Correlation Analysis

We will begin our analysis of the 2024 NBA player statistics by exploring the correlation among the variables we have within this data set.

We begin by loading in our cleaned dataset that we have saved within our repository.

We first take a quick look at the structure of our cleaned dataset.

```
dim(clean_data)
```

```
## [1] 457  32
```

```
names(clean_data)
```

```
##  [1] "Rk"                "Player"            "Age"
##  [4] "Team"              "Pos"               "G"
##  [7] "GS"                "MP"                "FG"
## [10] "FGA"               "FG_pct"            "X3P"
## [13] "X3PA"              "X3P_pct"           "X2P"
## [16] "X2PA"              "X2P_pct"           "eFG."
## [19] "FT"                "FTA"               "FT_pct"
## [22] "ORB"               "DRB"               "TRB"
## [25] "AST"               "STL"               "BLK"
## [28] "TOV"               "PF"                "PTS"
## [31] "Awards"            "Player.additional"
```

```
str(clean_data)
```

```
## 'data.frame':    457 obs. of  32 variables:
##  $ Rk               : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Player           : chr  "Shai Gilgeous-Alexander" "Giannis Antetokounmpo" "Nikola Jokić" "Luka Do
##  $ Age              : int  26 30 29 25 23 26 36 24 23 28 ...
##  $ Team             : chr  "OKC" "MIL" "DEN" "2TM" ...
##  $ Pos              : chr  "PG" "PF" "C" "PG" ...
##  $ G                : int  76 67 70 50 79 72 62 52 70 65 ...
##  $ GS               : int  76 67 70 50 79 72 62 52 70 65 ...
##  $ MP               : num  34.2 34.2 36.7 35.4 36.3 36.4 36.5 37.7 35 35.4 ...
##  $ FG               : num  11.3 11.8 11.2 9.2 9.1 9.2 9.5 9.2 9.8 9 ...
##  $ FGA              : num  21.8 19.7 19.5 20.5 20.4 20.3 18.1 21 20.8 18.5 ...
##  $ FG_pct           : num  0.519 0.601 0.576 0.45 0.447 0.452 0.527 0.437 0.469 0.488 ...
##  $ X3P              : num  2.1 0.2 2 3.5 4.1 3.5 2.6 3.1 2.1 2.3 ...
##  $ X3PA             : num  5.7 0.9 4.7 9.6 10.3 10.1 6 9.2 6 6.1 ...
##  $ X3P_pct          : num  0.375 0.222 0.417 0.368 0.395 0.343 0.43 0.337 0.356 0.383 ...
##  $ X2P              : num  9.2 11.6 9.3 5.7 5.1 5.7 7 6.1 7.6 6.7 ...
##  $ X2PA             : num  16.1 18.7 14.8 10.9 10.1 10.2 12.1 11.8 14.8 12.4 ...
##  $ X2P_pct          : num  0.571 0.62 0.627 0.522 0.501 0.559 0.574 0.515 0.515 0.539 ...
##  $ eFG.              : num  0.569 0.607 0.627 0.536 0.547 0.537 0.598 0.511 0.521 0.551 ...
##  $ FT                : num  7.9 6.5 5.2 6.2 5.3 5 4.9 4.9 4.5 5.7 ...
##  $ FTA               : num  8.8 10.6 6.4 7.9 6.3 6.1 5.8 5.6 5.3 6.9 ...
##  $ FT_pct            : num  0.898 0.617 0.8 0.782 0.837 0.814 0.839 0.879 0.846 0.821 ...
##  $ ORB               : num  0.9 2.2 2.9 0.8 0.8 0.7 0.4 0.3 0.8 0.4 ...
##  $ DRB               : num  4.1 9.7 9.9 7.4 4.9 8 5.7 3.1 5.3 2.5 ...
##  $ TRB               : num  5 11.9 12.7 8.2 5.7 8.7 6 3.3 6.1 2.9 ...
##  $ AST               : num  6.4 6.5 10.2 7.7 4.5 6 4.2 6.1 9.1 7.3 ...
##  $ STL               : num  1.7 0.9 1.8 1.8 1.2 1.1 0.8 1.8 1 0.9 ...
##  $ BLK               : num  1 1.2 0.6 0.4 0.6 0.5 1.2 0.4 0.8 0.1 ...
##  $ TOV               : num  2.4 3.1 3.3 3.6 3.2 2.9 3.1 2.4 4.4 2.5 ...
##  $ PF                : num  2.2 2.3 2.3 2.5 1.9 2.2 1.7 2.2 2.8 2.1 ...
##  $ PTS               : num  32.7 30.4 29.6 28.2 27.6 26.8 26.6 26.3 26.1 26 ...
##  $ Awards            : chr  "MVP-1DPOY-10CPOY-8ASNBA1" "MVP-3DPOY-8ASNBA1" "MVP-2CPOY-2ASNBA1" "" ...
##  $ Player.additional: chr  "gilgesh01" "antetgi01" "jokicni01" "doncilu01" ...
```

