Organic Waste Wrangling & Analysis

Joaquín Rodrigo Ponce de León Conconi

6/7/2021

I decided to create a list of ten different organic products, and measured the weight of each product that became organic waste in grams for fifteen weeks. The data was stored on pdf, just to put my data wrangling skills to the challenge. The wrangling process is as follows:

WRANGLING

-STEP 1:

- Let's call the libraries needed for the wrangling process:

```
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------------------- tidyverse 1
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.0
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ---------------------------------------------------------------------- tidyverse_confli
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(pdftools)
```

```
## Using poppler version 20.12.1
```

-STEP 2:

- Let's upload the pdf file:

```
pdf <- pdf_text("Organic waste data .pdf") %>% read_lines()
str(pdf)
```

```
##  chr [1:177] "Producto Desperdicios en 1ra semana (g)   Desperdicios en 2da semana (g)" ...
```

```
pdf
```

```
##   [1] "Producto Desperdicios en 1ra semana (g)   Desperdicios en 2da semana (g)"
##   [2] "Huevo       0                             0"
##   [3] ""
##   [4] "Tortillas   0                             42.78"
##   [5] ""
##   [6] "Tomates     0                             0"
##   [7] ""
##   [8] "Lechuga 0                                 278.51"
##   [9] ""
##  [10] "Perejil     0                             161.66"
##  [11] ""
##  [12] "Cilantro    10.8                          0"
##  [13] ""
##  [14] "Plátano     0                             0"
##  [15] ""
##  [16] "Naranjas 0                                0"
##  [17] ""
##  [18] "Limones 127.65                            107.23"
##  [19] ""
##  [20] "Queso       0                             0"
##  [21] ""
##  [22] ""
##  [23] ""
##  [24] ""
##  [25] "Desperdicios en 3ra semana (g)      Desperdicios en 4ta semana (g)"
##  [26] "50.45                               0"
##  [27] ""
##  [28] "0                                   0"
##  [29] ""
##  [30] "0                                   0"
##  [31] ""
##  [32] "0                                   0"
##  [33] ""
##  [34] "0                                   0"
##  [35] ""
##  [36] "0                                   0"
##  [37] ""
##  [38] "0                                   0"
##  [39] ""
##  [40] "0                                   0"
##  [41] ""
##  [42] "0                                   0"
##  [43] ""
##  [44] "0                                   0"
##  [45] ""
##  [46] "Desperdicios en 5ta semana (g)      Desperdicios en 6ta semana (g)"
##  [47] "0                                   0"
##  [48] ""
##  [49] "47.31                               0"
##  [50] ""
##  [51] "0                                   0"
##  [52] ""
##  [53] "0                                   33.9"
##  [54] ""
##  [55] "0                                   118.26"
##  [56] ""
##  [57] "0                                   0"
##  [58] ""
##  [59] "0                                   0"
##  [60] ""
##  [61] "0                                   0"
##  [62] ""
##  [63] "0                                   0"
##  [64] ""
##  [65] "94.12                               0"
##  [66] ""
##  [67] ""
##  [68] ""
##  [69] "Desperdicios en la 7ma semana (g)   Desperdicios en la 8va semana (g)"
##  [70] "0                                   0"
##  [71] ""
##  [72] "0                                   0"
##  [73] ""
##  [74] "0                                   0"
##  [75] ""
##  [76] "0                                   0"
##  [77] ""
##  [78] "0                                   0"
##  [79] ""
##  [80] "0                                   0"
##  [81] ""
##  [82] "0                                   0"
##  [83] ""
##  [84] "0                                   0"
##  [85] ""
##  [86] "0                                   0"
##  [87] ""
##  [88] "0                                   0"
##  [89] ""
##  [90] "Desperdicios en la 9na semana (g)    Desperdicios en la 10ma semana (g)"
##  [91] "0                                    0"
##  [92] ""
##  [93] "0                                    0"
##  [94] ""
##  [95] "0                                    0"
##  [96] ""
##  [97] "0                                    0"
##  [98] ""
##  [99] "0                                    0"
## [100] ""
## [101] "0                                    0"
## [102] ""
## [103] "0                                    6.2"
## [104] ""
## [105] "0                                    0"
## [106] ""
## [107] "0                                    0"
## [108] ""
## [109] "0                                    0"
## [110] ""
## [111] ""
## [112] ""
## [113] "Desperdicios en la 11va semana (g)       Desperdicios en la 12va semana (g)"
## [114] "0                                        0"
## [115] ""
## [116] "0                                        0"
## [117] ""
## [118] "0                                        0"
## [119] ""
## [120] "30.57                                    284.37"
## [121] ""
## [122] "0                                        154.21"
## [123] ""
## [124] "123.8                                    0"
## [125] ""
## [126] "0                                        0"
## [127] ""
## [128] "0                                        0"
## [129] ""
## [130] "111.81                                   0"
## [131] ""
## [132] "0                                        0"
## [133] ""
## [134] "Desperdicios en la 13va semana (g)   Desperdicios en la 14va semana (g)"
## [135] "0                                    0"
## [136] ""
## [137] "0                                    0"
## [138] ""
## [139] "0                                    0"
## [140] ""
## [141] "0                                    0"
## [142] ""
## [143] "0                                    0"
## [144] ""
## [145] "0                                    0"
## [146] ""
## [147] "0                                    0"
## [148] ""
## [149] "0                                    0"
## [150] ""
## [151] "0                                    0"
## [152] ""
## [153] "0                                    0"
## [154] ""
## [155] ""
## [156] ""
## [157] ""
## [158] "Desperdicios en la 15va semana (g)"
## [159] "0"
## [160] ""
## [161] "0"
## [162] ""
## [163] "0"
## [164] ""
## [165] "0"
## [166] ""
## [167] "0"
## [168] ""
## [169] "0"
## [170] ""
## [171] "0"
## [172] ""
## [173] "0"
## [174] ""
## [175] "138.32"
## [176] ""
## [177] "0"
```

-STEP 3:

- We see that the rows are separated by "", so let's filter those out and save the result in pdf_t:

```
pdf_t <- pdf[str_detect(pdf, "\\w|\\d+")]
pdf_t
```

```
##  [1] "Producto Desperdicios en 1ra semana (g)   Desperdicios en 2da semana (g)"
##  [2] "Huevo       0                             0"
##  [3] "Tortillas   0                             42.78"
##  [4] "Tomates     0                             0"
##  [5] "Lechuga 0                                 278.51"
##  [6] "Perejil     0                             161.66"
##  [7] "Cilantro    10.8                          0"
##  [8] "Plátano     0                             0"
##  [9] "Naranjas 0                                0"
## [10] "Limones 127.65                            107.23"
## [11] "Queso       0                             0"
## [12] "Desperdicios en 3ra semana (g)      Desperdicios en 4ta semana (g)"
## [13] "50.45                               0"
## [14] "0                                   0"
## [15] "0                                   0"
## [16] "0                                   0"
## [17] "0                                   0"
## [18] "0                                   0"
## [19] "0                                   0"
## [20] "0                                   0"
## [21] "0                                   0"
## [22] "0                                   0"
## [23] "Desperdicios en 5ta semana (g)      Desperdicios en 6ta semana (g)"
## [24] "0                                   0"
## [25] "47.31                               0"
## [26] "0                                   0"
## [27] "0                                   33.9"
## [28] "0                                   118.26"
## [29] "0                                   0"
## [30] "0                                   0"
## [31] "0                                   0"
## [32] "0                                   0"
## [33] "94.12                               0"
## [34] "Desperdicios en la 7ma semana (g)   Desperdicios en la 8va semana (g)"
## [35] "0                                   0"
## [36] "0                                   0"
## [37] "0                                   0"
## [38] "0                                   0"
## [39] "0                                   0"
## [40] "0                                   0"
## [41] "0                                   0"
## [42] "0                                   0"
## [43] "0                                   0"
## [44] "0                                   0"
## [45] "Desperdicios en la 9na semana (g)    Desperdicios en la 10ma semana (g)"
## [46] "0                                    0"
## [47] "0                                    0"
## [48] "0                                    0"
## [49] "0                                    0"
## [50] "0                                    0"
## [51] "0                                    0"
## [52] "0                                    6.2"
## [53] "0                                    0"
## [54] "0                                    0"
## [55] "0                                    0"
## [56] "Desperdicios en la 11va semana (g)       Desperdicios en la 12va semana (g)"
## [57] "0                                        0"
## [58] "0                                        0"
## [59] "0                                        0"
## [60] "30.57                                    284.37"
## [61] "0                                        154.21"
## [62] "123.8                                    0"
## [63] "0                                        0"
## [64] "0                                        0"
## [65] "111.81                                   0"
## [66] "0                                        0"
## [67] "Desperdicios en la 13va semana (g)   Desperdicios en la 14va semana (g)"
## [68] "0                                    0"
## [69] "0                                    0"
## [70] "0                                    0"
## [71] "0                                    0"
## [72] "0                                    0"
## [73] "0                                    0"
## [74] "0                                    0"
## [75] "0                                    0"
## [76] "0                                    0"
## [77] "0                                    0"
## [78] "Desperdicios en la 15va semana (g)"
## [79] "0"
## [80] "0"
## [81] "0"
## [82] "0"
## [83] "0"
## [84] "0"
## [85] "0"
## [86] "0"
## [87] "138.32"
## [88] "0"
```

-STEP 4:

- We now have all the needed data, but in an inadequate format. However, we also we also see that there are some column names that have the word “Desperdicios”, which means wastes in Spanish. Let’s show the column names:

```
col_names <- pdf_t[str_detect(pdf_t, "Desperdicios")]
col_names <- str_split(col_names, "\\s{2,}") %>% unlist() #Split if there are 2 spaces or more and unli
col_names
```

```
##  [1] "Producto Desperdicios en 1ra semana (g)"
##  [2] "Desperdicios en 2da semana (g)"
##  [3] "Desperdicios en 3ra semana (g)"
##  [4] "Desperdicios en 4ta semana (g)"
##  [5] "Desperdicios en 5ta semana (g)"
##  [6] "Desperdicios en 6ta semana (g)"
##  [7] "Desperdicios en la 7ma semana (g)"
##  [8] "Desperdicios en la 8va semana (g)"
##  [9] "Desperdicios en la 9na semana (g)"
## [10] "Desperdicios en la 10ma semana (g)"
## [11] "Desperdicios en la 11va semana (g)"
## [12] "Desperdicios en la 12va semana (g)"
## [13] "Desperdicios en la 13va semana (g)"
## [14] "Desperdicios en la 14va semana (g)"
## [15] "Desperdicios en la 15va semana (g)"
```

-STEP 5:

- Based on the data in STEP 4, we now now there are 15 weeks. However, we want the column names to be in English not Spanish, so let’s create a variable (raw_dig) to store data without the column names:

```
raw_dig <- pdf_t[!str_detect(pdf_t, "Desperdicios")] #raw digits
raw_dig
```

```
##  [1] "Huevo       0                             0"
##  [2] "Tortillas   0                             42.78"
##  [3] "Tomates     0                             0"
##  [4] "Lechuga 0                                 278.51"
##  [5] "Perejil     0                             161.66"
##  [6] "Cilantro    10.8                          0"
##  [7] "Plátano     0                             0"
##  [8] "Naranjas 0                                0"
##  [9] "Limones 127.65                            107.23"
## [10] "Queso       0                             0"
## [11] "50.45                               0"
## [12] "0                                   0"
## [13] "0                                   0"
## [14] "0                                   0"
## [15] "0                                   0"
## [16] "0                                   0"
## [17] "0                                   0"
## [18] "0                                   0"
## [19] "0                                   0"
## [20] "0                                   0"
## [21] "0                                   0"
## [22] "47.31                               0"
## [23] "0                                   0"
## [24] "0                                   33.9"
## [25] "0                                   118.26"
## [26] "0                                   0"
## [27] "0                                   0"
## [28] "0                                   0"
## [29] "0                                   0"
## [30] "94.12                               0"
## [31] "0                                   0"
## [32] "0                                   0"
## [33] "0                                   0"
## [34] "0                                   0"
## [35] "0                                   0"
## [36] "0                                   0"
## [37] "0                                   0"
## [38] "0                                   0"
## [39] "0                                   0"
## [40] "0                                   0"
## [41] "0                                    0"
## [42] "0                                    0"
## [43] "0                                    0"
## [44] "0                                    0"
## [45] "0                                    0"
## [46] "0                                    0"
## [47] "0                                    6.2"
## [48] "0                                    0"
## [49] "0                                    0"
## [50] "0                                    0"
## [51] "0                                        0"
## [52] "0                                        0"
## [53] "0                                        0"
## [54] "30.57                                    284.37"
## [55] "0                                        154.21"
## [56] "123.8                                    0"
## [57] "0                                        0"
## [58] "0                                        0"
## [59] "111.81                                   0"
## [60] "0                                        0"
## [61] "0                                    0"
## [62] "0                                    0"
## [63] "0                                    0"
## [64] "0                                    0"
## [65] "0                                    0"
## [66] "0                                    0"
## [67] "0                                    0"
## [68] "0                                    0"
## [69] "0                                    0"
## [70] "0                                    0"
## [71] "0"
## [72] "0"
## [73] "0"
## [74] "0"
## [75] "0"
## [76] "0"
## [77] "0"
## [78] "0"
## [79] "138.32"
## [80] "0"
```

-STEP 6:

- From STEP 5 we see that the week columns are stored in pairs except for week 15, and since the data for each week is stored in 10 rows we can store the data for each one of the in different variables to facilitate the wrangling:

```
raw_dig12 <- raw_dig[1:10]#First and second week with words
raw_dig34 <- raw_dig[11:20]#3rd and 4th
raw_dig56 <- raw_dig[21:30]#5th and 6th
raw_dig78 <- raw_dig[31:40]#7th and 8th
raw_dig910 <- raw_dig[41:50]#9th and 10th
raw_dig1112 <- raw_dig[51:60]#11th and 12th
raw_dig1314 <- raw_dig[61:70]#13th and 14th
raw_dig15 <- raw_dig[71:80]#15th
```

-STEP 7:

- A way to confirm the data storage was successful is by looking at the length of each variable, which should be 10 for all:

```
sapply(list(raw_dig12, raw_dig34, raw_dig56, raw_dig78, raw_dig910, raw_dig1112, raw_dig1314, raw_dig15
```

```
## [1] 10 10 10 10 10 10 10 10
```

-STEP 8:

- Since weeks 1 and 2 are attached to the product names (STEP 5) we need to split the data and arrange it properly. n_raw_dig12 contains the numeric data for weeks 1 and 2:

```
s_raw_dig12 <- str_split(raw_dig12, "(?<=[a-zA-Z])\\s*(?=[0-9])") %>% unlist()#split the waste elements

#See waste names
s_raw_dig12[str_detect(s_raw_dig12, "[a-zA-Z]")]
```

```
##  [1] "Huevo"     "Tortillas" "Tomates"   "Lechuga"   "Perejil"   "Cilantro"
##  [7] "Plátano"   "Naranjas"  "Limones"   "Queso"
```

```
#Translate manually Spanish names to English names
waste_n <- c("Egg", "Tortillas", "Tomatoes", "Lettuce", "Parsley", "Cilantro", "Bananas", "Oranges", "L
#Vector with waste numbers
n_raw_dig12 <- s_raw_dig12[!str_detect(s_raw_dig12, "[a-zA-Z]")] %>% str_split("\\s+")
n_raw_dig12
```

```
## [[1]]
## [1] "0" "0"
##
## [[2]]
## [1] "0"     "42.78"
##
## [[3]]
## [1] "0" "0"
##
## [[4]]
## [1] "0"      "278.51"
##
## [[5]]
## [1] "0"      "161.66"
##
## [[6]]
## [1] "10.8" "0"
##
## [[7]]
## [1] "0" "0"
##
## [[8]]
## [1] "0" "0"
##
## [[9]]
## [1] "127.65" "107.23"
##
## [[10]]
## [1] "0" "0"
```

-STEP 9:

- We see the 10 rows for both weeks, so we apply a for loop to extract the values using indices [[1-10]] (because it's a list) and [1-2] (because there are 2 columns). This way, we're going to be able to organize the data from weeks 1 and 2 into 2 different data frames (week1 and week2):

```
week1 <- c()
week2 <- c()
c <- 1
for (i in 1:10){
  week1 <- append(week1, n_raw_dig12[[i]][c])
  c <- 2
  week2 <- append(week2, n_raw_dig12[[i]][c])
  c <- 1
}
week1 <- data.frame("Name"=waste_n, "n"=as.numeric(week1), "Week"=1)#create dataframe for week1
```

```
week2 <- data.frame("Name"=waste_n, "n"=as.numeric(week2), "Week"=2)#create dataframe for week2
str(week1)
```

```
## 'data.frame':    10 obs. of  3 variables:
##  $ Name: chr  "Egg" "Tortillas" "Tomatoes" "Lettuce" ...
##  $ n   : num  0 0 0 0 0 ...
##  $ Week: num  1 1 1 1 1 1 1 1 1 1
```

```
str(week2)
```

```
## 'data.frame':    10 obs. of  3 variables:
##  $ Name: chr  "Egg" "Tortillas" "Tomatoes" "Lettuce" ...
##  $ n   : num  0 42.8 0 278.5 161.7 ...
##  $ Week: num  2 2 2 2 2 2 2 2 2 2
```

-STEP 10:

- Now that we have both dataframes for week1 and week2, we can proceed to wrangle the data for the rest of the weeks by creating a function "wrangle" that splits the data into weekn and week(n+1) for the raw_dig## variables:

```
wrangle <- function(raw, waste_n, wn){
  s <- str_split(raw,"\\s+")
  week_a <- c()
  week_b <- c()
  c <- 1
  for (i in 1:10){
    week_a <- append(week_a, s[[i]][c])
    c <- 2
    week_b <- append(week_b, s[[i]][c])
    c <- 1
    }
  week_a <- data.frame("Name"=waste_n, "n"=as.numeric(week_a), "Week"=wn[1])#create dataframe for week1
  week_b <- data.frame("Name"=waste_n, "n"=as.numeric(week_b), "Week"=wn[2])#create dataframe for week2
  return(list(week_a, week_b))
}
```

- Let us apply the function for all the weeks and make a dataframe for week 15 separately:

```
week3 <- wrangle(raw_dig34, waste_n, c(3,4))[[1]]#dataframe for week 3
week4 <- wrangle(raw_dig34, waste_n, c(3,4))[[2]]#dataframe for week 4

week5 <- wrangle(raw_dig56, waste_n, c(5,6))[[1]]#dataframe for week 5
week6 <- wrangle(raw_dig56, waste_n, c(5,6))[[2]]#dataframe for week 6

week7 <- wrangle(raw_dig78, waste_n, c(7,8))[[1]]#dataframe for week 7
week8 <- wrangle(raw_dig78, waste_n, c(7,8))[[2]]#dataframe for week 8

week9 <- wrangle(raw_dig910, waste_n, c(9,10))[[1]]#dataframe for week 9
week10 <- wrangle(raw_dig910, waste_n, c(9,10))[[2]]#dataframe for week 10
```

```
week11 <- wrangle(raw_dig1112, waste_n, c(11, 12))[[1]]#dataframe for week 11
week12 <- wrangle(raw_dig1112, waste_n, c(11, 12))[[2]]#dataframe for week 12

week13 <- wrangle(raw_dig1314, waste_n, c(13, 14))[[1]]#dataframe for week 13
week14 <- wrangle(raw_dig1314, waste_n, c(13, 14))[[2]]#dataframe for week 14

week15 <- data.frame("Name"=waste_n, "n"=as.numeric(raw_dig15), "Week"=15) #we don't have to apply the
```

-STEP 11:

- Now that we have all the waste from the different weeks, we can make a data frame that contains all the weeks vertically:

```
organic_waste <- rbind(week1, week2, week3, week4, week5, week6, week7, week8, week9, week10, week11, w
str(organic_waste)
```

```
## 'data.frame':    150 obs. of  3 variables:
##  $ Name: chr  "Egg" "Tortillas" "Tomatoes" "Lettuce" ...
##  $ n   : num  0 0 0 0 0 ...
##  $ Week: num  1 1 1 1 1 1 1 1 1 1 ...
```

-STEP 12:

- To confirm that organic_waste is the proper data frame, let us compare it with the Excel file:

```
library(readxl)

exl <- read_excel("Desperdicios.xlsx")
exl <- data.frame("Name"=waste_n, exl) %>% select(-Producto)
exl <- gather(exl, key = "Week", value = "n", -Name)
#Let's change the spanish names for weekly waste to week numbers
df <- exl %>% mutate(Week = recode(Week,
                             "Desperdicios.en.1ra.semana..g."=1,
                             "Desperdicios.en.2da.semana..g."=2,
                             "Desperdicios.en.3ra.semana..g."=3,
                             "Desperdicios.en.4ta.semana..g."=4,
                             "Desperdicios.en.5ta.semana..g."=5,
                             "Desperdicios.en.6ta.semana..g."=6,
                             "Desperdicios.en.la.7ma.semana..g."=7,
                             "Desperdicios.en.la.8va.semana..g."=8,
                             "Desperdicios.en.la.9na.semana..g."=9,
                             "Desperdicios.en.la.10ma.semana..g."=10,
                             "Desperdicios.en.la.11va.semana..g."=11,
                             "Desperdicios.en.la.12va.semana..g."=12,
                             "Desperdicios.en.la.13va.semana..g."=13,
                             "Desperdicios.en.la.14va.semana..g."=14,
                             "Desperdicios.en.la.15va.semana..g."=15)) #Let's change the week names to
df <- df %>% select(-Week) %>% cbind("Week"=df$Week)#Let's place Week to the left for df and organic_wa
identical(df,organic_waste) #Let's see if they are identical
```

```
## [1] TRUE
```

- Wee see that in fact they're identical. Thus, we can proceed with the analysis:

ANALYSIS

-STEP 1:

- Let's see what waste and in which week we had the highest mass (n represents the amount of grams):

```
#Let's see what waste and in which week we had the highest mass (n is the mass of the organic waste in
organic_waste %>% filter(n==max(n))
```

```
##      Name      n Week
## 1 Lettuce 284.37   12
```

- Lettuce on week 12 was the organic waste that was heaviest.

