Main R Markdown Document For NCAA Basketball Analysis

Ben Wasserman and Connor Kennedy

In this document we clean the data, join the day, add some columns, create factors, and provide some summary statistics around the data. We are also interested in adding TV ratings and/or revenue to see how upsets and seasons affect TV ratings and revenue generated

```
# get and set working directory

# read in files
tourney_games <- read.csv("ConferenceTourneyGames.csv")
tourney_compact_results <- read.csv("NCAATourneyCompactResults.csv")
regular_season_compact_results <- read.csv("RegularSeasonCompactResults.csv")
seasons <- read.csv("Seasons.csv")
team_coaches <- read.csv("TeamCoaches.csv")
teams <- read.csv("Teams.csv")
team_conferences <- read.csv("TeamConferences.csv")
tourney_seeds <- read.csv("NCAATourneySeeds.csv")
tourney_seeds_round_slots <- read.csv("NCAATourneySeedRoundSlots.csv")
tourney_detailed_result <- read.csv("NCAATourneyDetailedResults.csv")
tourney_slots <- read.csv("NCAATourneySlots.csv")
regular_season_detailed_result <- read.csv("RegularSeasonDetailedResults.csv")
```

```
########### BELOW ARE A VARIETY OF DATA CLEANING AND FEATURE ENGINEERING #################
tourney_slots <- tourney_slots %>% rename(GameSlot = Slot)

# merge rounds, teams, and seeds
rounds_df <- left_join(x = tourney_slots,
                       y = tourney_seeds_round_slots,
                       by = "GameSlot")
```

```
## Warning: Column `GameSlot` joining factors with different levels, coercing
## to character vector
```

```
# one unique game slot
rounds_df <- distinct(rounds_df, GameSlot, Season)

# join by slot and season
rounds_df <- left_join(rounds_df,
                       tourney_slots,
                       by = c("GameSlot", "Season"))
```

```
## Warning: Column `GameSlot` joining character vector and factor, coercing
## into character vector
```

```
# team conferences to tourney results
team_conferences <- team_conferences %>% rename(WTeamID = TeamID)
```

```r
# join with conferences
tourney_detailed_result <- left_join(x = tourney_detailed_result,
                                     y = team_conferences,
                                     by = c("WTeamID", "Season"))

team_conferences <- team_conferences %>% rename(LTeamID = WTeamID)

tourney_detailed_result <- left_join(x = tourney_detailed_result,