We will now select the variables of interest we would like to explore. To reduce visual clutter and multicollinearity, variables that were highly redundant or not performance-based were removed from the correlation analysis. Offensive and defensive rebounding metrics were replaced by total rebounds, individual shot types were summarized by total field goal attempts, and only two efficiency metrics were retained (FG% and eFG%)

```r
# Variables we want to examine (edit names as needed)
vars <- c(
  # Performance
  "PTS", "TRB", "AST", "STL", "BLK", "TOV", "PF",

  # Volume
  "MP", "FGA", "FTA",

  # Efficiency
  "FG_pct", "X3P_pct", "X2P_pct", "FT_pct", "eFG.",

  # Age(While not a performance statistics, would still like to explore if there is any correlation bet
  "Age"
)
```

We now will subset these variables of interest and ensure that we only keep numeric columns so that it is ready to compute a correlation matrix.

```r
stats <- clean_data[ , vars, drop = FALSE]

is_num <- sapply(stats, is.numeric)
stats_num <- stats[ , is_num, drop = FALSE]
str(stats_num)
```

```
## 'data.frame':    457 obs. of  16 variables:
##  $ PTS    : num  32.7 30.4 29.6 28.2 27.6 26.8 26.6 26.3 26.1 26 ...
##  $ TRB    : num  5 11.9 12.7 8.2 5.7 8.7 6 3.3 6.1 2.9 ...
##  $ AST    : num  6.4 6.5 10.2 7.7 4.5 6 4.2 6.1 9.1 7.3 ...
##  $ STL    : num  1.7 0.9 1.8 1.8 1.2 1.1 0.8 1.8 1 0.9 ...
##  $ BLK    : num  1 1.2 0.6 0.4 0.6 0.5 1.2 0.4 0.8 0.1 ...
##  $ TOV    : num  2.4 3.1 3.3 3.6 3.2 2.9 3.1 2.4 4.4 2.5 ...
##  $ PF     : num  2.2 2.3 2.3 2.5 1.9 2.2 1.7 2.2 2.8 2.1 ...
##  $ MP     : num  34.2 34.2 36.7 35.4 36.3 36.4 36.5 37.7 35 35.4 ...
##  $ FGA    : num  21.8 19.7 19.5 20.5 20.4 20.3 18.1 21 20.8 18.5 ...
##  $ FTA    : num  8.8 10.6 6.4 7.9 6.3 6.1 5.8 5.6 5.3 6.9 ...
##  $ FG_pct : num  0.519 0.601 0.576 0.45 0.447 0.452 0.527 0.437 0.469 0.488 ...
##  $ X3P_pct: num  0.375 0.222 0.417 0.368 0.395 0.343 0.43 0.337 0.356 0.383 ...
##  $ X2P_pct: num  0.571 0.62 0.627 0.522 0.501 0.559 0.574 0.515 0.515 0.539 ...
##  $ FT_pct : num  0.898 0.617 0.8 0.782 0.837 0.814 0.839 0.879 0.846 0.821 ...
##  $ eFG.   : num  0.569 0.607 0.627 0.536 0.547 0.537 0.598 0.511 0.521 0.551 ...
##  $ Age    : int  26 30 29 25 23 26 36 24 23 28 ...
```