-STEP 2:

- Let's calculate the average organic waste per week:

```
organic_waste %>% group_by(Week) %>% summarise(mean(n))
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

```
## # A tibble: 15 x 2
##     Week ‘mean(n)‘
##    <dbl>     <dbl>
##  1     1     13.8
##  2     2     59.0
##  3     3      5.04
##  4     4      0
##  5     5     14.1
##  6     6     15.2
##  7     7      0
##  8     8      0
##  9     9      0
## 10    10      0.62
## 11    11     26.6
## 12    12     43.9
## 13    13      0
## 14    14      0
## 15    15     13.8
```

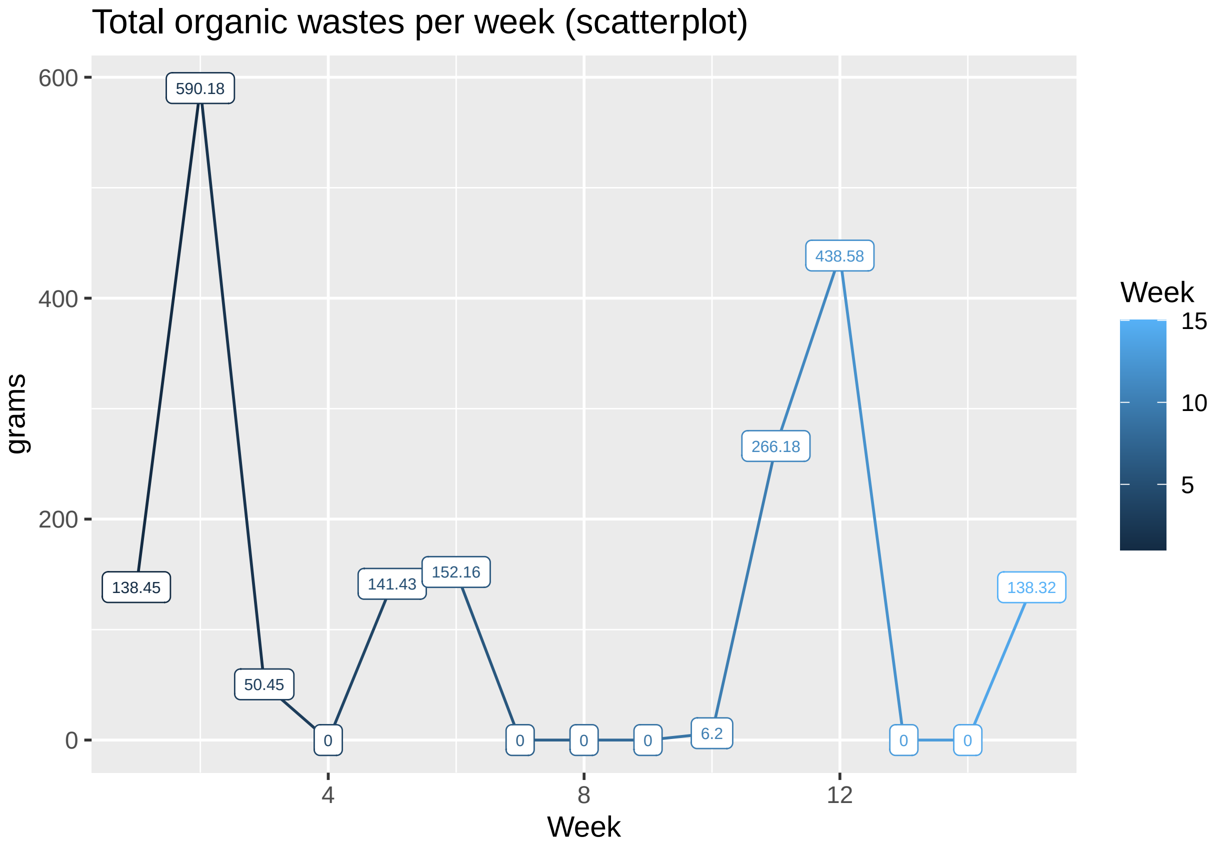
- Wee see that on average week 2 had the most wastes.

-STEP 3:

- Let's see how much waste was produced per week:

```
#Let's make a scatter plot to visualize better
organic_waste %>% group_by(Week) %>% summarise(S=sum(n)) %>%
  ggplot(aes(Week,S, col=Week))+geom_point()+
  ggtitle("Total organic wastes per week (scatterplot)")+
  ylab("grams")+
  geom_line()+geom_label(aes(label=S), size = 2)
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

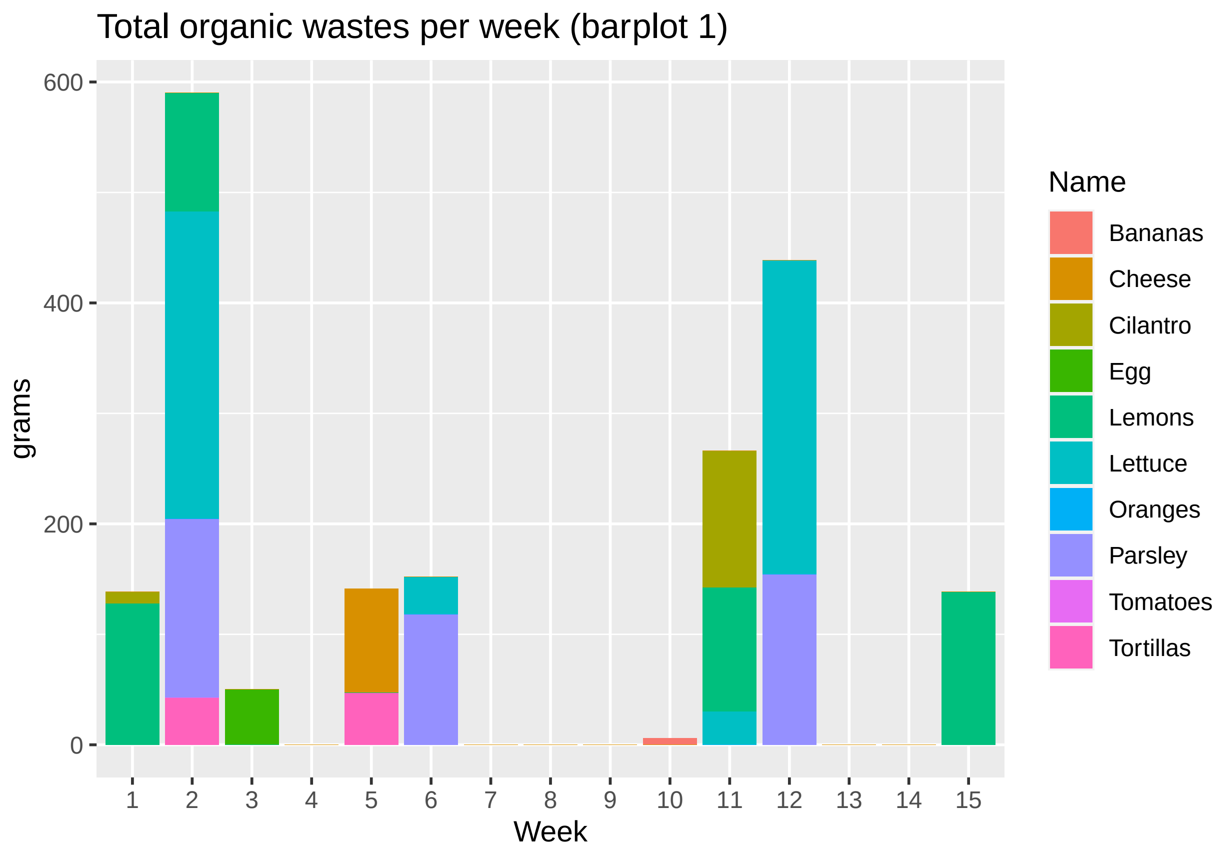


- Wee see 590.18g were organic waste on week 2, which makes us assume it was the week with most wastes.

