PRACTICE_baseball_data

Andrew Crena

2023-05-28

R Markdown

I prefer to set wd's through the user interface by either using the upper left hand quadrant, or the bottom right.

```
## [1] "C:/Users/adcre/OneDrive/Documents/Desktop_RStudio"
```

Baseball Data

Example of me importing a baseball data set, cleaning anything that needs it, and performing some exploratory analysis. Hopefully we will be able to use visualization, but my priority with this file is just to practice importing!

Import from loacal files, still zipped

```r
unzip_baseball <- unzip("C:\\Users\\adcre\\OneDrive\\Documents\\Desktop_RStudio\\baseball_data_zipped.zi
baseball_data <- read.csv(unzip_baseball)

View(baseball_data)
```

Import from URL

```r
library(utils)

# assign a variable to your url for future ref
url <- "https://www.kaggle.com/datasets/mathchi/hitters-baseball-data/download?datasetVersionNumber=1"
download.file(url, destfile = "dataset.zip")
unzpd_baseball <- unzip("dataset.zip")
baseball_data_url <- read.csv("C:\\Users\\adcre\\OneDrive\\Documents\\Desktop_RStudio\\Hitters.csv")

View(baseball_data_url)
```

Clean/Subset

```
baseball_data <- read.csv("Hitters.csv")
Column_indexes <- colnames(baseball_data)
Column_indexes
```

Let's see how we can subset this data. But first, take a look at your variables !

```
##  [1] "AtBat"     "Hits"      "HmRun"     "Runs"      "RBI"       "Walks"
##  [7] "Years"     "CAtBat"    "CHits"     "CHmRun"    "CRuns"     "CRBI"
## [13] "CWalks"    "League"    "Division"  "PutOuts"   "Assists"   "Errors"
## [19] "Salary"    "NewLeague"
```

```
# It looks like we have to create a subject column, here is one way of doing this
baseball_data$subject_id <- row.names(baseball_data)
View(baseball_data)

# Let's subset players with over 200 at-bats
manyabs_data <- baseball_data[baseball_data$AtBat > 200, ]
View(manyabs_data)

# After looking through the data, there appears to be some missing values, especially in certain column
noNA_baseball <- subset(baseball_data, !is.na(baseball_data$Salary))
View(noNA_baseball)

# Let's combine these two conditions for our next subset, just to show how specialized you can make you
new_baseball <- baseball_data[baseball_data$AtBat>250 & baseball_data$HmRun<5, ]
View(new_baseball)
```