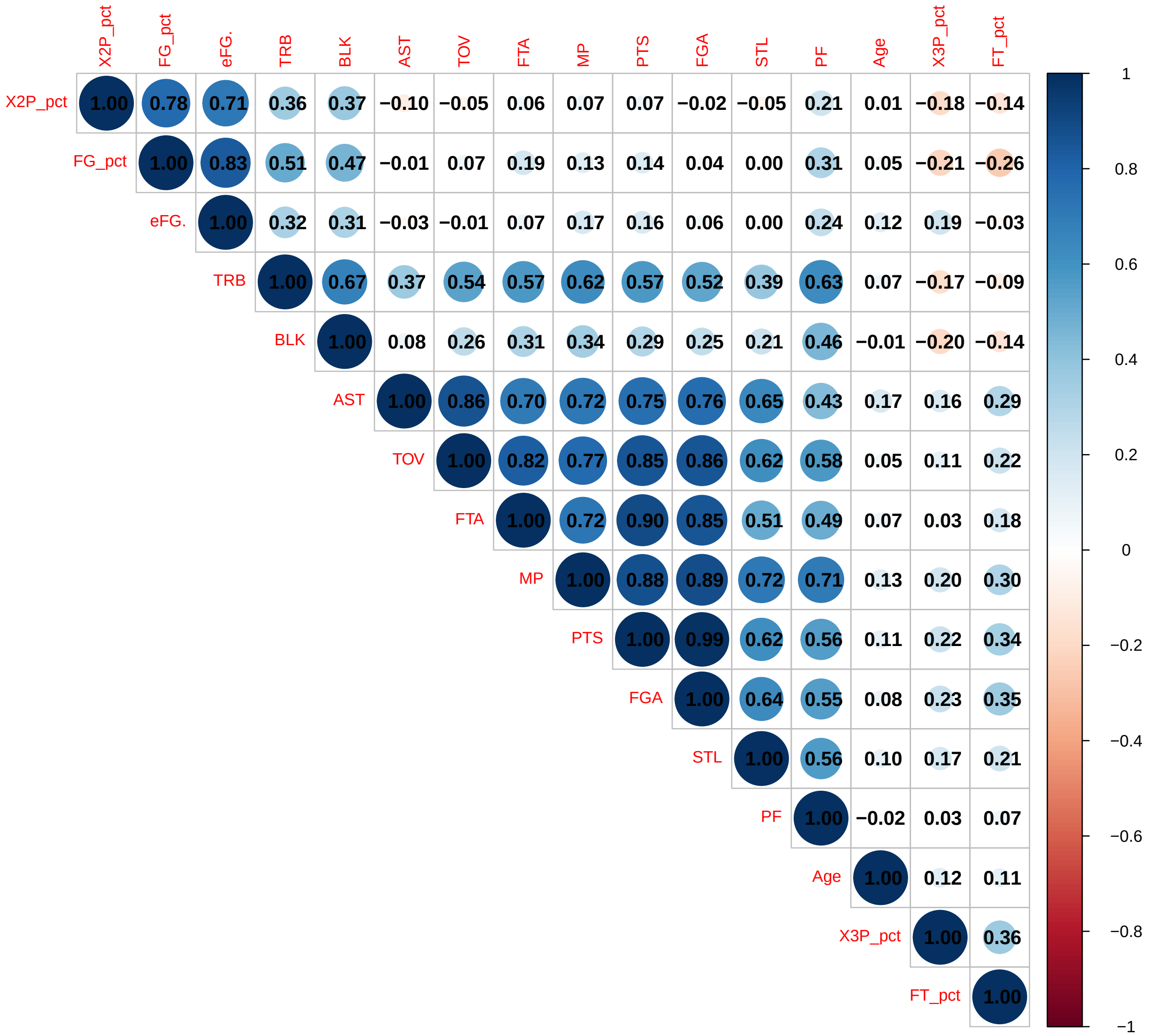
We then compute the correlation matrix for these selected variables.

```
cor_mat <- cor(stats_num, use = "pairwise.complete.obs")
round(cor_mat, 2)
```

```
##          PTS   TRB   AST   STL   BLK   TOV    PF   MP   FGA  FTA FG_pct X3P_pct
## PTS     1.00  0.57  0.75  0.62  0.29  0.85  0.56 0.88  0.99 0.90   0.14    0.22
## TRB     0.57  1.00  0.37  0.39  0.67  0.54  0.63 0.62  0.52 0.57   0.51   -0.17
## AST     0.75  0.37  1.00  0.65  0.08  0.86  0.43 0.72  0.76 0.70  -0.01    0.16
## STL     0.62  0.39  0.65  1.00  0.21  0.62  0.56 0.72  0.64 0.51   0.00    0.17
## BLK     0.29  0.67  0.08  0.21  1.00  0.26  0.46 0.34  0.25 0.31   0.47   -0.20
## TOV     0.85  0.54  0.86  0.62  0.26  1.00  0.58 0.77  0.86 0.82   0.07    0.11
## PF      0.56  0.63  0.43  0.56  0.46  0.58  1.00 0.71  0.55 0.49   0.31    0.03
## MP      0.88  0.62  0.72  0.72  0.34  0.77  0.71 1.00  0.89 0.72   0.13    0.20
## FGA     0.99  0.52  0.76  0.64  0.25  0.86  0.55 0.89  1.00 0.85   0.04    0.23
## FTA     0.90  0.57  0.70  0.51  0.31  0.82  0.49 0.72  0.85 1.00   0.19    0.03
## FG_pct  0.14  