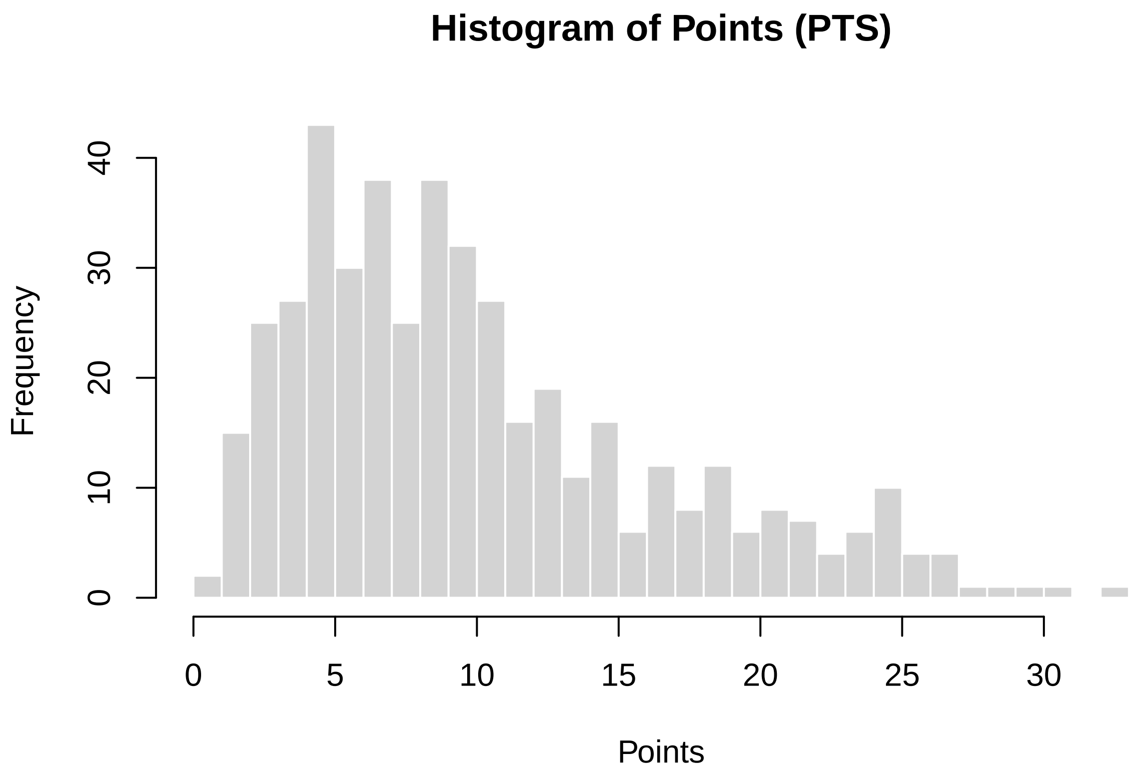
Regression Modeling

2025-12-05

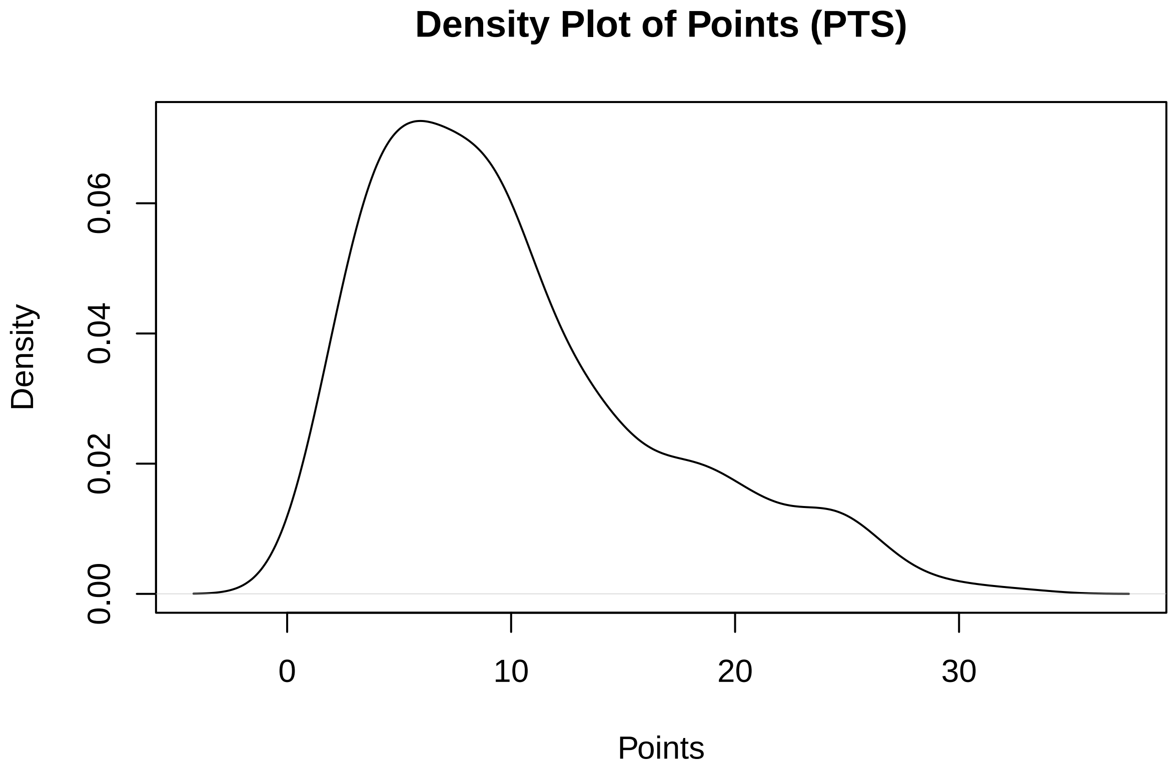
We will continue our analysis of the 2024 NBA player statistics by looking to implement both Linear and logistic Regression models to predict player's points per game (PTS) as well as the overall ranking of the players respectively based on other statistics. We will use our insights from our correlation analysis in helping choose appropriate parameters within these models.

