

**DATA SCIENCE**

**SEMESTER 1, 2020/2021**

**CSC 3305, SECTION: 01**

**LECTURER: DR. RAINI HASSAN**

**Prepared by:**

|  |  |
| --- | --- |
| Name | Matric number |
| TASNIM RAFIA | 1725826 |
| HASAN TANVEER MAHMOOD | 1725413 |
| BILLAH SYED MASHKUR | 1723387 |
| K. M. ZUBAIR | 1722931 |

Date: 16- December - 2020

**#Question 1 (R Code)**

> library(readxl)

> library(tidyverse)

>

> exc <- read\_excel("C:/Users/Mashkur/Desktop/Flowers Dataset.xlsx")

> exc

# A tibble: 20 x 1

Begonia

<chr>

1 Carnation

2 RoSe

3 Begonia

4 Carnation

5 tulip

6 ROSE

7 begonia

8 carnation

9 rose

10 tulip

11 begonia

12 TULIP

13 BEGONIA

14 beGONia

15 caRNAtiON

16 CARNATION

17 carnation

18 carnation

19 BEGONIA

20 TULIP

> #removing extra white spaces

> exc$Begonia <- gsub(' ','',exc$Begonia)

> exc

# A tibble: 20 x 1

Begonia

<chr>

1 Carnation

2 RoSe

3 Begonia

4 Carnation

5 tulip

6 ROSE

7 begonia

8 carnation

9 rose

10 tulip

11 begonia

12 TULIP

13 BEGONIA

14 beGONia

15 caRNAtiON

16 CARNATION

17 carnation

18 carnation

19 BEGONIA

20 TULIP

> exc$Begonia

[1] "Carnation" "RoSe" "Begonia" "Carnation" "tulip" "ROSE"

[7] "begonia" "carnation" "rose" "tulip" "begonia" "TULIP"

[13] "BEGONIA" "beGONia" "caRNAtiON" "CARNATION" "carnation" "carnation"

[19] "BEGONIA" "TULIP"

> #Now lets deal with inconsistant capitalization

> result <- str\_to\_sentence(exc$Begonia)

> result

[1] "Carnation" "Rose" "Begonia" "Carnation" "Tulip" "Rose"

[7] "Begonia" "Carnation" "Rose" "Tulip" "Begonia" "Tulip"

[13] "Begonia" "Begonia" "Carnation" "Carnation" "Carnation" "Carnation"

[19] "Begonia" "Tulip"

> #Now write the result list and update it back to the excel file by replacing it

> library("writexl")

> #Create a dataframe to make it organised and write it

> df <- data.frame(flowers = result)

