Opening a Dataset

Fan Wang

2020-05-02

Contents

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.

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            g
##  WILLIKE01 :   27   Min.   :1946   John   :  502   Williams:  441   TOT    : 1611   A: 1247   Min.
##  EDWARJA01 :   25   1st Qu.:1974   Bob    :  404   Johnson :  400   NYK    :  947   N:20712   1st Qu
##  CORBITY01 :   24   Median :1988   Mike   :  395   Smith   :  317   BOS    :  919             Median
##  JACKSJI01 :   24   Mean   :1986   Jim    :  344   Jones   :  298   DET    :  799             Mean
##  CASSESA01 :   23   3rd Qu.:1999   Chris  :  267   Davis   :  254   PHI    :  736             3rd Qu
##  MALONMO01 :   23   Max.   :2009   Tom    :  260   Brown   :  209   LAL    :  712             Max.
##  (Other)   :21813                  (Other):19787   (Other) :20040   (Other):16235
##       dreb             reb              asts             stl              blk           turnover
##  Min.   :   0.0   Min.   :   0.0   Min.   :   0.0   0      : 5939   0      : 6855   0      : 5696   M
##  1st Qu.:   1.0   1st Qu.:  44.0   1st Qu.:  20.0   1      :  607   1      : 1126   1      :  349   1
##  Median :  60.0   Median : 160.0   Median :  71.0   2      :  479   2      :  860   2      :  347   M
##  Mean   : 117.8   Mean   : 229.7   Mean   : 118.1   3      :  445   3      :  754   NULL   :  294   M
##  3rd Qu.: 180.0   3rd Qu.: 333.0   3rd Qu.: 167.0   5      :  345   4      :  647   3      :  287   3
##  Max.   :1538.0   Max.   :2149.0   Max.   :1164.0   4      :  343   5      :  608   5      :  246   M
##                                                     (Other):13801   (Other):11109   (Other):14740
##       ftm              tpa              tpm
##  Min.   :  0.0   Min.   :  0.00   Min.   :  0.0
##  1st Qu.: 20.0   1st Qu.:  0.00   1st Qu.:  0.0
##  Median : 70.0   Median :  2.00   Median :  0.0
##  Mean   :109.6   Mean   : 38.07   Mean   : 13.1
##  3rd Qu.:161.0   3rd Qu.: 27.00   3rd Qu.:  7.0
##  Max.   :840.0   Max.   :678.00   Max.   :269.0
##
```