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#htoc2 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:

```
install.packages("tidyverse")
```

```
## Installing package into '/home/swarag/R/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```
## also installing the dependencies 'gargle', 'curl', 'systemfonts', 'textshaping', 'googledrive', 'googlesheets
4', 'httr', 'ragg', 'rvest', 'xml2'
```

```
## Warning in install.packages("tidyverse"): installation of package 'curl' had
## non-zero exit status
```

```
## Warning in install.packages("tidyverse"): installation of package 'systemfonts'
## had non-zero exit status
```

```
## Warning in install.packages("tidyverse"): installation of package 'xml2' had
## non-zero exit status
```

```
## Warning in install.packages("tidyverse"): installation of package 'textshaping'
## had non-zero exit status
```

```
## Warning in install.packages("tidyverse"): installation of package 'httr' had
## non-zero exit status
```

```
## Warning in install.packages("tidyverse"): installation of package 'gargle' had
## non-zero exit status
```

```
## Warning in install.packages("tidyverse"): installation of package 'ragg' had
## non-zero exit status
```

```
## Warning in install.packages("tidyverse"): installation of package 'rvest' had
## non-zero exit status
```

```
## Warning in install.packages("tidyverse"): installation of package 'googledrive'
## had non-zero exit status
```

```
## Warning in install.packages("tidyverse"): installation of package
## 'googlesheets4' had non-zero exit status
```

```
## Warning in install.packages("tidyverse"): installation of package 'tidyverse'
## had non-zero exit status
```

```
library(tidyr)
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" ...
##  $ categoryFullName         : chr  "The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobe
l" "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 630000 10000000 1150000 800000 9000000 9000000 10000000 10000000
148822 ...
##  $ prizeAmountAdjusted      : int  12295082 3404179 11762861 3102518 2988048 9000000 9000000 10958504 1054555
7 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 disco
very 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://mast
erdataapi.nobelprize.org/2/nobelPrize/phy/1975" "https://masterdataapi.nobelprize.org/2/nobelPrize/che/2004" "htt
ps://masterdataapi.nobelprize.org/2/nobelPrize/che/1982" ...
##  $ id                       : int  745 102 779 259 114 982 981 843 866 199 ...
##  $ 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://masterdataap
i.nobelprize.org/2/laureate/102" "http://masterdataapi.nobelprize.org/2/laureate/779" "http://masterdataapi.nobel
prize.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, Copenhage
n, Denmark" "Technion - Israel Institute of Technology, Haifa, Israel" "MRC Laboratory of Molecular Biology, Camb
ridge, 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
##
##
##
```

8. Accessing A Specific Column

```
data$awardYear
```

```
##   [1] 2001 1975 2004 1982 1979 2019 2019 2009 2011 1939 1905 1928 1980 1999 2010
##  [16] 2019 2007 2000 1963 2000 1907 1957 1974 1921 2007 1902 1960 1952 1937 1910
##  [31] 1964 1970 2003 1912 1982 1969 1911 1994 1966 1913 2013 1979 1911 1907 1982
##  [46] 2012 1998 1947 1977 1921 1956 2010 1947 1965 1975 1963 1977 2006 2015 2003
##  [61] 1974 1978 1952 1922 2013 1926 1948 1978 2018 2015 1927 1929 1934 1959 1981
##  [76] 1945 2000 1920 1909 1991 2004 2015 2009 1983 2017 2005 2001 1976 1980 1975
##  [91] 2016 1982 2016 1947 1991 1905 1977 1994 1950 1976 1903 2016 1958 1973 2012
## [106] 2011 2011 1984 1976 1927 1906 1989 1931 1947 1936 1919 1935 2001 1984 1996
## [121] 1936 2009 1950 1984 1966 1920 1925 1917 1964 1987 2009 1928 1913 1957 1929
## [136] 1972 1974 1921 1995 2010 2011 1997 1985 1994 1937 2003 1945 1938 2006 1980
## [151] 1976 1961 2010 2011 1957 1998 2002 2000 1978 1997 1975 1981 2004 2016 2012
## [166] 1996 1998 2018 1971 1969 1992 1984 1956 2019 1960 1987 2018 2007 1964 1996
## [181] 1993 1986 1990 1952 1971 1932 1992 2006 1907 2014 1943 1995 1950 2004 1958
## [196] 1947 1992 1951 1949 2010 1974 2004 1981 1990 1902 1986 