mod3-dataframes.R

randomfluff

2024-09-15

```
# Mod 3: Data Frames & Matrices
# Author: Amy Holt
# Data: 9/15/2024

# Assignment 3:
# This data set is based on the presidential election during 2016, where
# it outlined the name of the candidate, the source of the poll (ABC vs,
# CBS).  Discuss your result in your blog. Important note, I made up this
# data, so this data does not reflect what really happened in the election

# Vectors containing the data
Name <- c("Jeb", "Donald", "Ted", "Marco", "Carly", "Hillary", "Bernie")
ABC.poll.results <- c(4, 62, 51, 21, 2, 14, 15)
CBS.poll.results <- c(12, 75, 43, 19, 1, 21, 19)

# Create a data frame from named vectors
poll_df <- data.frame(Name = Name, ABC=ABC.poll.results, CBS = CBS.poll.results)
poll_df
```

```
##      Name ABC CBS
## 1     Jeb   4  12
## 2  Donald  62  75
## 3     Ted  51  43
## 4   Marco  21  19
## 5   Carly   2   1
## 6 Hillary  14  21
## 7  Bernie  15  19
```

```
# Create a data frame by importing a file
poll.results_df <- read.csv("poll_data.csv", header = TRUE)
poll.results_df
```

```
##      Name ABC.poll.results CBS.poll.results
## 1     Jeb                4               12
## 2  Donald               62               75
## 3     Ted               51               43
## 4   Marco               21               19
## 5   Carly                2                1
## 6 Hillary               14               21
## 7  Bernie               15               19
```

```
row.names(poll.results_df) # view row names of data frame
```

```
## [1] "1" "2" "3" "4" "5" "6" "7"
```

```r
row.names(poll.results_df) <- Name # assign candidate names as row names
row.names(poll.results_df) # check row names assignment
```

```
## [1] "Jeb"     "Donald"  "Ted"     "Marco"   "Carly"   "Hillary" "Bernie"
```

```r
# index columns 2 thru 3, # reassign to poll.results_df; "deletes" Name
# column
poll.results_df <- poll.results_df[,2:3]
poll.results_df
```

```
##         ABC.poll.results CBS.poll.results
## Jeb                    4               12
## Donald                62               75
## Ted                   51               43
## Marco                 21               19
## Carly                  2                1
## Hillary               14               21
## Bernie                15               19
```

```r
# candidate results can be indexed by name and/or network
poll.results_df["Jeb",] # index by row (candidate) name
```

```
##     ABC.poll.results CBS.poll.results
## Jeb                4               12
```

```r
poll.results_df["Hillary", "CBS.poll.results"] # index by candidate and network
```

```
## [1] 21
```

```r
poll.results_df[,"ABC.poll.results"] # index by network
```

```
## [1]  4 62 51 21  2 14 15
```

```r
# basic statistical functions on the data frame
rowMeans(poll.results_df["Marco",]) # mean of Marco's results
```

```
## Marco
##    20
```

```r
mean(poll.results_df[, "CBS.poll.results"])
```

```
## [1] 27.14286
```

```r
sd(poll.results_df[, "CBS.poll.results"])
```

```
## [1] 24.57932
```

```r
max(poll.results_df)
```

```
## [1] 75
```

```r
# Filtering
which(poll.results_df > 50) # get indices where poll results are > 50
```

```
## [1] 2 3 9
```

```r
poll.results_df[poll.results_df > 50,] # subset df for results > 50
```

```
##        ABC.poll.results CBS.poll.results
## Donald               62               75
## Ted                  51               43
```

```
## NA                   NA                   NA
```