DATA607 Assignment 1

Alice Ding

2023-01-26

Overview

This data set contains one row per match for the 2022 World Cup. The specific data includes the chance that each team will win, lose or tie every one of their matches, and each team's SPI (soccer performance index) as well as a projected score. The table also holds information regarding non-shot expected goals (xG) and then adjusted forecast numbers based on things that happened during the game.

```
wc_matches <- read.csv("wc_matches.csv")
summary(wc_matches)
```

```
##      date              league_id       league             team1
##  Length:64          Min.   :1908   Length:64          Length:64
##  Class :character   1st Qu.:1908   Class :character   Class :character
##  Mode  :character   Median :1908   Mode  :character   Mode  :character
##                     Mean   :1908
##                     3rd Qu.:1908
##                     Max.   :1908
##     team2                spi1            spi2           prob1
##  Length:64          Min.   :48.16   Min.   :48.46   Min.   :0.0363
##  Class :character   1st Qu.:68.75   1st Qu.:66.05   1st Qu.:0.2851
##  Mode  :character   Median :78.72   Median :74.46   Median :0.4460
##                     Mean   :77.32   Mean   :74.30   Mean   :0.4432
##                     3rd Qu.:87.23   3rd Qu.:79.50   3rd Qu.:0.6070
##                     Max.   :93.66   Max.   :93.48   Max.   :0.8261
##      prob2           probtie         proj_score1     proj_score2
##  Min.   :0.0595   Min.   :0.0000   Min.   :0.310   Min.   :0.440
##  1st Qu.:0.2039   1st Qu.:0.1081   1st Qu.:0.985   1st Qu.:0.820
##  Median :0.3121   Median :0.2575   Median :1.315   Median :1.055
##  Mean   :0.3583   Mean   :0.1985   Mean   :1.325   Mean   :1.139
##  3rd Qu.:0.5047   3rd Qu.:0.2912   3rd Qu.:1.620   3rd Qu.:1.367
##  Max.   :0.8112   Max.   :0.3371   Max.   :2.600   Max.   :2.550
##      score1          score2           xg1             xg2
##  Min.   :0.000   Min.   :0.000   Min.   :0.070   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.600   1st Qu.:0.5075
##  Median :1.000   Median :1.000   Median :0.885   Median :0.9400
##  Mean   :1.578   Mean   :1.109   Mean   :1.075   Mean   :1.1089
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:1.430   3rd Qu.:1.4350
##  Max.   :7.000   Max.   :4.000   Max.   :3.100   Max.   :4.4100
##      nsxg1           nsxg2          adj_score1      adj_score2
##  Min.   :0.240   Min.   :0.0900   Min.   :0.000   Min.   :0.000
##  1st Qu.:0.760   1st Qu.:0.6475   1st Qu.:0.000   1st Qu.:0.000
```

```
##  Median :1.185   Median :0.9350   Median :1.050   Median :1.050
##  Mean   :1.194   Mean   :1.1553   Mean   :1.572   Mean   :1.122
##  3rd Qu.:1.433   3rd Qu.:1.4625   3rd Qu.:2.100   3rd Qu.:2.100
##  Max.   :3.100   Max.   :5.9000   Max.   :6.220   Max.   :3.720
```

Subset of Data

I'm curious about South Korea in particular, so I'd like to see their journey throughout the world cup this past year. I'm only curious about what was predicted and the eventual outcome, so details on unexpected goals alongside the adjustment in predictions isn't relevant to me.

```
# import libraries
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(scales)

# choose relevant columns
relevant_data <- subset(
  select(
    wc_matches, c('date', 'team1', 'team2', 'spi1', 'spi2', 'prob1'
                  , 'prob2', 'probtie', 'proj_score1', 'proj_score2'
                  , 'score1', 'score2'))
  , team1=="South Korea" | team2=="South Korea")

# rename spi to something more intuitive for folks who don't know about soccer
colnames(relevant_data)[colnames(relevant_data) == "spi1"] = "soccer_power_index_team1"
colnames(relevant_data)[colnames(relevant_data) == "spi2"] = "soccer_power_index_team2"

# renaming the probabilities for better readability
colnames(relevant_data)[colnames(relevant_data) == "prob1"] = "win_probability_team1"
colnames(relevant_data)[colnames(relevant_data) == "prob2"] = "win_probability_team2"
colnames(relevant_data)[colnames(relevant_data) == "probtie"] = "tie_probability"

# renaming the scores for better readability
colnames(relevant_data)[colnames(relevant_data) == "proj_score1"] = "projected_score_team1"
colnames(relevant_data)[colnames(relevant_data) == "proj_score2"] = "projected_score_team2"
colnames(relevant_data)[colnames(relevant_data) == "score1"] = "actual_score_team1"
colnames(relevant_data)[colnames(relevant_data) == "score2"] = "actual_score_team2"

# format the probability columns to percentages to make them easier to read
relevant_data[6:8] <- sapply(relevant_data[6:8], function(x) percent(x, accuracy=0.01))
relevant_data
```

```
##          date       team1       team2 soccer_power_index_team1
## 14 2022-11-24     Uruguay South Korea                    80.90
## 30 2022-11-28 South Korea       Ghana                    66.44
## 46 2022-12-02 South Korea    Portugal                    66.93
## 54 2022-12-05      Brazil South Korea                    92.90
##    soccer_power_index_team2 win_probability_team1 win_probability_team2
## 14                    66.12                52.56%                19.06%
## 30                    60.03                44.30%                23.48%
## 46                    87.55                16.66%                58.66%
## 54                    69.40                82.61%                17.39%
##    tie_probability projected_score_team1 projected_score_team2
## 14          28.38%                  1.52                   0.80
## 30          32.22%                  1.23                   0.81
## 46          24.68%                  0.86                   1.84
## 54           0.00%                  1.89                   0.58
##    actual_score_team1 actual_score_team2
## 14                  0                  0
## 30                  2                  3
## 46                  2                  1
## 54                  4                  1
```

Conclusion

We can see that South Korea only had a higher SPI than one of its competitors throughout the tournament, however they performed unexpectedly against the 4 other teams that they played. Their first game ended in a tie even though it was predicted for the other team to win, they ended up losing their second game despite having a higher SPI and a higher probability of winning, they ended up winning at only a 16.66% probability for their third match, and they ultimately lost to Brazil as predicted in their last game.

With this only being the 2022 match history, I'd be curious to expand this data set to include other years and also just work on seeing how accurate these predictions are overall. For South Korea in particular, only one of these games ended with the highest probability prediction being correct and that was the last game; is South Korea an outlier here or did all other teams also see unlikely wins/losses/ties?