reading_csv

October 4, 2020

Reading CSV File to DataFrame

1 1. Reading Data from CSV File

- CSV File is "comma-separated values"
- The , separator is conventional, but not mandatory
- The | character is also common

1.1 1.1 Providence Police Dept. Data

- We will be looking at public data regarding arrests and case

```
In [1]: # The line below reads the CSV file and creates a dataframe

        arrests_df <- read.csv("data/pvd_arrests_2020-10-03.csv")
```

1.2 1.2 Exploring the Data

```
In [2]: head(arrests_df)          # show first few lines of the dataframe
```

A data.frame: 6 Œ 18

	arrest_date <chr>	year <int>	month <int>	gender <chr>	race <chr>	ethnicity <chr>	year_of <int>
1	2019-08-24T02:23:00.0	2019	8	Male	White	NonHispanic	1981
2	2019-08-24T02:02:00.0	2019	8				1994
3	2019-08-24T02:02:00.0	2019	8	Female	Black	NonHispanic	1984
4	2019-08-24T02:02:00.0	2019	8	Female	Black	NonHispanic	1984
5	2019-08-24T02:02:00.0	2019	8	Female	Black	Unknown	2001
6	2019-08-24T02:02:00.0	2019	8	Female	Black	Unknown	2001

1.2.1 1.2.1 More Data Exploring

```
In [3]: dim(arrests_df)             # get dimensions of the dataframe
```

1. 8755 2. 18

```
In [4]: nrow(arrests_df)            # get number of rows
```

8755

```
In [5]: ncol(arrests_df)            # get the number of columns
```

18

```
In [6]: colnames(arrests_df)        # get the column names
```

1. 'arrest_date' 2. 'year' 3. 'month' 4. 'gender' 5. 'race' 6. 'ethnicity' 7. 'year_of_birth' 8. 'age' 9. 'from_address' 10. 'from_city' 11. 'from_state' 12. 'statute_type' 13. 'statute_code' 14. 'statute_desc' 15. 'counts' 16. 'case_number' 17. 'arresting_officers' 18. 'id'

2 2. Summaries from `data.frame`

```
In [7]: str(arrests_df)              # the str() function shows the structure of dataframe
```

```
'data.frame':        8755 obs. of  18 variables:
 $ arrest_date       : chr  "2019-08-24T02:23:00.0" "2019-08-24T02:02:00.0" "2019-08-24T02:02:0
 $ year              : int  2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 ...
 $ month             : int  8 8 8 8 8 8 8 8 8 8 ...
 $ gender            : chr  "Male" "" "Female" "Female" ...
 $ race              : chr  "White" "" "Black" "Black" ...
 $ ethnicity         : chr  "NonHispanic" "" "NonHispanic" "NonHispanic" ...
 $ year_of_birth     : int  1981 1994 1984 1984 2001 2001 2001 1991 1991 1991 ...
 $ age               : int  37 25 34 34 18 18 18 28 28 28 ...
 $ from_address      : chr  "No Permanent Address" "SUMMER AVE" "DOUGLAS AVE" "DOUGLAS AVE" ..
 $ from_city         : chr  "providence" "Cranston" "Providence" "Providence" ...
 $ from_state        : chr  "Rhode Island" "Rhode Island" "Rhode Island" "Rhode Island" ...
 $ statute_type      : chr  "" "RI Statute Violation" "RI Statute Violation" "RI Statute Violat
 $ statute_code      : chr  "" "31-11-18" "12-7-10" "11-45-1" ...
 $ statute_desc      : chr  "" "Driving after Denial, Suspension or Revocation of License" "RES
 $ counts            : int  NA 1 1 1 1 1 1 1 1 1 ...
 $ case_number       : chr  "2019-00084142" "2019-00084127" "2019-00084126" "2019-00084126" ..
 $ arresting_officers: chr  " YGonzalez, LTaveras" " NManfredi" " MPlace, JPerez, ASantos" " MI
 $ id                : chr  "pvd2218242150382148273" "pvd15166785558364246202" "pvd31429177062
```

