zhao.3738_Project2_4194

Megan Zhao

2025-02-16

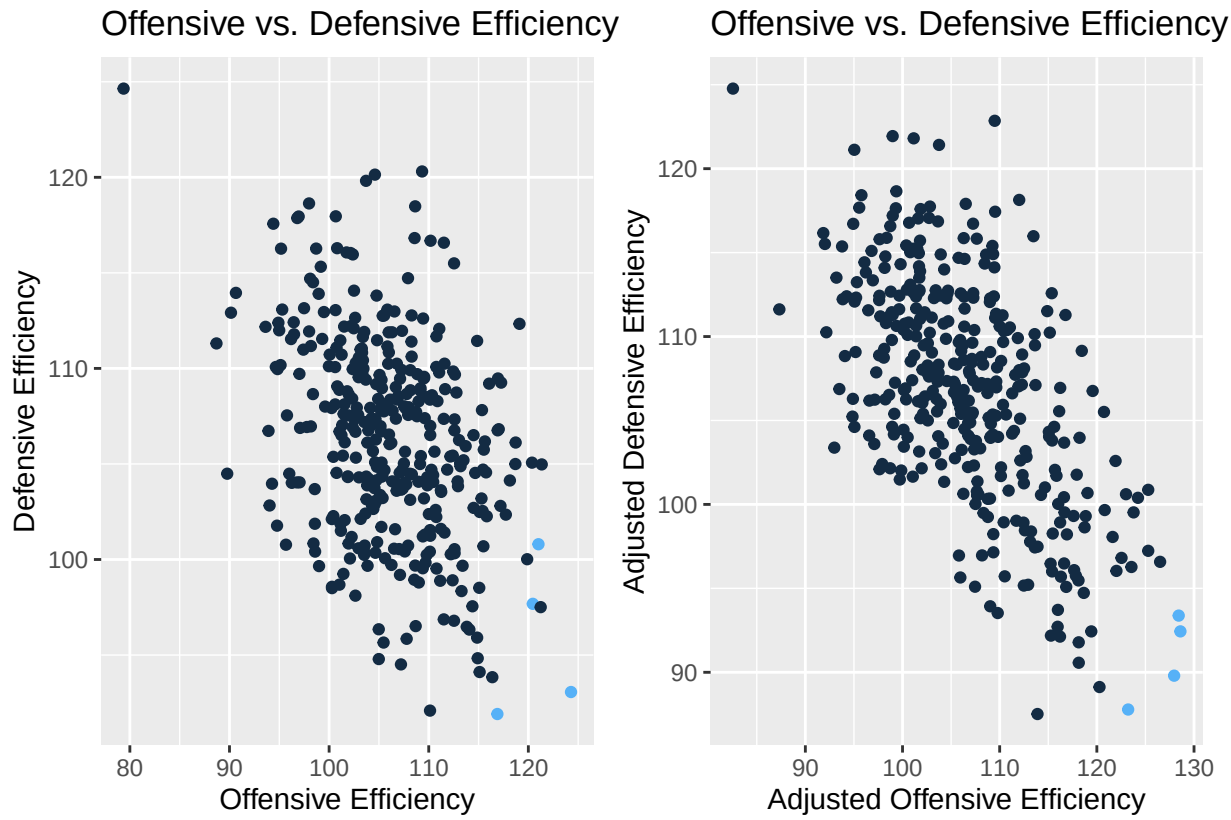
(a) Data Selection

EDA

The main goal of this project is to use supervised and unsupervised learning to model the success of various NCAA teams in the March Madness bracket and evaluate my picks.

```
final4.2025 <- c("Auburn", "Duke", "Florida", "Houston")
kpsum25 <- kpsummary %>%
  filter(Season == "2025") %>%
  mutate(final4 = ifelse(TeamName %in% final4.2025, 1, 0))
```

```
oede <- ggplot(kpsum25, aes(x = OE, y = DE)) +
  geom_point(aes(color = final4)) +
  labs(title = "Offensive vs. Defensive Efficiency",
       x = "Offensive Efficiency",
       y = "Defensive Efficiency") +
  theme(legend.position = "none")

a.oede<- ggplot(kpsum25, aes(x = AdjOE, y = AdjDE)) +
  geom_point(aes(color = final4)) +
  labs(title = "Offensive vs. Defensive Efficiency (Adj)",
       x = "Adjusted Offensive Efficiency",
       y = "Adjusted Defensive Efficiency") +
  theme(legend.position = "none")

oede + a.oede
```

