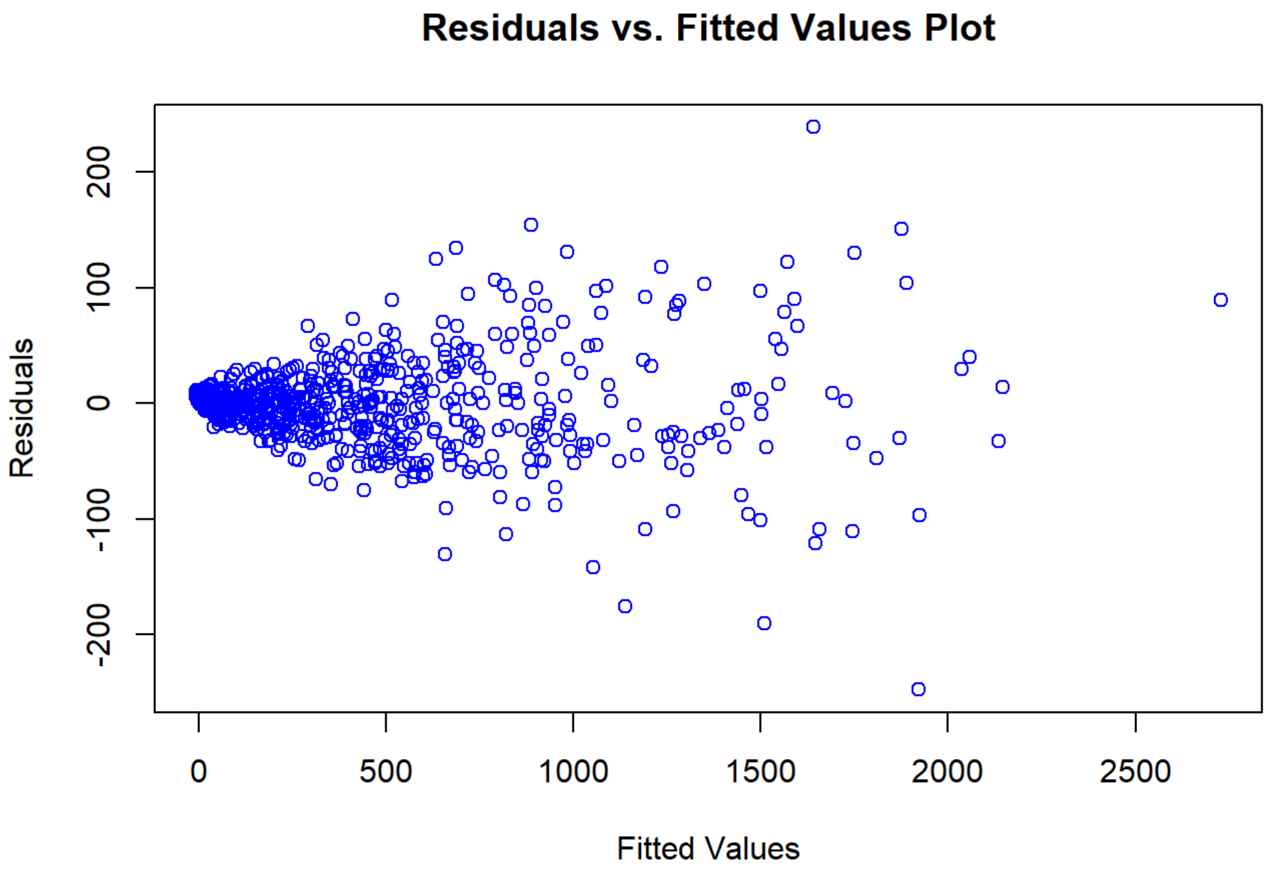
```
# Merging player and team statisitcs

# Merge player_stats and team_stats_2 datasets
team_stats <- full_join(team_stats1, team_stats2, by = "Team")

team_payroll <- subset(team_payroll, select = -team)
# team_payroll <- rename(team_payroll, Team = team_full_name)

team_stats_salary<-full_join( team_stats ,team_payroll, by = "Team")
player_stats_salaries <- rename(player_stats_salaries, Team = Tm)

# Merging the player and team stats as a master dataset

# master_player_team<- full_join(player_stats_salaries,team_stats_salary, by = "Team")
#
# colSums(is.na(master_player_team))
#
# master_player_team_omit<-na.omit(master_player_team)
#
# colSums(is.na(master_player_team_omit))
```

```
# 4.4. Data modelling and results


# Create linear regression model to predict PTS based on salary
lm_model <- lm(PTS ~ salary +  FGA + FTA, data = player_stats_salaries)

# Display model summary
summary(lm_model)
```

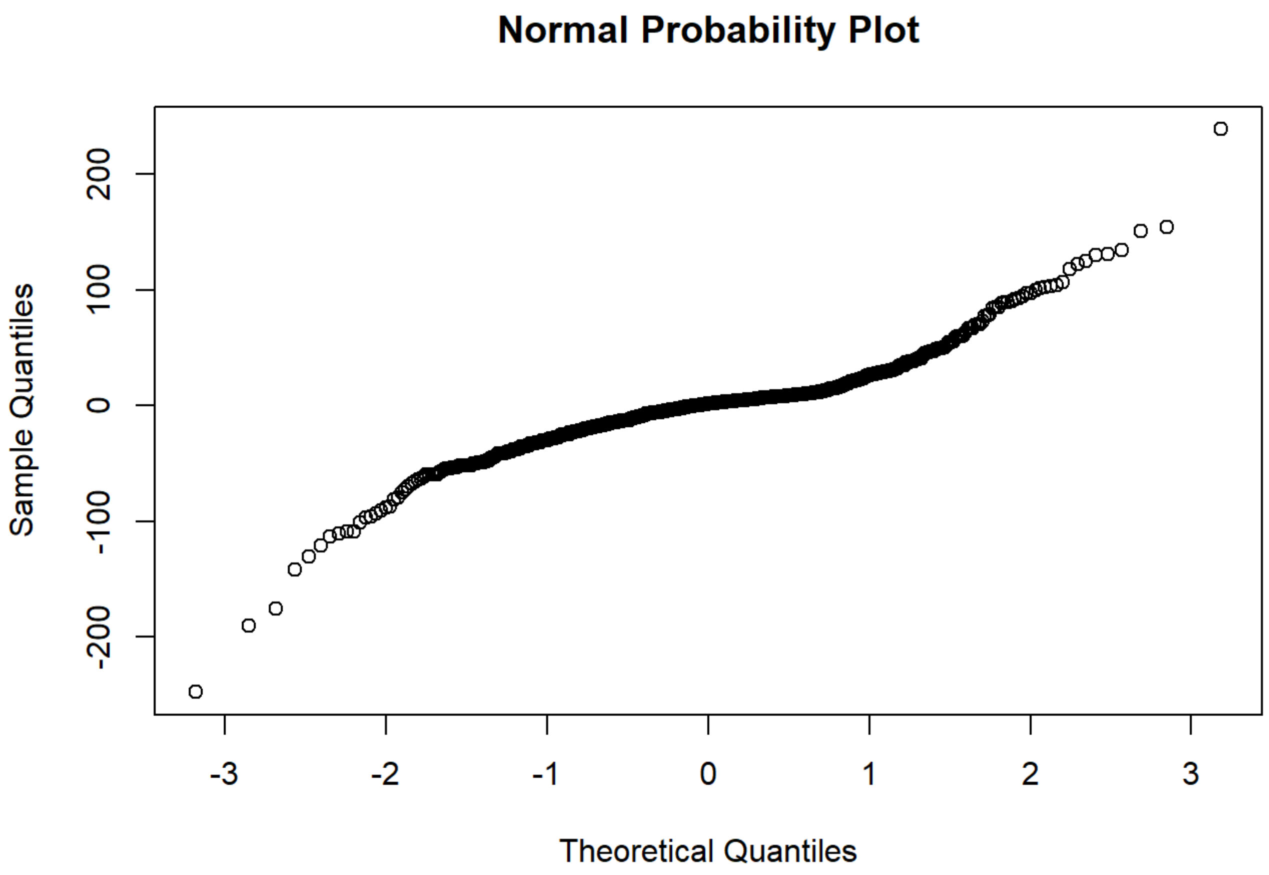
```
##
## Call:
## lm(formula = PTS ~ salary + FGA + FTA, data = player_stats_salaries)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -247.408  -16.939    1.845   11.649  239.043
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.121e+00  2.321e+00  -3.499 0.000497 ***
## salary       2.533e-07  2.382e-07   1.063 0.288113
## FGA          1.038e+00  9.361e-03 110.853  < 2e-16 ***
## FTA          8.713e-01  2.982e-02  29.215  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.32 on 682 degrees of freedom
## Multiple R-squared:  0.9923, Adjusted R-squared:  0.9922
## F-statistic: 2.915e+04 on 3 and 682 DF,  p-value: < 2.2e-16
```

```
# PTS = -8.121 + 2.533e-07(salary) + 1.038(FGA) + 0.8713(FTA)




#***************** Model Interpretation *************************************
# In the linear regression model with PTS as the dependent variable and salary, FGA, and FTA
as independent # variables, the coefficients for FGA and FTA are significant with p-values <
2e-16, indicating that these # variables are strongly associated with PTS. However, the coeff
icient for salary is not significant with a # p-value of 0.288, suggesting that salary is not
a good predictor of PTS. The adjusted R-squared value of # 0.9922 indicates that the model ex
