White-Sox-Question.R

kenji

2023-12-01

```
install.packages("readxl")
```

```
## Error in install.packages : Updating loaded packages
```

```
library(readxl)
library(tidyverse)
library(dplyr)

pitch_data <- read_csv("/Users/kenji/Desktop/2024 PD Internship Questionnaire Data/pitch_data.csv")
```

```
## Rows: 10429 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): p_throws, pitch_name
## dbl (11): pitcher_id, release_speed, release_pos_x, release_pos_z, release_extension, pl...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#Q.1 How many pitches do not have a recorded spin axis?
missing_spin <- sum(is.na(pitch_data$spin_axis))

missing_spin
```

```
## [1] 74
```

```
#Q.2 What is the ID of the pitcher who threw the highest percentage of fastballs
#(4-seam fastball + sinker) with a minimum of 30 total pitches? What is that percentage?

#filter pitchers with atleast 30 pitches
P_30 <- pitch_data %>%
  group_by(pitcher_id) %>%
  filter(n() >= 30)

# Calculating percentage of fastballs for each pitcher
fastballs <- P_30 %>%
  filter(pitch_name %in% c("4-Seam Fastball", "Sinker")) %>%
  group_by(pitcher_id) %>%
  summarize(percentage = sum(!is.na(pitch_name)) / n() * 100) %>%
```

```
  arrange(desc(percentage))

# Pitcher ID with the highest percentage of fastballs
top_fastball_pitcher <- fastballs$pitcher_id[1]
percentage <- fastballs$percentage[1]

top_fastball_pitcher
```

```
## [1] 10348
```

```
percentage
```

```
## [1] 100
```

```
#Q.3 What is the ID of the pitcher who on average has the furthest break from pitching
#hand to glove side on a slider or sweeper? Which pitch type is it?
#What is the average pitching hand to glove side break on that pitcher's pitch?
#Please group sliders and sweepers separately for pitchers who throw both.
# Filtering sliders or sweepers
sliders_sweepers <- pitch_data %>%
  filter(pitch_name %in% c("Slider", "Sweeper"))


# Calculating average break for each pitcher
average_break <- sliders_sweepers %>%
  group_by(pitcher_id, pitch_name) %>%
  summarize(avg_break = mean(abs(pfx_x), na.rm = TRUE)) %>%
  filter(!is.na(avg_break)) %>%
  group_by(pitcher_id) %>%
  summarize(max_avg_break = max(avg_break))
```

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

```
# Pitcher ID with the furthest break
max_break_pitcher <- average_break$pitcher_id[which.max(average_break$max_avg_break)]
max_break_value <- max(average_break$max_avg_break)

max_break_pitcher
```

```
## [1] 22327
```

```
max_break_value
```

```
## [1] 1.8175
```