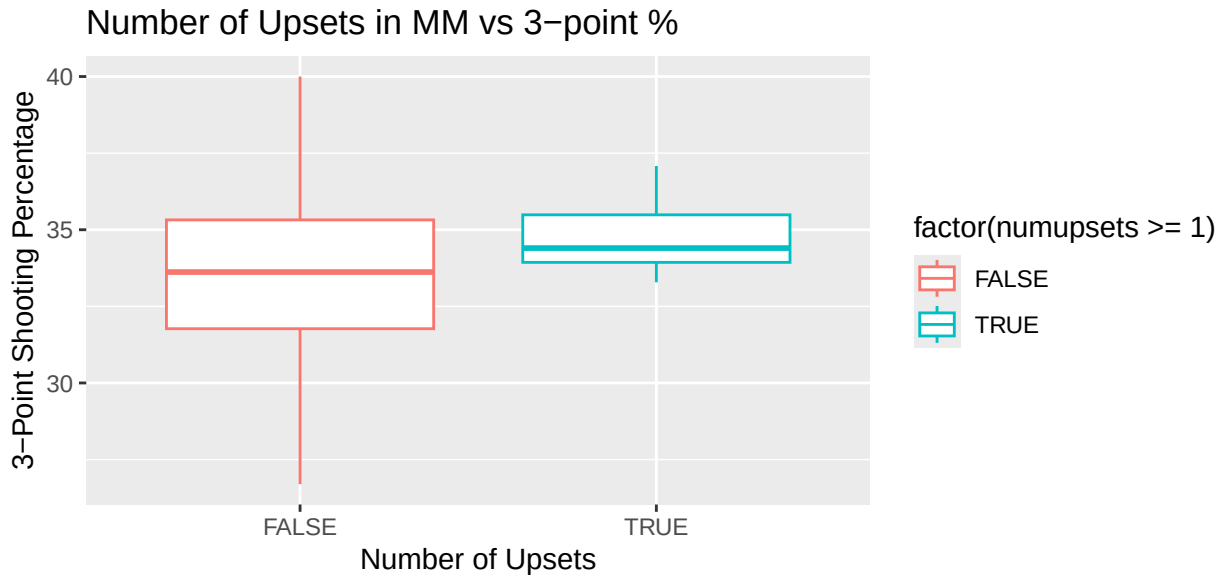


Because offensive and defensive strength are obvious contributors to the success of a basketball team, I wanted to see the relationship between offensive efficiency and defensive efficiency for this 2025 season. The teams who made the final 4 this year are marked with light blue points. Because offensive/defensive efficiency is calculated with points scored/allowed per 100 games, obviously teams that advance further play more games and will thus have more opportunities to score and concede more points. Therefore, this exploratory plot can't say much about AdjOE or AdjDE's impact on a team's success in the tournament because a team's success impacts AdjOE and AdjDE (to my understanding), but we can see an overall negative trend between offensive efficiency and defensive efficiency against the average D1 team, which means that teams that score more tend to concede less.

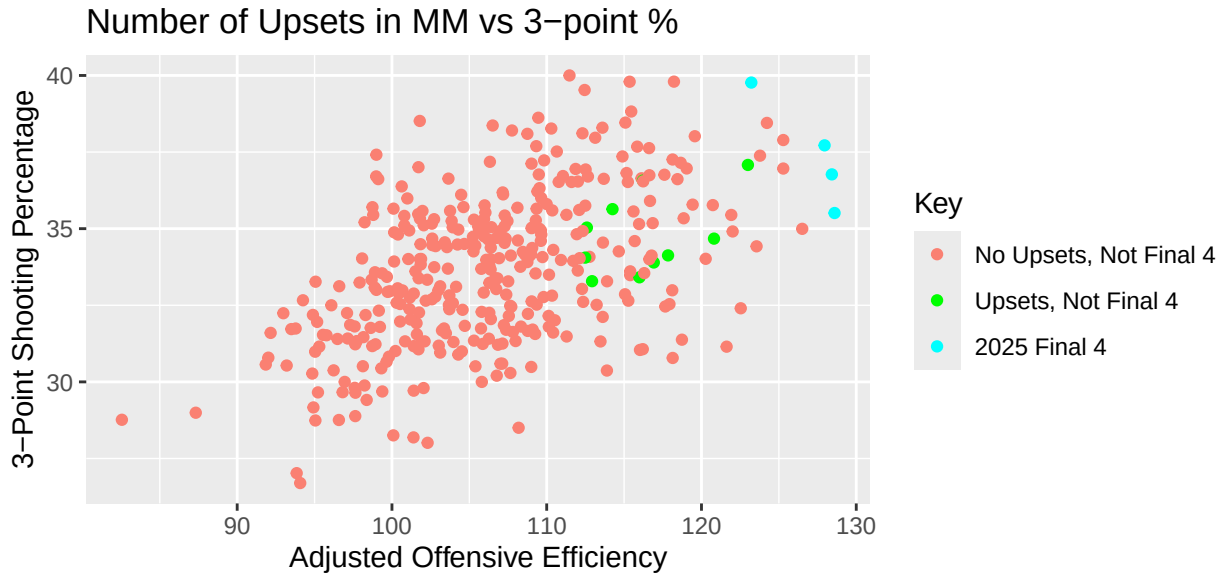
```
kpsum25 <- left_join(kpsum25,
                     misc,
                     by = c("TeamName", "Season"))

upsetters1 <- c("Creighton", "New Mexico", "Baylor", "McNeese", "Arkansas", "Drake", "Colorado St.") #7
upsetters2 <- c("Michigan", "Mississippi", "Arkansas", "BYU") #4

kpsum25 <- kpsum25 %>%
  mutate(upset1rd = ifelse(TeamName %in% upsetters1, 1, 0)) %>%
  mutate(upset2rd = ifelse(TeamName %in% upsetters2, 1, 0)) %>%
  mutate(numupsets = upset1rd + upset2rd)
```

```
# kpsum25 %>% filter(upset2rd ==1)
ggplot(kpsum25, aes(x = factor(numupsets>0), y = FG3Pct, color = factor(numupsets>=1))) +
  geom_boxplot() +
  labs(title = "Number of Upsets in MM vs 3-point %",
```

```
       x = "Number of Upsets",
       y = "3-Point Shooting Percentage")
```




```
ggplot(kpsum25, aes(x = AdjOE, y = FG3Pct, color = interaction(numupsets >= 1, final4))) +
  geom_point() +
  labs(title = "Number of Upsets in MM vs 3-point %",
       x = "Adjusted Offensive Efficiency",
       y = "3-Point Shooting Percentage") +
  scale_color_manual(
    values = c("TRUE.1" = "white", "FALSE.1" = "cyan", "TRUE.0" = "green", "FALSE.0" = "salmon"),  # Cu
    labels = c("No Upsets, Not Final 4", "Upsets, Not Final 4","2025 Final 4", "")) +
  guides(color = guide_legend(title = "Key"))
```