plains a high proportion of the variance in PTS, and the F-statistic of # 2.915e+04 with a p-
value < 2.2e-16 indicates that the overall model is significant. The residual standard # erro
r of 40.32 suggests that the model has a moderate level of error in predicting PTS, and the n
ormal Q-Q # plot and residual vs. fitted plot do not show any major departures from normality
or homoscedasticity # assumptions
```

```
# Assumption checking

# 1. homoscedasticity

# Plot residuals vs. fitted values
plot(lm_model$fitted.values, lm_model$residuals, type = "p", col = "blue",
     xlab = "Fitted Values", ylab = "Residuals",
     main = "Residuals vs. Fitted Values Plot")
```



```
# 2. Normality plot

#Normal probability plot
qqnorm(lm_model$residuals, main = "Normal Probability Plot")
```



```
# ***** 5. Player recommendation in each position ***************************

# Top 5 player with regards to position based on Cost effectiveness

# Calculate cost-effectiveness score
player_stats_salaries$cost_effectiveness <- player_stats_salaries$PTS / player_stats_salaries
$salary

# Select top player for each position
point_guard <- player_stats_salaries %>%
  filter(Pos == "PG") %>%
  slice_max(cost_effectiveness)

shooting_guard <- player_stats_salaries %>%
  filter(Pos == "SG") %>%
  slice_max(cost_effectiveness)

small_forward <- player_stats_salaries %>%
  filter(Pos == "SF") %>%
  slice_max(cost_effectiveness)

power_forward <- player_stats_salaries %>%
  filter(Pos == "PF") %>%
  slice_max(cost_effectiveness)

center <- player_stats_salaries %>%
  filter(Pos == "C") %>%
  slice_max(cost_effectiveness)

# Combine selected players into final output
top_five <- bind_rows(
  point_guard %>% select(player_name, Team, Pos, Age, G, PTS, salary, cost_effectiveness),
  shooting_guard %>% select(player_name,Team, Pos, Age, G, PTS, salary, cost_effectiveness),
  small_forward %>% select(player_name, Team, Pos, Age, G, PTS, salary, cost_effectiveness),
  power_forward %>% select(player_name, Team, Pos, Age, G, PTS, salary, cost_effectiveness),
  center %>% select(player_name, Team, Pos, Age, G, PTS, salary, cost_effectiveness)
)

top_five_budget <- sum(top_five$salary)

top_five_budget
```

```
## [1] 606827
```

```
# Output starting five
top_five
```

```
