

Reading and data wrangling in R

Andrea Marcela Huerfano Barbosa

August 14, 2019

Description

In this file you can find some tips to:

- Reading data from different formats (txt, csv, excel...)
- Cleaning data
- Creation of new variables
- Merging datasets
- Dealing with NA

All of the task above are related with how to clean and tidy our data, that is an inevitable phase when you work with data. Some terms for these activities are: data cleaning, data wrangling and data manipulation.

1. Reading data

There are many ways to import datasets depending on the file characteristics as separator, decimals, head, etc. The easy way is using the bottom Import Dataset in the R-Studio enviroment, however you have to copy the code into your script because the lines just run in the console. To know some of the fuctions that appear throw the bottom you are going to find some examples.

- read.csv: comma separated values with period as decimal separator.
- read.csv2: semicolon separated values with comma as decimal separator.
- read.delim: tab-delimited files with period as decimal separator.
- read.delim2 tab-delimited files with comma as decimal separator.
- read.fwf data with a predetermined number of bytes per column.

Some functions to inspect the data are: colnames(), srt(), head(), tail()

```
pigeon <- read.delim("C:/Users/Andrea/Desktop/pigeon-racing.txt")
colnames(pigeon)
```

```
## [1] "Pos"      "Breeder"  "Pigeon"   "Name"     "Color"    "Sex"
## [7] "Ent"      "Arrival"  "Speed"    "To.Win"   "Eligible"
```

```
str(pigeon)
```

```
## 'data.frame': 400 obs. of 11 variables:
## $ Pos : int 1 2 3 4 5 6 7 8 9 10 ...
## $ Breeder : Factor w/ 90 levels "4-Birds","7-11 Syndicate",...: 83 49 47 4 40 24 40 64 9 83 ...
## $ Pigeon : Factor w/ 400 levels "0001-AU15-RTEX",...: 272 99 101 283 381 40 383 184 191 271 ...
## $ Name : Factor w/ 21 levels "", "\"the Duck\"",...: 1 1 18 1 1 1 1 1 1 1 ...
## $ Color : Factor w/ 29 levels "BB", "BBPD", "BBPI",...: 9 26 1 4 6 6 5 6 1 6 ...
## $ Sex : Factor w/ 2 levels "C", "H": 2 2 2 2 2 2 1 2 2 2 ...
## $ Ent : int 1 1 1 1 1 1 2 1 1 2 ...
## $ Arrival : Factor w/ 355 levels "00:03.0", "00:04.0",...: 166 183 184 185 186 188 189 190 191 192 ..
## $ Speed : num 172 164 163 163 163 ...
## $ To.Win : Factor w/ 365 levels "0:00:00", "0:05:21",...: 1 2 3 4 5 6 7 8 9 10 ...
## $ Eligible: Factor w/ 1 level "Yes": 1 1 1 1 1 1 1 1 1 1 ...
```

The summary function give you a view about distribution for cuantitative variables and the levels of each factor.

```
summary(pigeon)
```

```
##          Pos          Breeder          Pigeon
## Min.      : 1.0    Jb & D      : 13    0001-AU15-RTEX: 1
## 1st Qu.:100.8    A P C Loft : 12    0001-IF15-POWS: 1
## Median :200.5    Family Loft: 12    0002-AU15-RTEX: 1
## Mean     :200.4    Redtex      : 12    0002-IF15-PJLO: 1
## 3rd Qu.:300.2    Alias-Alias: 11    0003-IF15-POWS: 1
## Max.      :400.0    Andy Skwiat: 10    0005-AU15-NPL : 1
##              (Other)      :330    (Other)      :394
##              Name          Color      Sex      Ent      Arrival
##              :380    BB      :177    C: 9    Min.      : 1.000    12:20.0: 3
## "the Duck"    : 1    BC      : 92    H:391    1st Qu.: 2.000    54:26.0: 3
## Alice        : 1    BBWF     : 36              Median : 3.000    56:10.0: 3
## BATTLE BORN 27: 1    RC      : 16              Mean   : 3.533    05:03.0: 2
## Bella        : 1    DC      : 10              3rd Qu.: 5.000    07:54.0: 2
## BLACK NIGTH 9 : 1    BCWF     : 8              Max.    :13.000    12:03.0: 2
## (Other)      : 15    (Other): 61              (Other):385
##      Speed          To.Win    Eligible
## Min.      : 76.68    0:13:56: 3    Yes:400
## 1st Qu.:104.43    0:05:48: 2
## Median :131.66    0:05:57: 2
## Mean     :128.71    0:06:02: 2
## 3rd Qu.:151.18    0:06:41: 2
## Max.      :172.16    0:06:48: 2
##              (Other):387
```

excel

The functions explained above don't require intallation of any library because they are in the R core, however to read excel files it is necessary to load the library readxl

```
library(readxl)
spanish_silver <- read_excel("C:/Users/Andrea/Desktop/spanish-silver.xls",
  sheet = "spanish-silver")
```

website

Subsets

Tibble

In all of the examples above the data were loaded as data_frame. However to display a sample of them and their visualization is more easy when the data is convert into a tibble

```
library(tibble)
pigeon_tb <- as_data_frame(pigeon)
pigeon_tb
```

```
## # A tibble: 400 x 11
##       Pos Breeder Pigeon Name Color Sex      Ent Arrival Speed To.Win
##   <int> <fct>    <fct>    <fct> <fct> <fct> <int> <fct>    <dbl> <fct>
## 1     1     1 Texas ~ 19633~ ""     BCWF  H           1 42:14.0  172. 0:00:~
```

```
## 2      2 Junior~ 0402~~ ""      SIWF H          1 47:36.0 164. 0:05:~
## 3      3 Jerry ~ 0404~~ Perc~ BB   H          1 47:41.0 163. 0:05:~
## 4      4 Alias~~ 2013~~ ""      BBSP H          1 47:43.0 163. 0:05:~
## 5      5 Greg G~ 5749~~ ""      BC   H          1 47:44.0 163. 0:05:~
## 6      6 Dal-Te~ 0032~~ ""      BC   H          1 47:51.0 163. 0:05:~
## 7      7 Greg G~ 5768~~ ""      BBWF C          2 47:53.0 163. 0:05:~
## 8      8 N C Sy~ 1067~~ ""      BC   H          1 47:57.0 163. 0:05:~
## 9      9 Baldwi~ 1194~~ ""      BB   H          1 48:02.0 163. 0:05:~
## 10     10 Texas ~ 19632~~ ""     BC   H          2 48:03.0 163. 0:05:~
## # ... with 390 more rows, and 1 more variable: Eligible <fct>
```

This sort of view is obtained directly into the original dataframe with the function head.

```
head(pigeon, n=4)
```

```
##   Pos      Breeder      Pigeon      Name Color Sex Ent Arrival
## 1   1      Texas Outlaws 19633-AU15-FOYS      BCWF  H   1 42:14.0
## 2   2      Junior Juanich 0402-AU15-JRL      SIWF  H   1 47:36.0
## 3   3 Jerry Allensworth 0404-AU15-VITA Perch Potato  BB   H   1 47:41.0
## 4   4      Alias-Alias 2013-AU15-ALIA      BBSP  H   1 47:43.0
##   Speed To.Win Eligible
## 1 172.155 0:00:00      Yes
## 2 163.569 0:05:21      Yes
## 3 163.442 0:05:27      Yes
## 4 163.392 0:05:28      Yes
```

In this script most of the data will be used in tibbles.