> df

flowers

1 Carnation

2 Rose

3 Begonia

4 Carnation

5 Tulip

6 Rose

7 Begonia

8 Carnation

9 Rose

10 Tulip

11 Begonia

12 Tulip

13 Begonia

14 Begonia

15 Carnation

16 Carnation

17 Carnation

18 Carnation

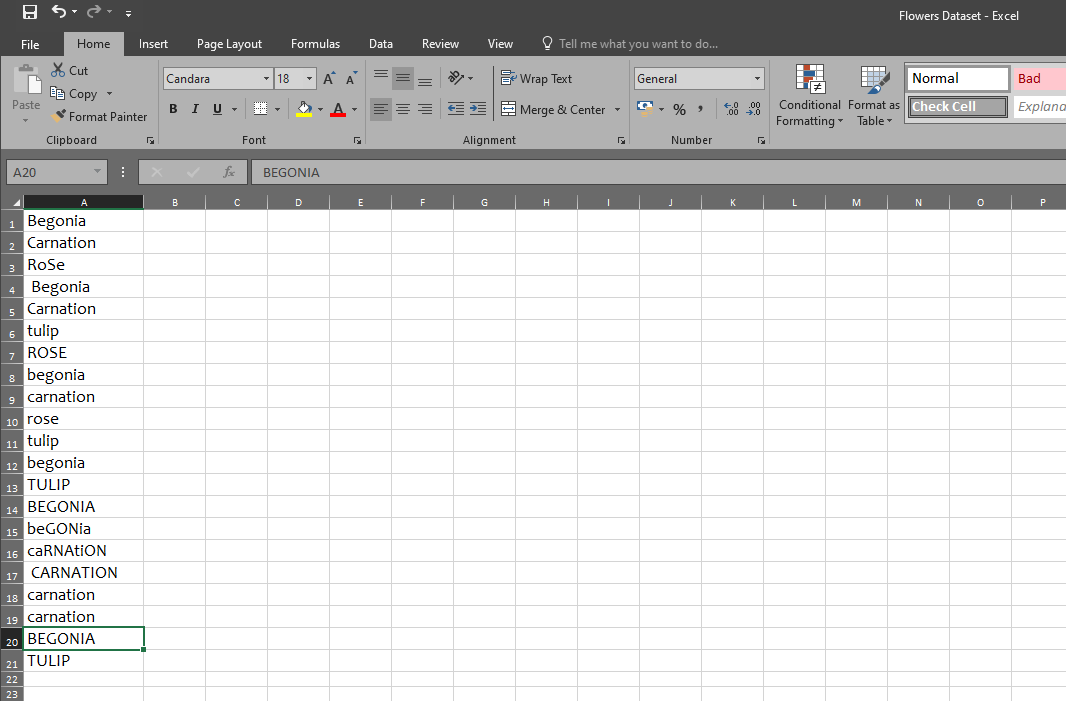
19 Begonia

20 Tulip

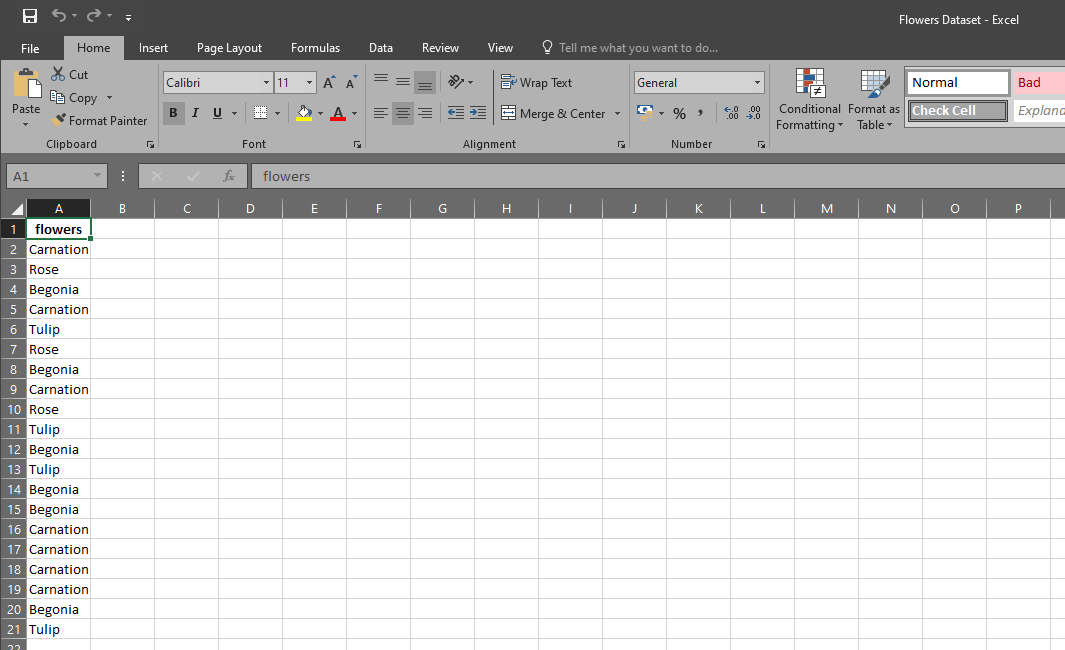
> #Finally write and update the data

> write\_xlsx(df,"C:\\Users\\Mashkur\\Desktop\\Flowers Dataset.xlsx")

**#Before Data Wrangling:**



**#After Data Wrangling:**



**#Question 2 (Map-reduce)**

N.B: Zoom in to view clearly.