                                     y = team_conferences,
                                     by = c("LTeamID", "Season"))

tourney_detailed_result <- tourney_detailed_result %>% rename(
  WTeamConf = ConfAbbrev.x,
  LTeamConf = ConfAbbrev.y
)

# get rid of all seeding prior to 2003
tourney_seeds <- tourney_seeds %>% mutate(as.numeric(Season) > 2002)
View(tourney_seeds)

# join with seeds
tourney_seeds <- subset(tourney_seeds, Season > 2002)
tourney_seeds <- tourney_seeds %>% rename(WTeamID = TeamID)

tourney_detailed_result <-left_join(x = tourney_detailed_result,
                                    y = tourney_seeds,
                                    by = c("WTeamID", "Season"))

tourney_seeds <- tourney_seeds %>% rename(LTeamID = WTeamID)

tourney_detailed_result <-left_join(x = tourney_detailed_result,
                                    y = tourney_seeds,
                                    by = c("LTeamID", "Season"))

tourney_detailed_result <- tourney_detailed_result %>% rename(
  WTeamSeed = Seed.x,
  LTeamSeed = Seed.y
)

# replace team id's with team names
tourney_detailed_result[["WTeamID"]] <- teams[ match(tourney_detailed_result[['WTeamID']], teams[['Team
tourney_detailed_result[["LTeamID"]] <- teams[ match(tourney_detailed_result[['LTeamID']], teams[['Team

# factoring
tourney_detailed_result %<>% mutate(Season = factor(Season),
                                    DayNum = factor(DayNum),
                                    WTeamID = factor(WTeamID),
                                    LTeamID = factor(LTeamID),
                                    NumOT = factor(NumOT))

# renaming
tourney_detailed_result %<>% rename(WTeamName = WTeamID,
                                    LTeamName = LTeamID)
```

```r
# add score difference
tourney_detailed_result <- tourney_detailed_result %>% mutate(scoreDiff = WScore - LScore)
```

```r
# summary statistics
summary(tourney_detailed_result)
```

```
##      Season        DayNum             WTeamName       WScore
##  2011   : 67   136    :240   North Carolina: 42   Min.   : 47.00
##  2012   : 67   137    :240   Kansas        : 38   1st Qu.: 68.00
##  2013   : 67   138    :120   Kentucky      : 35   Median : 75.00
##  2014   : 67   139    :120   Duke          : 33   Mean   : 75.08
##  2015   : 67   143    : 60   Michigan St   : 31   3rd Qu.: 82.00
##  2016   : 67   144    : 60   Florida       : 30   Max.   :121.00
##  (Other):579   (Other):141   (Other)       :772
##        LTeamName       LScore       WLoc    NumOT        WFGM
##  Gonzaga     : 15   Min.   : 29.00   N:981   0:923   Min.   :13.00
##  Michigan St: 15   1st Qu.: 57.00           1: 46   1st Qu.:23.00
##  Wisconsin   : 15   Median : 63.00           2: 12   Median :26.00
##  Kansas      : 14   Mean   : 63.66                   Mean   :26.17
##  Arizona     : 13   3rd Qu.: 71.00                   3rd Qu.:29.00
##  Duke        : 13   Max.   :105.00                   Max.   :44.00
##  (Other)     :896
##       WFGA           WFGM3            WFGA3            WFTM
##  Min.   :34.00   Min.   : 0.000   Min.   : 4.00   Min.   : 0.00
##  1st Qu.:50.00   1st Qu.: 5.000   1st Qu.:14.00   1st Qu.:12.00
##  Median :55.00   Median : 7.000   Median :17.00   Median :15.00
##  Mean   :54.98   Mean   : 6.788   Mean   :17.62   Mean   :15.94
##  3rd Qu.:59.00   3rd Qu.: 9.000   3rd Qu.:21.00   3rd Qu.:20.00
##  Max.   :84.00   Max.   :16.000   Max.   :35.00   Max.   :38.00
##
##       WFTA            WOR              WDR             WAst
##  Min.   : 1.00   Min.   : 0.00   Min.   :13.00   Min.   : 3.00
##  1st Qu.:17.00   1st Qu.: 8.00   1st Qu.:22.00   1st Qu.:11.00
##  Median :21.00   Median :10.00   Median :25.00   Median :14.00
##  Mean   :21.91   Mean   :10.74   Mean   :25.82   Mean   :14.23
##  3rd Qu.:27.00   3rd Qu.:13.00   3rd Qu.:29.00   3rd Qu.:17.00
##  Max.   :48.00   Max.   :26.00   Max.   :43.00   Max.   :29.00
##
##       WTO             WStl             WBlk             WPF
##  Min.   : 3.00   Min.   : 0.000   Min.   : 0.000   Min.   : 5.00
##  1st Qu.: 9.00   1st Qu.: 4.000   1st Qu.: 2.000   1st Qu.:14.00
##  Median :11.00   Median : 6.000   Median : 4.000   Median :16.00
##  Mean   :11.68   Mean   : 6.461   Mean   : 3.969   Mean   :16.42
##  3rd Qu.:14.00   3rd Qu.: 8.000   3rd Qu.: 5.000   3rd Qu.:19.00
##  Max.   :28.00   Max.   :20.000   Max.   :15.000   Max.   :29.00
##
##       LFGM            LFGA           LFGM3            LFGA3
##  Min.   :11.00   Min.   :37.00   Min.   : 0.000   Min.   : 5.00
##  1st Qu.:20.00   1st Qu.:53.00   1st Qu.: 4.000   1st Qu.:16.00
##  Median :23.00   Median :57.00   Median : 6.000   Median :20.00
##  Mean   :22.91   Mean   :57.61   Mean   : 6.124   Mean   :19.96
##  3rd Qu.:26.00   3rd Qu.:62.00   3rd Qu.: 8.000   3rd Qu.:23.00
##  Max.   :36.00   Max.   :85.00   Max.   :18.000   Max.   :42.00
```