Sampling

After loaded the dataset is useful sampling to know their data and identify steps to clean them.

```
library(dplyr)
pigeon_tb%>%sample_n(4)
```

```
## # A tibble: 4 x 11
##   Pos Breeder Pigeon Name Color Sex Ent Arrival Speed To.Win
##   <int> <fct> <fct> <fct> <fct> <fct> <int> <fct> <dbl> <fct>
## 1   300 Family~ 0065~~ ""     BB   C     10 48:30.0 104. 1:06:~
## 2   373 Stelli~ 0254~~ ""     BCH  H      3 12:24.0  91.5 1:30:~
## 3   236 Rick B~ 0305~~ ""     BB   H      3 27:11.0 120. 0:44:~
## 4   148 Loizzi~ 1741~~ ""     BC   H      2 59:25.0 147. 0:17:~
## # ... with 1 more variable: Eligible <fct>
```

Extracting a percentage in the data set

```
pigeon_tb%>%sample_frac(0.01, replace=FALSE)
```

```
## # A tibble: 4 x 11
##   Pos Breeder Pigeon Name Color Sex Ent Arrival Speed To.Win
##   <int> <fct> <fct> <fct> <fct> <fct> <int> <fct> <dbl> <fct>
## 1   280 Stenma~ 5263~~ ""     BC   H      8 42:13.0 108. 0:59:~
## 2   312 Woodse~ 1038~~ ""     BB   H      9 51:00.0 103. 1:08:~
## 3   372 Captai~ 1669~~ ""     BB   H      4 12:23.0  91.5 1:30:~
## 4   238 Flying~ 9518~~ ""     DC   H      3 27:21.0 119. 0:45:~
## # ... with 1 more variable: Eligible <fct>
```

Selecting columns

```
pigeon_tb%>%select(Pigeon, Color, Sex)
```

```
## # A tibble: 400 x 3
##   Pigeon      Color Sex
##   <fct>      <fct> <fct>
## 1 19633-AU15-FOYS BCWF  H
## 2 0402-AU15-JRL  SIWF  H
## 3 0404-AU15-VITA BB    H
## 4 2013-AU15-ALIA BBSP  H
## 5 5749-AU15-SLI  BC    H
## 6 0032-AU15-DRPC BC    H
## 7 5768-AU15-SLI  BBWF  C
## 8 1067-AU15-TXHC BC    H
## 9 1194-AU15-TENT BB    H
## 10 19632-AU15-FOYS BC    H
## # ... with 390 more rows
```

Filters

- And &
- Or |

```
pigeon_tb%>%filter(Color=='BB' | Sex=='H')
```

```
## # A tibble: 396 x 11
##   Pos Breeder Pigeon Name Color Sex Ent Arrival Speed To.Win
##   <int> <fct>   <fct> <fct> <fct> <fct> <int> <fct>   <dbl> <fct>
## 1     1 Texas ~ 19633~ ""    BCWF  H     1 42:14.0 172. 0:00::~
## 2     2 Junior~ 0402-- ""    SIWF  H     1 47:36.0 164. 0:05::~
## 3     3 Jerry ~ 0404-- Perc~ BB    H     1 47:41.0 163. 0:05::~
## 4     4 Alias-- 2013-- ""    BBSP  H     1 47:43.0 163. 0:05::~
## 5     5 Greg G~ 5749-- ""    BC    H     1 47:44.0 163. 0:05::~
## 6     6 Dal-Te~ 0032-- ""    BC    H     1 47:51.0 163. 0:05::~
## 7     8 N C Sy~ 1067-- ""    BC    H     1 47:57.0 163. 0:05::~
## 8     9 Baldwi~ 1194-- ""    BB    H     1 48:02.0 163. 0:05::~
## 9    10 Texas ~ 19632~ ""    BC    H     2 48:03.0 163. 0:05::~
## 10   10 Redtex 0024-- ""    RED   H     1 48:03.0 163. 0:05::~
## # ... with 386 more rows, and 1 more variable: Eligible <fct>
```

```
pigeon_tb%>%filter(Color=='BB' & Sex=='H')
```

```
## # A tibble: 172 x 11
##   Pos Breeder Pigeon Name Color Sex Ent Arrival Speed To.Win
##   <int> <fct>   <fct> <fct> <fct> <fct> <int> <fct>   <dbl> <fct>
## 1     3 Jerry ~ 0404-- Perc~ BB    H     1 47:41.0 163. 0:05::~
## 2     9 Baldwi~ 1194-- ""    BB    H     1 48:02.0 163. 0:05::~
## 3    14 Goshen~ 5834-- ""    BB    H     1 48:12.0 163. 0:05::~
## 4    16 Flyhom~ 1531-- ""    BB    H     1 48:15.0 163. 0:06::~
## 5    24 Jb & D 1214-- ""    BB    H     1 48:36.0 162. 0:06::~
## 6    30 Churn ~ 9216-- ""    BB    H     1 48:48.0 162. 0:06::~
## 7    32 Alias-- 2049-- ""    BB    H     3 48:56.0 162. 0:06::~
## 8    35 Clear ~ 0263-- ""    BB    H     1 49:06.0 161. 0:06::~
## 9    38 Clear ~ 0235-- ""    BB    H     2 49:17.0 161. 0:07::~
## 10   40 Skip's~ 5302-- ""    BB    H     2 49:28.0 161. 0:07::~
## # ... with 162 more rows, and 1 more variable: Eligible <fct>
```

