An earlier version of Joe’s code, adapted to take advantage of the functions library script

Anton Hung

2022-11-03

```r
# setwd('/users/joegyorda/Desktop/wranglinghub')
setwd('/Volumes/GoogleDrive/Mon disque/wrangling/project/wranglinghub')
football_data = read.csv('Merged_Stadium.csv')
```

```r
setwd('/Volumes/GoogleDrive/Mon disque/wrangling/project/wranglinghub')
source('functions_library/functions_library.R')
```

```r
# for now, we'll only focus on games where there was a spread

# a good next wrangling task would be to replace old team names with new team names
# e.g., Baltimore Colts --> Indianapolis Colts
# ^^or maybe not, depends on the question
sort(unique(football_data$team_home))
```

```
##  [1] "Arizona Cardinals"        "Atlanta Falcons"
##  [3] "Baltimore Colts"          "Baltimore Ravens"
##  [5] "Boston Patriots"          "Buffalo Bills"
##  [7] "Carolina Panthers"        "Chicago Bears"
##  [9] "Cincinnati Bengals"       "Cleveland Browns"
## [11] "Dallas Cowboys"           "Denver Broncos"
## [13] "Detroit Lions"            "Green Bay Packers"
## [15] "Houston Oilers"           "Houston Texans"
## [17] "Indianapolis Colts"       "Jacksonville Jaguars"
## [19] "Kansas City Chiefs"       "Las Vegas Raiders"
## [21] "Los Angeles Chargers"     "Los Angeles Raiders"
## [23] "Los Angeles Rams"         "Miami Dolphins"
## [25] "Minnesota Vikings"        "New England Patriots"
## [27] "New Orleans Saints"       "New York Giants"
## [29] "New York Jets"            "Oakland Raiders"
## [31] "Philadelphia Eagles"      "Phoenix Cardinals"
## [33] "Pittsburgh Steelers"      "San Diego Chargers"
## [35] "San Francisco 49ers"      "Seattle Seahawks"
## [37] "St. Louis Cardinals"      "St. Louis Rams"
## [39] "Tampa Bay Buccaneers"     "Tennessee Oilers"
## [41] "Tennessee Titans"         "Washington Commanders"
## [43] "Washington Football Team" "Washington Redskins"
```

```r
# remove missing values! just remove all for now
# football_data_filter = football_data[complete.cases(football_data),]
football_data_filter = football_data %>% drop_na(spread_favorite)

sum(is.na(football_data$spread_favorite))
```

```
## [1] 2735
```

```r
