

Geometric mean for results

```
library(dplyr)
library(ggplot2)
library(tidyr)
library(knitr)

load("data/HDA.RData")
load("~/GitHub/soccer-live-predictions/soccer-live-predictions/scrape/data/reds.RData")

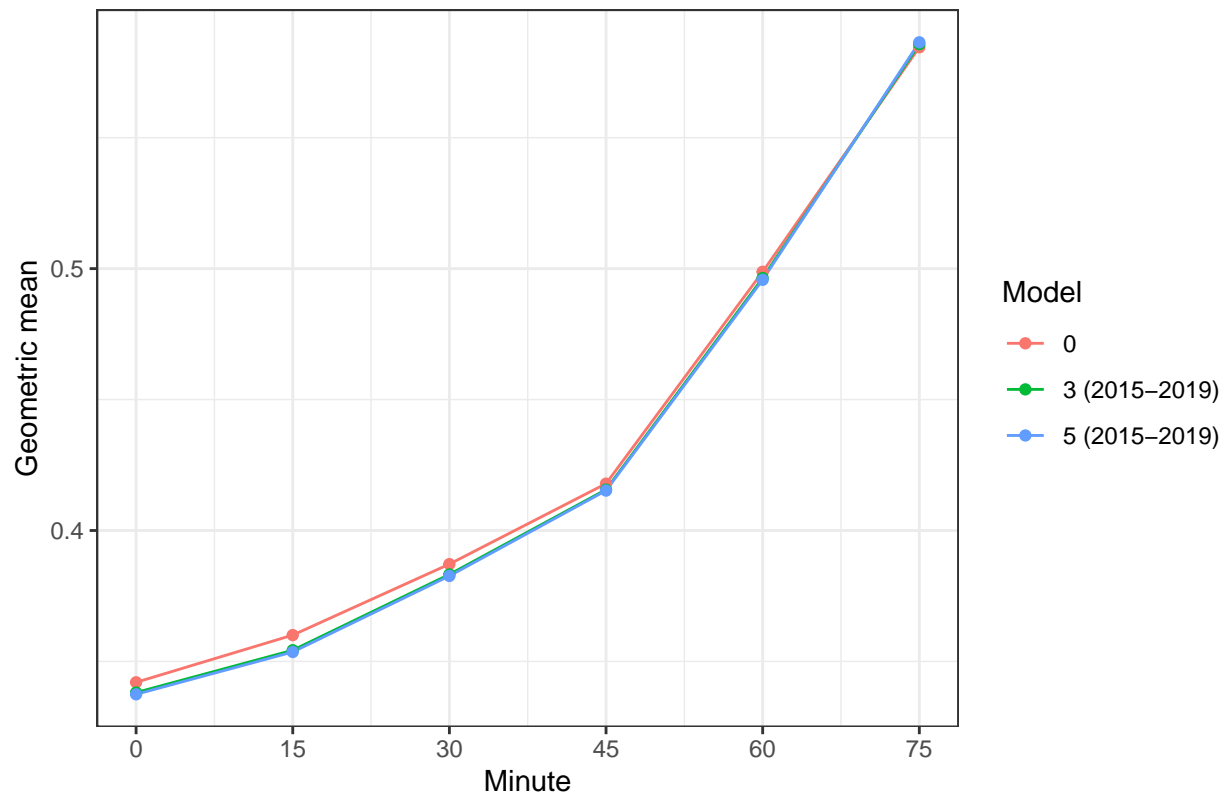
nrow(HDA)
```

```
## [1] 333
```

```
all = tibble(GeoMean = apply(HDA[,c(159:164, 177:188)], 2,
                             EnvStats::geoMean),
             Minute = as.integer(rep(c(0, 15, 30, 45, 60, 75), 3)),
             Model = factor(c(rep("0", 6), rep("3 (2015-2019)", 6),
                               rep("5 (2015-2019)", 6)),
                             levels = c("0", "3 (2015-2019)", "5 (2015-2019)")))

all %>%
  ggplot(aes(x = Minute, y = GeoMean, col = Model)) +
  geom_line() +
  geom_point() +
  scale_x_continuous(breaks = c(0, 15, 30, 45, 60, 75)) +
  theme_bw() +
  ggtitle("All predicted matches") +
  ylab("Geometric mean")
```