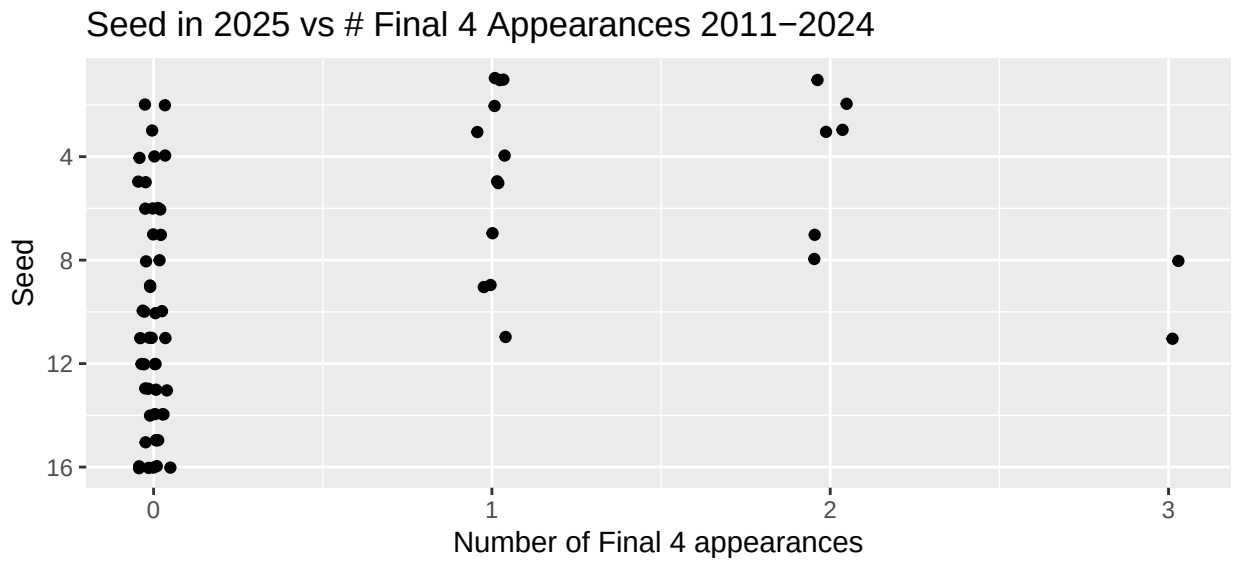
3-pointers can more quickly turn the tide of a game because they are high-risk high reward shots, and upsets may rely on the underdogs using 3-pointers to glose and overcome gaps.

Based on the boxplots, we can see that the median 3 point percentage in upsetter teams is higher than that of teams who won against lower-seeded teams. The spread of "non-upsetting" teams is wider and obviously contains many more teams of both good and bad performance, but we can see that upsetting teams may have a slightly better 3 point percentage than an average team.

We see that the final 4 teams in 2025 are in the upper group of 3 point shooting percentages, which makes sense as a bracket favors consistent top performers, and the upsets had between around 33-37% 3 point shooting, which is a range dense with other teams of similar performance, but the upsetters had higher offensive effiency.

```
pslast5 <- postseason %>% filter(Season %in% 2014:2024) %>%
  filter(`Final Four?` == "Yes") %>%
  group_by(`Team Name`) %>%
  summarise(nfinal4 = n())
ps2025 <- postseason %>% filter(Season == 2025) %>%
  left_join(pslast5, by = "Team Name") %>%
  filter(`Post-Season Tournament Sorting Index`==1) %>%
  mutate(nfinal4 = ifelse(is.na(nfinal4), 0, nfinal4))
```

```
ggplot(ps2025, aes(y = Seed, x = nfinal4)) +
  geom_jitter(width = 0.05, height = 0.05) +
  labs(title = "Seed in 2025 vs # Final 4 Appearances 2011-2024",
       y = "Seed",
       x = "Number of Final 4 appearances") +
  scale_y_reverse()
```



One might think that past program performance/reputation can impact a team's success in the upcoming tournament. We know that the teams were well seeded into the final 4, but based on this scatterplot there is not much of a relationship between a team's projected performance and their recent successes in the tournament.

Unsupervised Learning

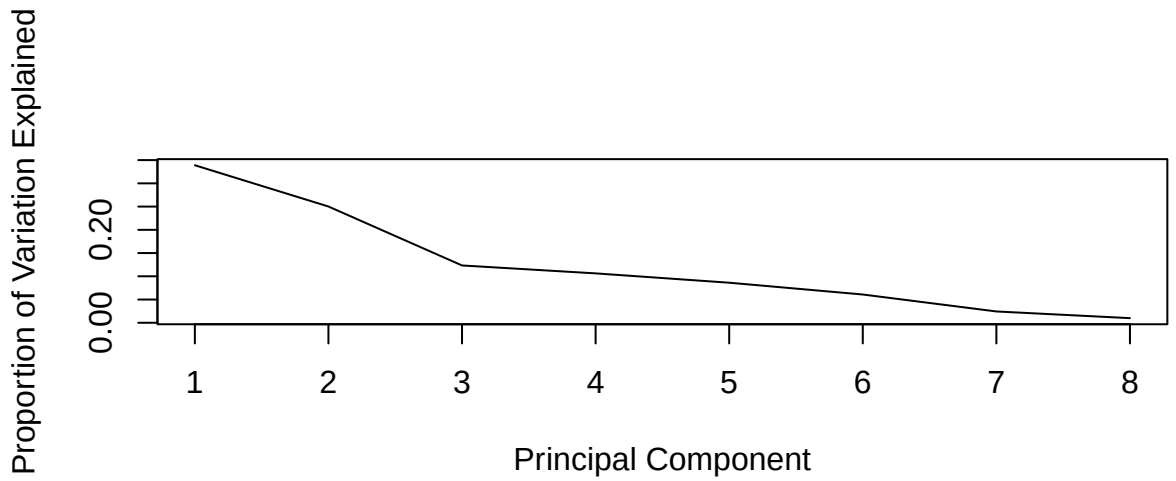
We will use the performance variables Adjusted Offensive Efficiency, Adjusted Defensive Efficiency, 2 point percentage, 3 point percentage, 3 pointer rate, Steal rate, opponent non steal turnover rate, and seed. These

