Datathon 2024

Eleazar Martin

2024-01-20

```
baseball_data <- read.csv("/Users/eleazarmartin/Desktop/datathon2024_data_w_dist.csv")
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

```
##
## 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
```

```
baseball_data <- baseball_data %>%
  mutate(
    home_time_zone_diff = abs(home_last_time_zone - time_zone),
    away_time_zone_diff = abs(away_last_time_zone - time_zone)
  )

head(baseball_data)
```

```
##   index  game_date home_team away_team is_day_game home_score away_score venue
## 1  2429 2001-04-01       TOR       TEX        True          8          1 SJU01
## 2  2430 2001-04-02       BAL       BOS        True          2          1 BAL12
## 3  2431 2001-04-02       CLE       CWS        True          4          7 CLE08
## 4  2432 2001-04-02       NYY        KC        True          7          3 NYC16
## 5  2433 2001-04-02       SEA       OAK       False          5          4 SEA03
## 6  2434 2001-04-02       CHC       MON        True          4          5 CHI11
##                    venue_name      city state away_pa away_1b away_2b away_3b
## 1       Estadio Hiram Bithorn  San Juan    PR      37       5       3       1
## 2 Oriole Park at Camden Yards Baltimore    MD      40       4       0       0
## 3           Progressive Field Cleveland    OH      36       5       3       0
## 4            Yankee Stadium I  New York    NY      36       5       1       0
## 5                Safeco Field   Seattle    WA      40       5       3       0
## 6               Wrigley Field   Chicago    IL      41       8       1       0
##   away_hr away_fo away_so away_bb away_hbp home_pa home_1b home_2b home_3b
## 1       0      16      10       1        1      39       8       3       0
```

```
## 2       1      25       8       2        0      37       4       2       0
## 3       1      14       9       4        0      35       4       0       0
## 4       1      22       5       2        0      39       9       1       0
## 5       0      18       8       6        0      41       7       0       0
## 6       1      23       6       2        0      44       6       3       0
##   home_hr home_fo home_so home_bb home_hbp      lat        lng time_zone year
## 1       2      17       7       2        0 18.46710  -66.11850        -4 2001
## 2       0      19       9       2        1 39.29040  -76.61220        -5 2001
## 3       3      20       6       2        0 41.49930  -81.69440        -5 2001
## 4       3      20       3       3        0 40.71280  -74.00600        -5 2001
## 5       0      16       8      10        0 47.60610 -122.33280        -8 2001
## 6       0      17      12       6        0 41.88183  -87.62318        -6 2001
##   home_win home_last_time_zone away_last_time_zone home_time_between_games
## 1        1                 -5                 -8                      182
## 2        1                 -5                 -5                      183
## 3        0                 -5                 -6                      183
## 4        1                 -5                 -6                      183
## 5        1                 -8                 -8                      183
## 6        0                 -5                 -5                      183
##   away_time_between_games home_time_since_last_series
## 1                     182                        182
## 2                     183                        183
## 3                     183                        183
## 4                     183                        183
## 5                     183                        183
## 6                     183                        183
##   away_time_since_last_series home_distance_travelled_game
## 1                         182                    2957.1915
## 2                         183                       0.0000
## 3                         183                       0.0000
## 4                         183                     272.5524
## 5                         183                    1575.0353
## 6                         183                     658.0650
##   away_distance_travelled_game home_distance_travelled_series
## 1                    5815.0941                       2957.1915
## 2                    1396.3365                          0.0000
## 3                     494.0283                          0.0000
## 4                    1143.7591                        272.5524
## 5                    1089.9123                       1575.0353
## 6                    1143.7591                        658.0650
##   away_distance_travelled_series home_time_zone_diff away_time_zone_diff
## 1                      5815.0941                   1                   4
## 2                      1396.3365                   0                   0
## 3                       494.0283                   0                   1
## 4                      1143.7591                   0                   1
## 5                      1089.9123                   0                   0
## 6                      1143.7591                   1                   1
```

