lab3_jupyter_data_frame

January 29, 2022

List and Dataframe

Estimated time needed: **15** minutes

0.1 Objectives

After completing this lab you will be able to:

- Understand the list and how list correlated with dataframe
- Understand dataframe
- Operate on dataframes

0.2 Table of Contents

About the Dataset

Imagine you got many movie recomendations from your friends and compiled all of the recomendations in a table, with specific info about each movie.

The table has one row for each movie and several columns

- **name** - The name of the movie
- **year** - The year the movie was released
- **length_min** - The lenght of the movie in minutes
- **genre** - The genre of the movie
- **average_rating** - Average rating on Imdb
- **cost_millions** - The movie's production cost in millions
- **sequences** - The amount of sequences
- **foreign** - Indicative of whether the movie is foreign (1) or domestic (0)
- **age_restriction** - The age restriction for the movie

Part of the dataset can be seen below

Lists

First of all, we're gonna take a look at lists in R.

A list is a sequenced collection of different objects of R, like vectors, numbers, characters, other lists and so on. You can consider a list as a container of correlated information, well structured, and easy to read with a context.

A list accepts items of different types, but a vector (or a matrix, which is a multidimensional vector) doesn't. To create a list just type **list()** with your content inside the parenthesis and separated by commas. Let's try it!

[2]:
```
movie <- list("Toy Story", 1995, c("Animation", "Adventure", "Comedy"))
```

In the code above, the variable movie contains a list of 3 objects, which are a string, a numeric value, and a vector of strings.

Easy, eh? Now let's print the content of the list. We just need to call its name.

[3]:
```
movie
```

1. 'Toy Story'
2. 1995
3. (a) 'Animation' (b) 'Adventure' (c) 'Comedy'

0.2.1 Accessing items in a list

It is possible to retrieve only a part of a list using the **single square** bracket operator "[ ]". This operator can be also used to get a single element in a specific position. Take a look at the next example:

The index number 2 returns the second element of a list, if that element exists:

[4]:
```
movie[2]
```

1. 1995

Or you can select a part or interval of elements of a list. In our next example we are retrieving the 1st, 2nd, and 3rd elements:

[5]:
```
movie[2:3]
```

1. 1995
2. (a) 'Animation' (b) 'Adventure' (c) 'Comedy'

It looks a little confusing, but lists can also store names for its elements.

0.2.2 Named lists

The following list is a named list:

[6]:
```
movie <- list(name = "Toy Story",
              year = 1995,
              genre = c("Animation", "Adventure", "Comedy"))
```

Let me explain that: the list **movie** has some named objects within it. **name**, for example, is an object of type **character**, **year** is an object of type **number**, and **genre** is a vector with objects of type **character**.

Now take a look at this list. This time, it's full of information and well organized. It's clear what each element means. You can see that the elements have different types, and that's ok because it's a list.

[7]:
```
movie
```

$name 'Toy Story'

$year 1995

$genre 1. 'Animation' 2. 'Adventure' 3. 'Comedy'

You can also get separated information from the list. You can use **listName***selectorName* ∗∗.*The* ∗ *dollar* − *signoperator* ∗ ∗∗ will give you the block of data that is related to selectorName.

Let's get the genre part of our movies list, for example.

[8]:
```
movie$genre
```

1. 'Animation' 2. 'Adventure' 3. 'Comedy'

Another way of selecting the genre column:

[9]:
```
movie["genre"]
```

$genre = 1. 'Animation' 2. 'Adventure' 3. 'Comedy'

Coding Exercise: in the code cell below, get the second element in the genre column

[11]:
```
# Write your code below. Don't forget to press Shift+Enter to execute the cell
movie$genre[2]
```

'Adventure'

Click here for the solution

```
movie$genre[2]
```

You can also use numerical selectors like an array. Here we are selecting elements ranged from 2 to 3.

[12]:
```
movie[2:3]
```

$year 1995

$genre 1. 'Animation' 2. 'Adventure' 3. 'Comedy'

The function **class()** returns the type of a object. You can use that function to retrieve the type of specific elements of a list:

[13]:
```
class(movie$name)
```

'character'

0.2.3 Adding, modifying, and removing items

Adding a new element is also very easy. The code below adds a new field named **age** and puts the numerical value 5 into it.

In this case we use the double square brackets operator, because we are directly referencing a list member (and we want to change its content).

[14]:
```
movie[["age"]] <- 5
movie
```

$name 'Toy Story'

$year 1995

$genre 1. 'Animation' 2. 'Adventure' 3. 'Comedy'

$age 5

In order to modify, you just need to refer a list member that already exists, then change its content.