statistics are likely to be correlated especially in higher-performing and more consistent teams (which we are trying to draw out), but not necessarily, as team's offensive strategies can be nuanced and can be imbalanced with their defensive strategies.

```
mm25 <- kpsum25 %>% filter(TeamName %in% mmteams) %>%
  left_join(postseason %>% select(Season, TeamName, Seed, Region),
            by = c("Season", "TeamName"))

mm25 <- as.data.frame((mm25))
rownames(mm25) <- mm25$TeamName
mm25 <- mm25 %>% select(-TeamName, -DFP) %>% select(-Season)

mm25.sub <- mm25 %>% select(AdjOE, AdjDE, FG2Pct, FG3Pct, FG3Rate, StlRate, OppNSTRate, Seed)
mm25.sub.noseed <- mm25 %>% select(AdjOE, AdjDE, FG2Pct, FG3Pct, FG3Rate, StlRate, OppNSTRate)
```

Principal Components Analysis

```
pr_out <- prcomp(mm25.sub, center = TRUE, scale. = TRUE)

plot(1:ncol(mm25.sub), summary(pr_out)$importance[2,], xlab="Principal Component",
ylab="Proportion of Variation Explained", type="l")
```

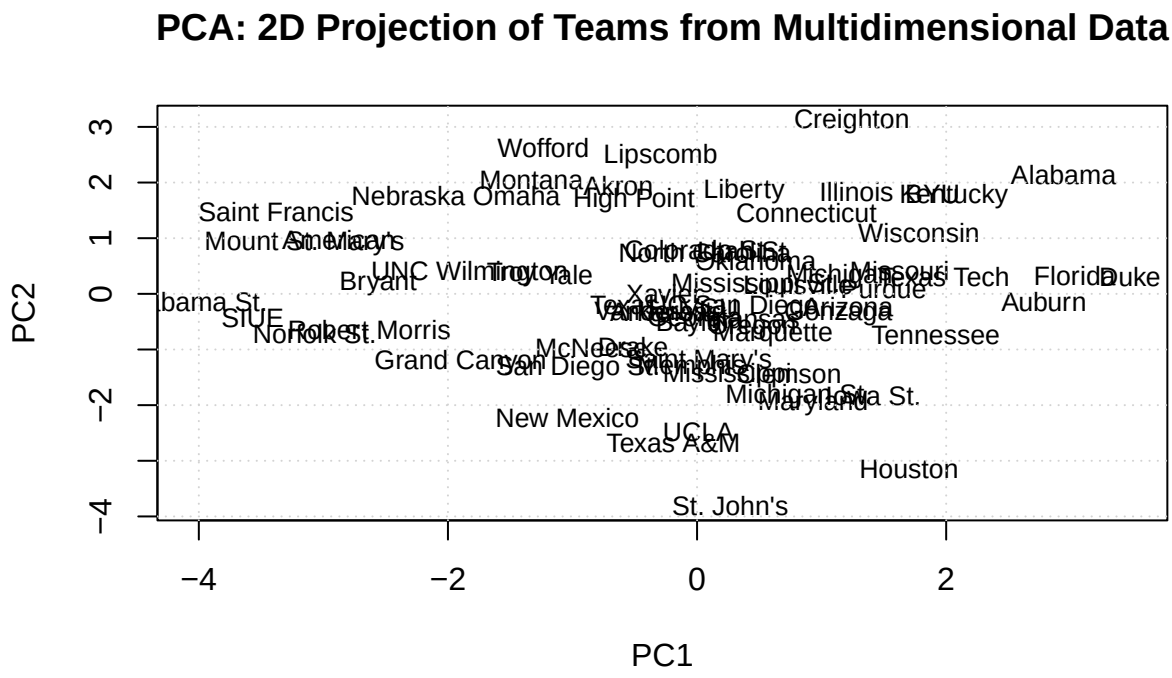



```
summary(pr_out)$importance[3,]
```

