SAL 608 Assignment 3

Andrew Fish

2025-11-16

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.2
## v ggplot2   4.0.0     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(readr)
library(rpart)
library(rpart.plot)
library(ggplot2)
library(DescTools)
```

```
## Warning: package 'DescTools' was built under R version 4.5.2
```

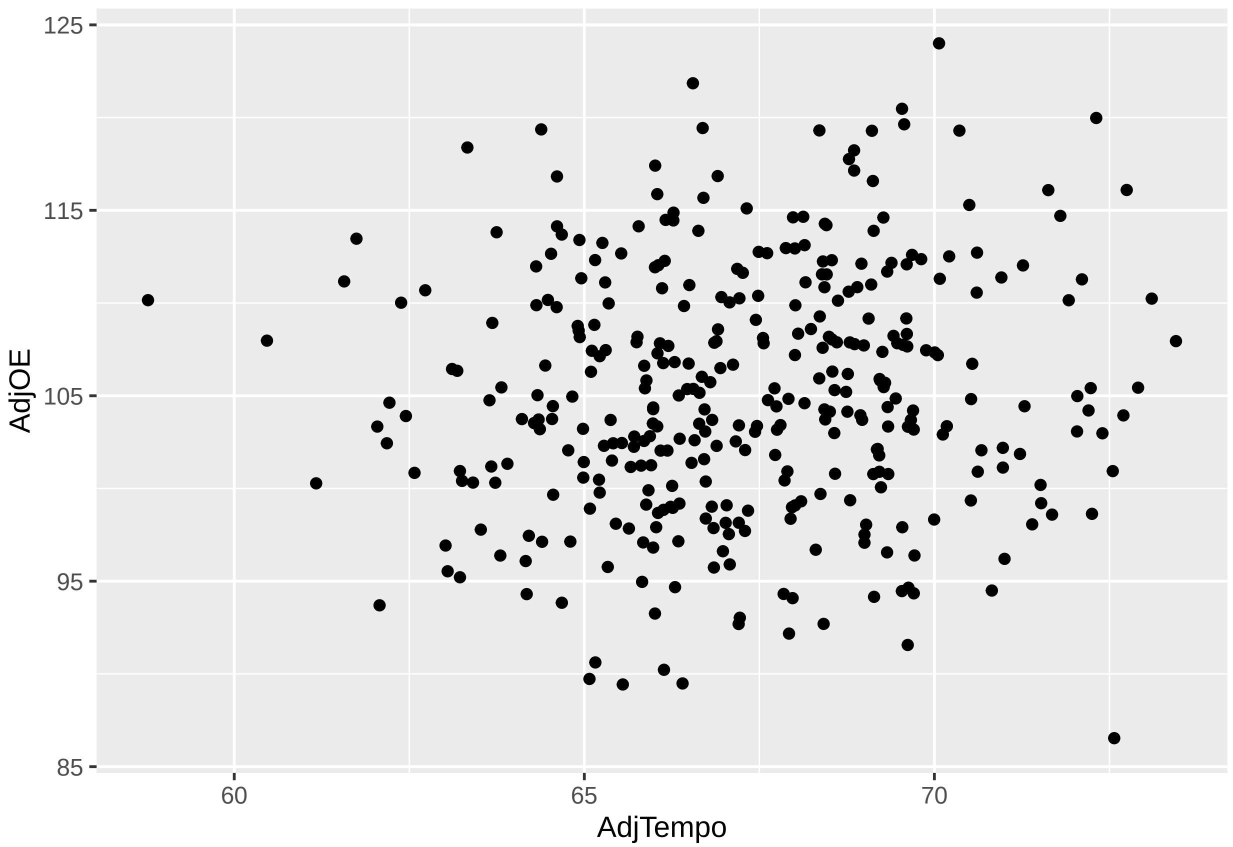
#1.