Order by

The “-” makes the order from the greatest to the shortest.

```
pigeon_tb%>%arrange(-Speed)
```

```
## # A tibble: 400 x 11
##       Pos Breeder Pigeon Name Color Sex      Ent Arrival Speed To.Win
##   <int> <fct>   <fct> <fct> <fct> <fct> <int> <fct>   <dbl> <fct>
## 1     1     1 Texas ~ 19633~ ""    BCWF  H        1 42:14.0  172. 0:00:~
## 2     2     2 Junior~ 0402-- ""    SIWF  H        1 47:36.0  164. 0:05:~
## 3     3     3 Jerry ~ 0404-- Perc~ BB    H        1 47:41.0  163. 0:05:~
## 4     4     4 Alias-- 2013-- ""    BBSP  H        1 47:43.0  163. 0:05:~
## 5     5     5 Greg G~ 5749-- ""    BC    H        1 47:44.0  163. 0:05:~
## 6     6     6 Dal-Te~ 0032-- ""    BC    H        1 47:51.0  163. 0:05:~
## 7     7     7 Greg G~ 5768-- ""    BBWF  C        2 47:53.0  163. 0:05:~
## 8     8     8 N C Sy~ 1067-- ""    BC    H        1 47:57.0  163. 0:05:~
## 9     9     9 Baldwi~ 1194-- ""    BB    H        1 48:02.0  163. 0:05:~
## 10    10    10 Texas ~ 19632~ ""    BC    H        2 48:03.0  163. 0:05:~
## # ... with 390 more rows, and 1 more variable: Eligible <fct>
```

2.Cleaning data

Creation of new variables

New variable

```
pigeon_tb%>%mutate(NewSpeed=Speed/2)
```

```
## # A tibble: 400 x 12
##       Pos Breeder Pigeon Name Color Sex      Ent Arrival Speed To.Win
##   <int> <fct>   <fct> <fct> <fct> <fct> <int> <fct>   <dbl> <fct>
## 1     1     1 Texas ~ 19633~ ""    BCWF  H        1 42:14.0  172. 0:00:~
## 2     2     2 Junior~ 0402-- ""    SIWF  H        1 47:36.0  164. 0:05:~
## 3     3     3 Jerry ~ 0404-- Perc~ BB    H        1 47:41.0  163. 0:05:~
## 4     4     4 Alias-- 2013-- ""    BBSP  H        1 47:43.0  163. 0:05:~
## 5     5     5 Greg G~ 5749-- ""    BC    H        1 47:44.0  163. 0:05:~
## 6     6     6 Dal-Te~ 0032-- ""    BC    H        1 47:51.0  163. 0:05:~
## 7     7     7 Greg G~ 5768-- ""    BBWF  C        2 47:53.0  163. 0:05:~
## 8     8     8 N C Sy~ 1067-- ""    BC    H        1 47:57.0  163. 0:05:~
## 9     9     9 Baldwi~ 1194-- ""    BB    H        1 48:02.0  163. 0:05:~
## 10    10    10 Texas ~ 19632~ ""    BC    H        2 48:03.0  163. 0:05:~
## # ... with 390 more rows, and 2 more variables: Eligible <fct>,
## #       NewSpeed <dbl>
```

Split

Split a string by an specific separator

```
library(dplyr)
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 3.5.3
```

```
pigeon_tb%>%separate(Pigeon, sep='-', c('Num', 'id', 'det'))
```

```
## # A tibble: 400 x 13
##       Pos Breeder Num   id   det Name Color Sex      Ent Arrival Speed
```

```
##      <int> <fct>      <chr> <chr> <chr> <fct> <fct> <fct> <int> <fct>      <dbl>
## 1      1 Texas ~ 19633 AU15 FOYS "" BCWF H      1 42:14.0 172.
## 2      2 Junior~ 0402 AU15 JRL  "" SIWF H      1 47:36.0 164.
## 3      3 Jerry ~ 0404 AU15 VITA Perc~ BB H      1 47:41.0 163.
## 4      4 Alias~~ 2013 AU15 ALIA "" BBSP H      1 47:43.0 163.
## 5      5 Greg G~ 5749 AU15 SLI  "" BC H      1 47:44.0 163.
## 6      6 Dal-Te~ 0032 AU15 DRPC "" BC H      1 47:51.0 163.
## 7      7 Greg G~ 5768 AU15 SLI  "" BBWF C      2 47:53.0 163.
## 8      8 N C Sy~ 1067 AU15 TXHC "" BC H      1 47:57.0 163.
## 9      9 Baldwi~ 1194 AU15 TENT "" BB H      1 48:02.0 163.
## 10     10 Texas ~ 19632 AU15 FOYS "" BC H      2 48:03.0 163.
## # ... with 390 more rows, and 2 more variables: To.Win <fct>,
## # Eligible <fct>
```

Concatenate

```
pigeon_tb%>%unite_('new', c('Pos','Sex'), sep = '-')
```

```
## # A tibble: 400 x 10
##   new Breeder Pigeon Name Color Ent Arrival Speed To.Win Eligible
##   <chr> <fct>      <fct>      <fct> <fct> <int> <fct>      <dbl> <fct>      <fct>
## 1 1-H Texas Ou~ 19633~~ "" BCWF      1 42:14.0 172. 0:00:~ Yes
## 2 2-H Junior J~ 0402-A~ "" SIWF      1 47:36.0 164. 0:05:~ Yes
## 3 3-H Jerry Al~ 0404-A~ Perch~ BB      1 47:41.0 163. 0:05:~ Yes
## 4 4-H Alias-Al~ 2013-A~ "" BBSP      1 47:43.0 163. 0:05:~ Yes
## 5 5-H Greg Gla~ 5749-A~ "" BC      1 47:44.0 163. 0:05:~ Yes
## 6 6-H Dal-Tex ~ 0032-A~ "" BC      1 47:51.0 163. 0:05:~ Yes
## 7 7-C Greg Gla~ 5768-A~ "" BBWF      2 47:53.0 163. 0:05:~ Yes
## 8 8-H N C Synd~ 1067-A~ "" BC      1 47:57.0 163. 0:05:~ Yes
## 9 9-H Baldwin ~ 1194-A~ "" BB      1 48:02.0 163. 0:05:~ Yes
## 10 10-H Texas Ou~ 19632~~ "" BC      2 48:03.0 163. 0:05:~ Yes
## # ... with 390 more rows
```

New variable base on levels of another one

```
levels(pigeon_tb$Color)
```

```
## [1] "BB" "BBPD" "BBPI" "BBSP" "BBWF" "BC" "BCH" "BCSP" "BCWF" "BKWF"
## [11] "BLCK" "BLK" "DC" "DCWF" "GRIZ" "GRZL" "OPAL" "OPWF" "PENC" "RC"
## [21] "RCSP" "RCWF" "RED" "SIL" "SILV" "SIWF" "WGRZ" "WHGR" "WHT"
```

```
B_I<-c("BB", "BBPD", "BBPI", "BBSP", "BBWF", "BC", "BCH", "BCSP", "BCWF", "BKWF", "BLCK", "BLK")
```

```
D_I<-c("DC", "DCWF")
```

```
G_I<-c("GRIZ", "GRZL")
```

```
for (i in 1:length(pigeon_tb$Color)){
  if (pigeon_tb$Color[i] %in% B_I){pigeon_tb$Ini[i]='B_I'}else{
    if (pigeon_tb$Color[i] %in% D_I){pigeon_tb$Ini[i]='D_I'}else{
      if(pigeon_tb$Color[i] %in% G_I){pigeon_tb$Ini[i]='G_I'}else{pigeon_tb$Ini[i]='Another'}
    }
  }
}
```

```
## Warning: Unknown or uninitialised column: 'Ini'.
```

```
as_data_frame(data.frame(pigeon_tb$Color,pigeon_tb$Ini))
```

```
## # A tibble: 400 x 2
##   pigeon_tb.Color pigeon_tb.Ini
##   <fct>          <fct>
## 1 BCWF          B_I
## 2 SIWF          Another
## 3 BB            B_I
## 4 BBSP          B_I
## 5 BC            B_I
## 6 BC            B_I
## 7 BBWF          B_I
## 8 BC            B_I
## 9 BB            B_I
## 10 BC           B_I
## # ... with 390 more rows
```

