Opening a Dataset

Fan Wang

2020-05-02

Contents

1 Opening a Dataset

> Go to the **RMD**, **R**, **PDF**, or **HTML** version of this file. Go back to fan's REconTools Package, R Code Examples Repository (bookdown site), or Intro Stats with R Repository (bookdown site).

We have a dataset on basketball teams. The dataset, *Basketball.csv*, can be downloaded here.

We will load in the dataset and do some analysis with it.

1.1 Paths to Data

Relative Path

The dataset is stored in a csv file. The folder structure for this file we are working inside and the data file is:

- main folder: Stat4Econ
  - subfolder: data
    - file: Basketball.csv
  - subfolder: descriptive
    - file: DataBasketball.ipynb (the jupyter notebook file)
    - file: DataBasetball.html (the html version of the jupyter notebook file

overall this means: - the csv file's location is: **'/Stat4Econ/descriptive/data/Basketball.csv'** - the working R code file's location is: **'/Stat4Econ/descriptive/data/DataBasketball.ipynb'**

Given this structure, to access the *Basketball.csv* dataset, we need to go one folder up from our current subfolder to the mainfolder, and then choose the data subfolder, and the Basketball.csv file in the subfolder.

Absolute Path

If these files are not in the same main folder but are in different locations on your computer, you can find the full path to the csv path and copy paste the path below in between the single quotes.

search on google to find out how to get the full path to file: - google search for find full path for file on mac + this might end up looking like: **'/Users/fan/Downloads/Basketball.csv'** - google search for find full path for file on PC + this might end up looking like: **'C:/Users/fan/Documents/Dropbox/Basketball.csv'**

Using Relative path to load in data

We will load in the data using base R read.csv function.

- For what the variables mean, see here
- For what NBA team names correspond to, see here.

```r
# We can load the dataset in first by setting our directory, then loading in the dataset
basetball_data <- read.csv('data/Basketball.csv')
```

```r
# Alternatively, we can just use one line
basetball_data <- read.csv('data/Basketball.csv')
```

```r
# Summarize all variables in data frame
summary(basetball_data)
```

```
##     ilkid                year       firstname           lastname            team               leag
##  Length:21959       Min.   :1946   Length:21959       Length:21959       Length:21959       Length:21
##  Class :character   1st Qu.:1974   Class :character   Class :character   Class :character   Class :ch
##  Mode  :character   Median :1988   Mode  :character   Mode  :character   Mode  :character   Mode  :ch
##                     Mean   :1986
##                     3rd Qu.:1999
##                     Max.   :2009
##       pts              oreb             dreb             reb              asts             stl
##  Min.   :   0.0   Min.   :  0.00   Min.   :   0.0   Min.   :   0.0   Min.   :   0.0   Length:21959
##  1st Qu.: 113.0   1st Qu.:  0.00   1st Qu.:   1.0   1st Qu.:  44.0   1st Qu.:  20.0   Class :charact
##  Median : 386.0   Median : 22.00   Median :  60.0   Median : 160.0   Median :  71.0   Mode  :charact
##  Mean   : 531.1   Mean   : 49.79   Mean   : 117.8   Mean   : 229.7   Mean   : 118.1
##  3rd Qu.: 811.0   3rd Qu.: 75.00   3rd Qu.: 180.0   3rd Qu.: 333.0   3rd Qu.: 167.0
##  Max.   :4029.0   Max.   :587.00   Max.   :1538.0   Max.   :2149.0   Max.   :1164.0
##        pf              fga              fgm              fta              ftm              tpa
##  Min.   :  0.0   Min.   :   0.0   Min.   :   0.0   Min.   :   0.0   Min.   :  0.0   Min.   :  0.00
##  1st Qu.: 43.0   1st Qu.: 106.0   1st Qu.:  43.0   1st Qu.:  30.0   1st Qu.: 20.0   1st Qu.:  0.00
##  Median :118.0   Median : 345.0   Median : 148.0   Median :  99.0   Median : 70.0   Median :  2.00
##  Mean   :123.6   Mean   : 452.5   Mean   : 204.2   Mean   : 146.9   Mean   :109.6   Mean   : 38.07
##  3rd Qu.:193.0   3rd Qu.: 696.0   3rd Qu.: 313.0   3rd Qu.: 218.0   3rd Qu.:161.0   3rd Qu.: 27.00
##  Max.   :386.0   Max.   :3159.0   Max.   :1597.0   Max.   :1363.0   Max.   :840.0   Max.   :678.00
```