0.51 -0.01  0.00  0.47  0.07  0.31 0.13  0.04 0.19   1.00   -0.21
## X3P_pct 0.22 -0.17  0.16  0.17 -0.20  0.11  0.03 0.20  0.23 0.03  -0.21    1.00
## X2P_pct 0.07  0.36 -0.10 -0.05  0.37 -0.05  0.21 0.07 -0.02 0.06   0.78   -0.18
## FT_pct  0.34 -0.09  0.29  0.21 -0.14  0.22  0.07 0.30  0.35 0.18  -0.26    0.36
## eFG.    0.16  0.32 -0.03  0.00  0.31 -0.01  0.24 0.17  0.06 0.07   0.83    0.19
## Age     0.11  0.07  0.17  0.10 -0.01  0.05 -0.02 0.13  0.08 0.07   0.05    0.12
##         X2P_pct FT_pct  eFG.   Age
## PTS        0.07   0.34  0.16  0.11
## TRB        0.36  -0.09  0.32  0.07
## AST       -0.10   0.29 -0.03  0.17
## STL       -0.05   0.21  0.00  0.10
## BLK        0.37  -0.14  0.31 -0.01
## TOV       -0.05   0.22 -0.01  0.05
## PF         0.21   0.07  0.24 -0.02
## MP         0.07   0.30  0.17  0.13
## FGA       -0.02   0.35  0.06  0.08
## FTA        0.06   0.18  0.07  0.07
## FG_pct     0.78  -0.26  0.83  0.05
## X3P_pct   -0.18   0.36  0.19  0.12
## X2P_pct    1.00  -0.14  0.71  0.01
## FT_pct    -0.14   1.00 -0.03  0.11
## eFG.       0.71  -0.03  1.00  0.12
## Age        0.01   0.11  0.12  1.00
```