Variable type conversion

Suppose that Ent is a factor variable not a numeric one.

```
pigeon_tb$Ent<- as.factor(pigeon_tb$Ent)
pigeon_tb
```

```
## # A tibble: 400 x 12
##   Pos Breeder Pigeon Name Color Sex Ent Arrival Speed To.Win
##   <int> <fct>   <fct> <fct> <fct> <fct> <fct> <fct> <dbl> <fct>
## 1     1   Texas ~ 19633~ ""    BCWF H    1   42:14.0  172. 0:00:~
## 2     2   Junior~ 0402-- ""    SIWF H    1   47:36.0  164. 0:05:~
## 3     3   Jerry ~ 0404-- Perc~ BB   H    1   47:41.0  163. 0:05:~
## 4     4   Alias-- 2013-- ""    BBSP H    1   47:43.0  163. 0:05:~
## 5     5   Greg G~ 5749-- ""    BC   H    1   47:44.0  163. 0:05:~
## 6     6   Dal-Te~ 0032-- ""    BC   H    1   47:51.0  163. 0:05:~
## 7     7   Greg G~ 5768-- ""    BBWF C    2   47:53.0  163. 0:05:~
## 8     8   N C Sy~ 1067-- ""    BC   H    1   47:57.0  163. 0:05:~
## 9     9   Baldwi~ 1194-- ""    BB   H    1   48:02.0  163. 0:05:~
## 10    10   Texas ~ 19632~ ""    BC   H    2   48:03.0  163. 0:05:~
## # ... with 390 more rows, and 2 more variables: Eligible <fct>, Ini <chr>
```

if the variable is as string to convert them into numeric the function is as.numeric()

4. Merging datasets

It's common that you have to merge many files to obtain your final dataset. In R at the same that Python you need to have the same colname in the key variable.

Joins

R has the SQL functions to join files, the key to join the data sets must have the same name in the files.

```
library(readxl)
athlete_country <- read_excel("C:/Users/Andrea/Desktop/python-ml-course-master/datasets/athletes/athlete_
  sheet = "Athelete_Country_Map")

athlete_sport <- read_excel("C:/Users/Andrea/Desktop/python-ml-course-master/datasets/athletes/athlete_
  sheet = "Athelete")
```

```
athlete_country
```

```
## # A tibble: 6,970 x 2
##   Athlete      Country
##   <chr>        <chr>
## 1 Michael Phelps United States
## 2 Natalie Coughlin United States
## 3 Aleksey Nemov   Russia
## 4 Alicia Coutts   Australia
## 5 Missy Franklin United States
## 6 Ryan Lochte     United States
## 7 Allison Schmitt United States
## 8 Ian Thorpe       Australia
## 9 Dara Torres     United States
## 10 Cindy Klassen  Canada
## # ... with 6,960 more rows
```

```
athlete_sport
```

```
## # A tibble: 6,975 x 2
##   Athlete      Sport
##   <chr>        <chr>
## 1 Michael Phelps Swimming
## 2 Natalie Coughlin Swimming
## 3 Aleksey Nemov   Gymnastics
## 4 Alicia Coutts   Swimming
## 5 Missy Franklin Swimming
## 6 Ryan Lochte     Swimming
## 7 Allison Schmitt Swimming
## 8 Ian Thorpe       Swimming
## 9 Dara Torres     Swimming
## 10 Cindy Klassen  Speed Skating
## # ... with 6,965 more rows
```

For this example the key is the column called 'Athlete'

```
inner_join(athlete_country, athlete_sport, by='Athlete')
```

```
## # A tibble: 6,994 x 3
##   Athlete      Country      Sport
##   <chr>        <chr>        <chr>
## 1 Michael Phelps United States Swimming
## 2 Natalie Coughlin United States Swimming
## 3 Aleksey Nemov   Russia        Gymnastics
## 4 Alicia Coutts   Australia     Swimming
## 5 Missy Franklin United States Swimming
## 6 Ryan Lochte     United States Swimming
## 7 Allison Schmitt United States Swimming
## 8 Ian Thorpe       Australia     Swimming
## 9 Dara Torres     United States Swimming
## 10 Cindy Klassen  Canada        Speed Skating
## # ... with 6,984 more rows
```

the structure to reproduce left and right join is the same that the example above.

Matching strings

There are two ways to match strings, the first one is creating a list of all levels and defining the category when each one belongs, the second way is defining a distance between two strings based on how different they are.

Uppercase

The 'M' don't match with 'm', first of all it is necessary to homogenize the strings for example all of them in uppercase

```
library(R.utils)

## Warning: package 'R.utils' was built under R version 3.5.3
## Loading required package: R.oo
## Warning: package 'R.oo' was built under R version 3.5.2
## Loading required package: R.methodsS3
## Warning: package 'R.methodsS3' was built under R version 3.5.2
## R.methodsS3 v1.7.1 (2016-02-15) successfully loaded. See ?R.methodsS3 for help.
## R.oo v1.22.0 (2018-04-21) successfully loaded. See ?R.oo for help.
##
## Attaching package: 'R.oo'
## The following objects are masked from 'package:methods':
##
##   getClasses, getMethods
## The following objects are masked from 'package:base':
##
##   attach, detach, gc, load, save
## R.utils v2.9.0 successfully loaded. See ?R.utils for help.
##
## Attaching package: 'R.utils'
## The following object is masked from 'package:tidyr':
##
##   extract
## The following object is masked from 'package:utils':
##
##   timestamp
## The following objects are masked from 'package:base':
##
##   cat, commandArgs, getOption, inherits, isOpen, parse, warnings
pigeon_tb %>% mutate(Breeder=toupper(Breeder))

## # A tibble: 400 x 12
##       Pos Breeder Pigeon Name Color Sex Ent Arrival Speed To.Win
##   <int> <chr>   <fct> <fct> <fct> <fct> <fct> <fct> <dbl> <fct>
## 1     1     TEXAS ~ 19633~ ""    BCWF  H    1    42:14.0  172. 0:00:~
## 2     2    JUNIOR~ 0402~ ""    SIWF  H    1    47:36.0  164. 0:05:~
## 3     3     JERRY ~ 0404~ Perc~ BB    H    1    47:41.0  163. 0:05:~
## 4     4    ALIAS~ 2013~ ""    BBSP  H    1    47:43.0  163. 0:05:~
## 5     5     GREG G~ 5749~ ""    BC    H    1    47:44.0  163. 0:05:~
## 6     6    DAL-TE~ 0032~ ""    BC    H    1    47:51.0  163. 0:05:~
```

```
## 7      7 GREG G~ 5768~ ""      BBWF C      2      47:53.0 163. 0:05:~
## 8      8 N C SY~ 1067~ ""      BC      H      1      47:57.0 163. 0:05:~
## 9      9 BALDWI~ 1194~ ""      BB      H      1      48:02.0 163. 0:05:~
## 10     10 TEXAS ~ 19632~ ""      BC      H      2      48:03.0 163. 0:05:~
## # ... with 390 more rows, and 2 more variables: Eligible <fct>, Ini <chr>
```

```
gender <- c("M", "male ", "Female", "fem.", 'ma', 'Fe')
grepl("ma", gender)
```