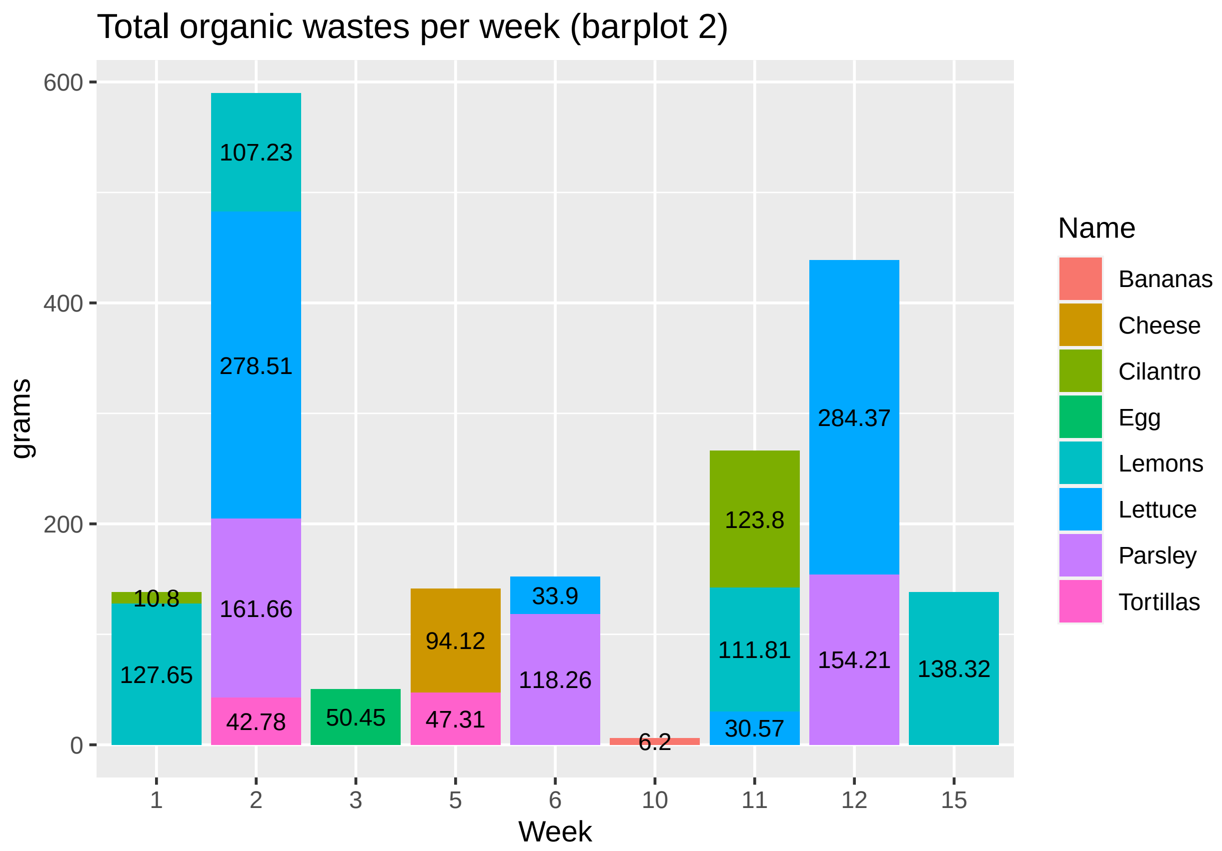
-STEP 4:

- Let’s make some bar plots to see which products contributed to the waste:

```
organic_waste %>%
  ggplot(aes(factor(Week), n, fill = Name))+
  ggtitle("Total organic wastes per week (barplot 1)")+
  ylab("grams")+xlab("Week")+
  geom_bar(stat = "identity")
```