1912 2009 2009 2011
## [211] 1902 1901 1959 1946 1938 2014 2001 1995 2000 2007 1931 1954 1939 1908 1951
## [226] 1907 1945 1973 1986 1991 1933 2019 2013 1936 1963 1975 2012 1974 2016 1995
## [241] 1993 1952 1964 1909 1927 1998 2004 2018 1962 1922 1985 2013 1965 1952 2008
## [256] 1929 2004 1939 1935 1904 1901 1954 1923 1995 1958 1921 1908 1922 1931 1974
## [271] 1947 1953 1918 1953 1923 1982 1908 1945 2000 1992 1973 1961 1979 2001 1994
## [286] 1958 1925 1953 1980 1943 1974 2009 2006 1988 1934 1982 2018 1937 1967 1934
## [301] 1967 1992 1984 1958 1972 1983 2018 1999 1986 1939 2007 1971 1912 1988 1947
## [316] 1963 1906 1963 1951 1979 2006 1926 2019 1909 1974 1999 1999 1912 1925 1926
## [331] 2004 1968 2002 1955 1978 1970 1967 1930 1989 1953 1935 1929 2008 1934 1989
## [346] 2005 1990 1974 1988 1913 1972 1986 1927 1902 1903 1927 1913 1906 1943 1917
## [361] 1901 1973 1983 1990 1905 1985 1979 2000 1944 1978 1946 1946 1953 2009 2000
## [376] 1949 2014 1921 1998 1975 1955 1958 1958 1908 1977 2002 1904 2007 2005 2017
## [391] 1997 1917 1969 1985 1935 1932 2004 1978 2014 1944 1933 1904 1973 1961 1987
## [406] 1963 2003 1989 2016 2005 1906 1922 2000 1988 2009 1901 2017 1965 1996 1946
## [421] 1935 1980 1977 2013 1925 2000 1986 2018 2019 1975 1981 1962 1969 1931 1959
## [436] 1984 1926 1980 2014 2008 1987 1964 2016 2017 1997 1990 1985 2002 2017 1997
## [451] 1988 1910 1926 1919 1944 2002 2019 1956 1994 1962 2006 1986 1951 1975 2002
## [466] 1997 1954 1994 1932 1946 1977 1998 2005 1923 2014 1998 1972 1946 1982 1962
## [481] 1904 1996 1998 1987 1990 2001 1944 1993 1985 1995 1958 2016 1956 2011 1919
## [496] 1965 1970 1927 1987 1981 2014 1917 1930 1973 1963 1993 2017 1967 1981 1982
## [511] 1972 1994 2000 2017 1908 1985 1920 2001 2002 1964 1973 2010 1950 2002 1968
## [526] 2013 1915 1980 1973 1963 1983 2001 1973 1920 1951 1988 1972 2007 1975 1939
## [541] 1957 1962 2011 2004 1962 2010 2012 1949 1904 1957 1929 1998 1970 1907 2008
## [556] 1927 1934 1968 1970 2019 1976 2008 2014 1967 1924 1963 1903 1995 2007 2010
## [571] 1968 2008 2013 1995 1964 1994 1974 1999 2008 2002 1988 1911 1962 1954 1969
## [586] 1962 1918 1951 1914 2014 1999 1961 1988 1978 1990 2019 2013 2017 1985 1993
## [601] 2017 2019 1967 1990 1965 1976 2012 2005 1979 2006 1969 1997 2018 1991 1988
## [616] 1938 1930 2015 1966 1993 1931 1981 1964 1922 1984 1903 1973 1956 1970 1989
## [631] 1937 1990 1969 1979 1954 2018 2009 2016 2007 2013 2006 2008 1987 1950 1944
## [646] 1936 1922 1943 1910 1931 1959 1928 1971 1951 1948 2014 1973 1970 1933 1980
## [661] 2003 1997 1908 2000 1909 1910 1995 1974 1937 2008 2018 2015 1948 1912 1958
## [676] 1938 1946 1910 2010 2003 1996 1936 2007 2019 2013 1960 1978 1966 1959 1950
## [691] 1977 1905 1993 1903 1991 1902 1955 1995 1978 1913 1969 1967 2017 1950 2011
## [706] 1903 2013 2002 1994 1975 1968 2002 2004 1996 1990 2010 2017 2017 1993 1938
## [721] 1952 1965 1991 2005 1984 1915 1925 1992 1986 1981 1923 1998 1965 1914 1997
## [736] 1996 1937 1995 1996 2003 1998 2010 2005 1961 1988 2005 2012 2013 1905 1987
## [751] 1999 1966 1972 1993 1968 1978 2003 1972 2007 2006 1977 1937 1981 2008 1996
## [766] 1915 1967 1991 1902 1977 2005 1908 1961 1992 1907 1993 2001 1960 1969 1959
## [781] 1969 1976 1906 2015 1976 2011 1995 1974 1909 1952 2012 1959 1979 1994 2012
## [796] 2003 1966 2014 1989 1928 1971 1984 1965 1930 1945 1979 1925 1970 1930 1932
## [811] 1956 1960 1929 2018 1996 1936 1945 2016 1988 2012 1963 2007 1977 1933 2001
## [826] 2019 2003 1947 1904 1972 1997 1986 2014 1997 1979 1983 1901 1982 1987 1903
## [841] 2015 2002 1948 1950 2015 2018 2011 1989 1926 1909 1902 2005 1906 1914 1979
## [856] 2009 2005 2013 1933 1954 2011 1929 1989 2001 1975 1911 2015 2011 1993 1981
## [871] 2008 1989 1957 2015 1970 2001 1965 1988 2001 1980 2009 1916 2002 1977 1936
## [886] 1912 1955 2003 1975 1980 1956 1949 1998 1954 1920 2004 1973 1946 1978 1956
## [901] 1932 1901 1909 1911 1960 2009 1924 1983 1956 1915 1923 2015 2018 1997 2014
## [916] 1949 1990 1949 2019 1983 1972 2001 1976 1934 1996 1955 1971 1953 1996 1924
## [931] 1986 2001 1989 1945 1919 1994 1968 1994 2008 2016 1986 2005 2000 1980 1944
## [946] 1972 1954 1911 1981 1963
```