Andres Izquierdo Data wrangling Extract, transform, and load your data

Andres Izquierdo

2/13/2022

Bringing in Data

```
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(ggplot2)
library(tsibble)
```

```
##
## Attaching package: 'tsibble'

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

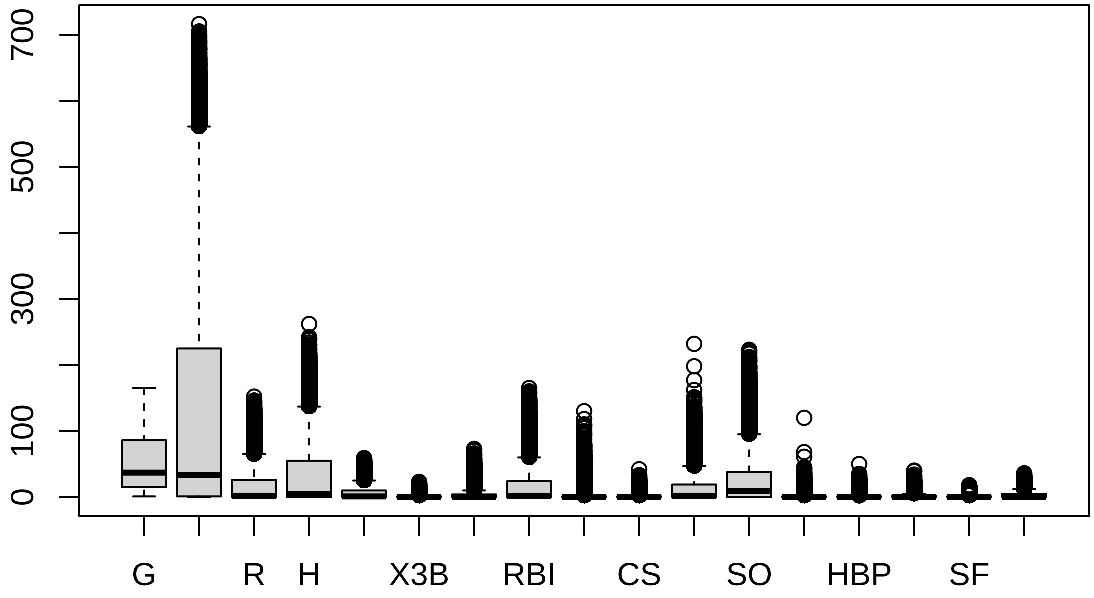
```
load("~/UVA SYS ME/SYS 5581 Time Series/Lahman-master/data/Batting.RData")
Batting.clean <- na.omit(Batting)
is_duplicated(Batting.clean, key = playerID, index = yearID)
```

```
## [1] TRUE
```

```
Batting.merge <- group_by(Batting.clean, playerID, yearID) %>% mutate(G = sum(G), AB = sum(AB), R = sum
Batting.merge <- Batting.merge [!duplicated(Batting.merge[c(1,2)]),] # Getting rid of duplicates
Batting_tsbl <- as_tsibble(Batting.merge, key = playerID, index = yearID)
Batting_tsbl
```

```
## # A tsibble: 66,626 x 22 [1Y]
## # Key:       playerID [12,285]
## # Groups:    playerID @ yearID [66,626]
##    playerID  yearID stint teamID lgID      G    AB     R     H   X2B   X3B    HR
##    <chr>      <int> <int> <fct>  <fct> <int> <int> <int> <int> <int> <int> <int>
##  1 aardsda01   2004     1 SFN    NL       11     0     0     0     0     0     0
##  2 aardsda01   2006     1 CHN    NL       45     2     0     0     0     0     0
##  3 aardsda01   2007     1 CHA    AL       25     0     0     0     0     0     0
##  4 aardsda01   2008     1 BOS    AL       47     1     0     0     0     0     0
##  5 aardsda01   2009     1 SEA    AL       73     0     0     0     0     0     0
##  6 aardsda01   2010     1 SEA    AL       53     0     0     0     0     0     0
##  7 aardsda01   2012     1 NYA    AL        1     0     0     0     0     0     0
##  8 aardsda01   2013     1 NYN    NL       43     0     0     0     0     0     0
##  9 aardsda01   2015     1 ATL    NL       33     1     0     0     0     0     0
## 10 aaronha01   1955     1 ML1    NL      153   602   105   189    37     9    27
## # ... with 66,616 more rows, and 10 more variables: RBI <int>, SB <int>,
## #   CS <int>, BB <int>, SO <int>, IBB <int>, HBP <dbl>, SH <int>, SF <int>,
## #   GIDP <int>
```

```
Batting.num <- Batting.merge[,c(6:22)]
boxplot(Batting.num)
```


G
R
H
X3B
RBI
CS
SO
HBP
SF