# how often is the spread correct (for each team)?
# comment out group_by for overall, otherwise gives each team's breakdown
filter_by_spread('Spread_Correct')
```

```
## # A tibble: 43 x 2
##    team_home          Spread_Correct
##    <chr>                       <dbl>
##  1 Arizona Cardinals            1.30
##  2 Atlanta Falcons              1.44
##  3 Baltimore Colts              5.26
##  4 Baltimore Ravens             4.61
##  5 Buffalo Bills                3.67
##  6 Carolina Panthers            3.12
##  7 Chicago Bears                3.10
##  8 Cincinnati Bengals           3.99
##  9 Cleveland Browns             2.80
## 10 Dallas Cowboys               2.18
## # ... with 33 more rows
```

```r
# how often does favored team outperform spread (for each team)?
# comment out group_by for overall, otherwise gives each team's breakdown
filter_by_spread('Over_Spread')
```

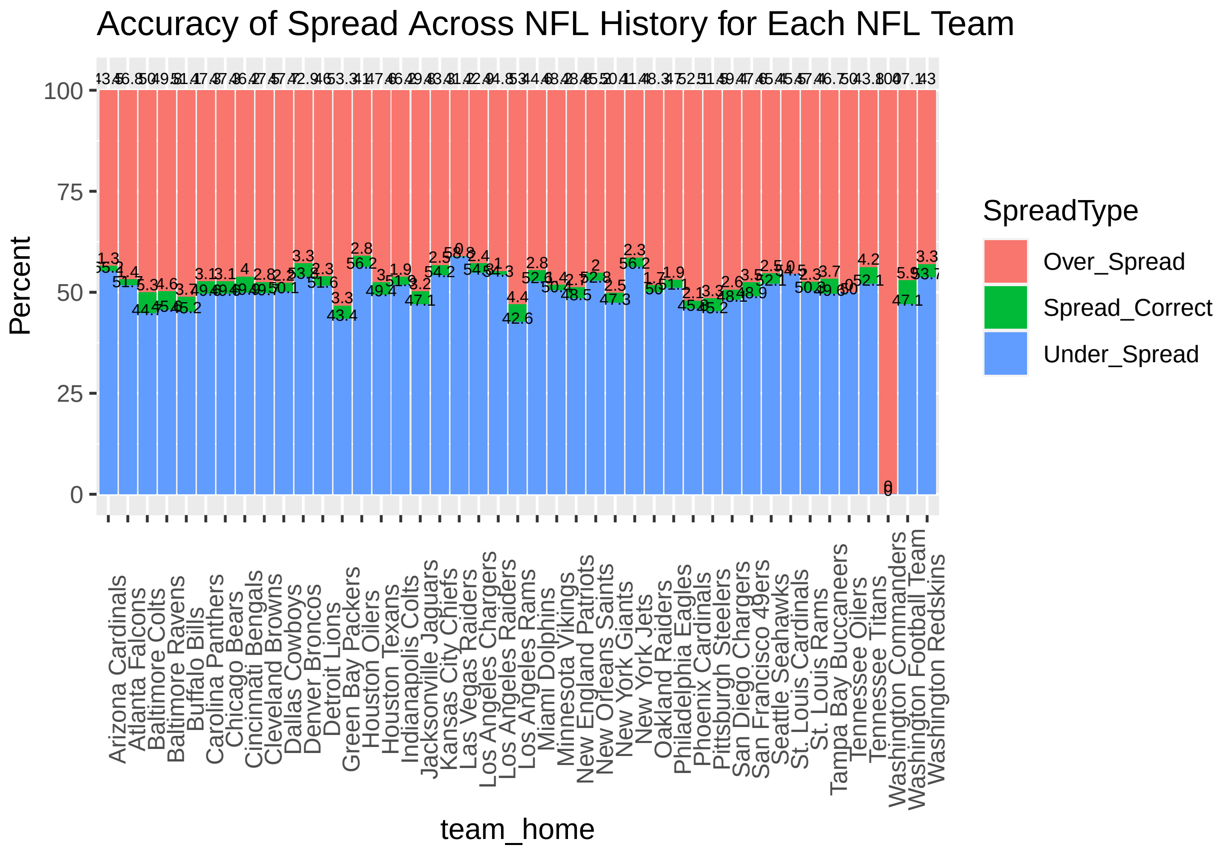
```
## # A tibble: 43 x 2
##    team_home          Spread_Correct
##    <chr>                       <dbl>
##  1 Arizona Cardinals            1.30
##  2 Atlanta Falcons              1.44
##  3 Baltimore Colts              5.26
##  4 Baltimore Ravens             4.61
##  5 Buffalo Bills                3.67
##  6 Carolina Panthers            3.12
##  7 Chicago Bears                3.10
##  8 Cincinnati Bengals           3.99
##  9 Cleveland Browns             2.80
## 10 Dallas Cowboys               2.18
## # ... with 33 more rows
```

```r
# how often does favored team underperform spread (for each team)?
# comment out group_by for overall, otherwise gives each team's breakdown
filter_by_spread('Under_Spread')
```

```
## # A tibble: 43 x 2
##    team_home          Spread_Correct
##    <chr>                       <dbl>
##  1 Arizona Cardinals            1.30
##  2 Atlanta Falcons              1.44
##  3 Baltimore Colts              5.26
##  4 Baltimore Ravens             4.61
##  5 Buffalo Bills                3.67
##  6 Carolina Panthers            3.12
##  7 Chicago Bears                3.10
##  8 Cincinnati Bengals           3.99
##  9 Cleveland Browns             2.80
## 10 Dallas Cowboys               2.18
## # ... with 33 more rows
```

```
# combine all into 1
spread_breakdown <- filter_by_spread_combined(football_data_filter)

