R-Ladies Bucharest

# Text mining with R

**Aurore Paligot,** Data Analyst Consultant

Tue, 20 October 2020

**Mișcare globală care promovează egalitatea de șanse în comunitatea de R prin întâlniri, networking și sesiuni de lucru aplicate într-un mediu sigur și prietenos.**


Logo creat de **Andra Garoi,** R-Ladies Bucharest team member ( Designer )

## Please read and respect our Global Code of Conduct!

# Urmărește @WeAreRLadies pe twitter

Bucharest

1

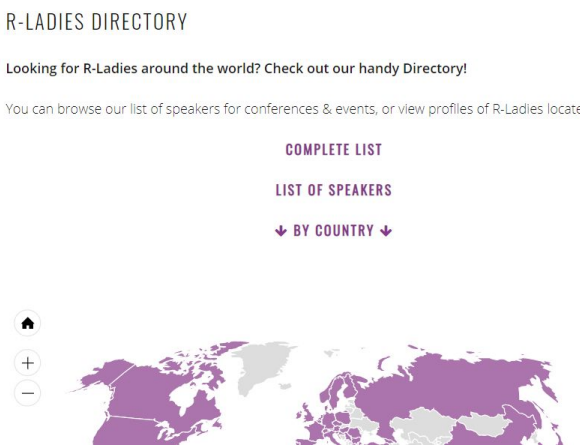


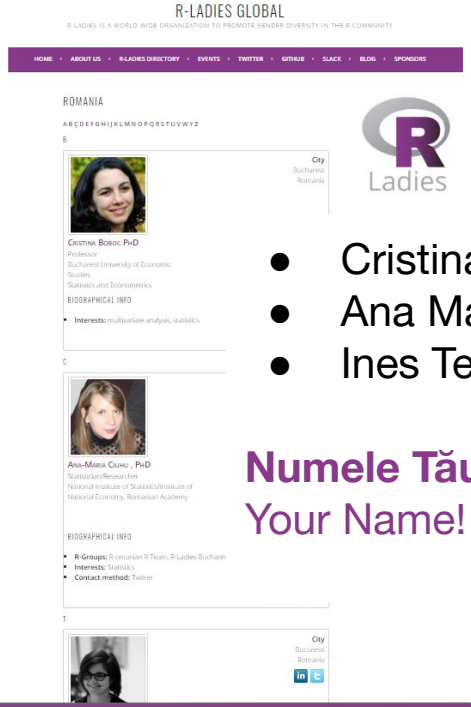
Share, blog or tweet: **#Rladies #RLadiesBucharest #rstats**

# Înscrie-te în directorul de membri

Bucharest

2 **Directorul R-Ladies global**: https://rladies.org/r-ladies-directory-form/




- Cristina Boboc
- Ana Maria Dobre
- Ines Teacă

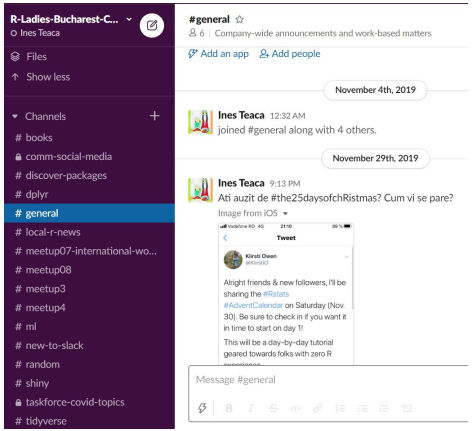
**Numele Tău**!
Your Name! ...