##          player_name               Team Pos Age  G PTS salary
## 1        Alex Caruso Los Angeles Lakers  PG  24 25 229  77250
## 2       Kadeem Allen    New York Knicks  SG  26 19 189  77250
## 3       Danuel House    Houston Rockets  SF  25 39 366 247827
## 4     Alex Poythress      Atlanta Hawks  PF  25 21 107  77250
## 5 Johnathan Williams Los Angeles Lakers   C  23 24 157 127250
##   cost_effectiveness
## 1        0.002964401
## 2        0.002446602
## 3        0.001476837
## 4        0.001385113
## 5        0.001233792
```

```
# Top 5 player with regards to position based on Cost effectiveness from Chicago Bulls team

chicago_players <- player_stats_salaries %>% filter(Team == "Chicago Bulls")



# Select top player for each position
point_guard <- chicago_players %>%
  filter(Pos == "PG") %>%
  slice_max(cost_effectiveness)

shooting_guard <- chicago_players %>%
  filter(Pos == "SG") %>%
  slice_max(cost_effectiveness)

small_forward <- chicago_players %>%
  filter(Pos == "SF") %>%
  slice_max(cost_effectiveness)

power_forward <- chicago_players %>%
  filter(Pos == "PF") %>%
  slice_max(cost_effectiveness)

center <- chicago_players %>%
  filter(Pos == "C") %>%
  slice_max(cost_effectiveness)

# Combine selected players into final output
chicago_top_five <- bind_rows(
  point_guard %>% select(player_name, Team, Pos, Age, G, PTS, salary, cost_effectiveness),
  shooting_guard %>% select(player_name,Team, Pos, Age, G, PTS, salary, cost_effectiveness),
  small_forward %>% select(player_name, Team, Pos, Age, G, PTS, salary, cost_effectiveness),
  power_forward %>% select(player_name, Team, Pos, Age, G, PTS, salary, cost_effectiveness),
  center %>% select(player_name, Team, Pos, Age, G, PTS, salary, cost_effectiveness)
)

# Output starting five
chicago_top_five
```

```
##        player_name          Team Pos Age  G PTS  salary cost_effectiveness
## 1 Ryan Arcidiacono Chicago Bulls  PG  24 81 544 1349383       0.0004031472
## 2  Brandon Sampson Chicago Bulls  SG  21 14  71   77250       0.0009190939
## 3   JaKarr Sampson Chicago Bulls  SF  25  4  80   85457       0.0009361433
## 4  Lauri Markkanen Chicago Bulls  PF  21 52 974 4536120       0.0002147210
## 5   Wendell Carter Chicago Bulls   C  19 44 455 4446840       0.0001023198
```

```
chicago_top_five_budget <- sum(chicago_top_five$salary)