```
# STAT SIGNIF > 50% WIN RATE FOR SMALLER TIME ZONE DIFF

library(dplyr)

proportion_by_team_and_year <- baseball_data %>%
  filter(home_team != 'MON') %>%
```

```
  mutate(
    year = as.numeric(format(as.Date(game_date), "%Y")),
    home_time_zone_diff = abs(home_last_time_zone - time_zone),
    away_time_zone_diff = abs(away_last_time_zone - time_zone),
    condition_met = (home_score > away_score & home_time_zone_diff < away_time_zone_diff) |
                    (home_score < away_score & home_time_zone_diff > away_time_zone_diff),
    total_condition = (home_score != away_score & home_time_zone_diff != away_time_zone_diff)
  ) %>%
  group_by(home_team, year) %>%
  summarise(
    success_games = sum(condition_met, na.rm = TRUE),
    total_games = sum(total_condition, na.rm = TRUE),
    proportion = success_games / total_games
  )
```

```
## `summarise()` has grouped output by 'home_team'. You can override using the
## `.groups` argument.
```

```
# Aggregating success and total counts across all teams and years
total_successes <- sum(proportion_by_team_and_year$success_games, na.rm = TRUE)
total_games <- sum(proportion_by_team_and_year$total_games, na.rm = TRUE)

# Performing the one-sample proportion test
test_result <- prop.test(total_successes, total_games, p = 0.50, alternative = "greater")

# Output the p-value
print(test_result$p.value)
```

```
## [1] 4.918904e-14
```

```
# PLOT SHOWING 'TIMEZONE ADVANTAGE'


library(ggplot2)
library(dplyr)

proportion_by_team_and_year_filtered <- na.omit(proportion_by_team_and_year)

# Calculate yearly mean proportions
yearly_means <- proportion_by_team_and_year_filtered %>%
  group_by(year) %>%
  summarise(mean_proportion = mean(proportion, na.rm = TRUE))

# Calculate the yearly home team win proportion
yearly_home_win_proportion <- baseball_data %>%
  mutate(year = as.numeric(format(as.Date(game_date), "%Y"))) %>%
  group_by(year) %>%
  summarise(home_win_proportion = mean(home_win == 1, na.rm = TRUE))

# Create the bar plot
proportion_plot <- ggplot(proportion_by_team_and_year_filtered, aes(x = factor(year), y = proportion)) +
  geom_bar(stat = "identity", position = position_dodge(), aes(fill = home_team)) +
```