```
##
##       LFTM            LFTA            LOR             LDR
##  Min.   : 0.00   Min.   : 2.00   Min.   : 1.00   Min.   : 8.00
##  1st Qu.: 8.00   1st Qu.:12.00   1st Qu.: 8.00   1st Qu.:18.00
##  Median :11.00   Median :16.00   Median :11.00   Median :21.00
##  Mean   :11.73   Mean   :16.79   Mean   :11.26   Mean   :21.14
##  3rd Qu.:15.00   3rd Qu.:21.00   3rd Qu.:14.00   3rd Qu.:24.00
##  Max.   :31.00   Max.   :39.00   Max.   :26.00   Max.   :42.00
##
##       LAst            LTO             LStl            LBlk
##  Min.   : 2.00   Min.   : 3.00   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 9.00   1st Qu.: 9.00   1st Qu.: 4.000   1st Qu.: 1.000
##  Median :11.00   Median :12.00   Median : 6.000   Median : 3.000
##  Mean   :11.45   Mean   :12.24   Mean   : 5.822   Mean   : 2.903
##  3rd Qu.:14.00   3rd Qu.:15.00   3rd Qu.: 7.000   3rd Qu.: 4.000
##  Max.   :23.00   Max.   :27.00   Max.   :19.000   Max.   :13.000
##
##       LPF             WTeamConf        LTeamConf     WTeamSeed
##  Min.   : 7.00   big_east  :147   big_east  : 98   X01    : 61
##  1st Qu.:16.00   acc       :141   big_ten   : 88   Y01    : 50
##  Median :19.00   big_ten   :128   big_twelve: 85   Z01    : 49
##  Mean   :19.12   big_twelve:119   acc       : 79   W01    : 40
##  3rd Qu.:22.00   sec       :106   sec       : 66   Y02    : 38
##  Max.   :33.00   pac_ten   : 52   a_ten     : 48   W04    : 36
##                  (Other)   :288   (Other)   :517   (Other):707
##  as.numeric(Season) > 2002.x   LTeamSeed   as.numeric(Season) > 2002.y
##  Mode:logical                  W01    : 15   Mode:logical
##  TRUE:981                      W02    : 15   TRUE:981
##                                W04    : 15
##                                W05    : 15
##                                W06    : 15
##                                W08    : 15
##                                (Other):891
##    scoreDiff
##  Min.   : 1.00
##  1st Qu.: 5.00
##  Median :10.00
##  Mean   :11.41
##  3rd Qu.:16.00
##  Max.   :56.00
##
```

```
# wins per team
tourney_wins_per_team <- tourney_detailed_result %>% group_by(WTeamName) %>%
  summarize(num_wins = n()) %>%
  arrange(desc(num_wins))
head(tourney_wins_per_team)
```

```