Using corrplot we will create a correlation heatmap for the selected variables of interest.

```
corrplot(
  cor_mat,
  method = "circle",       # color tiles
  type   = "upper",       # show upper triangle only
  order  = "hclust",      # cluster similar variables
  addCoef.col = "black", # add correlation coefficients
  tl.cex = 0.8            # label size
)
```



The Figure above shows a correlation heatmap for the selected performance variables.
As expected, points per game (PTS) exhibits a strong positive correlation with field-goal attempts (FGA) and free-throw attempts (FTA), reflecting that higher-volume shooters score more.
Total rebounds (TRB) are positively associated with blocks (BLK), consistent with the idea that rim-protecting players also secure more rebounds.

Assists (AST) show moderate positive relationships with both minutes played (MP) and points (PTS), indicating that playmaking tends to be concentrated among high-usage players.
In contrast, turnovers (TOV) are positively correlated with both PTS and AST, suggesting that players who handle the ball more often both create more offense and commit more turnovers. An interesting correlation to note is the lack of correlation between age and other variables. Indicating that increasing age in NBA players doesn't have any negative relationships with their performances which explains how players such as Lebron James are still so dominant despite being 40 years old.

```r
pair_vars <- c(
  "PTS",      # Scoring
  "AST",      # Playmaking
  "TRB",      # Rebounding
  "STL",      # On-ball defense
  "BLK",      # Rim defense
  "FG_pct",   # Shooting efficiency
  "eFG.",     # Adjusted efficiency
  "TOV"       # Turn overs
)

pair_vars <- intersect(pair_vars, colnames(stats_num))  # safety check

suppressWarnings(
  ggpairs(
    stats_num[ , pair_vars, drop = FALSE],
    progress = FALSE
  )
)
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (‘stat_density()‘).

## Warning: Removing 1 row that contained a missing value
## Removing 1 row that contained a missing value
## Removing 1 row that contained a missing value
## Removing 1 row that contained a missing value
## Removing 1 row that contained a missing value
## Removing 1 row that contained a missing value
## Removing 1 row that contained a missing value

## Warning: Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).

## Warning: Removed 1 row containing non-finite outside the scale range
## (‘stat_density()‘).

## Warning: Removing 1 row that contained a missing value
## Removing 1 row that contained a missing value
## Removing 1 row that contained a missing value
## Removing 1 row that contained a missing value
## Removing 1 row that contained a missing value
## Removing 1 row that contained a missing value
## Warning: Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
## Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (‘stat_density()‘).
```

```
## Warning: Removing 1 row that contained a missing value
## Removing 1 row that contained a missing value
## Removing 1 row that contained a missing value
## Removing 1 row that contained a missing value
## Removing 1 row that contained a missing value
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
## Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
## Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (‘stat_density()‘).
```