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

```
grepl("m", gender, ignore.case = TRUE)
```

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

```
#gender
#y=c()
#for (i in gender){
#  y[i]='ma' %in% gender[i]
#}
#y
#x<-c()
#if (y[i]==FALSE){x[i]='H'}
#x
```

5. Dealing with NA

Counting the na values

```
sapply(pigeon_tb, function(x) sum(is.na(x)))
```

```
##      Pos  Breeder   Pigeon   Name   Color   Sex   Ent  Arrival
##      0      0      0      0      0      0      0      0
##  Speed  To.Win Eligible   Ini
##      0      0      0      0
```

This is weird especially when I new that in name there are too many rows in blank, then one of the levels of the variable must be “”

```
levels(pigeon_tb$Name)
```

```
## [1] ""
## [5] "Bella"
## [9] "Christie"
## [13] "Gage"
## [17] "Lil Dat"
## [21] "SEMPER FI 11"
## [25] "\"the Duck\""
## [29] "BLACK NIGTH 9"
## [33] "Color Me Hot"
## [37] "Gypsy"
## [41] "Jack Frost"
## [45] "Pop's Pick"
## [49] "BATTLE BORN 27"
## [53] "Canned Heat"
## [57] "Edward"
## [61] "Kingston"
## [65] "Rogue Brew"
## [69] "Alice"
## [73] "Charlie"
## [77] "Elle"
## [81] "Gage"
## [85] "Lil Dat"
## [89] "SEMPER FI 11"
## [93] "BATTLE BORN 27"
## [97] "Canned Heat"
## [101] "Edward"
## [105] "Kingston"
## [109] "Rogue Brew"
## [113] "Alice"
## [117] "Charlie"
## [121] "Elle"
## [125] "Gage"
## [129] "Lil Dat"
## [133] "SEMPER FI 11"
## [137] "BATTLE BORN 27"
## [141] "Canned Heat"
## [145] "Edward"
## [149] "Kingston"
## [153] "Rogue Brew"
## [157] "Alice"
## [161] "Charlie"
## [165] "Elle"
## [169] "Gage"
## [173] "Lil Dat"
## [177] "SEMPER FI 11"
## [181] "BATTLE BORN 27"
## [185] "Canned Heat"
## [189] "Edward"
## [193] "Kingston"
## [197] "Rogue Brew"
## [201] "Alice"
## [205] "Charlie"
## [209] "Elle"
## [213] "Gage"
## [217] "Lil Dat"
## [221] "SEMPER FI 11"
## [225] "BATTLE BORN 27"
## [229] "Canned Heat"
## [233] "Edward"
## [237] "Kingston"
## [241] "Rogue Brew"
## [245] "Alice"
## [249] "Charlie"
## [253] "Elle"
## [257] "Gage"
## [261] "Lil Dat"
## [265] "SEMPER FI 11"
## [269] "BATTLE BORN 27"
## [273] "Canned Heat"
## [277] "Edward"
## [281] "Kingston"
## [285] "Rogue Brew"
## [289] "Alice"
## [293] "Charlie"
## [297] "Elle"
## [301] "Gage"
## [305] "Lil Dat"
## [309] "SEMPER FI 11"
## [313] "BATTLE BORN 27"
## [317] "Canned Heat"
## [321] "Edward"
## [325] "Kingston"
## [329] "Rogue Brew"
## [333] "Alice"
## [337] "Charlie"
## [341] "Elle"
## [345] "Gage"
## [349] "Lil Dat"
## [353] "SEMPER FI 11"
## [357] "BATTLE BORN 27"
## [361] "Canned Heat"
## [365] "Edward"
## [369] "Kingston"
## [373] "Rogue Brew"
## [377] "Alice"
## [381] "Charlie"
## [385] "Elle"
## [389] "Gage"
## [393] "Lil Dat"
## [397] "SEMPER FI 11"
## [401] "BATTLE BORN 27"
## [405] "Canned Heat"
## [409] "Edward"
## [413] "Kingston"
## [417] "Rogue Brew"
## [421] "Alice"
## [425] "Charlie"
## [429] "Elle"
## [433] "Gage"
## [437] "Lil Dat"
## [441] "SEMPER FI 11"
## [445] "BATTLE BORN 27"
## [449] "Canned Heat"
## [453] "Edward"
## [457] "Kingston"
## [461] "Rogue Brew"
## [465] "Alice"
## [469] "Charlie"
## [473] "Elle"
## [477] "Gage"
## [481] "Lil Dat"
## [485] "SEMPER FI 11"
## [489] "BATTLE BORN 27"
## [493] "Canned Heat"
## [497] "Edward"
## [501] "Kingston"
## [505] "Rogue Brew"
## [509] "Alice"
## [513] "Charlie"
## [517] "Elle"
## [521] "Gage"
## [525] "Lil Dat"
## [529] "SEMPER FI 11"
## [533] "BATTLE BORN 27"
## [537] "Canned Heat"
## [541] "Edward"
## [545] "Kingston"
## [549] "Rogue Brew"
## [553] "Alice"
## [557] "Charlie"
## [561] "Elle"
## [565] "Gage"
## [569] "Lil Dat"
## [573] "SEMPER FI 11"
## [577] "BATTLE BORN 27"
## [581] "Canned Heat"
## [585] "Edward"
## [589] "Kingston"
## [593] "Rogue Brew"
## [597] "Alice"
## [601] "Charlie"
## [605] "Elle"
## [609] "Gage"
## [613] "Lil Dat"
## [617] "SEMPER FI 11"
## [621] "BATTLE BORN 27"
## [625] "Canned Heat"
## [629] "Edward"
## [633] "Kingston"
## [637] "Rogue Brew"
## [641] "Alice"
## [645] "Charlie"
## [649] "Elle"
## [653] "Gage"
## [657] "Lil Dat"
## [661] "SEMPER FI 11"
## [665] "BATTLE BORN 27"
## [669] "Canned Heat"
## [673] "Edward"
## [677] "Kingston"
## [681] "Rogue Brew"
## [685] "Alice"
## [689] "Charlie"