2.1 2.1 Summarizing Numeric Data

```
In [8]: summary(arrests_df)
```

```
 arrest_date             year          month          gender
 Length:8755        Min.   :2019   Min.   : 1.00   Length:8755
 Class :character   1st Qu.:2019   1st Qu.: 3.00   Class :character
 Mode  :character   Median :2020   Median : 7.00   Mode  :character
                    Mean   :2020   Mean   : 6.67
                    3rd Qu.:2020   3rd Qu.:10.00
                    Max.   :2020   Max.   :12.00

    race            ethnicity         year_of_birth       age
```

```
Length:8755        Length:8755        Min.   :1943   Min.   :18.00
Class :character   Class :character   1st Qu.:1980   1st Qu.:24.00
Mode  :character   Mode  :character   Median :1989   Median :30.00
                                      Mean   :1986   Mean   :32.93
                                      3rd Qu.:1995   3rd Qu.:39.00
                                      Max.   :2002   Max.   :77.00

from_address        from_city          from_state        statute_type
Length:8755        Length:8755        Length:8755        Length:8755
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character



statute_code       statute_desc          counts       case_number
Length:8755        Length:8755        Min.   : 1.000   Length:8755
Class :character   Class :character   1st Qu.: 1.000   Class :character
Mode  :character   Mode  :character   Median : 1.000   Mode  :character
                                      Mean   : 1.083
                                      3rd Qu.: 1.000
                                      Max.   :15.000
                                      NA's   :1906
arresting_officers      id
Length:8755        Length:8755
Class :character   Class :character
Mode  :character   Mode  :character
```

2.1.1 2.1.1 Summarizing Numeric Data (cont.)

```
In [9]: numeric_vars <- c("month", "year", "age", "year_of_birth", "counts")
```

```
In [10]: summary(arrests_df[, numeric_vars])
```

```
     month            year           age        year_of_birth      counts
 Min.   : 1.00   Min.   :2019   Min.   :18.00   Min.   :1943   Min.   : 1.000
 1st Qu.: 3.00   1st Qu.:2019   1st Qu.:24.00   1st Qu.:1980   1st Qu.: 1.000
 Median : 7.00   Median :2020   Median :30.00   Median :1989   Median : 1.000
 Mean   : 6.67   Mean   :2020   Mean   :32.93   Mean   :1986   Mean   : 1.083
 3rd Qu.:10.00   3rd Qu.:2020   3rd Qu.:39.00   3rd Qu.:1995   3rd Qu.: 1.000
 Max.   :12.00   Max.   :2020   Max.   :77.00   Max.   :2002   Max.   :15.000
                                                               NA's   :1906
```

2.2 2.2 Summarizing String Variables

```
In [11]: table(arrests_df[, "gender"])    # show summary of "gender" column in `arrests_df`
```

```
         Female    Male    NULL Unknown
     21    1906    6804      20       4
```

```
In [12]: table(arrests_df$race)           # show summary of "race" column in `arrests_df`
```

```
                            American Indian/Alaskan Native
                        24                              18
        Asian/Pacific Islander                       Black
                        81                            3807
                      NULL                         Unknown
                        15                             385
                     White          ZHispanic (FD only)
                      4419                               6
```

3 3. Options when Reading CSV

- The `read.csv()` function has many optional arguments
- Critically, we can tell R the strings that ought to be considered missing

```
In [13]: help(read.csv)
```

```
In [14]: arrests_df2 <- read.csv("data/pvd_arrests_2020-10-03.csv",
                                 na.strings = c("NA", "", " ", "NULL", "Unknown"))
```

3.0.1 3.1 Effects of `na.strings`

```
In [15]: table(arrests_df$race)             # explore `race` in original dataframe
```

```
                            American Indian/Alaskan Native
                        24                              18
        Asian/Pacific Islander                       Black
                        81                            3807
                      NULL                         Unknown
                        15                             385
                     White          ZHispanic (FD only)
                      4419                               6
```

```
In [16]: table(arrests_df2$race)           # dataframe after setting `na.strings`
```

American Indian/Alaskan Native	Asian/Pacific Islander
18	81
Black	White
3807	4419
ZHispanic (FD only)	
6	