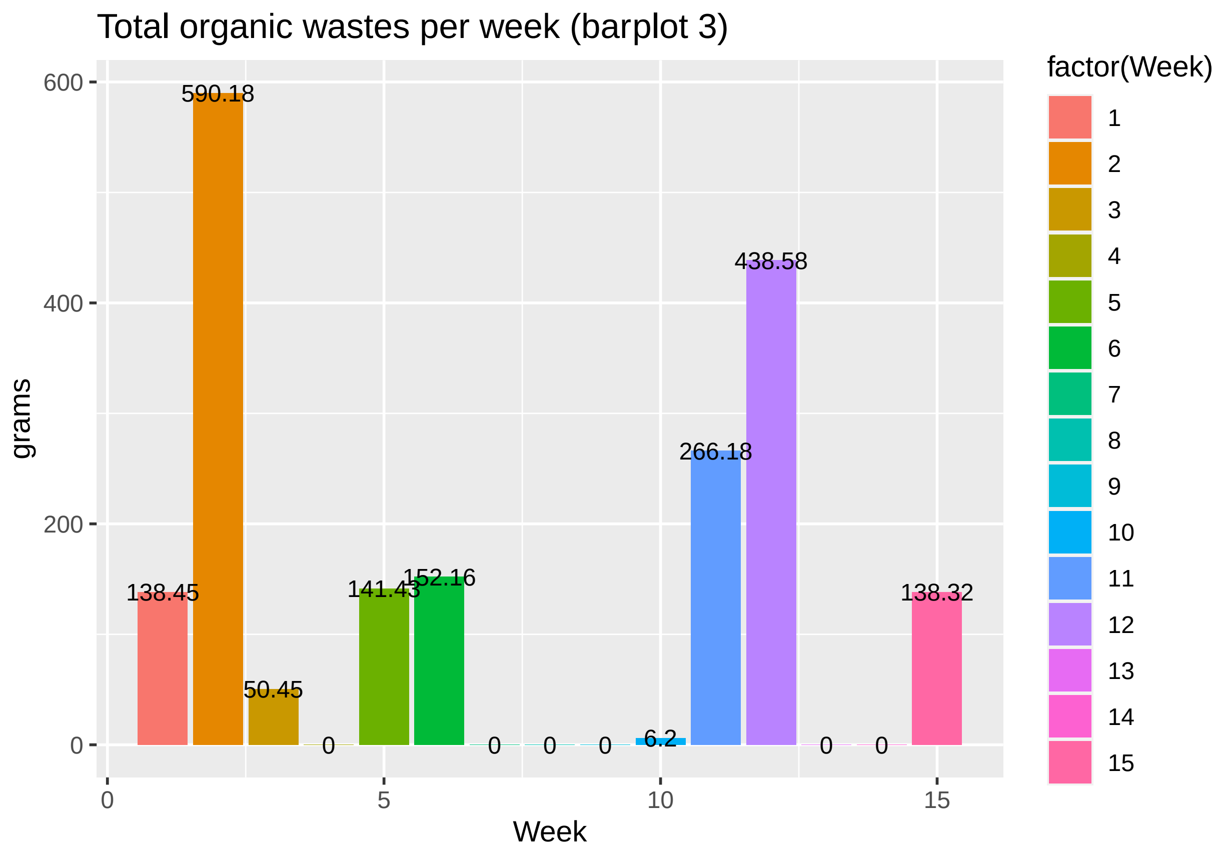


```
organic_waste %>% filter(n>0) %>%
  ggplot(aes(factor(Week), n, fill = Name, label=n))+
  ggtitle("Total organic wastes per week (barplot 2)")+
  ylab("grams")+xlab("Week")+
  geom_bar(stat = "identity")+
  geom_text(size = 3, position = position_stack(vjust = 0.5))
```



```
organic_waste %>% group_by(Week) %>% summarise(S=sum(n)) %>%
  ggplot(aes(Week,S, fill=factor(Week)))+
  ggtitle("Total organic wastes per week (barplot 3)")+
  ylab("grams")+geom_bar(stat = "identity")+
  geom_text(aes(label=S), size=3)
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

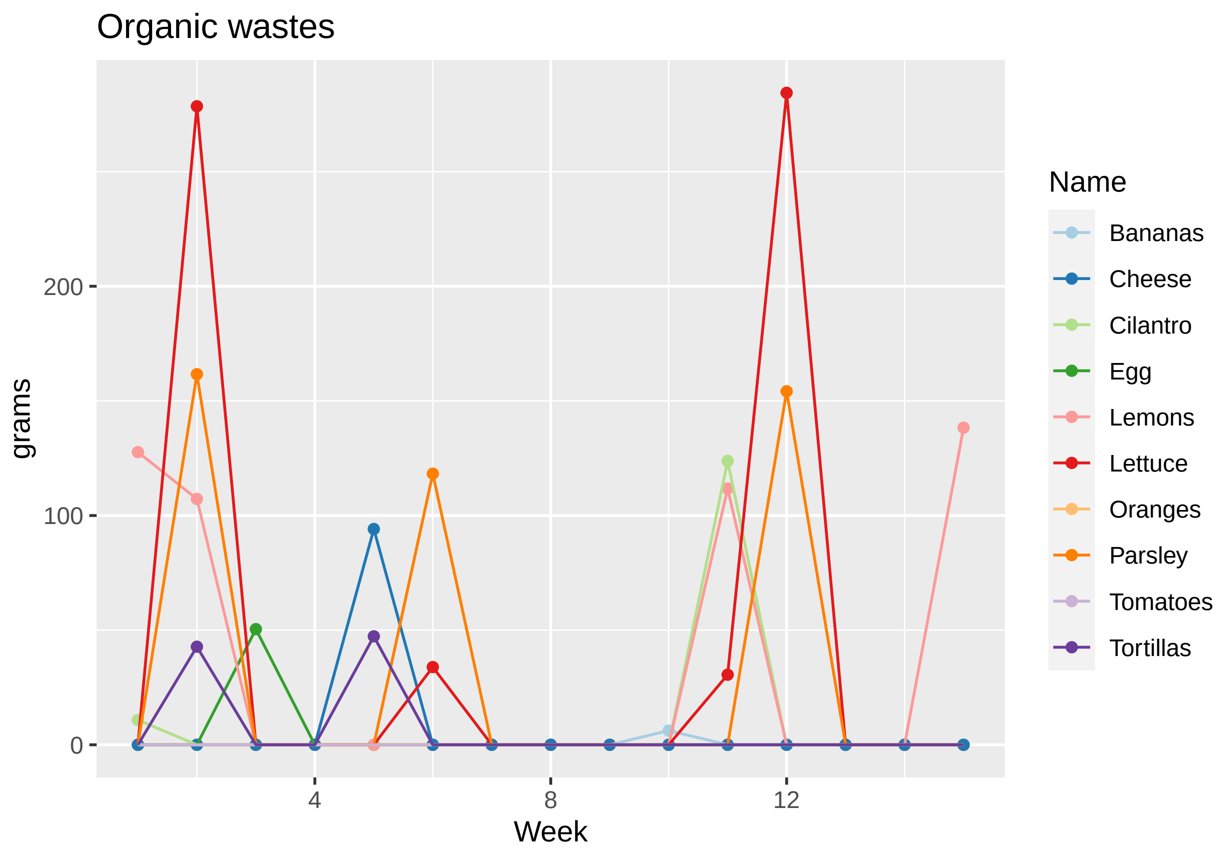


- barplot 2 makes us see that week 2 indeed had the greatest amount of waste, having 4 different types of organic waste.

-STEP 5:

- Let's analyze how each organic waste behaved throughout the 15 weeks:

```
organic_waste %>%
  ggplot(aes(Week, n, col = Name))+
  ggtitle("Organic wastes")+
  ylab("grams")+xlab("Week")+
  geom_point()+geom_line() + scale_color_brewer(palette="Paired")
```

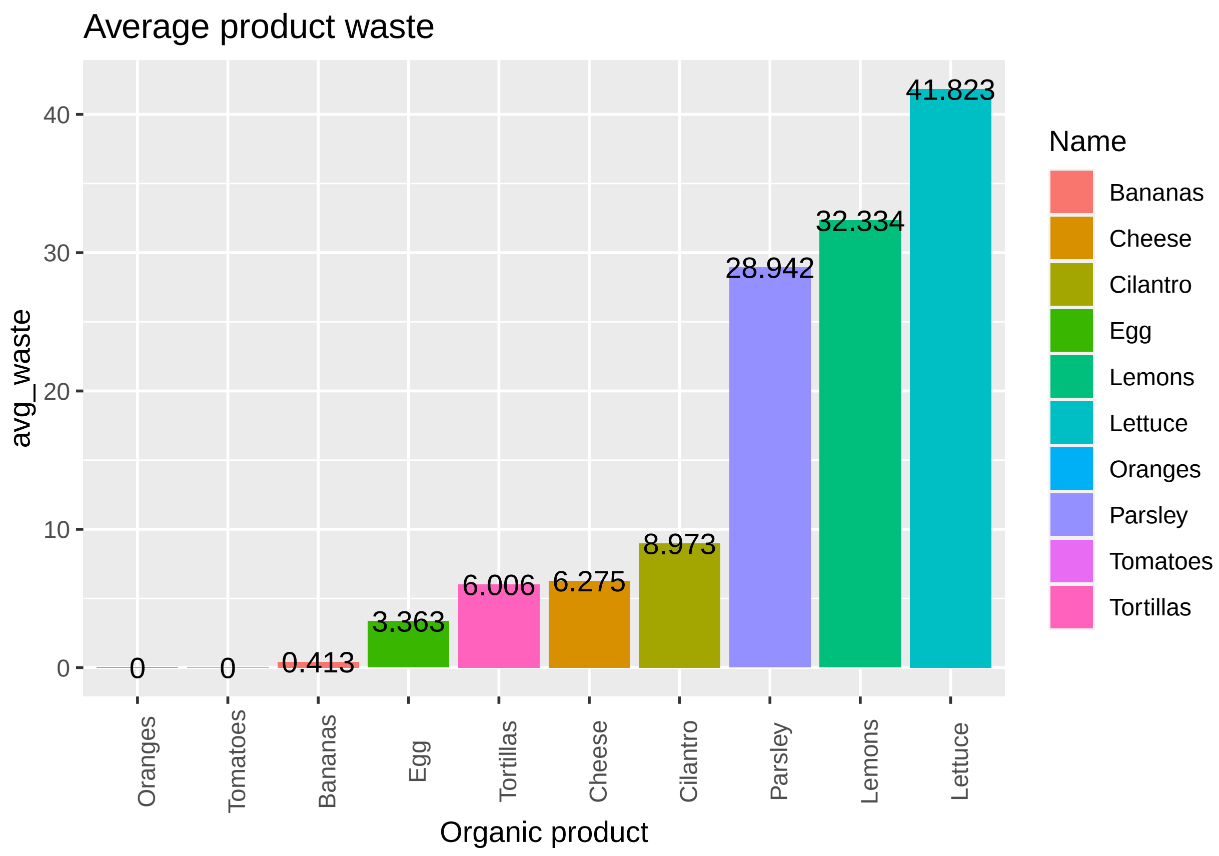


-STEP 6:

- Let's see on average how much of each product was waste throughout the 15 weeks:

```
organic_waste %>% group_by(Name) %>% summarise(avg_waste=mean(n)) %>% ggplot(aes(reorder(Name, avg_wast
  ggtitle("Average product waste")+ xlab("Organic product")+
  theme(axis.text.x = element_text(angle = 90))+
  geom_bar(stat = "identity")+geom_text(aes(label=round(avg_waste, 3)))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```



- As expected, the lettuce has the highest average waste per week.