## [693] "Elle"
## [697] "Gage"
## [701] "Lil Dat"
## [705] "SEMPER FI 11"
## [709] "BATTLE BORN 27"
## [713] "Canned Heat"
## [717] "Edward"
## [721] "Kingston"
## [725] "Rogue Brew"
## [729] "Alice"
## [733] "Charlie"
## [737] "Elle"
## [741] "Gage"
## [745] "Lil Dat"
## [749] "SEMPER FI 11"
## [753] "BATTLE BORN 27"
## [757] "Canned Heat"
## [761] "Edward"
## [765] "Kingston"
## [769] "Rogue Brew"
## [773] "Alice"
## [777] "Charlie"
## [781] "Elle"
## [785] "Gage"
## [789] "Lil Dat"
## [793] "SEMPER FI 11"
## [797] "BATTLE BORN 27"
## [801] "Canned Heat"
## [805] "Edward"
## [809] "Kingston"
## [813] "Rogue Brew"
## [817] "Alice"
## [821] "Charlie"
## [825] "Elle"
## [829] "Gage"
## [833] "Lil Dat"
## [837] "SEMPER FI 11"
## [841] "BATTLE BORN 27"
## [845] "Canned Heat"
## [849] "Edward"
## [853] "Kingston"
## [857] "Rogue Brew"
## [861] "Alice"
## [865] "Charlie"
## [869] "Elle"
## [873] "Gage"
## [877] "Lil Dat"
## [881] "SEMPER FI 11"
## [885] "BATTLE BORN 27"
## [889] "Canned Heat"
## [893] "Edward"
## [897] "Kingston"
## [901] "Rogue Brew"
## [905] "Alice"
## [909] "Charlie"
## [913] "Elle"
## [917] "Gage"
## [921] "Lil Dat"
## [925] "SEMPER FI 11"
## [929] "BATTLE BORN 27"
## [933] "Canned Heat"
## [937] "Edward"
## [941] "Kingston"
## [945] "Rogue Brew"
## [949] "Alice"
## [953] "Charlie"
## [957] "Elle"
## [961] "Gage"
## [965] "Lil Dat"
## [969] "SEMPER FI 11"
## [973] "BATTLE BORN 27"
## [977] "Canned Heat"
## [981] "Edward"
## [985] "Kingston"
## [989] "Rogue Brew"
## [993] "Alice"
## [997] "Charlie"
## [1001] "Elle"
## [1005] "Gage"
## [1009] "Lil Dat"
## [1013] "SEMPER FI 11"
## [1017] "BATTLE BORN 27"
## [1021] "Canned Heat"
## [1025] "Edward"
## [1029] "Kingston"
## [1033] "Rogue Brew"
## [1037] "Alice"
## [1041] "Charlie"
## [1045] "Elle"
## [1049] "Gage"
## [1053] "Lil Dat"
## [1057] "SEMPER FI 11"
## [1061] "BATTLE BORN 27"
## [1065] "Canned Heat"
## [1069] "Edward"
## [1073] "Kingston"
## [1077] "Rogue Brew"
## [1081] "Alice"
## [1085] "Charlie"
## [1089] "Elle"
## [1093] "Gage"
## [1097] "Lil Dat"
## [1101] "SEMPER FI 11"
## [1105] "BATTLE BORN 27"
## [1109] "Canned Heat"
## [1113] "Edward"
## [1117] "Kingston"
## [1121] "Rogue Brew"
## [1125] "Alice"
## [1129] "Charlie"
## [1133] "Elle"
## [1137] "Gage"
## [1141] "Lil Dat"
## [1145] "SEMPER FI 11"
## [1149] "BATTLE BORN 27"
## [1153] "Canned Heat"
## [1157] "Edward"
## [1161] "Kingston"
## [1165] "Rogue Brew"
## [1169] "Alice"
## [1173] "Charlie"
## [1177] "Elle"
## [1181] "Gage"
## [1185] "Lil Dat"
## [1189] "SEMPER FI 11"
## [1193] "BATTLE BORN 27"
## [1197] "Canned Heat"
## [1201] "Edward"
## [1205] "Kingston"
## [1209] "Rogue Brew"
## [1213] "Alice"
## [1217] "Charlie"
## [1221] "Elle"
## [1225] "Gage"
## [1229] "Lil Dat"
## [1233] "SEMPER FI 11"
## [1237] "BATTLE BORN 27"
## [1241] "Canned Heat"
## [1245] "Edward"
## [1249] "Kingston"
## [1253] "Rogue Brew"
## [1257] "Alice"
## [1261] "Charlie"
## [1265] "Elle"
## [1269] "Gage"
## [1273] "Lil Dat"
## [1277] "SEMPER FI 11"
## [1281] "BATTLE BORN 27"
## [1285] "Canned Heat"
## [1289] "Edward"
## [1293] "Kingston"
## [1297] "Rogue Brew"
## [1301] "Alice"
## [1305] "Charlie"
## [1309] "Elle"
## [1313] "Gage"
## [1317] "Lil Dat"
## [1321] "SEMPER FI 11"
## [1325] "BATTLE BORN 27"
## [1329] "Canned Heat"
## [1333] "Edward"
## [1337] "Kingston"
## [1341] "Rogue Brew"
## [1345] "Alice"
## [1349] "Charlie"
## [1353] "Elle"
## [1357] "Gage"
## [1361] "Lil Dat"
## [1365] "SEMPER FI 11"
## [1369] "BATTLE BORN 27"
## [1373] "Canned Heat"
## [1377] "Edward"
## [1381] "Kingston"
## [1385] "Rogue Brew"
## [1389] "Alice"
## [1393] "Charlie"
## [1397] "Elle"
## [1401] "Gage"
## [1405] "Lil Dat"
## [1409] "SEMPER FI 11"
## [1413] "BATTLE BORN 27"
## [1417] "Canned Heat"
## [1421] "Edward"
## [1425] "Kingston"
## [1429] "Rogue Brew"
## [1433] "Alice"
## [1437] "Charlie"
## [1441] "Elle"
## [1445] "Gage"
## [1449] "Lil Dat"
## [1453] "SEMPER FI 11"
## [1457] "BATTLE BORN 27"
## [1461] "Canned Heat"
## [1465] "Edward"
## [1469] "Kingston"
## [1473] "Rogue Brew"
## [1477] "Alice"
## [1481] "Charlie"
## [1485] "Elle"
## [1489] "Gage"
## [1493] "Lil Dat"
## [1497] "SEMPER FI 11"
## [1501] "BATTLE