Share, blog or tweet: **#Rladies #RLadiesBucharest #rstats**

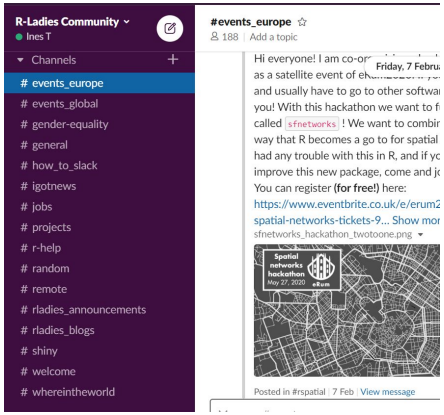
# Participă la discuțiile din slack

Bucharest

3 **R-Ladies-Bucharest-Community: http://tiny.cc/slack30days**




## Global Slack Rladies-Community




Share, blog or tweet: **#Rladies #RLadiesBucharest #rstats**

4


# SPREAD THE WORD!
# TELL YOUR FRIENDS

# agenda

- 19:15 Text mining with R, **Aurore Paligot**


**Be Safe! #wear-A-mask**

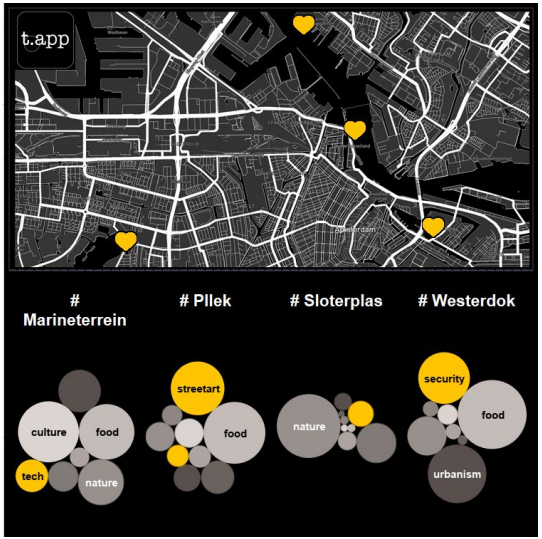
Bucharest

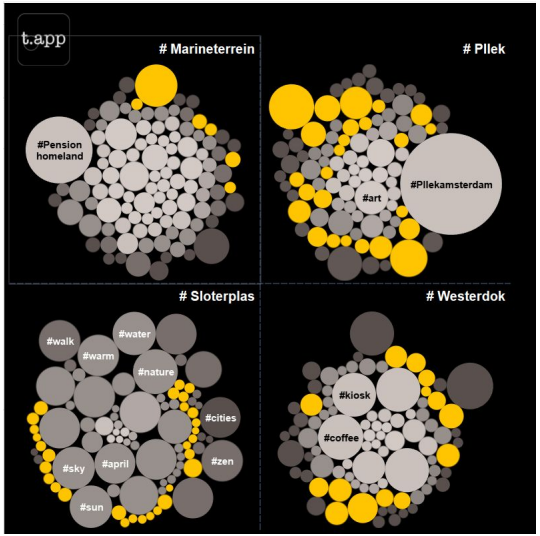
# Hello!

I'm Aurore Paligot and
I create stuff with data.
Visit my webpage here.