Before fitting a linear regression model to predict PTS, it's useful to look at how the response variable is distributed. This helps confirm that the values make sense, there are no extreme outliers, and that PTS behaves in a way that works well for linear modeling.

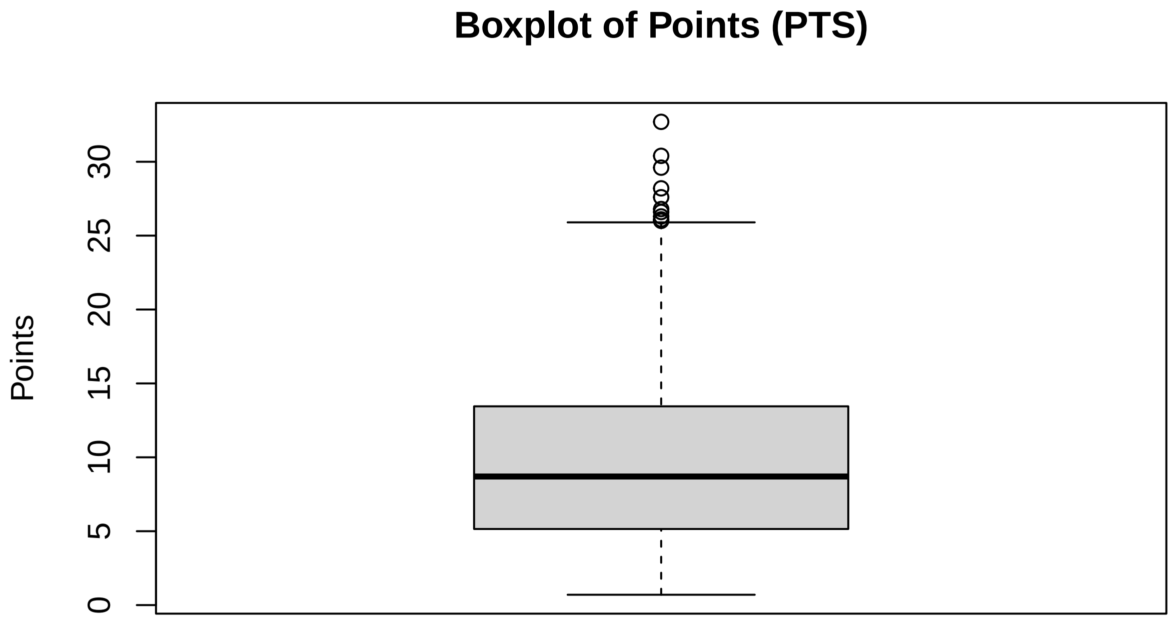
```r
hist(
  clean_data$PTS,
  breaks = 30,
  main = "Histogram of Points (PTS)",
  xlab = "Points",
  col = "lightgray",
  border = "white"
)
```



```r
plot(
  density(clean_data$PTS, na.rm = TRUE),
  main = "Density Plot of Points (PTS)",
  xlab = "Points"
)
```



```r
boxplot(
  clean_data$PTS,
  main = "Boxplot of Points (PTS)",
  ylab = "Points"
)
```



The histogram shows a right-skewed distribution, which is completely expected for NBA scoring data. Most players score modest amounts, while a smaller group scores very high this is natural and not a problem for linear regression. There is no severe skew, no long tail, and no shape that suggests the need for a transformation.

