Tutorial 1.5

2023-08-12

Data Frames

References other sources for learning: 1. *learnr* package 2. https://www.sas.upenn.edu/~baron/from_cattell/rpsych/rpsych.html#toc2 3. https://intro2r.com/basics_r.html 4. Discovering Statistics with R by Andy Field

NOTE: This worksheet is for you to get a hands-on experience of R. If you are unfamiliar with R or coding in general, this should help, but you must explore more from the references (1, 2, and 3) above to get a better hang of all things R.

There are also some OPTIONAL bits in this worksheet which you can skip.

Contents: * Introduction to tidyverse * IMPORTING AND LOADING LIBRARIES * Manipulating data frame
==

 # 1. Introduction to tidyverse

The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures. Tidyverse Packages in R following:

Data Visualization and Exploration *ggplot2* Data Wrangling and Transformation *dplyr tidyr stringr forcats* Data Import and Management *tibble readr* Functional Programming *purrr*

2. IMPORTING AND LOADING LIBRARIES

Installing makes the library available to your PC. Loading makes it available to the R environment. You need to install a package once but load it every time you want to run the script.

A package is a bundle of functions that you can use in your code. When you talk to these functions in the syntax they understand, these functions will save you tons of time and lines of complicated code. Best thing about them is you (most often) do not need to know how they are doing any of this. Just knowing the syntax is enough.

GUI for packages: bottom right pane has a tab for packages. You can install and then load (by checking off) packages from there. Install tidyverse as follows:

```
library(tidyr) # for plotting
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
```

what do other packages do? ?[package/function] is a help command to get more info ?dplyr
find out what other packages from the list above do by using ?

OPTIONAL: look up pacman for installing and loading multiple packages

 # 3. Import/Create data frame

```
data <- read.csv(file = "nobel_data.csv", header = TRUE, sep = ",")
```

sep = "," is used as it is a comm separated variable(csv) file. We can check what kind of object "my_data" is by:

```
class(data)
```

```
## [1] "data.frame"
```