```r
set.seed(20240320)
teams <- read_csv('data/kenpom_23_pre_tourney.csv',
                 show_col_types = FALSE)

summary(teams)
```

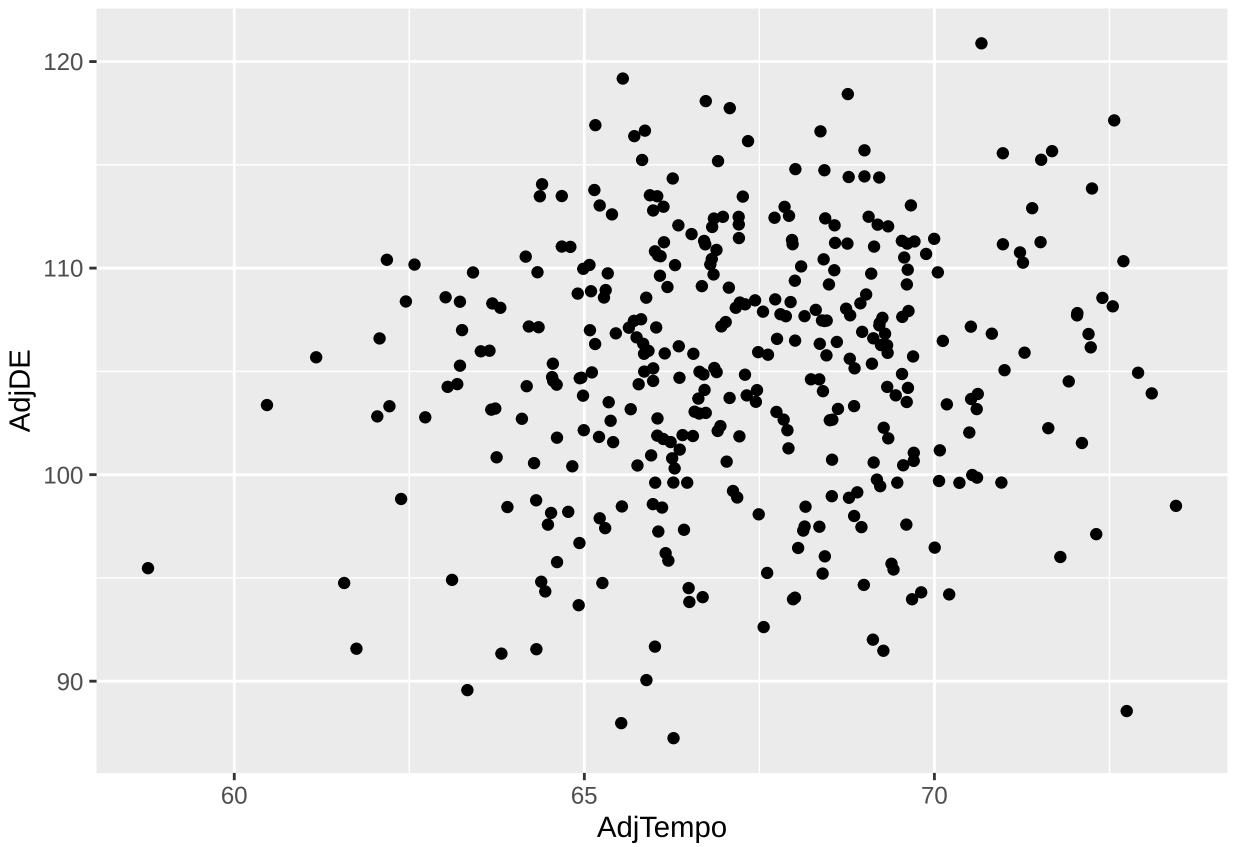
```
##      Season       TeamName             Tempo         RankTempo
##  Min.   :2023   Length:363         Min.   :59.27   Min.   :  1.0
##  1st Qu.:2023   Class :character   1st Qu.:66.23   1st Qu.: 91.5
##  Median :2023   Mode  :character   Median :67.89   Median :182.0
##  Mean   :2023                      Mean   :67.97   Mean   :182.0
##  3rd Qu.:2023                      3rd Qu.:69.75   3rd Qu.:272.5
##  Max.   :2023                      Max.   :73.98   Max.   :363.0
##
##     AdjTempo      RankAdjTempo          OE              RankOE
##  Min.   :58.77   Min.   :  1.0   Min.   : 87.46   Min.   :  1.0
##  1st Qu.:65.71   1st Qu.: 91.5   1st Qu.: 99.25   1st Qu.: 91.5  
##  Median :67.20   Median :182.0   Median :102.81   Median :182.0  
##  Mean   :67.30   Mean   :182.0   Mean   :103.07   Mean   :182.0  
##  3rd Qu.:69.11   3rd Qu.:272.5   3rd Qu.:107.21   3rd Qu.:272.5  
##  Max.   :73.45   Max.   :363.0   Max.   :120.41   Max.   :363.0  
##                                                                  
##      AdjOE          RankAdjOE           DE             RankDE     
##  Min.   : 86.54   Min.   :  1.0   Min.   : 87.42   Min.   :  1.0  
##  1st Qu.:100.45   1st Qu.: 91.5   1st Qu.: 99.70   1st Qu.: 91.5  
##  Median :104.76   Median :182.0   Median :103.36   Median :182.0  
##  Mean   :105.13   Mean   :182.0   Mean   :103.33   Mean   :182.0  
##  3rd Qu.:110.14   3rd Qu.:272.5   3rd Qu.:107.03   3rd Qu.:272.5  
##  Max.   :124.00   Max.   :363.0   Max.   :116.62   Max.   :363.0  
##                                                                   
##      AdjDE          RankAdjDE         AdjEM              RankAdjEM    
##  Min.   : 87.24   Min.   :  1.0   Min.   :-3.061e+01   Min.   :  1.0  
##  1st Qu.:101.00   1st Qu.: 91.5   1st Qu.:-8.728e+00   1st Qu.: 91.5  
##  Median :105.72   Median :182.0   Median :-9.292e-01   Median :182.0  
##  Mean   :105.13   Mean   :182.0   Mean   :-1.200e-06   Mean   :182.0  
##  3rd Qu.:109.80   3rd Qu.:272.5   3rd Qu.: 8.210e+00   3rd Qu.:272.5  
##  Max.   :120.88   Max.   :363.0   Max.   : 2.882e+01   Max.   :363.0  
##                                                                       
##       seed       
##  Min.   : 1.000  
##  1st Qu.: 5.000  
##  Median : 9.000  
##  Mean   : 8.794  
##  3rd Qu.:13.000  
##  Max.   :16.000  
##  NA's   :295
```

This data set includes college basketball data that helps determine seeds for March Madness. Factors include temp, offensive efficiency, defensive efficiency as well as ranks, adjusted, and adjusted ranks of the three factors.

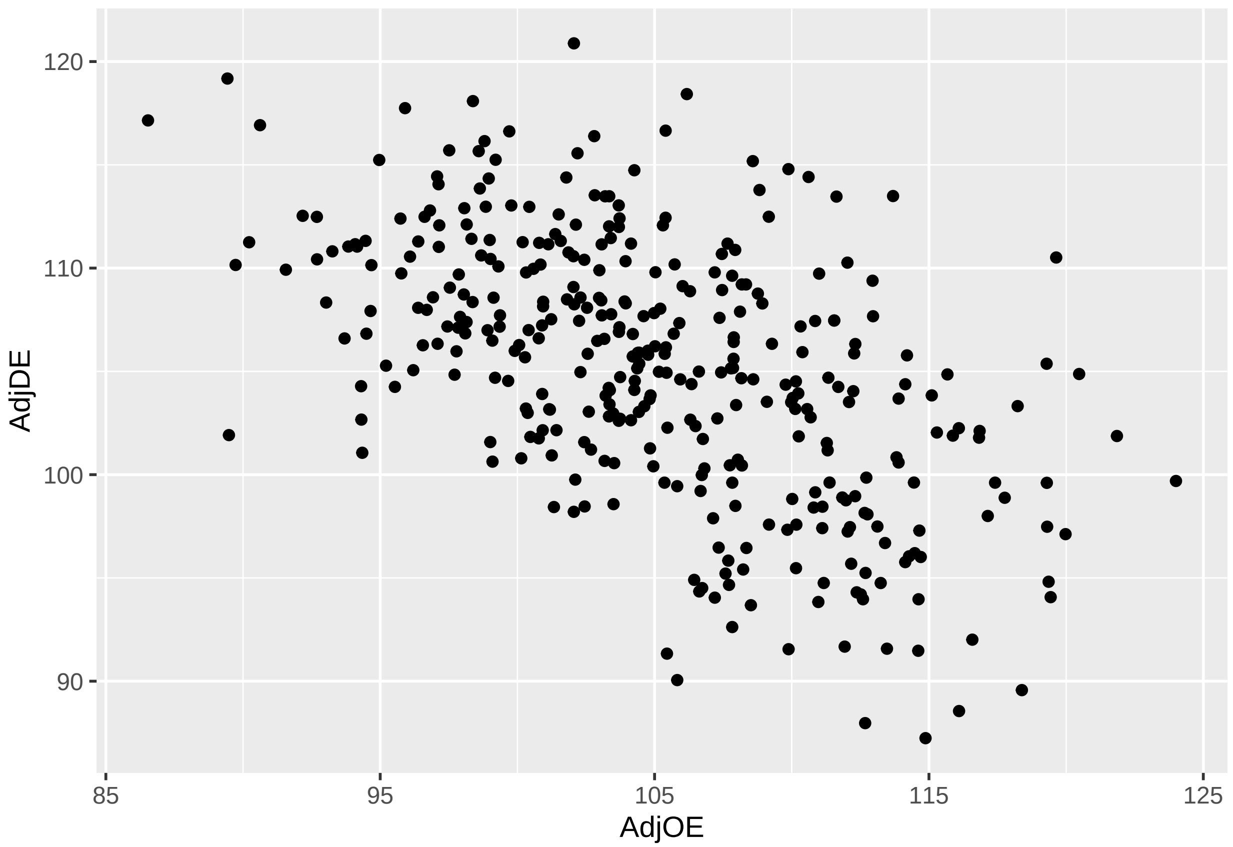
```r
ggplot(data = teams, aes(AdjTempo, AdjOE)) +
  geom_point()
```