# making a plot to visualize the history of spreads
plot_spreadtype(spread_breakdown)
```



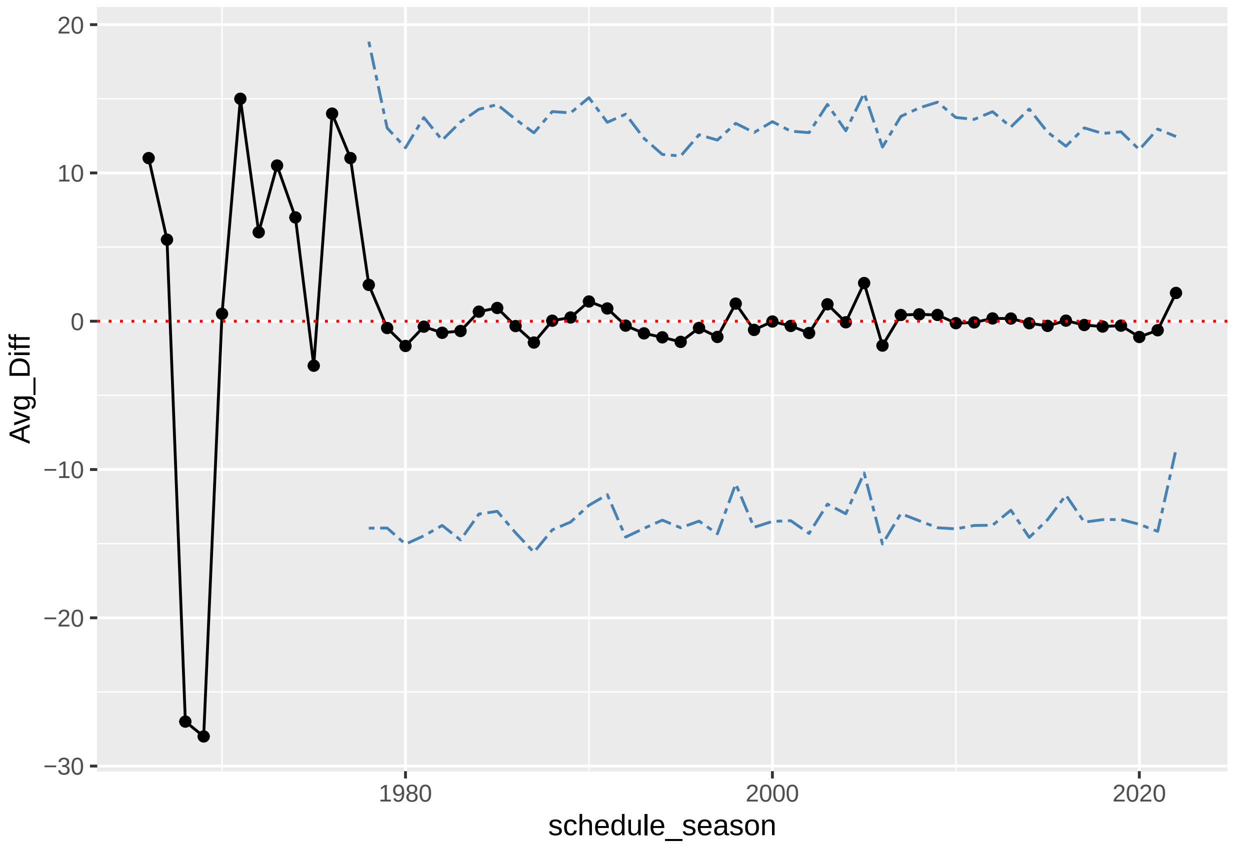
```
spread_score_diff_over_time <- view_spread_accuracy(football_data_filter)

spread_score_diff_over_time
```

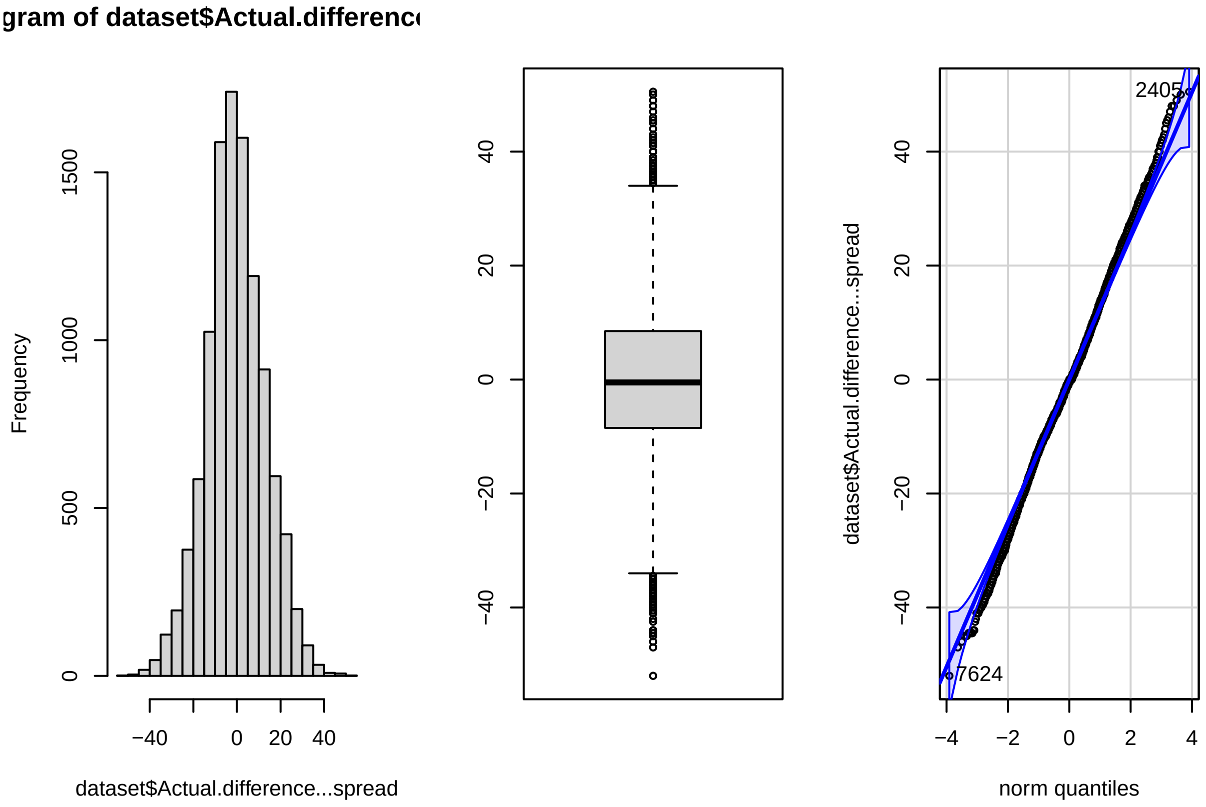
```
## # A tibble: 57 x 4
##    schedule_season Avg_Diff SD_Diff Med_Diff
##              <int>    <dbl>   <dbl>    <dbl>
##  1            1966     11         NA     11
##  2            1967      5.5       NA      5.5
##  3            1968    -27         NA    -27
##  4            1969    -28         NA    -28
##  5            1970      0.5       NA      0.5
##  6            1971     15         NA     15
##  7            1972      6         NA      6
##  8            1973     10.5       NA     10.5
##  9            1974      7         NA      7
## 10            1975     -3         NA     -3
## # ... with 47 more rows
```

```
plot_spread_accuracy(spread_score_diff_over_time)
```

```
## Warning: Removed 12 rows containing missing values (‘geom_line()‘).
## Removed 12 rows containing missing values (‘geom_line()‘).
```



```
assess_normality(football_data_filter)
```



```
## [1] 7624 2405
```

```
# outcome looks normal!
```

```
# how important are weather, location, field type, etc to covering the spread? how do
# these predictors differ by team?

# the outcome variable is actual difference - spread
# this variable takes the difference b/w the real game score difference, and
# the predicted difference (spread)
# positive value means favored team outperformed spread, negative means favored
# team underperformed the spread, and 0 means spread was correct

# makes it easier to generate predictions later
# football_data_complete = football_data_filter[complete.cases(football_data_filter),]
#
# mod1 = lmer(`Actual.difference...spread`~ weather_temperature + weather_wind_mph +
#                weather_humidity + schedule_season + schedule_week + weather_detail + schedule_playoff
#                stadium_type + stadium_weather_type + stadium_surface + Abs.value.of.spread
#              + as.numeric(ELEVATION) +
#                (1|schedule_season) + (schedule_season|team_favorite_id),
#              data=football_data_complete)
#
# sum1 = summary(mod1)
# sum1
```

