dataframes

October 4, 2020

The DataFrame Object in R

1 1. What is a `data.frame`?

- Tabular data structure (i.e., like Excel spreadsheet)
- Canonical data structure for data analysis
- Capable of storing heterogeneous data

1.1 1.1 What does it look like?

```
In [1]: idx <- 1:4
        score <- rnorm(4)
        vocals <- c(TRUE, TRUE, FALSE, FALSE)
        firstname <- c("john", "paul", "george", "ringo")

        dat <- data.frame(idx, firstname, score, vocals)

        dat
```

A data.frame: 4 Œ 4

idx <int>	firstname <chr>	score <dbl>	vocals <lgl>
1	john	-1.2070035	TRUE
2	paul	0.8243397	TRUE
3	george	0.9136364	FALSE
4	ringo	1.8258713	FALSE

1.2 1.2 Indexing and Slicing a `data.frame`

- Similar to `vector`, `matrix`, and `array` objects

```
In [2]: dat[1, 2]           # get element in first row, second column
```

'john'

```
In [3]: dat[, 2]            # get all of second column
```

1. 'john' 2. 'paul' 3. 'george' 4. 'ringo'

```
In [4]: dat[3, ]            # get third row
```

A data.frame: 1 Œ 4

	idx	firstname	score	vocals
	<int>	<chr>	<dbl>	<lgl>
3	3	george	0.9136364	FALSE

1.2.1 1.2.1 Indexing using Column Names

```
In [5]: dat[3, "score"]          # element from row 3 and "score" column
```

0.913636425821495

```
In [6]: dat[2:4, "firstname"]    # get elements 2, 3, and 4 from "firstname" column
```

1. 'paul' 2. 'george' 3. 'ringo'

1.3 1.3 The $ Operator and `data.frame` Objects

```
In [7]: dat$firstname            # get the "firstname" column
```

1. 'john' 2. 'paul' 3. 'george' 4. 'ringo'

2 2. Filter `data.frame` using Logical Indexing

```
In [8]: dat
```

A data.frame: 4 Œ 4

idx	firstname	score	vocals
<int>	<chr>	<dbl>	<lgl>
1	john	-1.2070035	TRUE
2	paul	0.8243397	TRUE
3	george	0.9136364	FALSE
4	ringo	1.8258713	FALSE

```
In [9]: dat[dat$vocals, ]
```

A data.frame: 2 Œ 4

	idx	firstname	score	vocals
	<int>	<chr>	<dbl>	<lgl>
1	1	john	-1.2070035	TRUE
2	2	paul	0.8243397	TRUE

2.1 2.1 Create New `data.frame` from Another

```
In [10]: dat2 <- dat[dat$vocals, ]       # create new dataframe, from subset of original

         head(dat2)
```

A data.frame: 2 Œ 4

	idx	firstname	score	vocals
	<int>	<chr>	<dbl>	<lgl>
1	1	john	-1.2070035	TRUE
2	2	paul	0.8243397	TRUE

2.1.1 2.1.1 Take Subset of `data.frame` Columns

```
In [11]: cols <- c("firstname", "score")     # columns we care about

         dat_namescore <- dat2[, cols]       # create new dataframe

         head(dat_namescore)
```

A data.frame: 2 Œ 2

	firstname <chr>	score <dbl>
1	john	-1.2070035
2	paul	0.8243397

3 3. Adding Columns to a `data.frame`

```
In [12]: dat
```

A data.frame: 4 Œ 4

idx <int>	firstname <chr>	score <dbl>	vocals <lgl>
1	john	-1.2070035	TRUE
2	paul	0.8243397	TRUE
3	george	0.9136364	FALSE
4	ringo	1.8258713	FALSE

```
In [13]: dat$food <- c("steak", "chicken", "potato", "rice")

         dat
```

A data.frame: 4 Œ 5

idx <int>	firstname <chr>	score <dbl>	vocals <lgl>	food <chr>
1	john	-1.2070035	TRUE	steak
2	paul	0.8243397	TRUE	chicken
3	george	0.9136364	FALSE	potato
4	ringo	1.8258713	FALSE	rice

3.1 3.1. Adding Columns (cont.)

```
In [14]: dat[, "drink"] <- c("water", "milk", "beer", "scotch")

         dat
```

A data.frame: 4 Œ 6

idx <int>	firstname <chr>	score <dbl>	vocals <lgl>	food <chr>	drink <chr>
1	john	-1.2070035	TRUE	steak	water
2	paul	0.8243397	TRUE	chicken	milk
3	george	0.9136364	FALSE	potato	beer
4	ringo	1.8258713	FALSE	rice	scotch

```
In [15]: dat
```

A data.frame: 4 Œ 6

idx <int>	firstname <chr>	score <dbl>	vocals <lgl>	food <chr>	drink <chr>
1	john	-1.2070035	TRUE	steak	water
2	paul	0.8243397	TRUE	chicken	milk
3	george	0.9136364	FALSE	potato	beer
4	ringo	1.8258713	FALSE	rice	scotch