-STEP 7:

- Let's calculate the number of times the organic waste was 0 and for which products:

```
sum(organic_waste$n==0)
```

```
## [1] 132
```

```
organic_waste[organic_waste$n==0,] %>% group_by(Name) %>% count(n)
```

```
## Storing counts in ‘nn‘, as ‘n‘ already present in input
## i Use ‘name = "new_name"‘ to pick a new name.
```

```
## # A tibble: 10 x 3
## # Groups:   Name [10]
##    Name          n    nn
##    <chr>     <dbl> <int>
##  1 Bananas       0    14
##  2 Cheese        0    14
##  3 Cilantro      0    13
##  4 Egg           0    14
##   5 Lemons        0    11
##   6 Lettuce       0    11
##   7 Oranges       0    15
##   8 Parsley       0    12
##   9 Tomatoes      0    15
##  10 Tortillas     0    13
```

- We see that oranges and tomatoes were never organic waste, since the number of times 0 appears is 15.

-STEP 8:

- Let's see if there's any product that was wasted every single week, and if there is no product like that, let's compute the one that repeats the most throughout the weeks:

```
#Mode function
g_mode <- function(v) {
   u_v <- unique(v)
   u_v[which.max(tabulate(match(v, u_v)))]
}
#Let's calculate the product that repeats the most
m <- organic_waste[organic_waste$n>0,] %>% summarise(g_mode(Name))
#We know that Lemons are the most repeated organic wastes
organic_waste[organic_waste$Name == m[1,],] %>% filter(n>0)
```

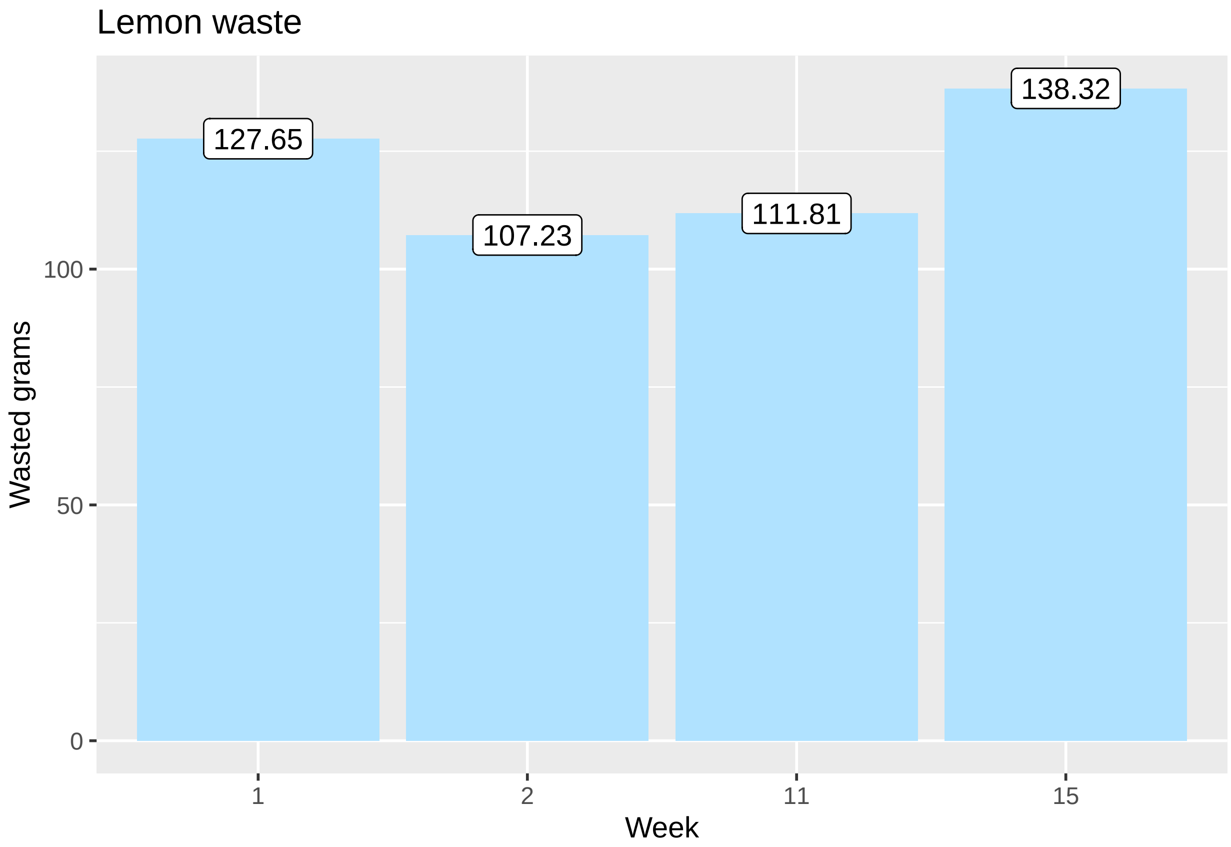
```
##     Name      n Week
## 1 Lemons 127.65    1
## 2 Lemons 107.23    2
## 3 Lemons 111.81   11
## 4 Lemons 138.32   15
```

```
organic_waste %>% filter(Name == m[1,], n>0) %>% summarise(count=n()) %>% cbind("name"=m[1,])
```

```
##   count   name
## 1     4 Lemons
```

- We see that lemons were the most frequent organic waste products, being disposed on weeks 1, 2, 11, and 15. Let's make a bar plot of this result:

```
organic_waste[organic_waste$Name == m[1,],] %>% filter(n>0) %>%
  ggplot(aes(factor(Week), n))+
  ylab("Wasted grams") + ggtitle("Lemon waste") + xlab("Week")+
  geom_bar(stat = "identity", fill="lightskyblue1") +
  geom_label(aes(label = n))
```



- However, let’s check the number of times each product became organic waste just to be sure that there are no other modes.