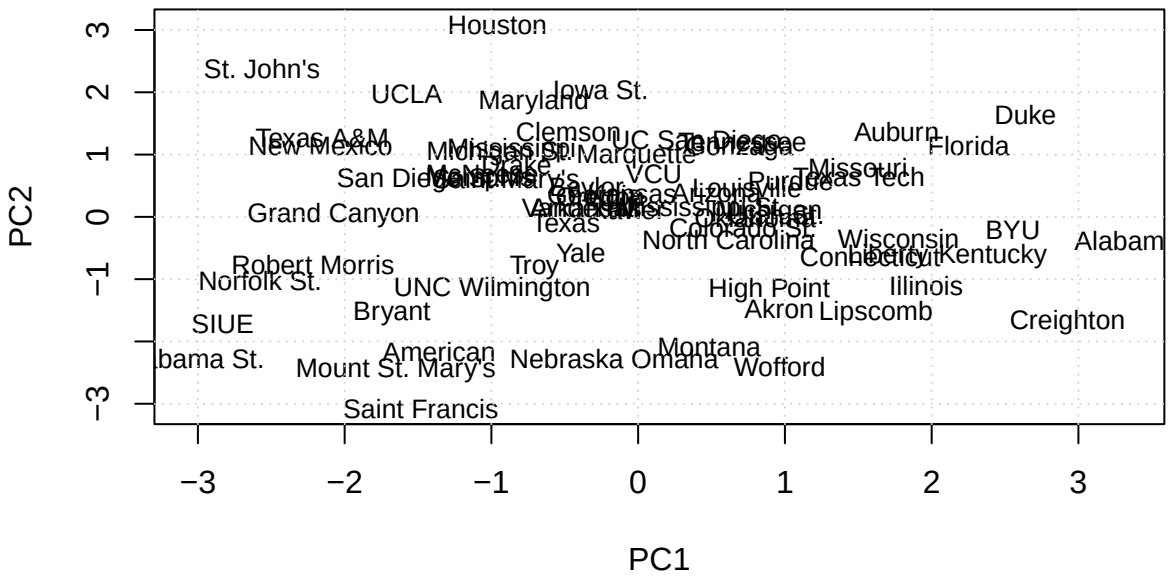
```
##     PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8
## 0.33893 0.58912 0.71242 0.81870 0.90489 0.96571 0.99004 1.00000
```

```
# pr_out$rotation
#
# #Scree
# plot(summary(pr_out)$importance[3,], type = "b",
#      xlab = "Principal Component", ylab = "Cumulative Proportion of Variance",
#      main = "Scree Plot")
```

```r
#PCA Scatterplot
plot(pr_out$x[,1], pr_out$x[,2],
     xlab = "PC1", ylab = "PC2",
     main = "PCA: 2D Projection of Teams from Multidimensional Data",
     col = "white", pch = 16)
text(pr_out$x[,1], pr_out$x[,2],
     labels = rownames(mm25.sub), col = "black", cex = 0.8)
grid()
```



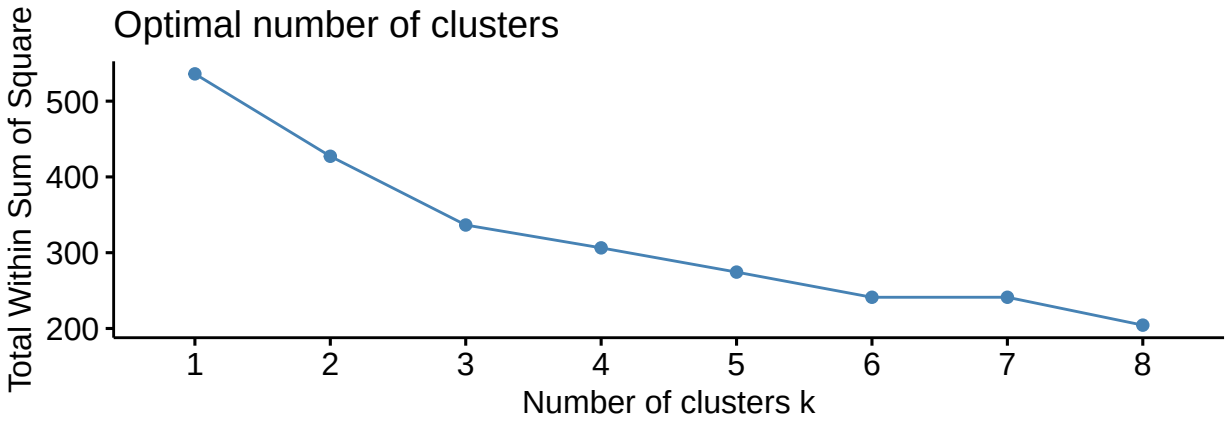

```r
#No Seed
pr_out2 <- prcomp(mm25.sub.noseed, center = TRUE, scale. = TRUE)

plot(pr_out2$x[,1], pr_out2$x[,2],
     xlab = "PC1", ylab = "PC2",
     main = "PCA: 2D Projection of Teams from Multidimensional Data (No Seed)",
     col = "white", pch = 16)
text(pr_out2$x[,1], pr_out2$x[,2],
     labels = rownames(mm25.sub.noseed), col = "black", cex = 0.8)
grid()
```