```r
# r_sq = r.squaredGLMM(mod1)
#
# # sum1$coefficients
# random_effects = ranef(mod1)
#
#
# plot(mod1)
#
# library(sjPlot)
# sjPlot::plot_model(mod1)
# sjPlot::tab_model(mod1)
# # sjPlot::plot_residuals(mod1)
#
# preds = predict(mod1)
#
# plot(football_data_complete$Actual.difference...spread, preds)
# summary(lm(football_data_complete$Actual.difference...spread~preds))
```

```r
cor(football_data_filter[,c(22,2,12,13,16,17,18)],use = "complete.obs")
```

```
##                            Actual.difference...spread schedule_season
## Actual.difference...spread                1.000000000      0.02614983
## schedule_season                           0.026149825      1.00000000
## spread_favorite                          -0.023598485     -0.04627298
## over_under_line                          -0.007113234      0.13971562
## weather_temperature                      -0.016519496      0.03069339
## weather_wind_mph                         -0.015238195     -0.22990594
## weather_humidity                         -0.011469718     -0.08116868
##                            spread_favorite over_under_line weather_temperature
## Actual.difference...spread    -0.023598485    -0.007113234         -0.01651950
## schedule_season               -0.046272981     0.139715619          0.03069339
## spread_favorite                1.000000000    -0.046835986          0.06115315
## over_under_line               -0.046835986     1.000000000          0.08114199
## weather_temperature            0.061153150     0.081141988          1.00000000
## weather_wind_mph              -0.029304444    -0.118451137         -0.18882236
## weather_humidity              -0.002407822    -0.067178990         -0.02173374
##                            weather_wind_mph weather_humidity
## Actual.difference...spread      -0.01523820     -0.011469718
## schedule_season                 -0.22990594     -0.081168681
## spread_favorite                 -0.02930444     -0.002407822
## over_under_line                 -0.11845114     -0.067178990
## weather_temperature             -0.18882236     -0.021733741
## weather_wind_mph                 1.00000000      0.034030578
## weather_humidity                 0.03403058      1.000000000
```