```
ggplot(data = teams, aes(AdjTempo, AdjDE)) +
  geom_point()
```



```
ggplot(data = teams, aes(AdjOE, AdjDE)) +
  geom_point()
```



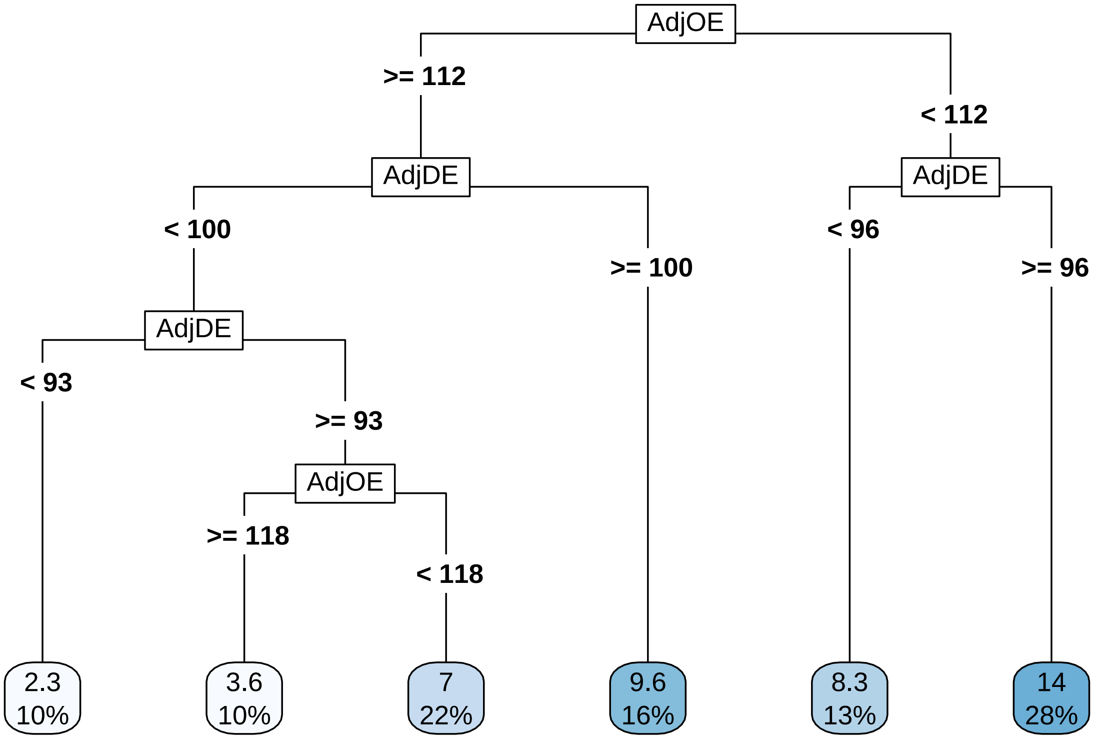
From the scatter plots we can see a clear relationship between AdjOE and AdjDE. As AdjOE increases AdjDE decreases.

#2.