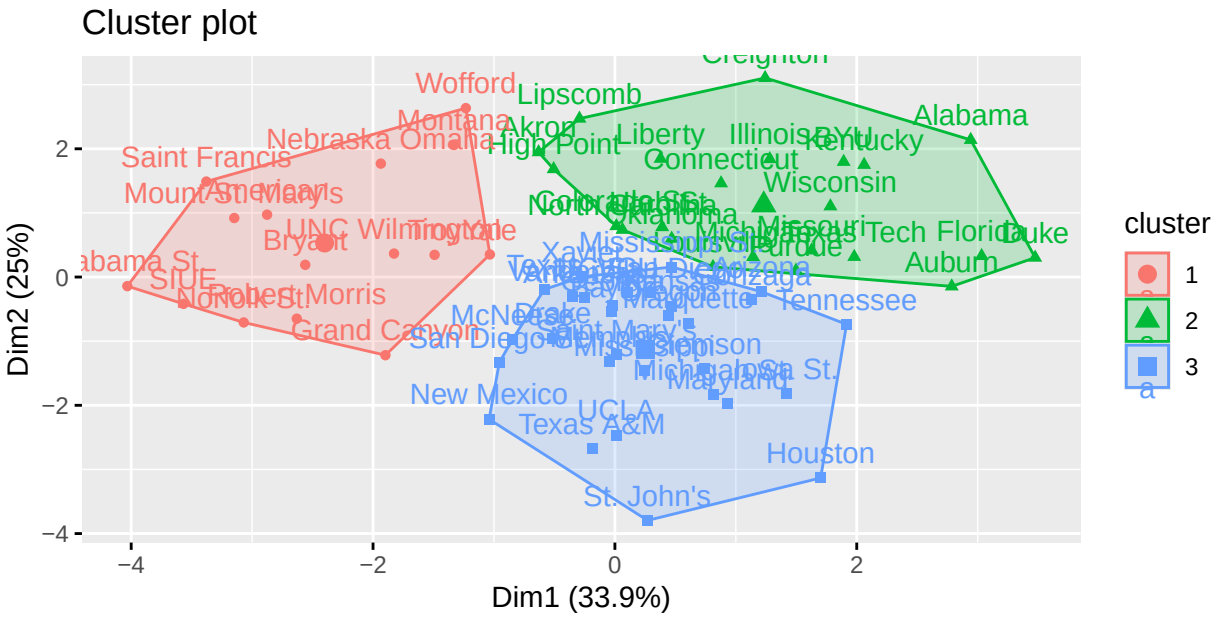
Principal components analysis can help us cluster teams based on performance metrics. The elbow plot shows us that between 3/4 PCs capture most of the variance within the dataset. In the 2D PCA plot, we see quite a few finer clusters with many in positions that are difficult to discern, which makes it difficult to judge which teams are more similar and more plausible to be successful given the variables considered. Offensive Efficiency, Seed, and Defensive Efficiency appear to contribute more than other variables to the first 2 principal components which comprise of most of the variability, but not to a largely significant degree. Notably, Florida, Auburn and Duke appear close to one another, but other teams with dissimilar projections and dissimilar results appear close and far from one another.

K-Means

```
sub_scaled <- scale(mm25.sub)
fviz_nbclust(sub_scaled, kmeans, method = "wss", k.max=8)
```




```
out <- kmeans(sub_scaled, centers=3, nstart=10)
fviz_cluster(out, data = sub_scaled)
```



Using K-means, I clustered the teams into 3 groups (based on the elbow plot, 3 clusters significantly lowered variance within clusters) based on their performance statistics, like similar offensive/defensive performance and may help identify what types of team profiles are more associated with tournament success. The red cluster for instance contains similar teams who weren't favored, seeded lower, and also performed lower. We can examine what variables the clusters have in common to help predict success in the tournament.

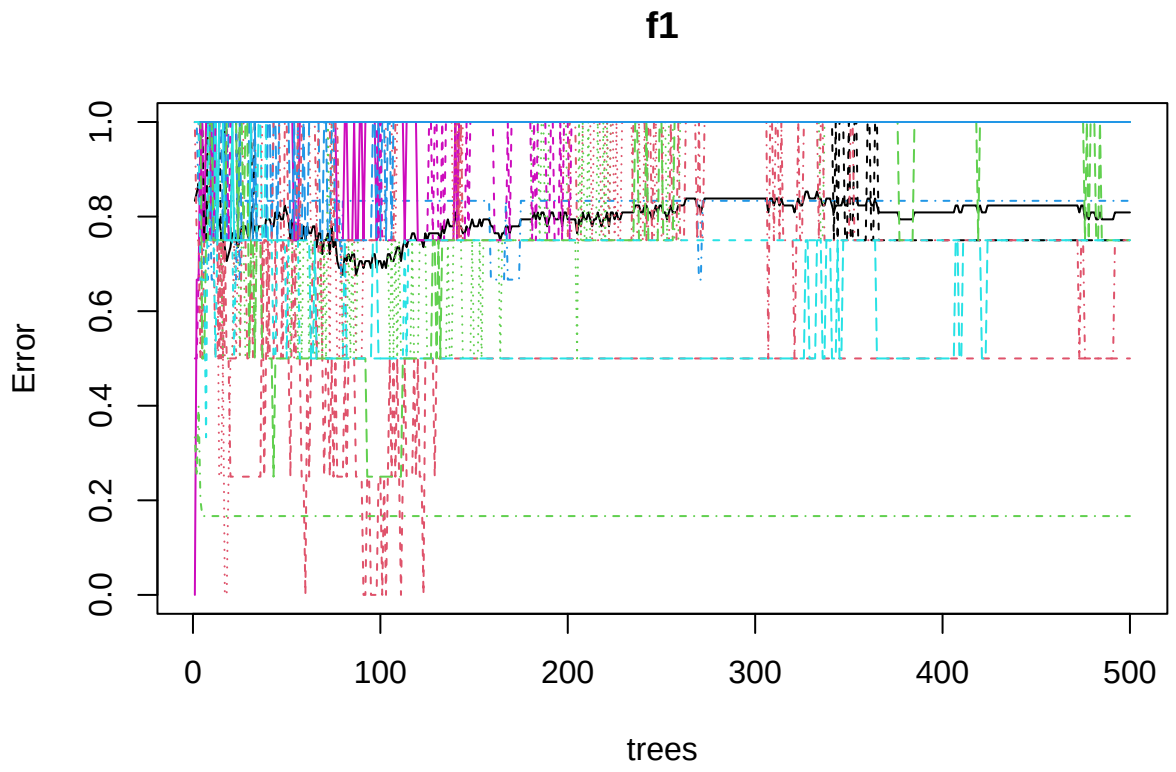
Supervised Strategies

For these supervised strategies, I chose to analyze Adjusted Efficiency Margin, which is the points scored vs conceded in the past 100 games (combining offensive/defensive efficiency since they were highly correlated from the EDA), 3 point percentages, Coaching Length (previous tournament performance didn't seem promising from EDA but possibly coaching experience could contribute to better performance), Opponent non-steal turnover rate (which could be unique as it is not directly cause by defensive pressure and can have more to do with mentality, pace, zoning, timing in the game), average height. I will use Seed as a proxy

variable for overall success in March Madness since we saw that this year the outcomes were fairly close to the seeding of teams, notably the final 4. These variables were chosen to minimize correlation.