[15]:
```
movie[["age"]] <- 6
# Now it's 6, not 5
movie
```

$name 'Toy Story'

$year 1995

$genre 1. 'Animation' 2. 'Adventure' 3. 'Comedy'

$age 6

And removing is also easy! You just put ***NULL***, which means missing value/data, into it.

[16]:
```
movie[["age"]] <- NULL
movie
```

$name 'Toy Story'

$year 1995

$genre 1. 'Animation' 2. 'Adventure' 3. 'Comedy'

Coding Exercise: in the code cell below, add a logical column with a name 'WillWatch' with value TRUE or T

[18]:
```
# Write your code below. Don't forget to press Shift+Enter to execute the cell
movie[["WillWatch"]]<-T
movie
```

$name 'Toy Story'

$year 1995

$genre 1. 'Animation' 2. 'Adventure' 3. 'Comedy'

$WillWatch TRUE

Click here for the solution

```
movie[["WillWatch"]] <- T
movie
```

0.2.4 Concatenating lists

Concatenation is the proccess of puting things together, in sequence. And yes, you can do it with lists. Just call the function ***c()***. Take a look at the next example:

[19]:
```
# We split our previous list in two sublists
movie_part1 <- list(name = "Toy Story")
movie_part2 <- list(year = 1995, genre = c("Animation", "Adventure", "Comedy"))

# Now we call the function c() to put everything together again
movie_concatenated <- c(movie_part1, movie_part2)

# Check it out
movie_concatenated
```

$name 'Toy Story'

$year 1995

$genre 1. 'Animation' 2. 'Adventure' 3. 'Comedy'

Lists are really handy for organizing different types of elements in R, and also easy to use.

Additionally, lists are also important since this type of data structure is essential to create data frames, our next covered topic.

DataFrames

A DataFrame is a structure that is used for storing data tables. Underneath it all, a data frame is a list of vectors of same length, exactly like a table (each vector is a column).

We can use the function **data.frame()** to create a data frame and pass vector, which are our columns, as arguments. It is required to name the columns that will compose the data frame.

[25]:
```
movies <- data.frame(name = c("Toy Story", "Akira", "The Breakfast Club", "The␣
↪Artist",
                              "Modern Times", "Fight Club", "City of God", "The␣
↪Untouchables"),
                     year = c(1995, 1998, 1985, 2011, 1936, 1999, 2002, 1987)
                     ,stringsAsFactors=F)
```

Let's print its content of our recently created data frame:

[26]:
```
movies
```

A data.frame: 8 × 2

name <chr>	year <dbl>
Toy Story	1995
Akira	1998
The Breakfast Club	1985
The Artist	2011
Modern Times	1936
Fight Club	1999
City of God	2002
The Untouchables	1987

It's very easy! You can note how it looks like a table.

We can also use the **"$"** selector to get some type of information. This operator returns the content of a specific column of a data frame (that's why we have to choose a name for each column).

[27]:
```
movies$name
```

1. 'Toy Story' 2. 'Akira' 3. 'The Breakfast Club' 4. 'The Artist' 5. 'Modern Times' 6. 'Fight Club' 7. 'City of God' 8. 'The Untouchables'

You retrieve data using numeric indexing, like in lists:

[28]:
```
# This returns the first (1st) column
movies[1]
```

A data.frame: 8 × 1

name <chr>
Toy Story
Akira
The Breakfast Club
The Artist
Modern Times
Fight Club
City of God
The Untouchables

Coding Exercise: in the code cell below, select the first row of movies data frame

[30]:
```
# Write your code below. Don't forget to press Shift+Enter to execute the cell
movies[1,]
```

A data.frame: 1 × 2

	name <chr>	year <dbl>
1	Toy Story	1995

Click here for the solution

```
movies[1, ]
```

Coding Exercise: in the code cell below, select the first and second rows but only with first column selected from the movies data frame

[31]:
```
# Write your code below. Don't forget to press Shift+Enter to execute the cell
movies[1:2,1]
```

1. 'Toy Story' 2. 'Akira'

Click here for the solution

```
movies[1:2, 1]
```

The function called **str()** is one of most useful functions in R. With this function you can obtain textual information about an object. In this case, it delivers information about the objects whitin a data frame. Let's see what it returns:

[32]:
```
str(movies)
```

```
'data.frame':   8 obs. of  2 variables:
 $ name: chr  "Toy Story" "Akira" "The Breakfast Club" "The Artist" …
 $ year: num  1995 1998 1985 2011 1936 …
```

It shows this data frame has 8 observations, for 2 columns, so called **name** and **year**. The "name" column is a factor with 8 levels and "year" is a numerical column.