# Introduction :
# #Instagram The Dam project

Bucharest

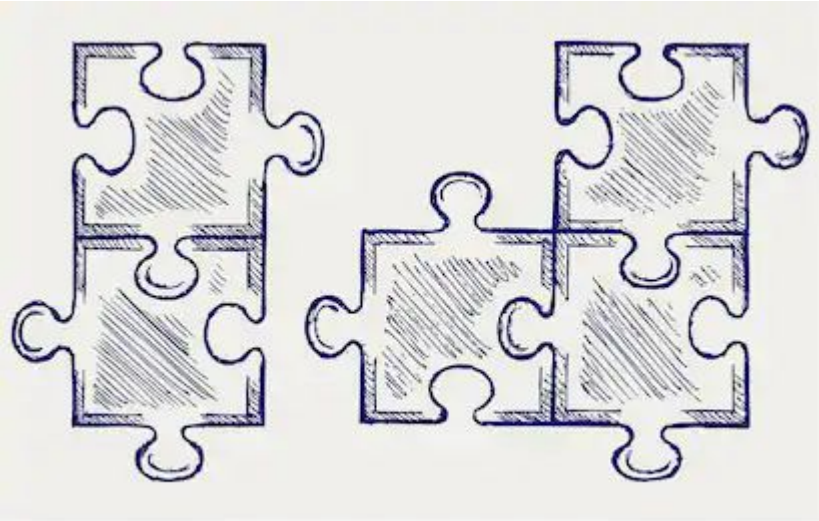



# #StrongTogether: the Covid19 crisis through the lenses of Instagramers

- Currently collecting data
- Amsterdam, Brussels, Bucharest, … ?
- Observation, documentation, research
- Open for collaborations !

# My Goal: take you through the research process

Bucharest



STEP 1 : Instagram Crawling

STEP 2 : Hashtags Frequency

STEP 3 : Hashtags Evolution

# STEP 1 : InstaCrawleR




*"**InstaCrawlR** is a collection of R scripts that can be used to **crawl public Instagram data without the need to have access to the official API**. Its functionality is limited compared to what is possible using the official API. However, it seems to be the only option for non-developers to gather and analyze Instagram data."*

Author : InstaCrawleR de Jonas Schröder (University of Mannheim, July 2018)

# 1.1: jsonReader.R

Bucharest

**Get InstaData using geolocalized (or any other) #hashtags**

*#amsterdam*

*#adam*

*#baneasa*

*#mokum*

*#lipscani*

*#amsterdamcentral*

*#bucharest*

# Open jsonReader.R


*Choose a hashtag*

```
#-------------------------------------------------------------
#Download JSON File from Instagram for a specific Hashtag
#-------------------------------------------------------------
hashtag <- "bucharest"
```

*Run the Script*

```

#Start the Madness
extractInfo(index)

```

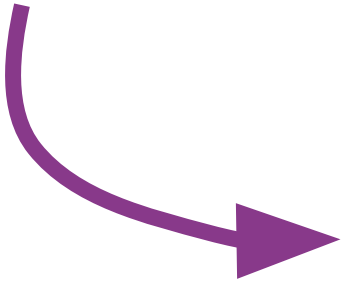
*Name and save your file*

```
filename <- str_glue("Bucharest.csv")
write.csv(table, filename, fileEncoding = "UTF-8")
```

# 1.2. Clean and anonymize data

*What data do you need? For which purpose?*

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| | | ID | Post_URL | Img_URL | Likes | Owner |
| 1 | | CGjAztMplmW | http://instagram.com/p/CGjAztMplmW | https://scontent-bru2-1.cdninstagram.com/v/t51.2885-15/e35/122044617_191727872504399_4( | 2 | 7094695825 |
| 2 | | CGi__0bpFb5 | http://instagram.com/p/CGi__0bpFb5 | https://scontent-bru2-1.cdninstagram.com/v/t51.2885-15/e35/122012113_425209508467887_5( | 3 | 7094695825 |
| 3 | | CGi_m_cB_Nm | http://instagram.com/p/CGi_m_cB_Nm | https://scontent-bru2-1.cdninstagram.com/v/t51.2885-15/e35/p1080x1080/121806171_6374195 | 1 | 5651260764 |
| 4 | | CGi-grVnR8L | http://instagram.com/p/CGi-grVnR8L | https://scontent-bru2-1.cdninstagram.com/v/t51.2885-15/e35/s1080x1080/122019583_3419595 | 1 | 36892690096 |
| 5 | | CGi9cpLJr0O | http://instagram.com/p/CGi9cpLJr0O | https://scontent-bru2-1.cdninstagram.com/v/t51.2885-15/e35/121981485_843030196439769_5( | 0 | 38027929492 |
| 6 | | CGi5JWIpcMW | http://instagram.com/p/CGi5JWIpcMW | https://scontent-bru2-1.cdninstagram.com/v/t51.2885-15/e35/s1080x1080/121970430_3520439 | 6 | 582727360 |
| 7 | | CGi4aoHBB6i | http://instagram.com/p/CGi4aoHBB6i | https://scontent-bru2-1.cdninstagram.com/v/t51.2885-15/e35/s1080x1080/122081896_1644606 | 21 | 14323258420 |



| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | X | Likes | Text | Date | |
| 2 | 1 | 2 | FondÃ© en 18 | 20-10-20 | |
| 3 | 2 | 3 | "Thankfulness | 20-10-20 | |
| 4 | 3 | 1 | #smile #gym # | 20-10-20 | |
| 5 | 4 | 1 | #bucharest | 20-10-20 | |
| 6 | 5 | 0 | #tnb #nationa | 20-10-20 | |
| 7 | 6 | 6 | River #romani | 20-10-20 | |
| 8 | 7 | 21 | Teschio nell'e | 20-10-20 | |

# 1.3. HashtagExtractor.R

Calculates the frequency of the #hashtags contained in the captions

*Locate your file*

```
18 htags <- data.frame()
19 data <- read_excel("Data/R Ladies Data/Bucharest.xlsx") #locate your data
20
```

*Check the column number (text)*

```
27 for(i in 1:maxrows){
28   text[i] <- as.character(data[i,2]) #the column number is hard coded
29   htemp <- str_extract_all(text[i], "#\\S+", TRUE)
```

*Run the script and save your file*

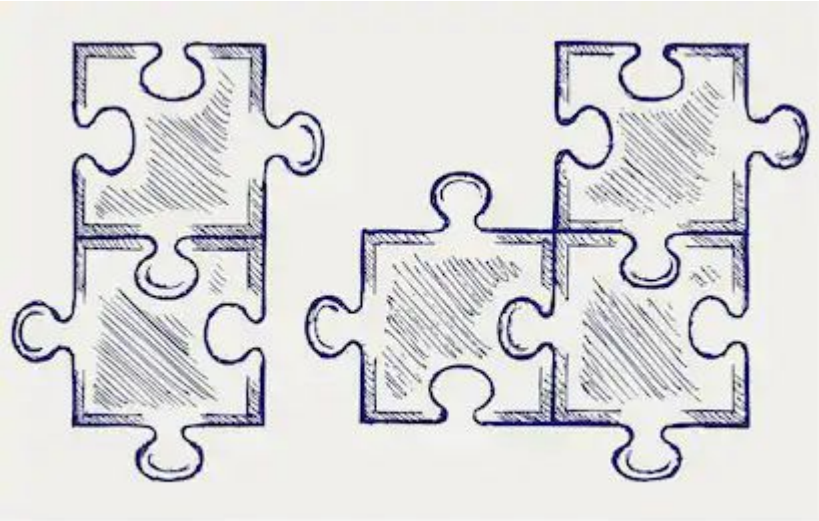
```
41
42 write.csv(df_htags, "Bucharest_corona_sort.csv") #sorted list with frequencies
43
```

# Create a corona variable manually

Take your last file “Bucharest_corona_sort.csv” and identify all the hashtags that are related to the coronavirus topic

| Hashtags | Frequency | Corona |
|---|---|---|
| #love | 1918 | no |
| #iamsterdam | 1806 | no |
| #amsterdam<U+0001F1F3><U+00 | 1731 | no |
| #art | 1692 | no |
| #photooftheday | 1680 | no |
| #travelphotography | 1649 | no |
| #instagood | 1498 | no |
| #Amsterdam | 1403 | no |
| #amsterdamworld | 1385 | no |
| #europe | 1368 | no |
| #amsterdamlife | 1363 | no |
| #amsterdamcanals | 1353 | no |
| #streetphotography | 1345 | no |
| #fashion | 1238 | no |
| #travelgram | 1088 | no |
| #picoftheday | 1055 | no |
| #visitamsterdam | 1041 | no |
| #igersamsterdam | 1030 | no |
| #amsterdamview | 990 | no |
| #stayhome | 970 | yes |
| #corona | 955 | yes |