```
(tree_fit <- rpart(seed ~ AdjTempo + AdjOE + AdjDE,
                   data = teams))
```

```
## n=68 (295 observations deleted due to missingness)
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 68 1479.11800  8.794118
##    2) AdjOE>=112.3415 40  546.40000  6.300000
##      4) AdjDE< 100.1414 29  250.96550  5.034483
##        8) AdjDE< 92.98855 7   15.42857  2.285714 *
##        9) AdjDE>=92.98855 22  165.81820  5.909091
##         18) AdjOE>=117.586 7   53.71429  3.571429 *
##         19) AdjOE< 117.586 15   56.00000  7.000000 *
##      5) AdjDE>=100.1414 11  126.54550  9.636364 *
##    3) AdjOE< 112.3415 28  328.42860 12.357140
##      6) AdjDE< 96.0681 9   68.00000  8.333333 *
##      7) AdjDE>=96.0681 19   45.68421 14.263160 *
```

```
##displaying our Decision Tree
##type 5 is just a visualization method
rpart.plot(x = tree_fit, type = 5)
```



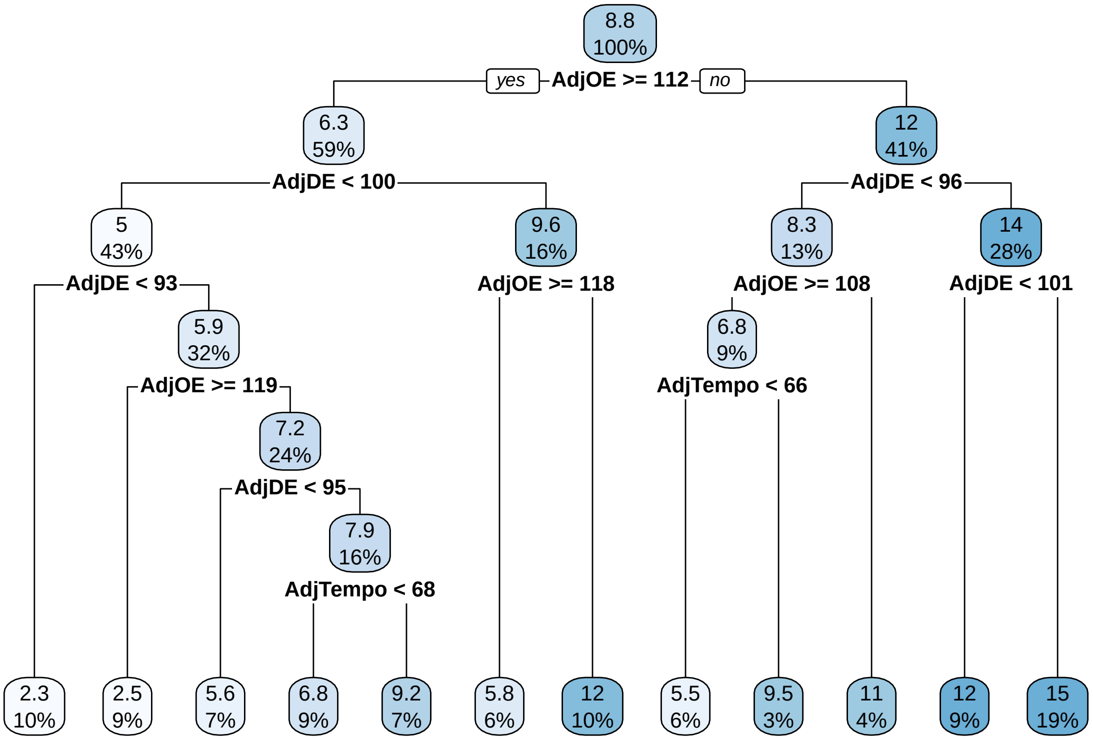
From this Decision Tree the order of importance in the predictors is AdjOE, AdjDE, and then AdjTempo. The only surprising part is that AdjTempo isn't included in the tree. Yes it is importance how efficient a team is on each side of the ball but the tempo should help determine how many possessions they are able to get in a game. It makes sense to have the AdjOE in front of the AdjDE since defense in basketball tends to rely on the opposing team missing shots and not rebounding the ball. So if your team can be more efficient on offense then you don't have play as good of defense.

#3.