 # 4. View Data Structure of the Data Frame

```
str(data)
```

```
## 'data.frame':    950 obs. of  52 variables:
##  $ awardYear                : int  2001 1975 2004 1982 1979 2019 2019 2009 2011 1939 ...
##  $ category                 : chr  "Economic Sciences" "Physics" "Chemistry" "Chemistry" "Chemistry" ...
##  $ categoryFullName         : chr  "The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel" "The Nobel Prize in Physics" "The Nobel Prize in Chemistry" "The Nobel Prize in Chemistry" ...
##  $ sortOrder                : int  2 1 1 1 2 1 1 3 3 1 ...
##  $ portion                  : chr  "1/3" "1/3" "1/3" "1" ...
##  $ prizeAmount              : int  10000000 6300000 10000000 1150000 8000000 9000000 9000000 10000000 10000000 140822 ...
##  $ prizeAmountAdjusted      : int  12295082 3404179 11762861 3102518 2988048 9000000 9000000 10958504 10545557 4227898 ...
##  $ dateAwarded              : chr  "2001-10-10" "1975-10-17" "2004-10-06" "1982-10-18" ...
##  $ prizeStatus              : chr  "received" "received" "received" "received" ...
##  $ motivation               : chr  "for their analyses of markets with asymmetric information" "for the discovery of the connection between collective motion and particle motion in atomic nuclei and the deve"| __truncated__ "for the discovery of ubiquitin-mediated protein degradation" "for his development of crystallographic electron microscopy and his structural elucidation of biologically impo"| __truncated__ ...
##  $ categoryTopMotivation    : chr  "" "" "" "" ...
##  $ award_link               : chr  "https://masterdataapi.nobelprize.org/2/nobelPrize/eco/2001" "https://masterdataapi.nobelprize.org/2/nobelPrize/phy/1975" "https://masterdataapi.nobelprize.org/2/nobelPrize/che/2004" "https://masterdataapi.nobelprize.org/2/nobelPrize/che/1982" ...
##  $ id                       : int  745 182 779 259 114 982 981 843 866 199 ...
##  $ name                     : chr  "A. Michael Spence" "Aage N. Bohr" "Aaron Ciechanover" "Aaron Klug" ...
##  $ knownName                : chr  "A. Michael Spence" "Aage N. Bohr" "Aaron Ciechanover" "Aaron Klug" ...
##  $ givenName                : chr  "A. Michael" "Aage N." "Aaron" "Aaron" ...
##  $ familyName               : chr  "Spence" "Bohr" "Ciechanover" "Klug" ...
##  $ fullName                 : chr  "A. Michael Spence" "Aage Niels Bohr" "Aaron Ciechanover" "Aaron Klug" ...
##  $ penName                  : chr  "" "" "" "" ...
##  $ gender                   : chr  "male" "male" "male" "male" ...
##  $ laureate_link            : chr  "http://masterdataapi.nobelprize.org/2/laureate/745" "http://masterdataapi.nobelprize.org/2/laureate/102" "http://masterdataapi.nobelprize.org/2/laureate/779" "http://masterdataapi.nobelprize.org/2/laureate/259" ...
##  $ birth_date               : chr  "1943-00-00" "1922-06-19" "1947-10-01" "1926-08-11" ...
##  $ birth_city               : chr  "Montclair, NJ" "Copenhagen" "Haifa" "Zelvas" ...
##  $ birth_cityNow            : chr  "Montclair, NJ" "Copenhagen" "Haifa" "Zelvas" ...
##  $ birth_continent          : chr  "North America" "Europe" "Asia" "Europe" ...
##  $ birth_country            : chr  "USA" "Denmark" "British Protectorate of Palestine" "Lithuania" ...
##  $ birth_countryNow         : chr  "USA" "Denmark" "Israel" "Lithuania" ...
##  $ birth_locationString     : chr  "Montclair, NJ, USA" "Copenhagen, Denmark" "Haifa, British Protectorate of Palestine (now Israel)" "Zelvas, Lithuania" ...
##  $ death_date               : chr  "" "2009-09-08" "" "2018-11-20" ...
##  $ death_city               : chr  "" "Copenhagen" "" "" ...
##  $ death_cityNow            : chr  "" "Copenhagen" "" "" ...
##  $ death_continent          : chr  "" "Europe" "" "" ...
##  $ death_country            : chr  "" "Denmark" "" "" ...
##  $ death_countryNow         : chr  "" "Denmark" "" "" ...
##  $ death_locationString     : chr  "" "Copenhagen, Denmark" "" "N/A" ...
##  $ orgName                  : chr  "" "" "" "" ...
##  $ nativeName               : chr  "" "" "" "" ...
##  $ acronym                  : chr  "" "" "" "" ...
##  $ org_founded_date         : chr  "" "" "" "" ...
##  $ org_founded_city         : chr  "" "" "" "" ...
##  $ org_founded_cityNow      : chr  "" "" "" "" ...
##  $ org_founded_continent    : chr  "" "" "" "" ...
##  $ org_founded_country      : chr  "" "" "" "" ...
##  $ org_founded_countryNow   : chr  "" "" "" "" ...
##  $ org_founded_locationString: chr  "" "" "" "" ...
##  $ ind_or_org               : chr  "Individual" "Individual" "Individual" "Individual" ...
##  $ residence_1              : chr  "" "" "" "" ...
##  $ residence_2              : chr  "" "" "" "" ...
##  $ affiliation_1            : chr  "Stanford University, Stanford, CA, USA" "Niels Bohr Institute, Copenhagen, Denmark" "Technion - Israel Institute of Technology, Haifa, Israel" "MRC Laboratory of Molecular Biology, Cambridge, United Kingdom" ...
##  $ affiliation_2            : chr  "" "" "" "" ...
##  $ affiliation_3            : chr  "" "" "" "" ...
##  $ affiliation_4            : chr  "" "" "" "" ...
```