chicago_top_five_budget
```

```
## [1] 10495050
```

```
# Top 5 players based on Points alone, neglecting salary

# Select top player for each position
point_guard_top <- player_stats_salaries %>%
  filter(Pos == "PG") %>%
  slice_max(PTS)

shooting_guard_top <- player_stats_salaries %>%
  filter(Pos == "SG") %>%
  slice_max(PTS)

small_forward_top <- player_stats_salaries %>%
  filter(Pos == "SF") %>%
  slice_max(PTS)

power_forward_top <- player_stats_salaries %>%
  filter(Pos == "PF") %>%
  slice_max(PTS)

center_top <- player_stats_salaries %>%
  filter(Pos == "C") %>%
  slice_max(PTS)

# Combine selected players into final output
top_five_cost <- bind_rows(
  point_guard_top %>% select(player_name, Team, Pos, Age, G, PTS, salary),
  shooting_guard_top %>% select(player_name,Team, Pos, Age, G, PTS, salary),
  small_forward_top %>% select(player_name, Team, Pos, Age, G, PTS, salary),
  power_forward_top %>% select(player_name, Team, Pos, Age, G, PTS, salary),
  center_top %>% select(player_name, Team, Pos, Age, G, PTS, salary)
)

top_five_cost_budget <- sum(top_five_cost$salary)

top_five_cost_budget
```

```
## [1] 118561701
```

```
# Output starting five not based on cost effectiveness
top_five_cost
```

```
##             player_name                    Team Pos Age  G  PTS   salary
## 1          James Harden        Houston Rockets  PG  29 78 2818 30570000
## 2          Bradley Beal     Washington Wizards  SG  25 82 2099 25434262
## 3           Paul George  Oklahoma City Thunder  SF  28 77 2159 30560700
## 4 Giannis Antetokounmpo        Milwaukee Bucks  PF  24 72 1994 24157304
## 5    Karl-Anthony Towns Minnesota Timberwolves   C  23 77 1880  7839435
```

```
# Top 5 irrespective of position

player_stats_salaries[order(-player_stats_salaries$salary),][1:5,] %>% select(player_name, Te
am, Pos, Age, G, PTS, salary)
```

```
##           player_name                  Team Pos Age  G  PTS   salary
## 160     Stephen Curry Golden State Warriors  PG  30 69 1881 37457154
## 673 Russell Westbrook Oklahoma City Thunder  PG  30 73 1675 35665000
## 333      LeBron James    Los Angeles Lakers  SF  34 55 1505 35654150
## 532        Chris Paul       Houston Rockets  PG  33 58  906 35654150
## 412        Kyle Lowry       Toronto Raptors  PG  32 65  926 32700000
```