Random Forest

```
supersub$Seed <- as.factor(supersub$Seed)
f1 <- randomForest(Seed ~ ., data=supersub)
predicted <- predict(f1)
plot(f1)
```

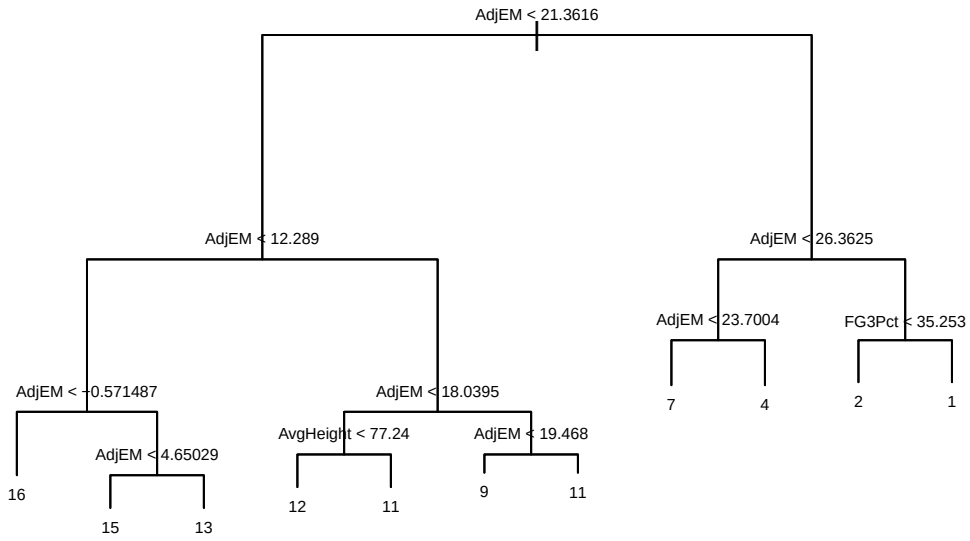



```
importance(f1)
```

```
##             MeanDecreaseGini
## AdjEM               22.099519
## FG3Pct              10.716711
## CoachLength          8.187216
## OppNSTRate          10.629398
## AvgHeight           11.082180
```

```
confmatrix <- confusionMatrix(predicted, supersub$Seed)
accuracy <- confmatrix$overall['Accuracy']
```

```
#tree no CoachLength
t1 <- tree(Seed~AdjEM+FG3Pct+OppNSTRate+AvgHeight, data=supersub)
plot(t1)
text(t1, cex = 0.5)
```



My random forest model had an overall accuracy fo 0.24, which is not good. However, we can discern from the importance and from the individual tree that Adjusted Efficiency Margin is the most important predictor for Seed in these variables.

Ridge Regression

```
supersub$Seed <- as.numeric(supersub$Seed)
set.seed(222)
n <- nrow(supersub)
trainrows <- sample(1:n, size=trunc(.7*n), replace=FALSE)
testrows <- (1:n)[-trainrows]

train <- supersub[trainrows,1:6]
test <- supersub[testrows,1:6]

out <- cv.glmnet(as.matrix(train[, 1:5]), train$Seed, nfolds=10, type.measure="mse", alpha=0)
preds_ridge <- predict(out, as.matrix(test[, 1:5]))
```

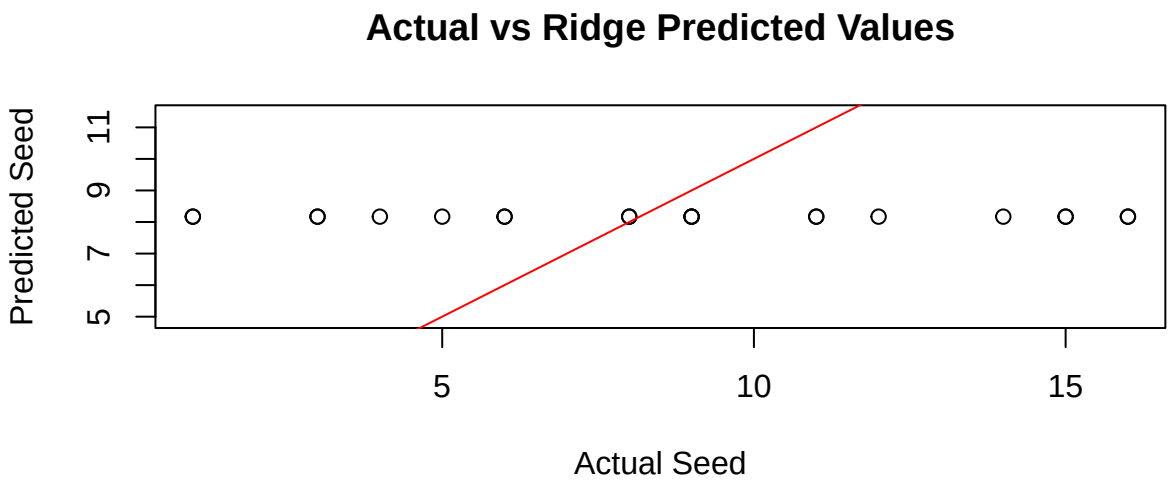
```
coef(out)
```

```
## 6 x 1 sparse Matrix of class "dgCMatrix"
##                        s1
## (Intercept)  8.170213e+00
## AdjEM        9.051422e-38
## FG3Pct      -3.426525e-37
## CoachLength  2.149835e-37
## OppNSTRate   4.862547e-37
## AvgHeight    4.688463e-37
```

```
(RMSPE <- sqrt(mean((preds_ridge-test$Seed)^2)))
```

```
## [1] 4.760164
```

```
plot(test$Seed, preds_ridge,
     main = "Actual vs Ridge Predicted Values",
     xlab = "Actual Seed",
     ylab = "Predicted Seed")
abline(a = 0, b = 1, col = "red") #Hopeful line for accurate predictions
```