The class() function works for data frames as well. You can use it to determine the type of a column of a data frame.

[33]:
```
class(movies$year)
```

'numeric'

You can use numerical selectors to reach information inside the table.

[35]:
```
movies[1,2] #1-Toy Story, 2-1995
```

1995

The ***head()*** function is very useful when you have a large table and you need to take a peek at the first elements. This function returns the first 6 values of a data frame (or event a list).

[34]:
```
head(movies)
```

A data.frame: 6 × 2

	name <chr>	year <dbl>
1	Toy Story	1995
2	Akira	1998
3	The Breakfast Club	1985
4	The Artist	2011
5	Modern Times	1936
6	Fight Club	1999

Similar to the previous function, ***tail()*** returns the last 6 values of a data frame or list.

[36]:
```
tail(movies)
```

A data.frame: 6 × 2

	name <chr>	year <dbl>
3	The Breakfast Club	1985
4	The Artist	2011
5	Modern Times	1936
6	Fight Club	1999
7	City of God	2002
8	The Untouchables	1987

Now, let's try to get the row with name "Toy Story"

[37]:
```
# Find the rows with name "Toy Story"
selected <- movies["name"] == "Toy Story"
# Get the selected row(s) with 'name' and 'year' columns
toy_story <- movies[selected, c("name", "year")]
toy_story
```

A data.frame: 1 × 2

	name <chr>	year <dbl>
1	Toy Story	1995

Now let's try to add a new column to our data frame with the length of each movie in minutes.

[38]:
```
movies['length'] <- c(81, 125, 97, 100, 87, 139, 130, 119)
movies
```

A data.frame: 8 × 3

name <chr>	year <dbl>	length <dbl>
Toy Story	1995	81
Akira	1998	125
The Breakfast Club	1985	97
The Artist	2011	100
Modern Times	1936	87
Fight Club	1999	139
City of God	2002	130
The Untouchables	1987	119

A new column was included into our data frame with just one line of code. We just needed to add a vector to data frame, then it will be our new column.

Now let's try to add a new movie to our data set.

[39]:
```
movies <- rbind(movies, c(name="Dr. Strangelove", year=1964, length=94))
movies
```

	name <chr>	year <chr>	length <chr>
	Toy Story	1995	81
	Akira	1998	125
	The Breakfast Club	1985	97
A data.frame: 9 × 3	The Artist	2011	100
	Modern Times	1936	87
	Fight Club	1999	139
	City of God	2002	130
	The Untouchables	1987	119
	Dr. Strangelove	1964	94

Remember, you can't add a list with more variables than the data frame, and vice-versa.

We don't need this movie anymore, so let's delete it. Here we are deleting row 12 by assigning to itself the movies dataframe without the 12th row.

[40]:
```
movies <- movies[-12,]
movies
```

		name <chr>	year <chr>	length <chr>
	1	Toy Story	1995	81
	2	Akira	1998	125
	3	The Breakfast Club	1985	97
A data.frame: 9 × 3	4	The Artist	2011	100
	5	Modern Times	1936	87
	6	Fight Club	1999	139
	7	City of God	2002	130
	8	The Untouchables	1987	119
	9	Dr. Strangelove	1964	94

To delete a column you can just set it as ***NULL***.

[41]:
```
movies[["length"]] <- NULL
movies
```

		name <chr>	year <chr>
	1	Toy Story	1995
	2	Akira	1998
	3	The Breakfast Club	1985
A data.frame: 9 × 2	4	The Artist	2011
	5	Modern Times	1936
	6	Fight Club	1999
	7	City of God	2002
	8	The Untouchables	1987
	9	Dr. Strangelove	1964

That is it! You learned a lot about data frames and how easy it is to work with them.

Scaling R with big data As you learn more about R, if you are interested in exploring platforms that can help you run analyses at scale, you might want to sign up for a free account on IBM Watson Studio, which allows you to run analyses in R with two Spark executors for free.

0.3 Authors

Hi! It's Thiago Felipe Correa Borges, the author of this notebook. I hope you found R easy to learn! There's lots more to learn about R but you're well on your way. Feel free to connect with me if you have any questions.

0.3.1 Other Contributors

Yan Luo

0.4 Change Log

Date (YYYY-MM-DD)	Version	Changed By	Change Description
2021-03-03	2.0	Yan	Added coding tasks

##