## # A tibble: 6 x 2
##   WTeamName      num_wins
##   <fct>             <int>
## 1 North Carolina       42
## 2 Kansas               38
## 3 Kentucky             35
```

```
## 4 Duke                 33
## 5 Michigan St          31
## 6 Florida              30
```

```
# wins per team per season
wins_per_team_per_season <- tourney_detailed_result %>% group_by(Season) %>%
  count(WTeamName)
head(wins_per_team_per_season)
```

```
## # A tibble: 6 x 3
## # Groups:   Season [1]
##   Season WTeamName      n
##   <fct>  <fct>      <int>
## 1 2003   Arizona        3
## 2 2003   Arizona St     1
## 3 2003   Auburn         2
## 4 2003   Butler         2
## 5 2003   C Michigan     1
## 6 2003   California     1
```

```
# summary details of score by seed
seed_details <- tourney_detailed_result %>% group_by(WTeamSeed) %>%
  summarize(avg_points_per_seed = mean(WScore),
            max_points_per_seed = max(WScore),
            min_points_per_seed = min(WScore))
seed_details
```

```
## # A tibble: 76 x 4
##    WTeamSeed avg_points_per_seed max_points_per_seed min_points_per_seed
##    <fct>                   <dbl>               <int>               <int>
##  1 W01                      80.6                 113                  58
##  2 W02                      76.4                 102                  60
##  3 W03                      76.2                  95                  59
##  4 W04                      71.6                  93                  53
##  5 W05                      79.8                  99                  68
##  6 W06                      70.2                  79                  61
##  7 W07                      73.9                 111                  47
##  8 W08                      74                    88                  61
##  9 W09                      74.8                  86                  61
## 10 W10                      69.2                  86                  58
## # ... with 66 more rows
```

```
# summary details of score by conference
conf_details <- tourney_detailed_result %>% group_by(WTeamConf) %>%
  summarize(avg_points_per_conf = mean(WScore),
            max_points_per_conf = max(WScore),
            min_points_per_conf = min(WScore))
conf_details
```

```
## # A tibble: 34 x 4
##    WTeamConf  avg_points_per_conf max_points_per_conf min_points_per_conf
##    <fct>                    <dbl>               <int>               <int>
```