The density curve reinforces this pattern. It is unimodal, smooth, and shows only mild right-skewness. Nothing in the distribution suggests instability or irregularity.

The boxplot shows a handful of high-scoring players that appear as “upper outliers,” but these represent real star players, not data errors. This is normal for sports stats and does not violate linear regression assumptions.

The distribution of PTS shows mild right-skewness and a small number of high-scoring players, which is expected in NBA data. These patterns do not violate linear regression assumptions, and the data appears appropriate for modeling without transformation. As such we can continue with implementation of our model.

We will begin with choosing variables we want to include in the model.

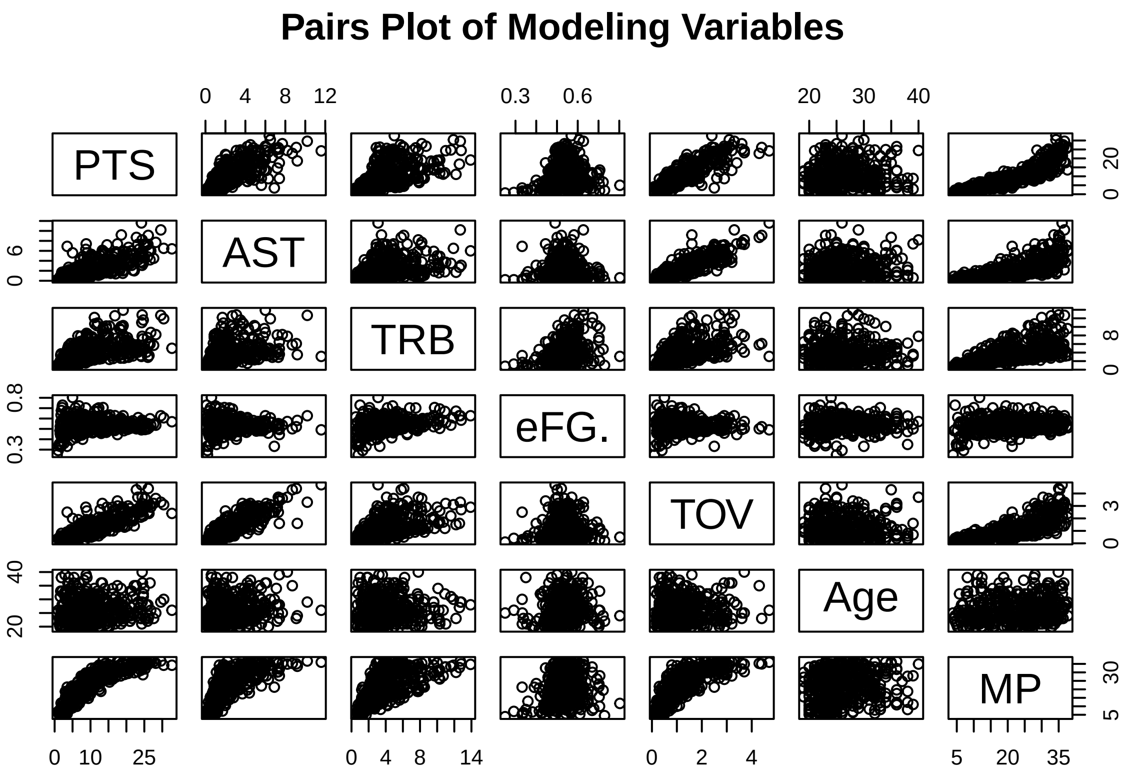
```r
model_vars <- c("PTS", "AST", "TRB", "eFG.", "TOV", "Age", "MP")
stats_model <- clean_data[, model_vars, drop = FALSE]
```

Remove rows with missing values only for modelling

```r
stats_model <- stats_model |> drop_na()
```

Exploratory Check of Predictors:

```r
pairs(stats_model, main = "Pairs Plot of Modeling Variables")
```



Before fitting the regression model, we take a quick look at the relationships between PTS and potential predictors. This helps confirm that the variables chosen are relevant and not too strongly correlated with each other. For example we would not want to include "FGA" as one of our variables as shown in our correlation analysis pts~fga had a correlation value of 0.99 making them too strongly correlated for the sake of our analysis as we would like to explore the predictive power of other variables rather then just creating the most accurate model. W

As we already know if we include FGA in our model it will make other variables insignificant we will look to create two seprate models. One without FGA to look explore the contribution of other statistics in predicting number of points. And another with FGA to create the most accurate model.