All predicted matches



```
all %>%
  pivot_wider(id_cols = "Model", values_from = "GeoMean", names_from = "Minute",
              names_prefix = "Minute ") %>%
  kable()
```

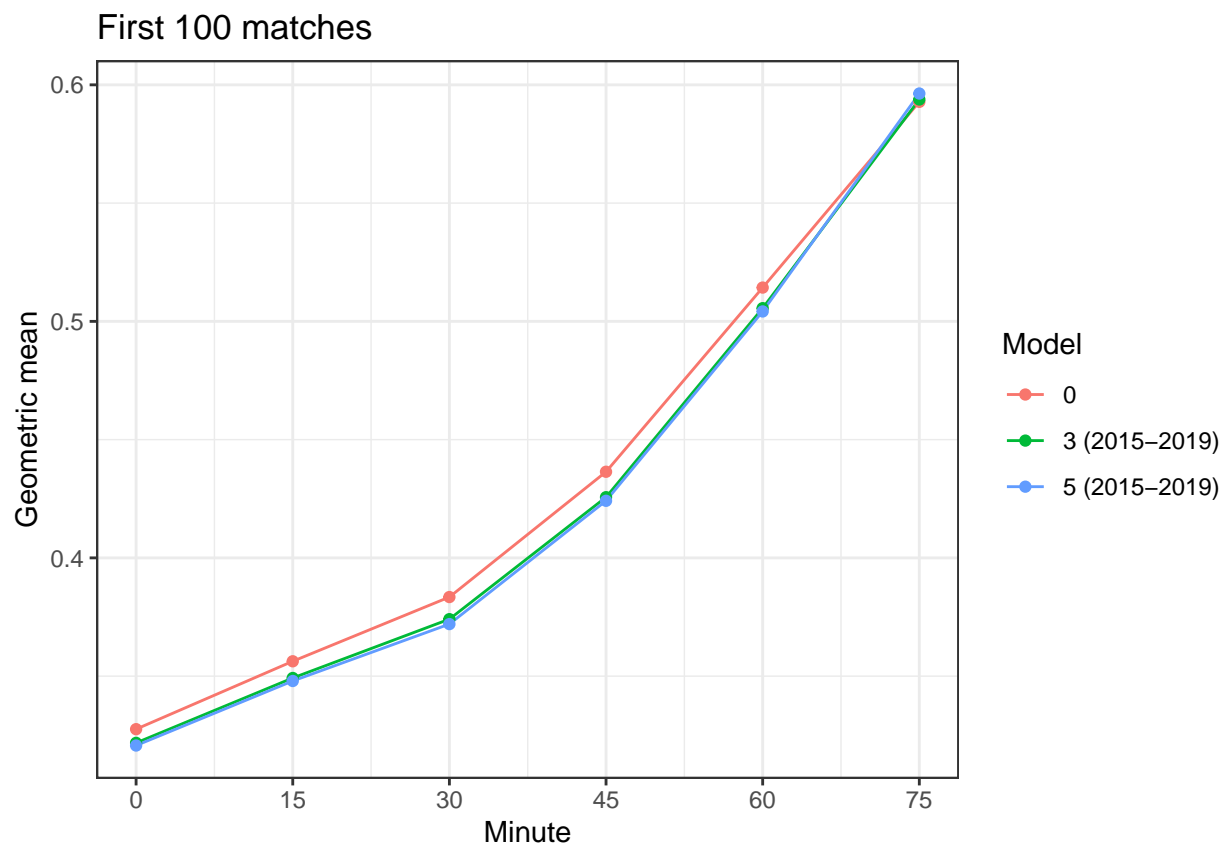
Model	Minute 0	Minute 15	Minute 30	Minute 45	Minute 60	Minute 75
0	0.3420235	0.3600504	0.3871622	0.4178882	0.4988071	0.5845338
3 (2015-2019)	0.3381131	0.3542723	0.3832209	0.4156023	0.4964127	0.5857147
5 (2015-2019)	0.3374428	0.3536051	0.3826444	0.4152420	0.4957400	0.5864195

```

first_100 = tibble(GeoMean = apply(HDA[c(1:100)], c(159:164, 177:188)], 2,
                             EnvStats::geoMean),
                  Minute = as.integer(rep(c(0, 15, 30, 45, 60, 75), 3)),
                  Model = factor(c(rep("0", 6), rep("3 (2015-2019)", 6),
                                rep("5 (2015-2019)", 6)),
                              levels = c("0", "3 (2015-2019)", "5 (2015-2019)")))

first_100 %>%
  ggplot(aes(x = Minute, y = GeoMean, col = Model)) +
  geom_line() +
  geom_point() +
  scale_x_continuous(breaks = c(0, 15, 30, 45, 60, 75)) +
  theme_bw() +
  ggtitle("First 100 matches") +
  ylab("Geometric mean")