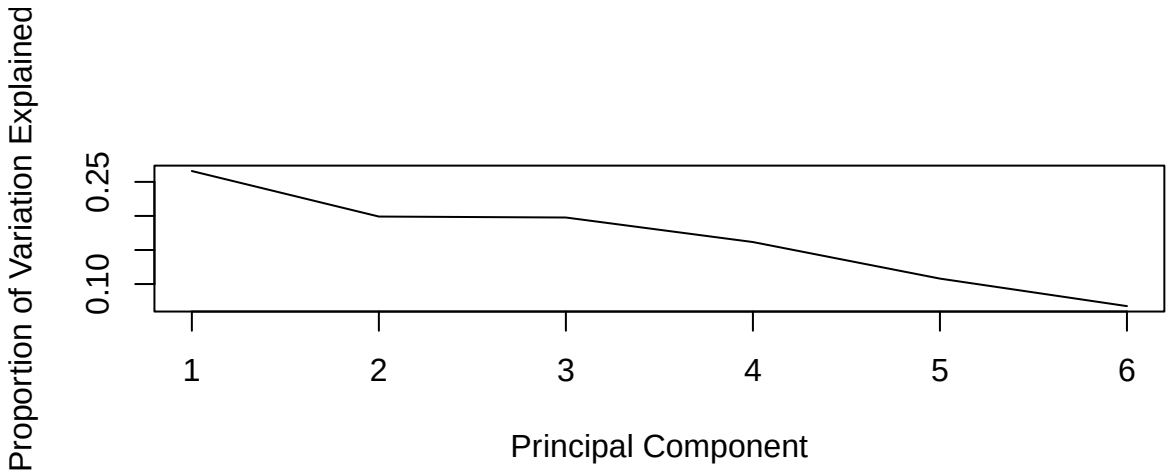


The Root Mean Squared Percentage Error is 4.76, which is okay. However, when we look at the ridge regression, which shrinks coefficients that aren't significantly contributing to the model in order to penalize complexity, all of the predictors are extremely small. Thus, These predictors aren't doing much to predict the seed or team performance at all, and looking at the plot of our actual vs ridge predicted seeds, we are quite off.

PCR

```
supersub$Seed <- as.numeric(supersub$Seed)
m1 <- pcr(Seed~., center=TRUE, scale=TRUE, data=supersub)
```

```
# determine number of PCs to use
pr_out2 <- prcomp(supersub, center=TRUE, scale.=TRUE)
plot(1:ncol(supersub), summary(pr_out2)$importance[2,], xlab="Principal Component",
ylab="Proportion of Variation Explained", type="l")
```




```
# Using 2 PCs
m1$coefficients[, , 1:2]
```

```
##                 1 comps     2 comps
## AdjEM        0.55240651  0.56736908
## FG3Pct      -0.16463778 -0.14101496
## CoachLength  0.06182313  0.08401746
## OppNSTRate  -0.04738603 -0.03219640
## AvgHeight    0.62339198  0.61532553
```

```
pc2 <- predict(m1)[,,2]
sqrt(mean((pc2-supersub$Seed)^2))
```

```
## [1] 4.453121
```

In the PCR, we see that Adjusted Efficiency Margin and Average Height have the strongest positive impact on the first two principal components which predict seed. Perhaps surprisingly 3 point percentage and opponent non steal turnovers contribute negatively, though quite small in magnitude. The RMSE of the principal components is 4.45, meaning my predicted seed differs from the actual seed by about 4.5. Considering there are only 16 seeds, this is not amazing, but since teams are seeded by 4 conferences, falling within a range of around can still be useful to predict what teams might have a chance of winning the championship/into the final 4 versus thouse that drop out earlier in the tournament.

```
michst <- supermm %>% filter (Name == "Michigan State") %>% select(-Name, -Region)
michst$Seed <- as.numeric(michst$Seed)
michst_seed_pred <- predict(m1, newdata = michst)
michst_seed_pred[,,1:2]
```

```
##  1 comps  2 comps
## 9.554197 9.607327
```

Final conclusions and findings

Overall, the PCR predicted that Michigan State would seed in around 9.6. Michigan finished in the Elite 8 and then lost. The model does a decent job of predicting that out of the 68 teams, Michigan State was a top team and would advance close to the final 8 teams. However, I picked Michigan state to win the whole thing based on vibes alone which was ultimately a poor decision based on the final results of the tournament and this prediction. In the unsupervised plots, we see in the K-means clusters that Michigan state was clustered with other mid-performing teams, and in the PCA plot not accounting for Seed, nothing about their statistics around offensive/defensive strategy set them apart or made them notably similar to other teams and indicated better chance of success.

Based on my supervised learning models, I did not do a good job at choosing predictors for Seed as a proxy for a team's success in March Madness due to variable predictions and low accuracy.