```
##  1 a_sun                    83.2                96                  78
##  2 a_ten                    73.1                89                  55
##  3 aac                      71.1                89                  60
##  4 acc                      76.5               113                  54
##  5 aec                      67.3                71                  60
##  6 big_east                 74.4               111                  53
##  7 big_sky                  87                  87                  87
##  8 big_south                82.3                92                  74
##  9 big_ten                  73.6                99                  47
## 10 big_twelve               77.8               108                  59
## # ... with 24 more rows
```

```
# summary details of three points by season
three_pt_details <- tourney_detailed_result %>% group_by(Season) %>%
  summarize(avg_three_pts = mean(WFGM3),
            max_three_pts = max(WFGM3),
            min_three_pts = min(WFGM3))
three_pt_details
```

```
## # A tibble: 15 x 4
##    Season avg_three_pts max_three_pts min_three_pts
##    <fct>          <dbl>         <int>         <int>
##  1 2003            6.64            14             1
##  2 2004            6.91            13             2
##  3 2005            6.95            16             2
##  4 2006            6.61            12             2
##  5 2007            7.42            14             1
##  6 2008            6.94            15             3
##  7 2009            6.45            14             2
##  8 2010            7.12            15             1
##  9 2011            7.18            16             1
## 10 2012            5.97            11             1
## 11 2013            6.70            14             0
## 12 2014            6.19            14             0
## 13 2015            6.34            12             2
## 14 2016            7.18            13             1
## 15 2017            7.24            16             2
```

```
# summay details of field goals by season
field_goals_details <- tourney_detailed_result %>% group_by(Season) %>%
  summarize(avg_field_goals = mean(WFGM),
            max_field_goals = max(WFGM),
            min_field_foals = min(WFGM))
field_goals_details
```

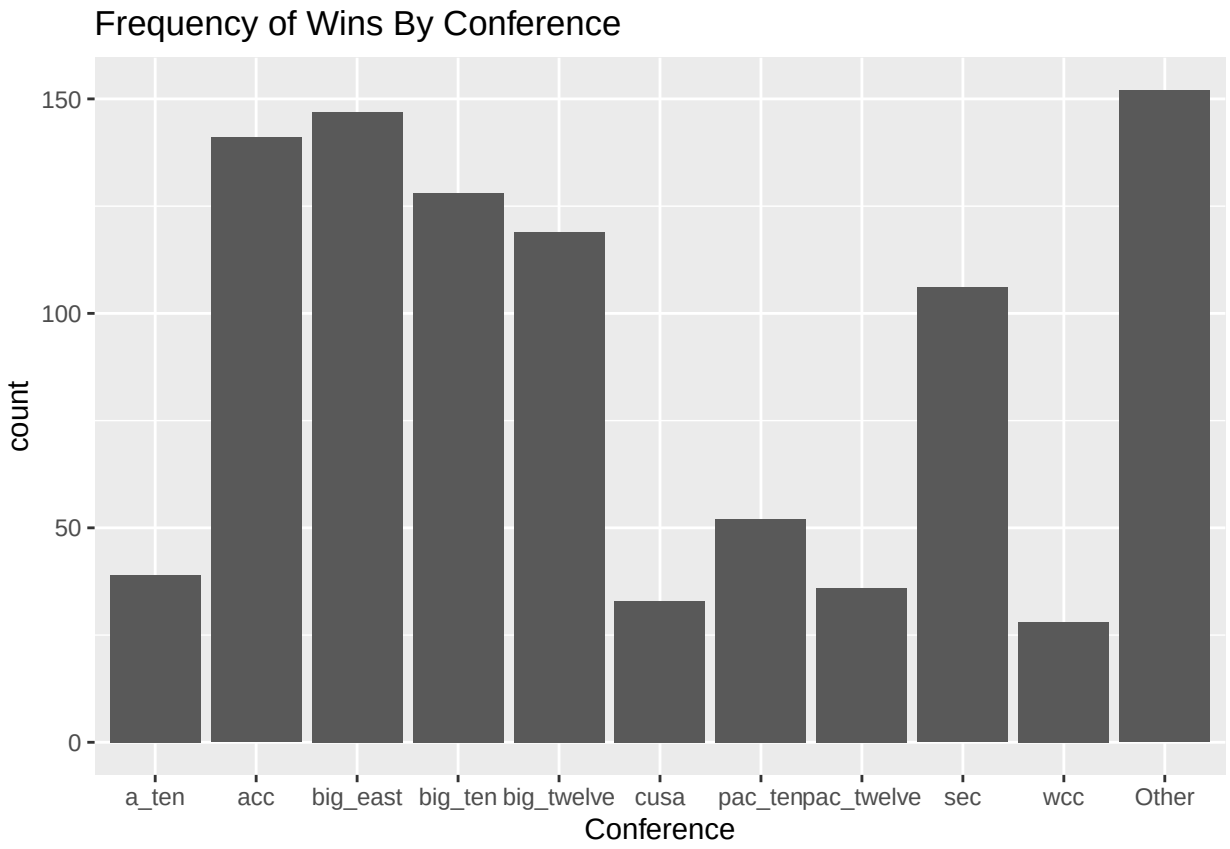
```