```



```

first_100 %>%
  pivot_wider(id_cols = "Model", values_from = "GeoMean", names_from = "Minute",
              names_prefix = "Minute ") %>%
  kable()

```

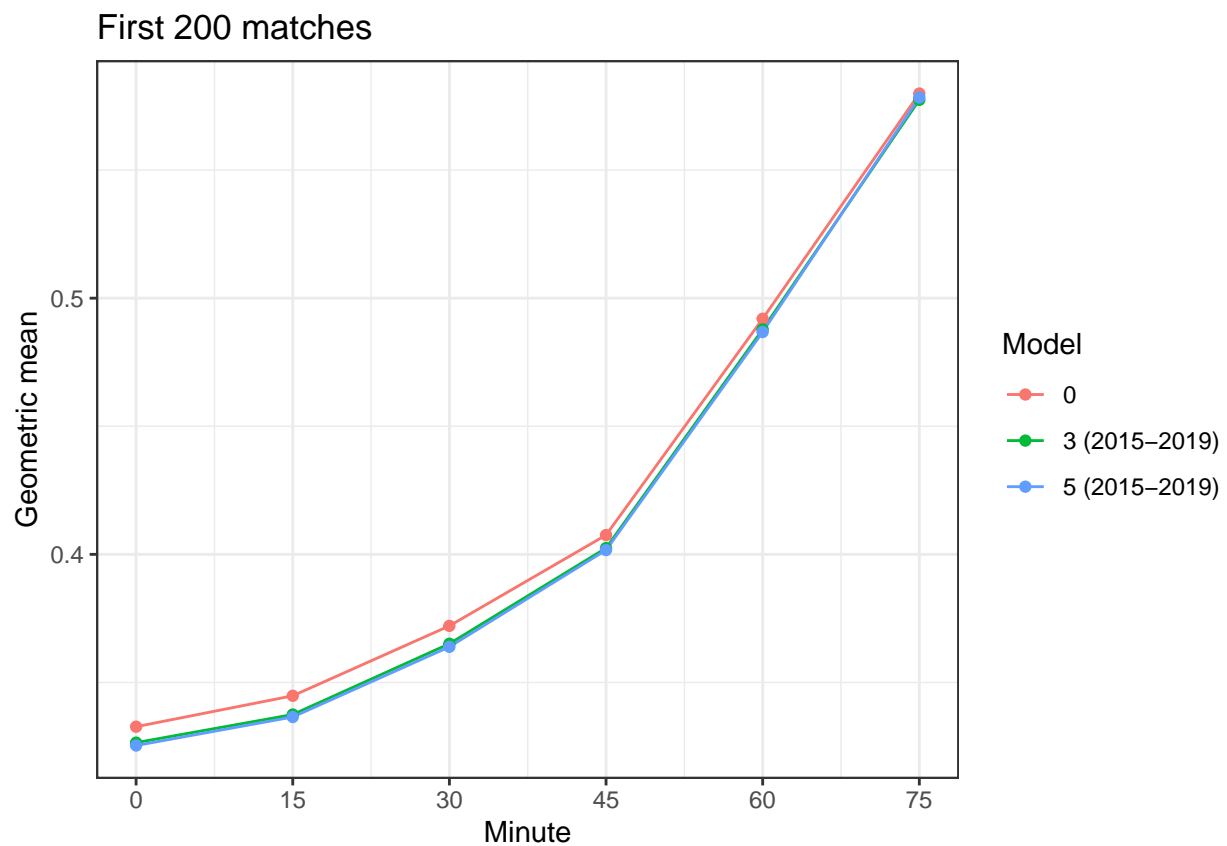
Model	Minute 0	Minute 15	Minute 30	Minute 45	Minute 60	Minute 75
0	0.3275872	0.3562892	0.3834474	0.4364133	0.5142821	0.5928162
3 (2015-2019)	0.3217835	0.3492214	0.3740952	0.4256627	0.5055736	0.5938135
5 (2015-2019)	0.3206727	0.3479583	0.3720119	0.4241408	0.5041992	0.5963165

```

first_200 = tibble(GeoMean = apply(HDA[c(1:200)], c(159:164, 177:188)], 2,
                             EnvStats::geoMean),
                  Minute = as.integer(rep(c(0, 15, 30, 45, 60, 75), 3)),
                  Model = factor(c(rep("0", 6), rep("3 (2015-2019)", 6),
                                rep("5 (2015-2019)", 6)),
                                levels = c("0", "3 (2015-2019)", "5 (2015-2019)")))

first_200 %>%
  ggplot(aes(x = Minute, y = GeoMean, col = Model)) +
  geom_line() +
  geom_point() +
  scale_x_continuous(breaks = c(0, 15, 30, 45, 60, 75)) +
  theme_bw() +
  ggtitle("First 200 matches") +
  ylab("Geometric mean")