# Now that we have some data, let's explore !



**STEP 1 : Instagram Crawling**

**STEP 2 : Hashtags Frequency**

**STEP 3 : Hashtags Evolution**

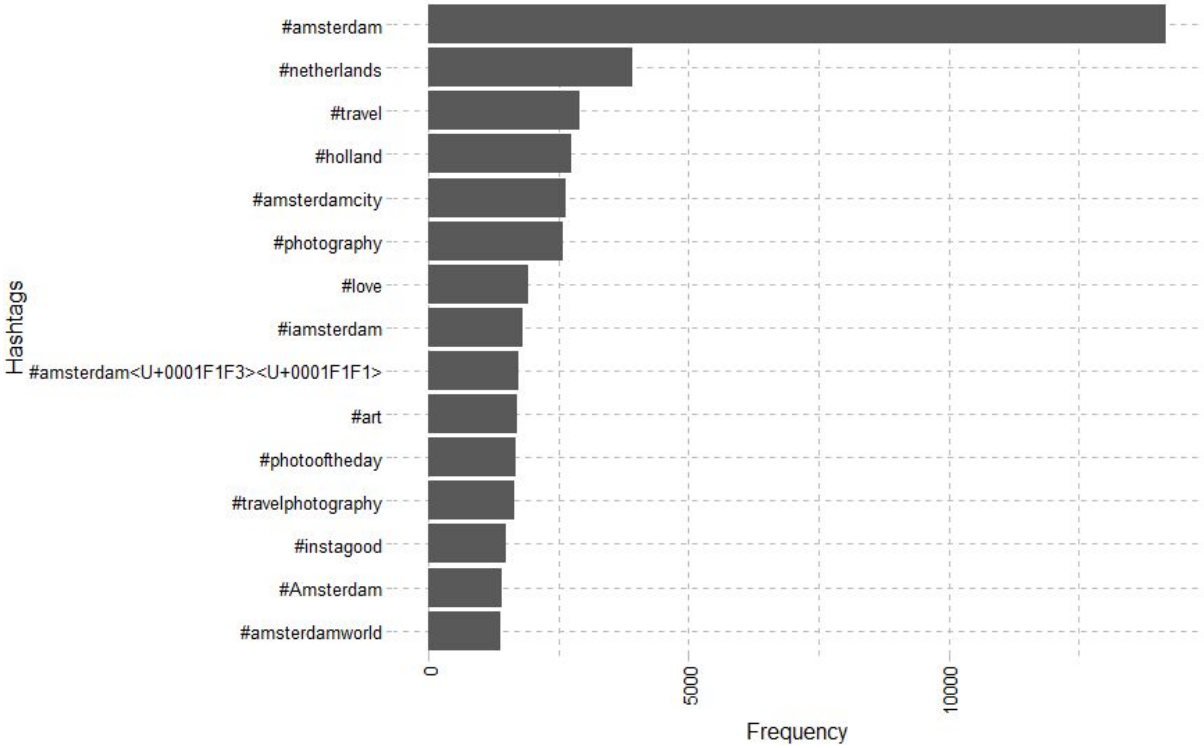
Bucharest

# Explore_Amsterdam.R


*Load your Frequency file and start visualizing some data*