BORN 27"
## [1505] "Canned Heat"
## [1509] "Edward"
## [1513] "Kingston"
## [1517] "Rogue Brew"
## [1521] "Alice"
## [1525] "Charlie"
## [1529] "Elle"
## [1533] "Gage"
## [1537] "Lil Dat"
## [1541] "SEMPER FI 11"
## [1545] "BATTLE BORN 27"
## [1549] "Canned Heat"
## [1553] "Edward"
## [1557] "Kingston"
## [1561] "Rogue Brew"
## [1565] "Alice"
## [1569] "Charlie"
## [1573] "Elle"
## [1577] "Gage"
## [1581] "Lil Dat"
## [1585] "SEMPER FI 11"
## [1589] "BATTLE BORN 27"
## [1593] "Canned Heat"
## [1597] "Edward"
## [1601] "Kingston"
## [1605] "Rogue Brew"
## [1609] "Alice"
## [1613] "Charlie"
## [1617] "Elle"
## [1621] "Gage"
## [1625] "Lil Dat"
## [1629] "SEMPER FI 11"
## [1633] "BATTLE BORN 27"
## [1637] "Canned Heat"
## [1641] "Edward"
## [1645] "Kingston"
## [1649] "Rogue Brew"
## [1653] "Alice"
## [1657] "Charlie"
## [1661] "Elle"
## [1665] "Gage"
## [1669] "Lil Dat"
## [1673] "SEMPER FI 11"
## [1677] "BATTLE BORN 27"
## [1681] "Canned Heat"
## [1685] "Edward"
## [1689] "Kingston"
## [1693] "Rogue Brew"
## [1697] "Alice"
## [1701] "Charlie"
## [1705] "Elle"
## [1709] "Gage"
## [1713] "Lil Dat"
## [1717] "SEMPER FI 11"
## [1721] "BATTLE BORN 27"
## [1725] "Canned Heat"
## [1729] "Edward"
## [1733] "Kingston"
## [1737] "Rogue Brew"
## [1741] "Alice"
## [1745] "Charlie"
## [1749] "Elle"
## [1753] "Gage"
## [1757] "Lil Dat"
## [1761] "SEMPER FI 11"
## [1765] "BATTLE BORN 27"
## [1769] "Canned Heat"
## [1773] "Edward"
## [1777] "Kingston"
## [1781] "Rogue Brew"
## [1785] "Alice"
## [1789] "Charlie"
## [1793] "Elle"
## [1797] "Gage"
## [1801] "Lil Dat"
## [1805] "SEMPER FI 11"
## [1809] "BATTLE BORN 27"
## [1813] "Canned Heat"
## [1817] "Edward"
## [1821] "Kingston"
## [1825] "Rogue Brew"
## [1829] "Alice"
## [1833] "Charlie"
## [1837] "Elle"
## [1841] "Gage"
## [1845] "Lil Dat"
## [1849] "SEMPER FI 11"
## [1853] "BATTLE BORN 27"
## [1857] "Canned Heat"
## [1861] "Edward"
## [1865] "Kingston"
## [1869] "Rogue Brew"
## [1873] "Alice"
## [1877] "Charlie"
## [1881] "Elle"
## [1885] "Gage"
## [1889] "Lil Dat"
## [1893] "SEMPER FI 11"
## [1897] "BATTLE BORN 27"
## [1901] "Canned Heat"
## [1905] "Edward"
## [1909] "Kingston"
## [1913] "Rogue Brew"
## [1917] "Alice"
## [1921] "Charlie"
## [1925] "Elle"
## [1929] "Gage"
## [1933] "Lil Dat"
## [1937] "SEMPER FI 11"
## [1941] "BATTLE BORN 27"
## [1945] "Canned Heat"
## [1949] "Edward"
## [1953] "Kingston"
## [1957] "Rogue Brew"
## [1961] "Alice"
## [1965] "Charlie"
## [1969] "Elle"
## [1973] "Gage"
## [1977] "Lil Dat"
## [1981] "SEMPER FI 11"
## [1985] "BATTLE BORN 27"
## [1989] "Canned Heat"
## [1993] "Edward"
## [1997] "Kingston"
## [2001] "Rogue Brew"
## [2005] "Alice"
## [2009] "Charlie"
## [2013] "Elle"
## [2017] "Gage"
## [2021] "Lil Dat"
## [2025] "SEMPER FI 11"
## [2029] "BATTLE BORN 27"
## [2033] "Canned Heat"
## [2037] "Edward"
## [2041] "Kingston"
## [2045] "Rogue Brew"
## [2049] "Alice"
## [2053] "Charlie"
## [2057] "Elle"
## [2061] "Gage"
## [2065] "Lil Dat"
## [2069] "SEMPER FI 11"
## [2073] "BATTLE BORN 27"
## [2077] "Canned Heat"
## [2081] "Edward"
## [2085] "Kingston"
## [2089] "Rogue Brew"
## [2093] "Alice"
## [2097] "Charlie"
## [2101] "Elle"
## [2105] "Gage"
## [2109] "Lil Dat"
## [2113] "SEMPER FI 11"
## [2117] "BATTLE BORN 27"
## [2121] "Canned Heat"
## [2125] "Edward"
## [2129] "Kingston"
## [2133] "Rogue Brew"
## [2137] "Alice"
## [2141] "Charlie"
## [2145] "Elle"
## [2149] "Gage"
## [2153] "Lil Dat"
## [2157] "SEMPER FI 11"
## [2161] "BATTLE BORN 27"
## [2165] "Canned Heat"
## [2169] "Edward"
## [2173] "Kingston"
## [2177] "Rogue Brew"
## [2181] "Alice"
## [2185] "Charlie"
## [2189] "Elle"
## [2193] "Gage"
## [2197] "Lil Dat"
## [2201] "SEMPER FI 11"
## [2205] "BATTLE BORN 27"
## [2209] "Canned Heat"
## [2213] "Edward"
## [2217] "Kingston"
## [2221] "Rogue Brew"
## [2225] "Alice"
## [2229] "Charlie"
## [2233] "Elle"
## [2237] "Gage"
## [2241] "Lil Dat"
## [2245] "SEMPER FI 11"
## [2249] "BATTLE BORN 27"
## [2253] "Canned Heat"
## [2257] "Edward"
## [2261] "Kingston"
## [2265] "Rogue Brew"
## [2269] "Alice"
## [2273] "Charlie"
## [2277] "Elle"
## [2281] "Gage"
## [2285] "Lil Dat"
## [2289] "SEMPER FI 11"
## [2293] "BATTLE BORN 27"