```



```

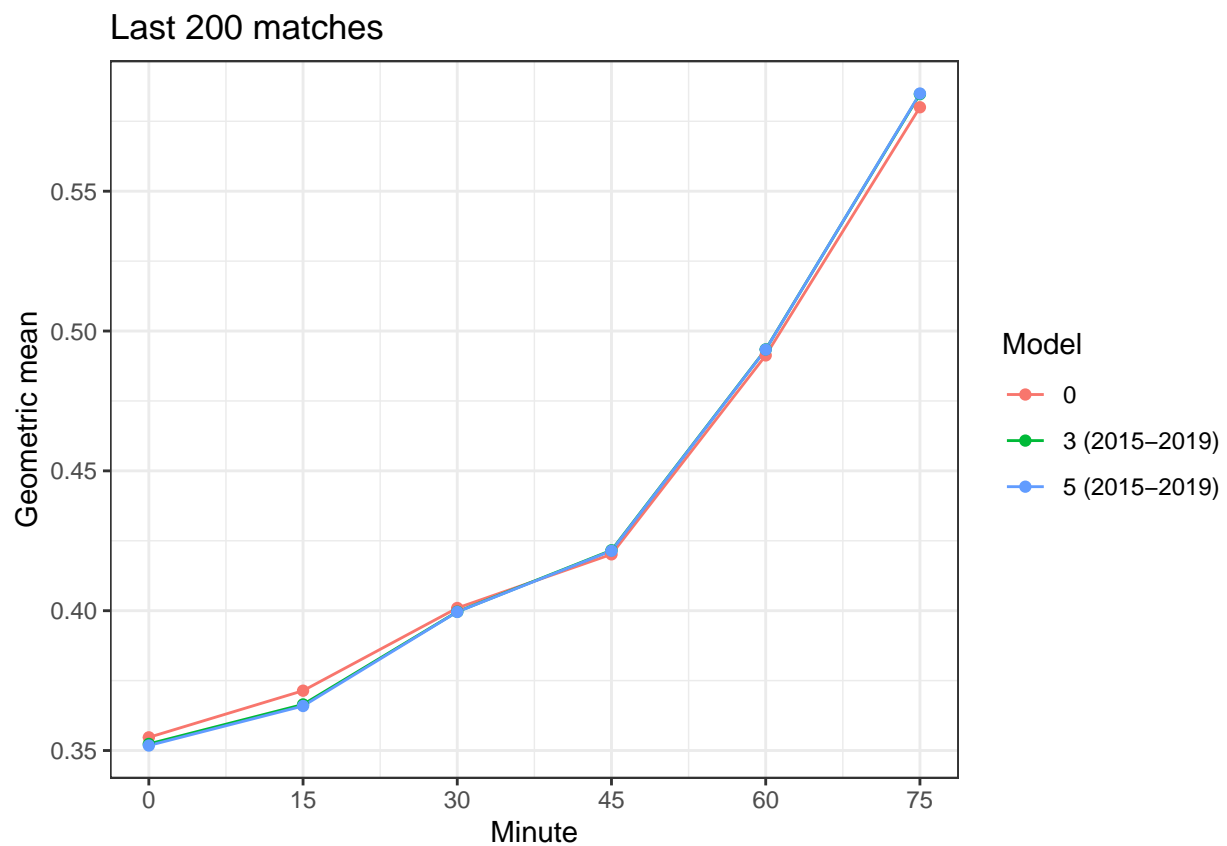
first_200 %>%
  pivot_wider(id_cols = "Model", values_from = "GeoMean", names_from = "Minute",
              names_prefix = "Minute ") %>%
  kable()