```r
##setting up train/test sets
prop <- .6
n <- nrow(teams)
train <- sample(n, n * prop)

seeded <- teams %>%
  filter(is.na(seed) == FALSE)

train_data <- seeded[train, ]
test_data <- seeded[-train, ]



find_min_term <- function(min_term) {
  min_term <- floor(min_term)
  mod <- rpart(
    seed ~ AdjTempo + AdjOE + AdjDE,
    data = train_data,
    control = rpart.control(minbucket = min_term)
  )
  preds <- predict(mod, test_data)
  ##rmse
  sqrt(mean((test_data$seed - preds) ^2))
}
```

```r
##minimum RMSE and using floor since did in the function find_min_term
minterm <- floor(min(sapply(1:16, find_min_term)))
```

```r
tuned <- rpart(
  seed ~ AdjTempo + AdjOE + AdjDE,
  data = teams,
  control = rpart.control(minbucket = minterm)
)

rpart.plot(tuned)
```



There have been very significant changes to the tree. Tempo has become an actual factor in the decision tree. As well there are more branches. AdjOE and AdjDE are even more present in the tree.

#4.

```
teams <- teams %>%
  mutate(predicted = predict(tuned, teams),
         seedDiff = seed - predicted)
```

```
##looking for snubs
snubs <- teams %>%
  filter(is.na(seed)) %>%
  select(TeamName, seed, predicted) %>%
  arrange(predicted)
head(snubs, 10)
```

```
## # A tibble: 10 x 3
##    TeamName      seed predicted
##    <chr>        <dbl>     <dbl>
##  1 North Texas     NA      5.5
##  2 Toledo          NA      5.75
##  3 Liberty         NA      6.83
##  4 Michigan        NA      6.83
##  5 Oregon          NA      6.83
##  6 UAB             NA      9.2
##  7 Florida         NA      9.5
##  8 Colorado        NA     11.3
##  9 Oklahoma St.    NA     11.3
## 10 Rutgers         NA     11.3
```

```
##in but maybe shouldn't be
reverse_snubs <- teams %>%
  filter(!is.na(seed)) %>%
  select(TeamName, seed, predicted) %>%
  arrange(desc(predicted))
head(reverse_snubs, 10)
```

```
## # A tibble: 10 x 3
##    TeamName                  seed predicted
##    <chr>                    <dbl>     <dbl>
##  1 Fairleigh Dickinson         16      15.1
##  2 Grand Canyon                14      15.1
##  3 Howard                      16      15.1
##  4 Kennesaw St.                14      15.1
##  5 Louisiana                   13      15.1
##  6 Northern Kentucky           16      15.1
##  7 Princeton                   15      15.1
##  8 Southeast Missouri St.      16      15.1
##  9 Texas A&M Corpus Chris      16      15.1
## 10 Texas Southern              16      15.1
```

```
over_seed <- teams %>%
  filter(!is.na(seed)) %>%
  select(TeamName, seed, predicted, seedDiff) %>%
  arrange(desc(seedDiff))
head(over_seed, 10)
```

```
## # A tibble: 10 x 4
##    TeamName      seed predicted seedDiff
##    <chr>        <dbl>     <dbl>    <dbl>
##  1 Colgate         15     11.9      3.14
##  2 Saint Mary’s     5      2.29     2.71
##  3 Arkansas         8      5.6      2.4
##  4 Iowa             8      5.75     2.25
##  5 Auburn           9      6.83     2.17
##  6 Tennessee        4      2.29     1.71
##  7 Connecticut      4      2.5      1.5
##  8 Montana St.     14     12.5      1.5
##  9 Northwestern     7      5.5      1.5
## 10 Missouri         7      5.75     1.25
```

```
under_seed <- teams %>%
  filter(!is.na(seed)) %>%
  select(TeamName, seed, predicted, seedDiff) %>%
  arrange(seedDiff)
head(under_seed, 10)
```

```
## # A tibble: 10 x 4
##    TeamName    seed predicted seedDiff
##    <chr>      <dbl>     <dbl>    <dbl>
##  1 Indiana        4      6.83    -2.83
##  2 Baylor         3      5.75    -2.75
##  3 Kansas St.     3      5.6     -2.6
##  4 Louisiana     13     15.1     -2.08
##  5 Penn St.      10     11.9     -1.86
##  6 Nevada        11     12.5     -1.5
##  7 Purdue         1      2.5     -1.5
##  8 Virginia       4      5.5     -1.5
##  9 Alabama        1      2.29    -1.29
## 10 Houston        1      2.29    -1.29
```