```
  geom_hline(yintercept = 0.5, linetype = "dashed", color = "black") +
  scale_y_continuous(breaks = seq(0, 1, 0.1), limits = c(0, 0.9)) +
  theme_minimal() +
  labs(x = "Year", y = "Win rate w/ Smaller Time zone Difference", fill = "Team") +
  ggtitle("Win Rate w/ Smaller Time zone Difference by Team and Year") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1), axis.title.y = element_text(size = 14)) +
  geom_line(data = yearly_means, aes(x = factor(year), y = mean_proportion, group = 1, color = "'Timezo
  geom_line(data = yearly_home_win_proportion, aes(x = factor(year), y = home_win_proportion, group = 1
```

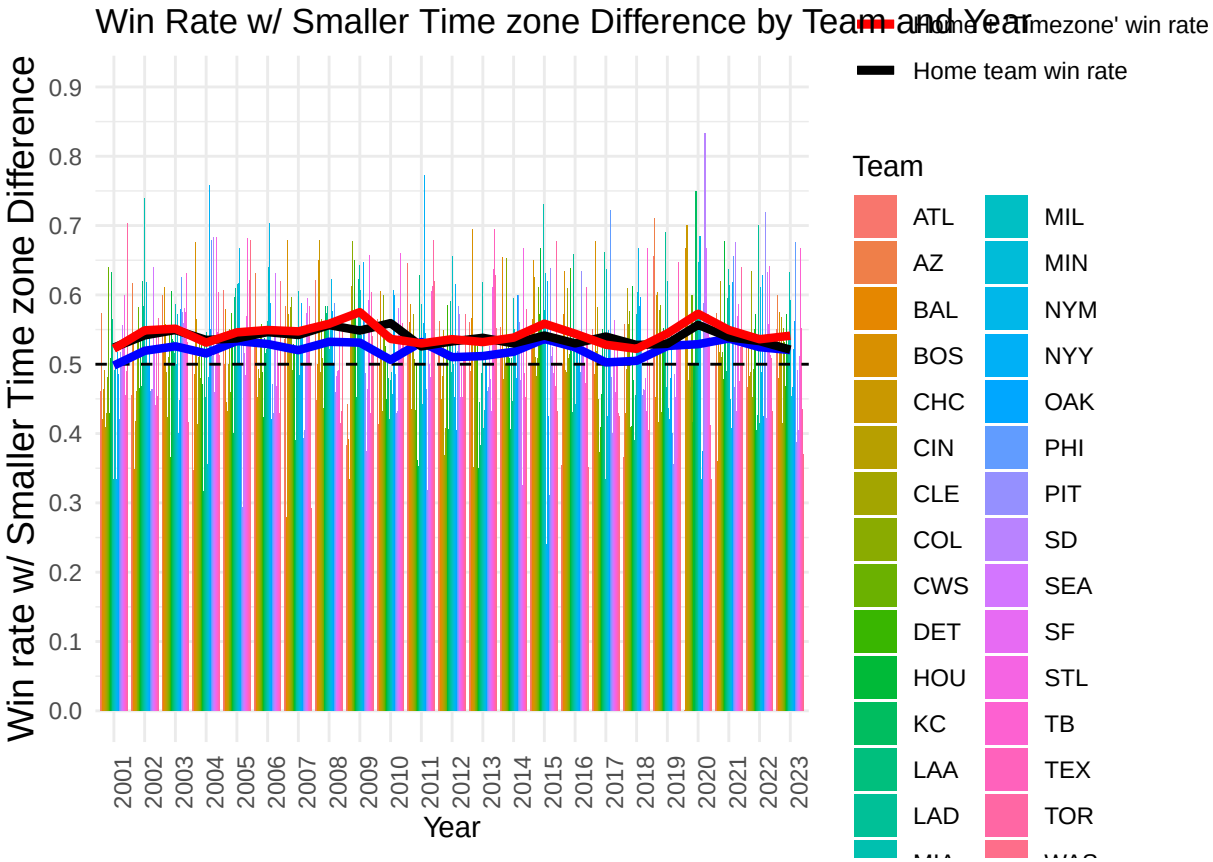
```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```

```
baseball_data <- na.omit(baseball_data)
# Calculate the specific yearly proportion
yearly_specific_proportion <- baseball_data %>%
  mutate(
    year = as.numeric(format(as.Date(game_date), "%Y")),
    specific_condition = (home_win == 1 & home_time_zone_diff < away_time_zone_diff)
  ) %>%
  group_by(year) %>%
  summarise(specific_proportion = sum(specific_condition, na.rm = TRUE)/
              sum(home_time_zone_diff < away_time_zone_diff))


proportion_plot <- proportion_plot +
  geom_line(data = yearly_specific_proportion, aes(x = factor(year), y = specific_proportion, group = 1

proportion_plot <- proportion_plot +
  scale_color_manual(values = c("Home team win rate" = "black",
                                "'Timezone' team win rate" = "blue",
                                "Home + 'Timezone' win rate" = "red"),
                     name = "",
                     guide = guide_legend(title.position = "top", title.hjust = 0.5))

print(proportion_plot)
```


Win Rate w/ Smaller Time zone Difference by Team and Year
Win rate w/ Smaller Time zone Difference
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
Year
'Home team timezone' win rate
Home team win rate
Team
ATL
AZ
BAL
BOS
CHC
CIN
CLE
COL
CWS
DET
HOU
KC
LAA
LAD
MIL
MIN
NYM
NYY
OAK
PHI
PIT
SD
SEA
SF
STL
TB
TEX
TOR