```

Model	Minute 0	Minute 15	Minute 30	Minute 45	Minute 60	Minute 75
0	0.3327333	0.3447984	0.3720881	0.4075370	0.4919810	0.5798640
3 (2015-2019)	0.3265095	0.3374801	0.3651082	0.4024520	0.4877935	0.5774073
5 (2015-2019)	0.3253885	0.3365300	0.3639059	0.4016872	0.4868064	0.5783199

```
last_200 = tibble(GeoMean = apply(HDA[c(134:333), c(159:164, 177:188)], 2,
                                EnvStats::geoMean),
                  Minute = as.integer(rep(c(0, 15, 30, 45, 60, 75), 3)),
                  Model = factor(c(rep("0", 6), rep("3 (2015-2019)", 6),
                                rep("5 (2015-2019)", 6)),
                                levels = c("0", "3 (2015-2019)", "5 (2015-2019)")))

last_200 %>%
  ggplot(aes(x = Minute, y = GeoMean, col = Model)) +
  geom_line() +
  geom_point() +
  scale_x_continuous(breaks = c(0, 15, 30, 45, 60, 75)) +
  theme_bw() +
  ggtitle("Last 200 matches") +
  ylab("Geometric mean")
```

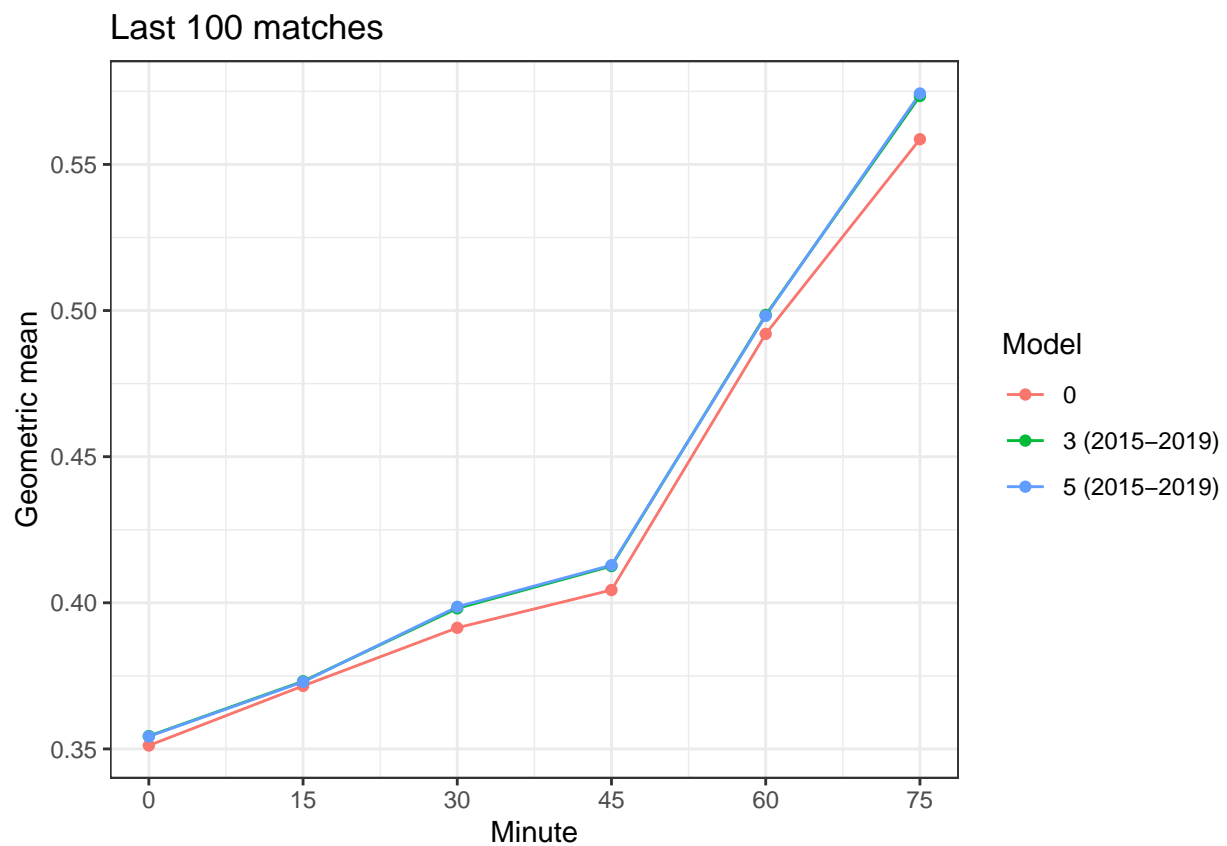


```
last_200 %>%
  pivot_wider(id_cols = "Model", values_from = "GeoMean", names_from = "Minute",
              names_prefix = "Minute ") %>%
  kable()
```