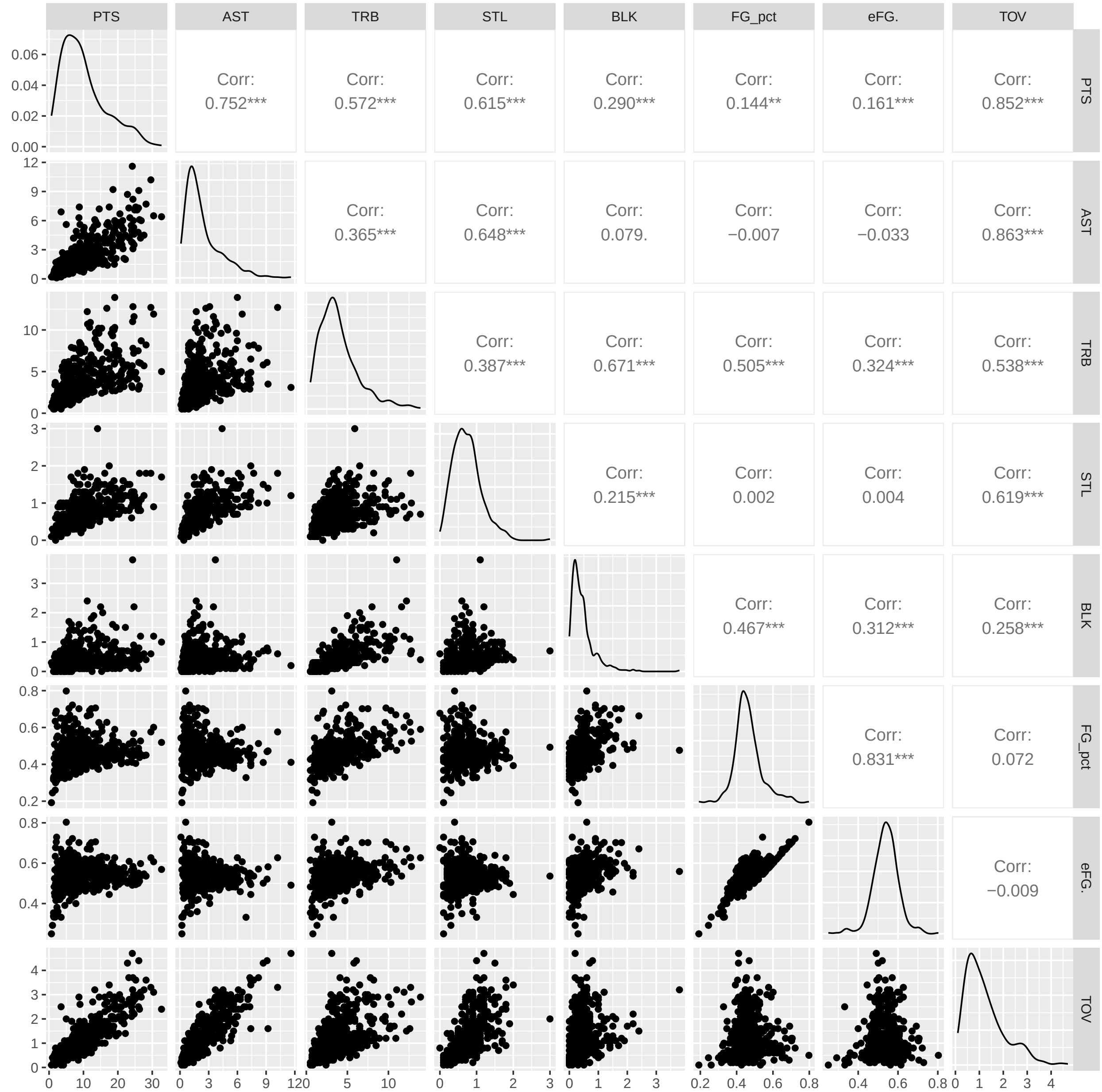
```
## Warning: Removing 1 row that contained a missing value
## Removing 1 row that contained a missing value
## Removing 1 row that contained a missing value
## Removing 1 row that contained a missing value
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
## Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
## Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
## Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (‘stat_density()‘).
```

```
## Warning: Removing 1 row that contained a missing value
## Removing 1 row that contained a missing value
## Removing 1 row that contained a missing value
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
## Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
## Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
## Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
## Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).

## Warning: Removed 1 row containing non-finite outside the scale range
## (‘stat_density()‘).

## Warning: Removing 1 row that contained a missing value
## Removing 1 row that contained a missing value

## Warning: Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
## Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
## Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
## Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
## Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
## Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).

## Warning: Removed 1 row containing non-finite outside the scale range
## (‘stat_density()‘).

## Warning: Removing 1 row that contained a missing value

## Warning: Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
## Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
## Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
## Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
## Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
## Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
## Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).

## Warning: Removed 1 row containing non-finite outside the scale range
## (‘stat_density()‘).
```



To further investigate the relationships among key performance variables, we produced pairwise scatterplots with marginal distributions using the ggpairs function from the GGally package.

The pair plots reveal clear positive associations between points and both field-goal efficiency measures, indicating that higher scorers tend to be more efficient rather than simply taking more shots.
Total rebounds and blocks also show a noticeable positive relationship, reflecting the shared role of interior players in both rim protection and rebounding.