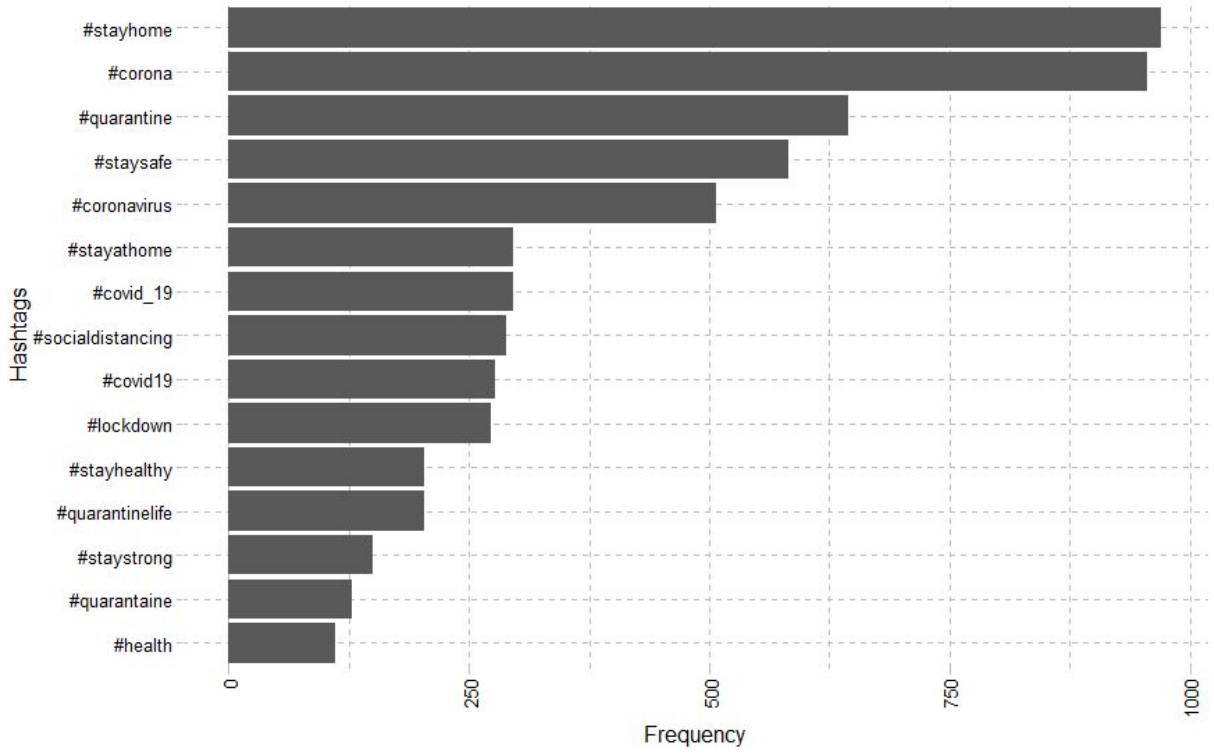
```
24 ########################## STEP 2 : Hashtags Frequency ######################
25
26 data <- read_excel("Data/R Ladies Data/Amsterdam_Frequencies.xlsx")
27
28 #1. explore & format data
29
30 summary(data)
31
```