## # A tibble: 15 x 4
##    Season avg_field_goals max_field_goals min_field_foals
##    <fct>           <dbl>           <int>           <int>
##  1 2003             27.7              40              17
##  2 2004             25.9              35              13
##  3 2005             26.3              37              15
##  4 2006             25.0              37              14
##  5 2007             25.5              43              15
```

```
##  6 2008              27.5              44              19
##  7 2009              27.1              38              17
##  8 2010              25.5              38              15
##  9 2011              25.5              37              18
## 10 2012              25.2              40              17
## 11 2013              25.0              35              16
## 12 2014              25.8              35              17
## 13 2015              25.3              36              16
## 14 2016              27.5              39              17
## 15 2017              27.9              38              18
```

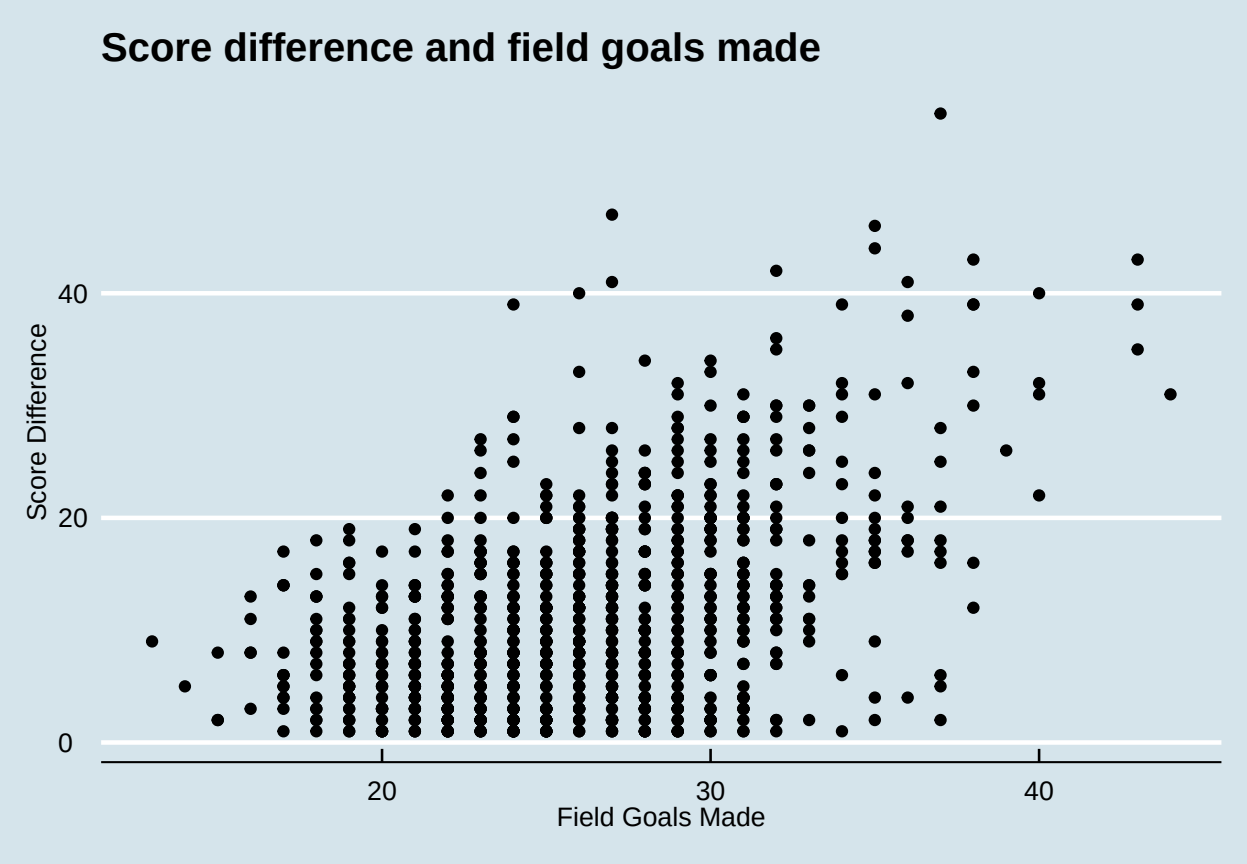
```
# summary details of score by winning
winner_seed_details <- tourney_detailed_result %>% group_by(Season, WTeamSeed) %>%
  summarize(avg_points_per_WTeamSeed = mean(WScore),
            max_points_per_WTeamSeed = max(WScore),
            min_points_per_WTeamSeed = min(WScore))
winner_seed_details
```

```
## # A tibble: 509 x 5
## # Groups:   Season [15]
##    Season WTeamSeed avg_points_per_WT~ max_points_per_WT~ min_points_per_W~
##    <fct>  <fct>                  <dbl>              <int>             <int>
##  1 2003   W01                     70                 74                65
##  2 2003   W02                     76                 76                76
##  3 2003   W03                     77                 95                63
##  4 2003   W04                     86                 86                86
##  5 2003   W06                     77                 77                77
##  6 2003   W08                     76                 76                76
##  7 2003   W10                     66.5               68                65
##  8 2003   W12                     63                 79                47
##  9 2003   X01                     81.5               85                77
## 10 2003   X02                     85                 85                85
## # ... with 499 more rows
```