Model	Minute 0	Minute 15	Minute 30	Minute 45	Minute 60	Minute 75
0	0.3547162	0.3713931	0.4009642	0.4201523	0.491283	0.5800195
3 (2015-2019)	0.3523572	0.3664708	0.3996298	0.4216039	0.493442	0.5847180
5 (2015-2019)	0.3517904	0.3659172	0.3995628	0.4214364	0.493374	0.5849180

```
last_100 = tibble(GeoMean = apply(HDA[c(234:333), c(159:164, 177:188)], 2,
                                EnvStats::geoMean),
                  Minute = as.integer(rep(c(0, 15, 30, 45, 60, 75), 3)),
                  Model = factor(c(rep("0", 6), rep("3 (2015-2019)", 6),
                                rep("5 (2015-2019)", 6)),
                                levels = c("0", "3 (2015-2019)", "5 (2015-2019)")))

last_100 %>%
  ggplot(aes(x = Minute, y = GeoMean, col = Model)) +
  geom_line() +
  geom_point() +
  scale_x_continuous(breaks = c(0, 15, 30, 45, 60, 75)) +
  theme_bw() +
  ggtitle("Last 100 matches") +
  ylab("Geometric mean")
```



```
last_100 %>%
  pivot_wider(id_cols = "Model", values_from = "GeoMean", names_from = "Minute",
              names_prefix = "Minute ") %>%
  kable()
```

Model	Minute 0	Minute 15	Minute 30	Minute 45	Minute 60	Minute 75
0	0.3511613	0.3715584	0.3914419	0.4043702	0.4920182	0.5586297
3 (2015-2019)	0.3544034	0.3731460	0.3980953	0.4125607	0.4984910	0.5734262
5 (2015-2019)	0.3542642	0.3729366	0.3986158	0.4128666	0.4981790	0.5742555

```

matches = reds %>%
  filter(Season == 2020, Half == 1) %>%
  .$Match
length(matches)