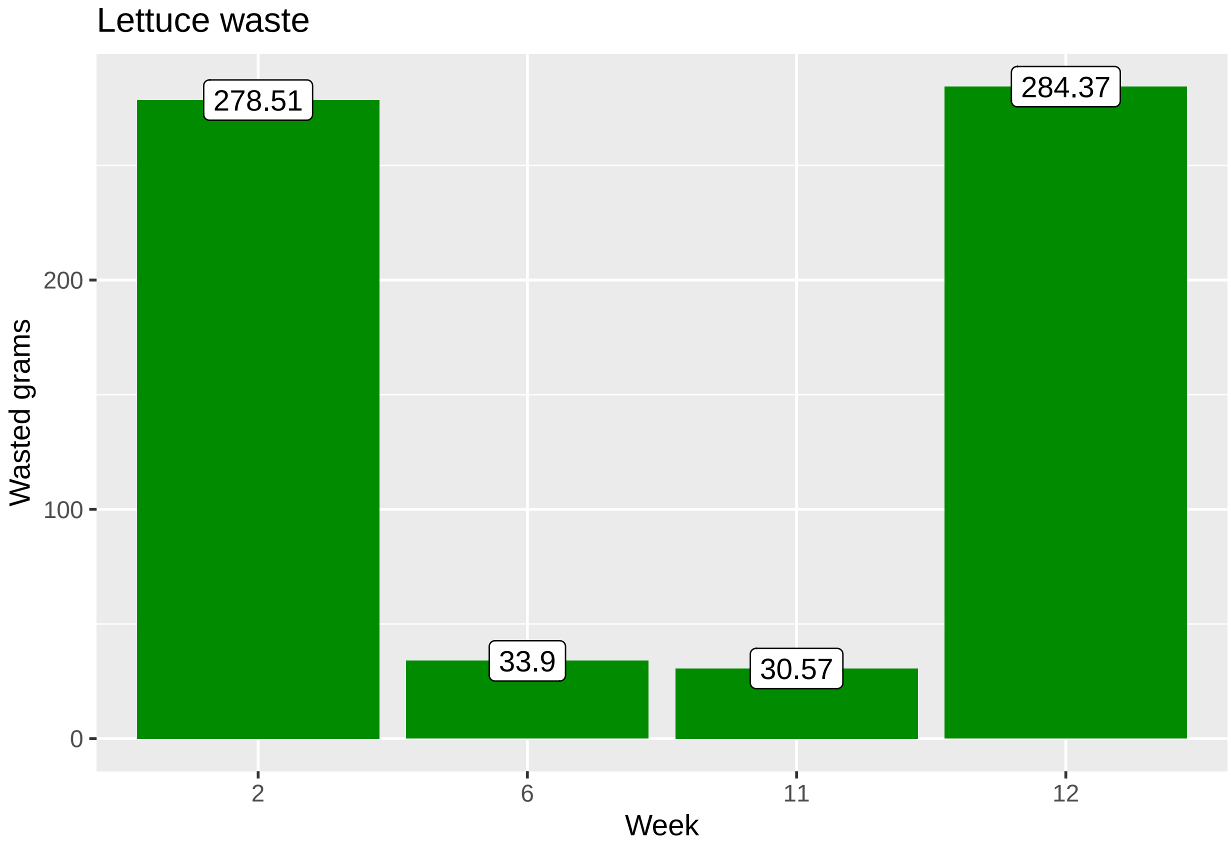
```
organic_waste %>% filter(n>0) %>% group_by(Name) %>% summarise(r=n())
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

```
## # A tibble: 8 x 2
##   Name          r
##   <chr>     <int>
## 1 Bananas       1
## 2 Cheese        1
## 3 Cilantro      2
## 4 Egg           1
## 5 Lemons        4
## 6 Lettuce       4
## 7 Parsley       3
## 8 Tortillas     2
```

- We see that indeed there are more modes. Both lettuce and lemons were wasted 4 times. So let’s make a plot for lettuce as well:

```
organic_waste[organic_waste$Name == "Lettuce",] %>% filter(n>0) %>%
  ggplot(aes(factor(Week), n))+
  ylab("Wasted grams") + ggtitle("Lettuce waste") + xlab("Week")+
  geom_bar(stat = "identity", fill="green4") +
  geom_label(aes(label = n))
```



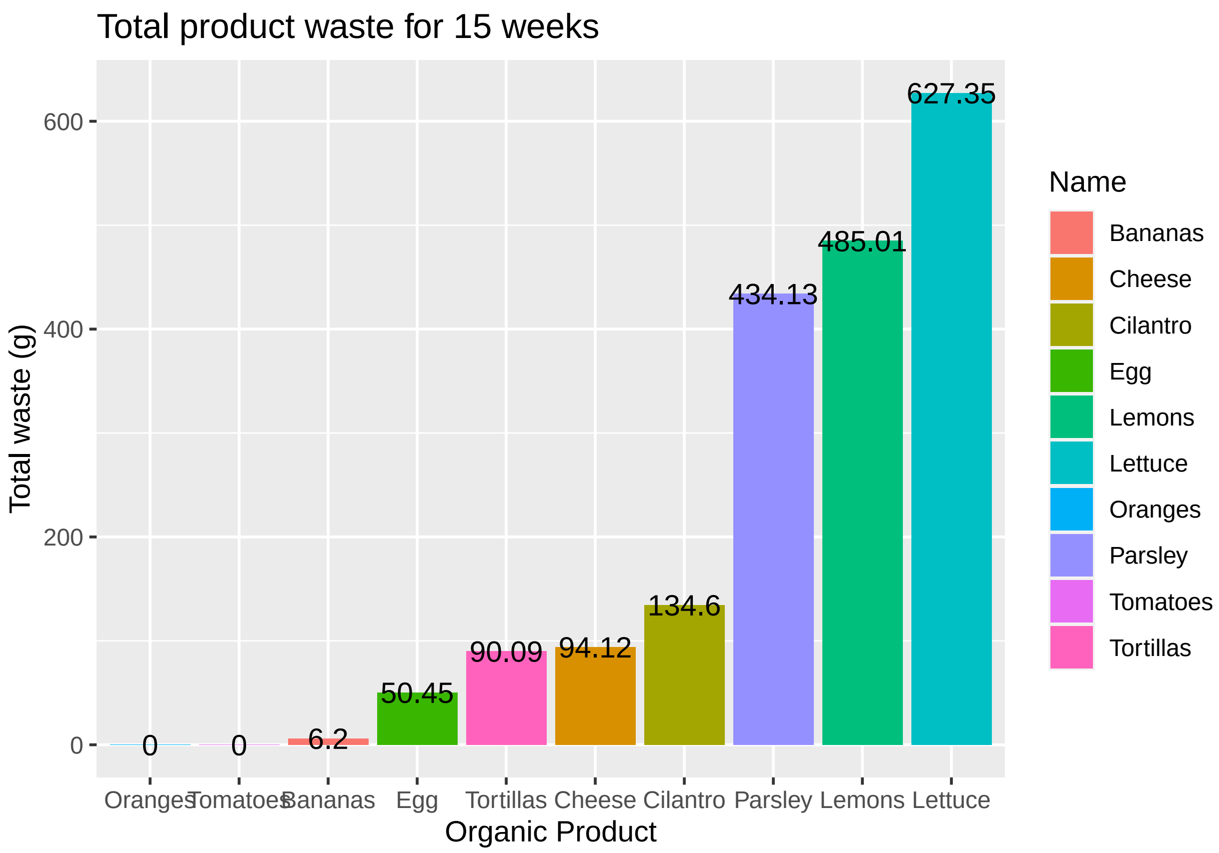
- We see that at the beginning of the 15 week period, and at the end of it, the greatest amount of grams of lettuce became organic waste.

-STEP 9:

- Let's see in total how many product grams were wasted per organic product:

```
organic_waste %>% group_by(Name) %>% summarise(tot_waste=sum(n)) %>% ggplot(aes(reorder(Name, tot_waste
  ggtitle("Total product waste for 15 weeks")+
  xlab("Organic Product")+
  ylab("Total waste (g)")+
  geom_bar(stat = "identity")+geom_text(aes(label=round(tot_waste, 3)))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```



```
organic_waste %>% group_by(Name) %>% summarise(tot_waste=sum(n)) %>% .$tot_waste %>% sum()
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## [1] 1921.95
```

- We see that in total, in the 15 week period, 1921.95 grams of food were wasted, lettuce being the organic product that was the most wasted. A way to optimize the organic waste could be by registering the amount of food bought (specifically in grams) at a certain time, and determining how much of the products we bought, at that certain time, became organic waste. If we did this for every time we buy the groceries, we could know on average how much of each product we buy and waste, and the difference between the average bought products and the average wasted products would tell us the average amount of food we need to buy in order to reduce our organic waste.