```
# plots to reveal interesting patterns

library('ggthemes')

tourney_detailed_result <- tourney_detailed_result %>% mutate(WTeamConfFactor = fct_lump(WTeamConf, 10))

# top ten conferences frequency of wins
p4 <- ggplot(tourney_detailed_result, aes(x=WTeamConfFactor)) +
  geom_bar() +
  labs(x = "Conference",
       title = "Frequency of Wins By Conference")

p4
```




```
# field goals made vs score difference
p1 <- ggplot(
  tourney_detailed_result,
  aes(x = WFGM, y = scoreDiff)) +
  geom_point() +
  labs(x = "Field Goals Made",
       y = "Score Difference",
       title = "Score difference and field goals made") +
  theme_economist()
p1
```



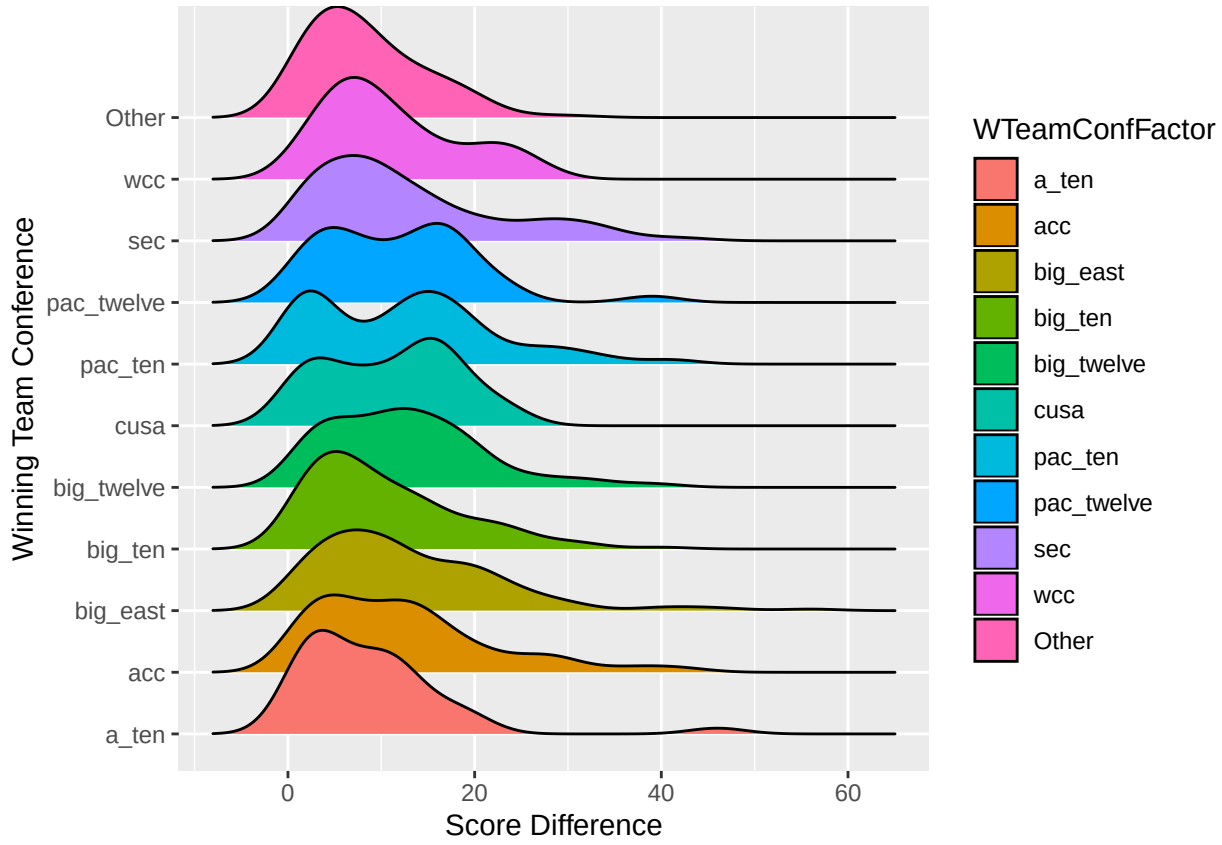

```
# density plot by conference
library(ggridges)
```

```
##
## Attaching package: 'ggridges'
```

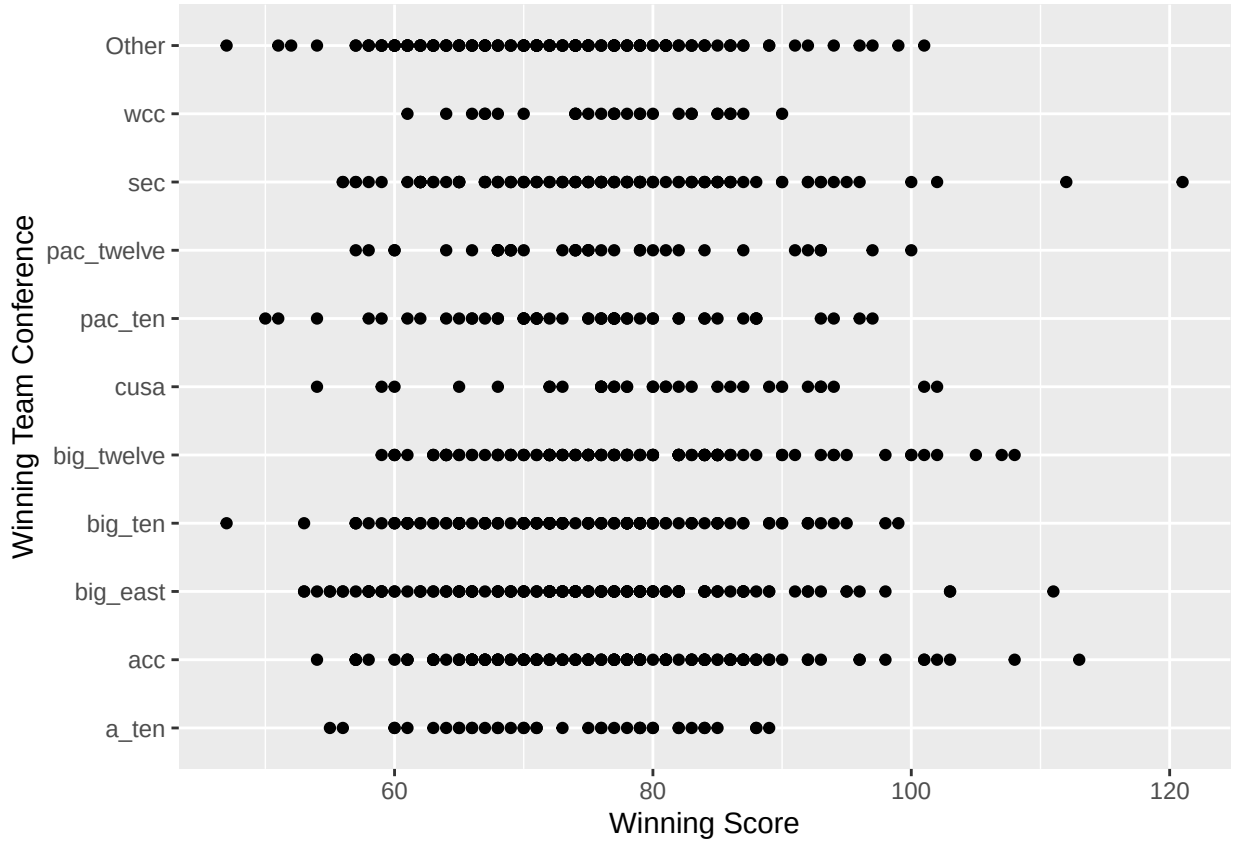
```
## The following object is masked from 'package:ggplot2':
##
##     scale_discrete_manual
```

```
p2 <- ggplot(tourney_detailed_result, aes(x = scoreDiff, y = WTeamConfFactor, fill = WTeamConfFactor)) +
  geom_density_ridges() +
  labs(x = "Score Difference",
       y = "Winning Team Conference")

p2
```

```
## Picking joint bandwidth of 3.01
```



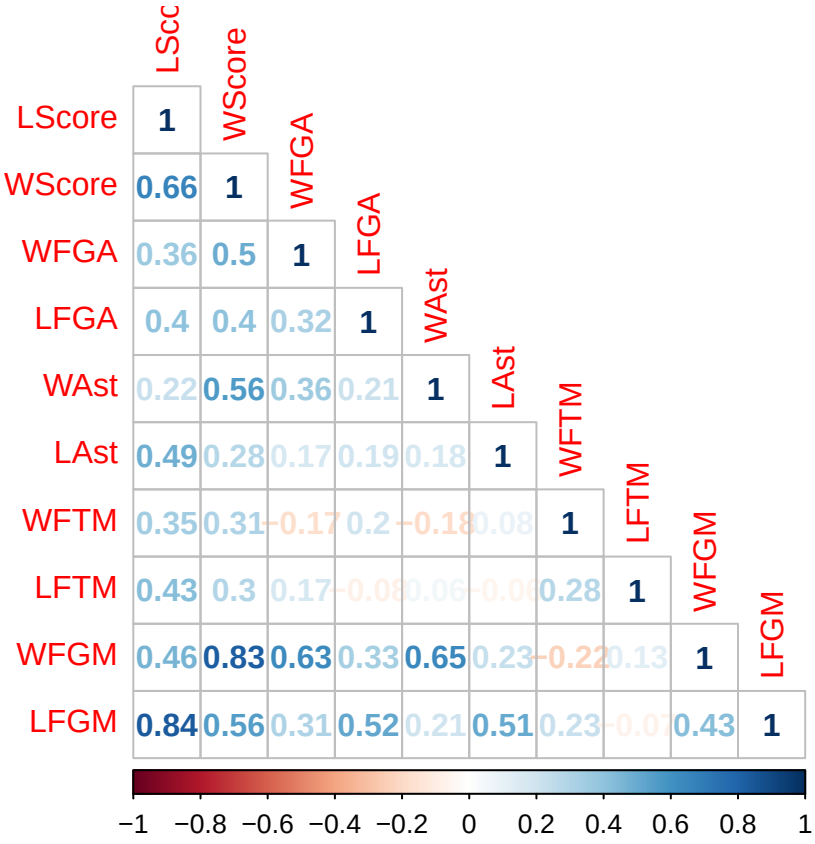

```
# scatter plot of winning scores and conferences
p3 <- ggplot(tourney_detailed_result, aes(x = WScore, y = WTeamConfFactor)) +
  geom_point() +
  labs(x = "Winning Score",
       y = "Winning Team Conference")

p3
```




```
# correlation plot of a variety of game stats
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
cormat <- cor(tourney_detailed_result %>% select(LScore, WScore,
                                                 WFGA, LFGA,
                                                WAst, LAst, WFTM, LFTM,
                                                 WFGM, LFGM) %>% drop_na())
corrplot(cormat, method = "number", type="lower")
```


LScc
WScore
WFGA
LFGA
WAst
LAst
WFTM
LFTM
WFGM
LFGM
LScore 1
WScore 0.66 1
WFGA 0.36 0.5 1
LFGA 0.4 0.4 0.32 1
WAst 0.22 0.56 0.36 0.21 1
LAst 0.49 0.28 0.17 0.19 0.18 1
WFTM 0.35 0.31 −0.17 0.2 −0.18 0.08 1
LFTM 0.43 0.3 0.17 −0.08 0.06 −0.06 0.28 1
WFGM 0.46 0.83 0.63 0.33 0.65 0.23 −0.22 0.13 1
LFGM 0.84 0.56 0.31 0.52 0.21 0.51 0.23 −0.07 0.43 1
−1 −0.8 −0.6 −0.4 −0.2 0 0.2 0.4 0.6 0.8 1