```

```
## [1] 23
```

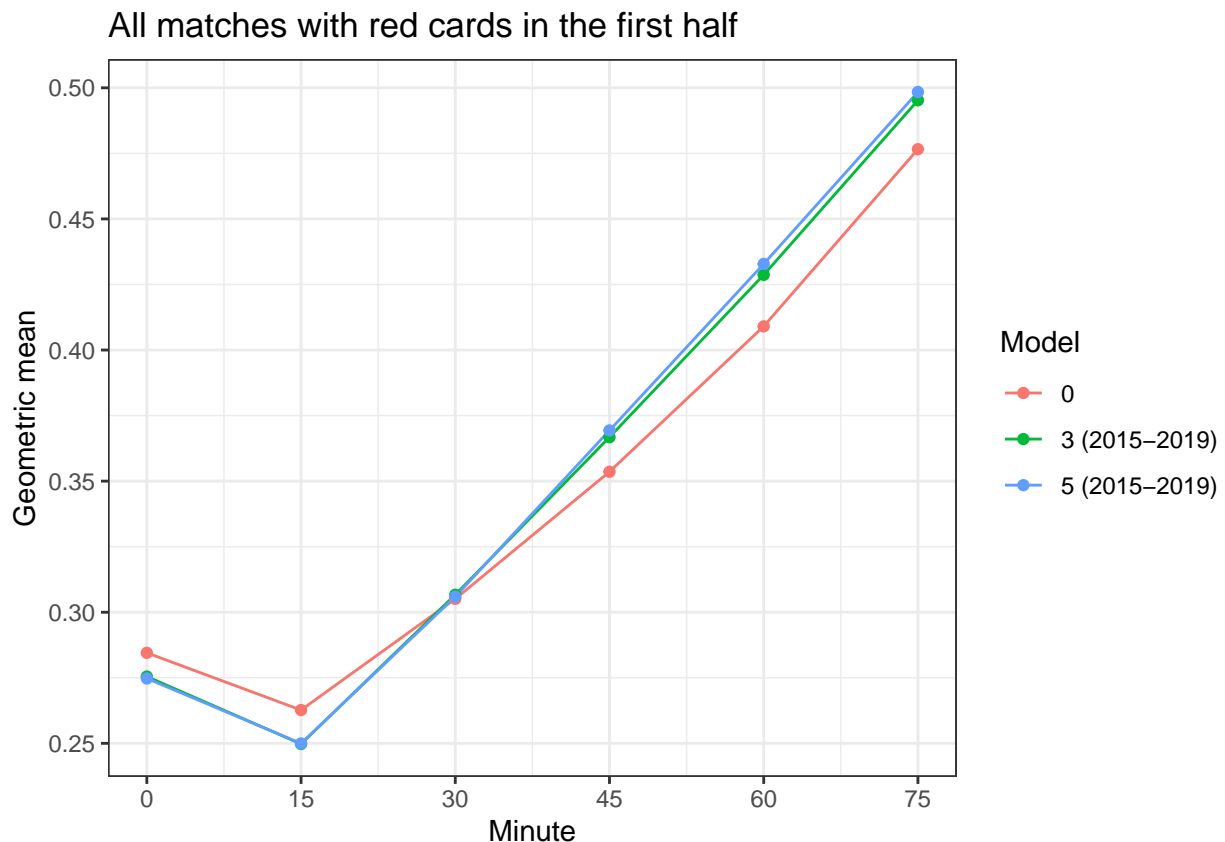
```

HDA_reds = HDA %>%
  filter(Match %in% matches)

all_reds = tibble(GeoMean = apply(HDA_reds[,c(159:164, 177:188)], 2,
  EnvStats::geoMean),
  Minute = as.integer(rep(c(0, 15, 30, 45, 60, 75), 3)),
  Model = factor(c(rep("0", 6), rep("3 (2015-2019)", 6),
    rep("5 (2015-2019)", 6)),
    levels = c("0", "3 (2015-2019)", "5 (2015-2019)")))

all_reds %>%
  ggplot(aes(x = Minute, y = GeoMean, col = Model)) +
  geom_line() +
  geom_point() +
  scale_x_continuous(breaks = c(0, 15, 30, 45, 60, 75)) +
  theme_bw() +
  ggtitle("All matches with red cards in the first half") +
  ylab("Geometric mean")

```



```
all_recs %>%
  pivot_wider(id_cols = "Model", values_from = "GeoMean", names_from = "Minute",
              names_prefix = "Minute ") %>%
  kable()
```

Model	Minute 0	Minute 15	Minute 30	Minute 45	Minute 60	Minute 75
0	0.2845144	0.2626938	0.3050977	0.3535781	0.4090376	0.4765909
3 (2015-2019)	0.2754792	0.2497547	0.3066738	0.3667610	0.4286659	0.4952394
5 (2015-2019)	0.2747454	0.2499620	0.3059432	0.3693354	0.4328354	0.4983925