 # 5. View column names

```
colnames(data)
```

```
##  [1] "awardYear"                "category"
##  [3] "categoryFullName"         "sortOrder"
##  [5] "portion"                  "prizeAmount"
##  [7] "prizeAmountAdjusted"      "dateAwarded"
##  [9] "prizeStatus"              "motivation"
## [11] "categoryTopMotivation"    "award_link"
## [13] "id"                       "name"
## [15] "knownName"                "givenName"
## [17] "familyName"               "fullName"
## [19] "penName"                  "gender"
## [21] "laureate_link"            "birth_date"
## [23] "birth_city"               "birth_cityNow"
## [25] "birth_continent"          "birth_country"
## [27] "birth_countryNow"         "birth_locationString"
## [29] "death_date"               "death_city"
## [31] "death_cityNow"            "death_continent"
## [33] "death_country"            "death_countryNow"
## [35] "death_locationString"     "orgName"
## [37] "nativeName"               "acronym"
## [39] "org_founded_date"         "org_founded_city"
## [41] "org_founded_cityNow"      "org_founded_continent"
## [43] "org_founded_country"      "org_founded_countryNow"
## [45] "org_founded_locationString" "ind_or_org"
## [47] "residence_1"              "residence_2"
## [49] "affiliation_1"            "affiliation_2"
## [51] "affiliation_3"            "affiliation_4"
```

 # 6. Count number of rows

```
count(data)
```

```
##     n
## 1 950
```

 # 7. Get summary of the data

```
summary(data)
```

```
##    awardYear      category         categoryFullName    sortOrder
##  Min.   :1901   Length:950         Length:950         Min.   :1.000
##  1st Qu.:1947   Class :character   Class :character   1st Qu.:1.000
##  Median :1977   Mode  :character   Mode  :character   Median :1.000
##  Mean   :1971                                         Mean   :1.483
##  3rd Qu.:2000                                         3rd Qu.:2.000
##  Max.   :2019                                         Max.   :3.000
##    portion          prizeAmount      prizeAmountAdjusted dateAwarded
##  Length:950         Min.   :  114935   Min.   : 2377268   Length:950
##  Class :character   1st Qu.:  170332   1st Qu.: 3052326   Class :character
##  Mode  :character   Median :  700000   Median : 4997406   Mode  :character
##                     Mean   : 3460596   Mean   : 6145681
##                     3rd Qu.: 8000000   3rd Qu.: 9044276
##                     Max.   :10000000   Max.   :12295082
##  prizeStatus         motivation        categoryTopMotivation  award_link
##  Length:950         Length:950         Length:950             Length:950
##  Class :character   Class :character   Class :character       Class :character
##  Mode  :character   Mode  :character   Mode  :character       Mode  :character
##
##
##
##
##        id             name           knownName          givenName
##  Min.   :  1.0   Length:950         Length:950         Length:950
##  1st Qu.:238.2   Class :character   Class :character   Class :character
##  Median :477.5   Mode  :character   Mode  :character   Mode  :character
##  Mean   :483.0
##  3rd Qu.:727.8
##  Max.   :984.0
##   familyName          fullName          penName            gender
##  Length:950         Length:950         Length:950         Length:950
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  laureate_link       birth_date         birth_city        birth_cityNow
##  Length:950         Length:950         Length:950         Length:950
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  birth_continent    birth_country      birth_countryNow   birth_locationString
##  Length:950         Length:950         Length:950         Length:950
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##   death_date         death_city        death_cityNow      death_continent
##  Length:950         Length:950         Length:950         Length:950
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  death_country      death_countryNow   death_locationString   orgName
##  Length:950         Length:950         Length:950           Length:950
##  Class :character   Class :character   Class :character     Class :character
##  Mode  :character   Mode  :character   Mode  :character     Mode  :character
##
##
##
##   nativeName          acronym         org_founded_date   org_founded_city
##  Length:950         Length:950         Length:950         Length:950
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  org_founded_cityNow org_founded_continent org_founded_country
##  Length:950          Length:950            Length:950
##  Class :character    Class :character      Class :character
##  Mode  :character    Mode  :character      Mode  :character
##
##
##
##  org_founded_countryNow org_founded_locationString  ind_or_org
##  Length:950             Length:950                  Length:950
##  Class :character       Class :character            Class :character
##  Mode  :character       Mode  :character            Mode  :character
##
##
##
##  residence_1        residence_2        affiliation_1      affiliation_2
##  Length:950         Length:950         Length:950         Length:950
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  affiliation_3      affiliation_4
##  Length:950         Length:950
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```