# STEP 2 : Hashtags Frequency




**Top 15 Hashtags of Amsterdam**

# STEP 2 : Hashtags Frequency



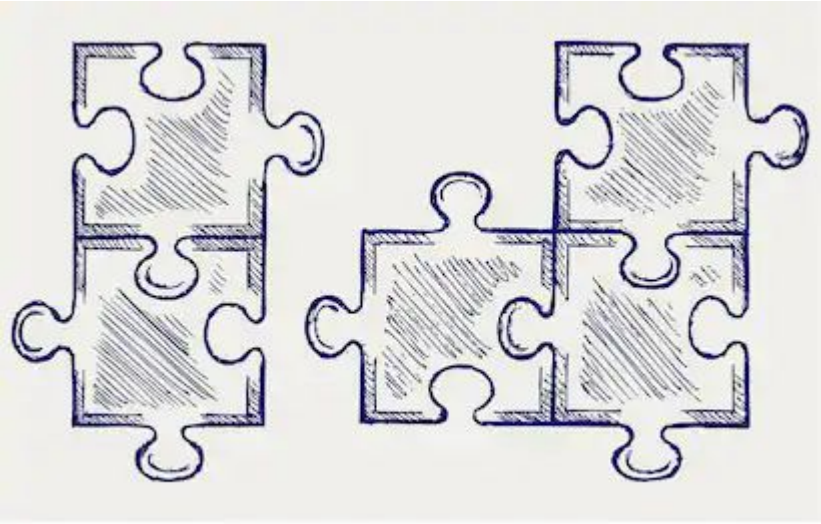

**Top 15 Hashtags with Corona**

# STEP 2 : Hashtags Frequency


**Proportion of hashtags with a corona related topic compared to whole dataset**

# Data Evolution

Bucharest



**STEP 1 : Instagram Crawling**

**STEP 2 : Hashtags Frequency**

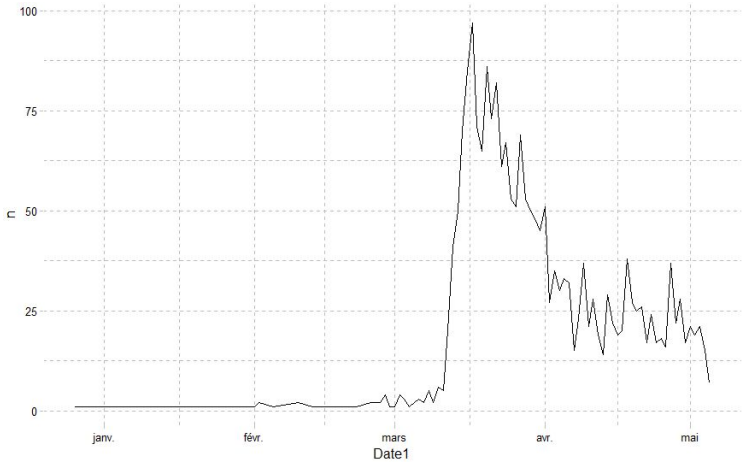
**STEP 3 : Hashtags Evolution**

# Explore_Amsterdam.R

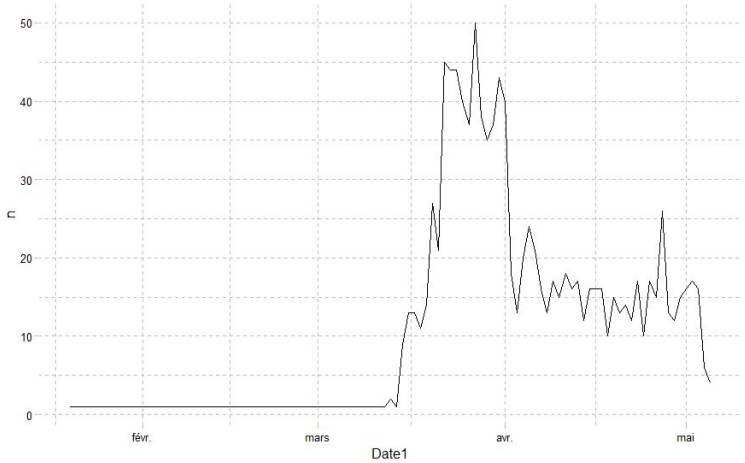
*Load your anonymized file and start visualizing some data*

```
79 ########################## STEP 3 : Hashtags Evolution ####################
80
81 adata <- read_excel("Data/R Ladies Data/Amsterdam_AnonymisedData.xlsx")
82
83 summary(adata)
84
85 #1. Identify captions that contain the hashtag or word corona
86
87 corona_vector <- stringr::str_detect(adata$captions, "corona")
88
```

# STEP 3 : Hashtags Evolution




#corona




#stayhome

*What are your project ideas?*
*What do you want to observe, create or analyse with text, data, and social media?*

Bucharest

# ending

Thank you | Multumesc | Merci

# Little guide

Bucharest

## EDIT IN GOOGLE SLIDES

Go to the ***File*** menu and select ***Make a copy***.

You will get a copy of this document on your Google Drive and will be able to edit, add or delete slides.

## COLOR PALETTE

| #181818 | #D3D3D3 | #88398A | #FFFFF | #562457 |
|---|---|---|---|---|