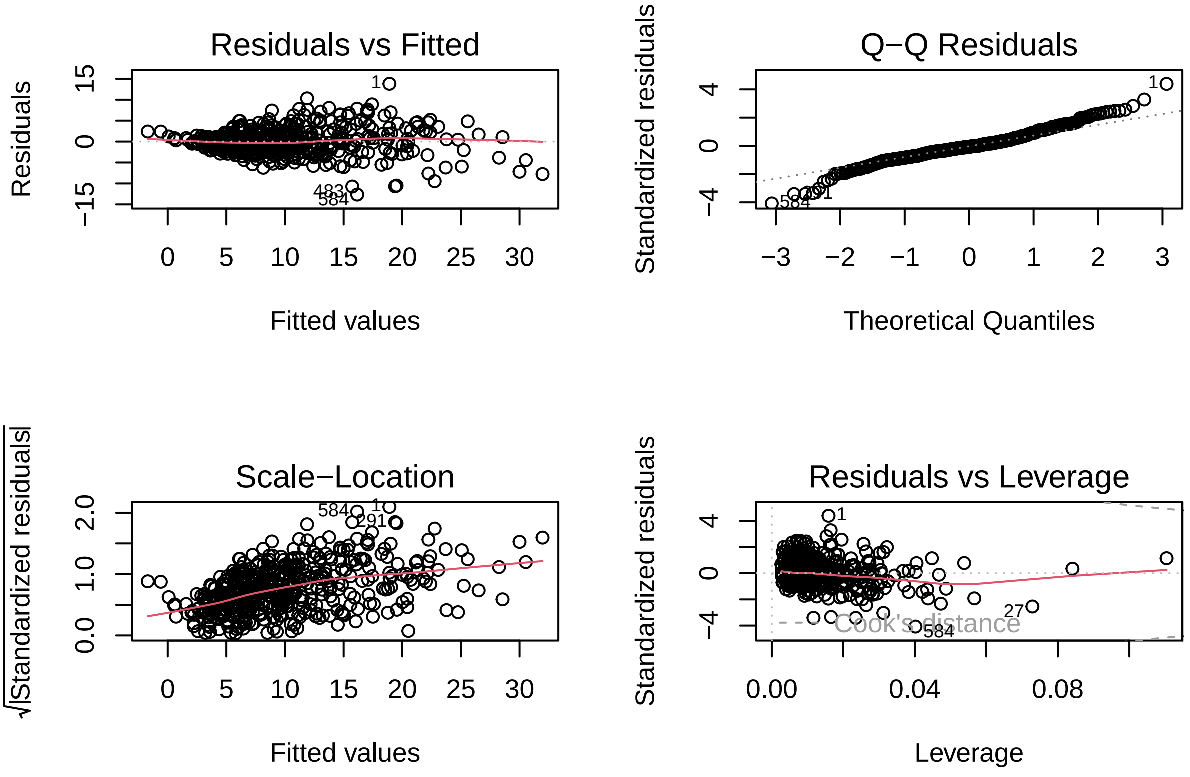
We will begin with the model excluding FGA.

```r
linear_model_1 <- lm(PTS ~ AST + TRB + eFG. + TOV + Age, data = stats_model)
summary(linear_model_1)
```

```
##
## Call:
## lm(formula = PTS ~ AST + TRB + eFG. + TOV + Age, data = stats_model)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.6543  -1.7210  -0.2434   1.5287  13.7962
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.98182    1.53232  -4.556 6.71e-06 ***
## AST          0.37809    0.16846   2.244 0.025292 *
## TRB          0.30345    0.08250   3.678 0.000263 ***
## eFG.        13.28761    2.52456   5.263 2.19e-07 ***
## TOV          5.52321    0.42108  13.117  < 2e-16 ***
## Age          0.04355    0.03593   1.212 0.226106
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.169 on 450 degrees of freedom
## Multiple R-squared:  0.7643, Adjusted R-squared:  0.7616
## F-statistic: 291.8 on 5 and 450 DF,  p-value: < 2.2e-16
```

```
#The diagnostic plots help us check whether the assumptions of linear regression are met, including lin
par(mfrow=c(2,2))
plot(linear_model_1)
```



```
par(mfrow=c(1,1))
```

This model predicts points per game using a combination of shooting volume (FGA), playmaking (AST), rebounding (TRB), shooting efficiency (eFG.), turnovers (TOV), and age. These variables were chosen to avoid multicollinearity and to reflect different aspects of player performance.