## [2297] "Canned Heat"
## [2301] "Edward"
## [2305] "Kingston"
## [2309] "Rogue Brew"
## [2313] "Alice"
## [2317] "Charlie"
## [2321] "Elle"
## [2325] "Gage"
## [2329] "Lil Dat"
## [2333] "SEMPER FI 11"
## [2337] "BATTLE BORN 27"
## [2341] "Canned Heat"
## [2345] "Edward"
## [2349] "Kingston"
## [2353] "Rogue Brew"
## [2357] "Alice"
## [2361] "Charlie"
## [2365] "Elle"
## [2369] "Gage"
## [2373] "Lil Dat"
## [2377] "SEMPER FI 11"
## [2381] "BATTLE BORN 27"
## [2385] "Canned Heat"
## [2389] "Edward"
## [2393] "Kingston"
## [2397] "Rogue Brew"
## [2401] "Alice"
## [2405] "Charlie"
## [2409] "Elle"
## [2413] "Gage"
## [2417] "Lil Dat"
## [2421] "SEMPER FI 11"
## [2425] "BATTLE BORN 27"
## [2429] "Canned Heat"
## [2433] "Edward"
## [2437] "Kingston"
## [2441] "Rogue Brew"
## [2445] "Alice"
## [2449] "Charlie"
## [2453] "Elle"
## [2457] "Gage"
## [2461] "Lil Dat"
## [2465] "SEMPER FI 11"
## [2469] "BATTLE BORN 27"
## [2473] "Canned Heat"
## [2477] "Edward"
## [2481] "Kingston"
## [2485] "Rogue Brew"
## [2489] "Alice"
## [2493] "Charlie"
## [2497] "Elle"
## [2501] "Gage"
## [2505] "Lil Dat"
## [2509] "SEMPER FI 11"
## [2513] "BATTLE BORN 27"
## [2517] "Canned Heat"
## [2521] "Edward"
## [2525] "Kingston"
## [2529] "Rogue Brew"
## [2533] "Alice"
## [2537] "Charlie"
## [2541] "Elle"
## [2545] "Gage"
## [2549] "Lil Dat"
## [2553] "SEMPER FI 11"
## [2557] "BATTLE BORN 27"
## [2561] "Canned Heat"
## [2565] "Edward"
## [2569] "Kingston"
## [2573] "Rogue Brew"
## [2577] "Alice"
## [2581] "Charlie"
## [2585] "Elle"
## [2589] "Gage"
## [2593] "Lil Dat"
## [2597] "SEMPER FI 11"
## [2601] "BATTLE BORN 27"
## [2605] "Canned Heat"
## [2609] "Edward"
## [2613] "Kingston"
## [2617] "Rogue Brew"
## [2621] "Alice"
## [2625] "Charlie"
## [2629] "Elle"
## [2633] "Gage"
## [2637] "Lil Dat"
## [2641] "SEMPER FI 11"
## [2645] "BATTLE BORN 27"
## [2649] "Canned Heat"
## [2653] "Edward"
## [2657] "Kingston"
## [2661] "Rogue Brew"
## [2665] "Alice"
## [2669] "Charlie"
## [2673] "Elle"
## [2677] "Gage"
## [2681] "Lil Dat"
## [2685] "SEMPER FI 11"
## [2689] "BATTLE BORN 27"
## [2693] "Canned Heat"
## [2697] "Edward"
## [2701] "Kingston"
## [2705] "Rogue Brew"
## [2709] "Alice"
## [2713] "Charlie"
## [2717] "Elle"
## [2721] "Gage"
## [2725] "Lil Dat"
## [2729] "SEMPER FI 11"
## [2733] "BATTLE BORN 27"
## [2737] "Canned Heat"
## [2741] "Edward"
## [2745] "Kingston"
## [2749] "Rogue Brew"
## [2753] "Alice"
## [2757] "Charlie"
## [2761] "Elle"
## [2765] "Gage"
## [2769] "Lil Dat"
## [2773] "SEMPER FI 11"
## [2777] "BATTLE BORN 27"
## [2781] "Canned Heat"
## [2785] "Edward"
## [2789] "Kingston"
## [2793] "Rogue Brew"
## [2797] "Alice"
## [2801] "Charlie"
## [2805] "Elle"
## [2809] "Gage"
## [2813] "Lil Dat"
## [2817] "SEMPER FI 11"
## [2821] "BATTLE BORN 27"
## [2825] "Canned Heat"
## [2829] "Edward"
## [2833] "Kingston"
## [2837] "Rogue Brew"
## [2841] "Alice"
## [2845] "Charlie"
## [2849] "Elle"
## [2853] "Gage"
## [2857] "Lil Dat"
## [2861] "SEMPER FI 11"
## [2865] "BATTLE BORN 27"
## [2869] "Canned Heat"
## [2873] "Edward"
## [2877] "Kingston"
## [2881] "Rogue Brew"
## [2885] "Alice"
## [2889] "Charlie"
## [2893] "Elle"
## [2897] "Gage"
## [2901] "Lil Dat"
## [2905] "SEMPER FI 11"
## [2909] "BATTLE BORN 27"
## [2913] "Canned Heat"
## [2917] "Edward"
## [2921] "Kingston"
## [2925] "Rogue Brew"
## [2929] "Alice"
## [2933] "Charlie"
## [2937] "Elle"
## [2941] "Gage"
## [2945] "Lil Dat"
## [2949] "SEMPER FI 11"
## [2953] "BATTLE BORN 27"
## [2957] "Canned Heat"
## [2961] "Edward"
## [2965] "Kingston"
## [2969] "Rogue Brew"
## [2973] "Alice"
## [2977] "Charlie"
## [2981] "Elle"
## [2985] "Gage"
## [2989] "Lil Dat"
## [2993] "SEMPER FI 11"
## [2997] "BATTLE BORN 27"
## [3001] "Canned Heat"
## [3005] "Edward"
## [3009] "Kingston"
## [3013] "Rogue Brew"
## [3017] "Alice"
## [3021] "Charlie"
## [3025] "Elle"
## [3029] "Gage"
## [3033] "Lil Dat"
## [3037] "SEMPER FI 11"
## [3041] "BATTLE BORN 27"
## [3045] "Canned Heat"
## [3049] "Edward"
## [3053] "Kingston"
## [3057] "Rogue Brew"
## [3061] "Alice"
## [3065] "Charlie"
## [3069] "Elle"
## [3073] "Gage"
## [3077] "Lil Dat"
## [3081] "SEMPER FI 11"
## [3085] "BATTLE BORN 27"
## [3089] "Canned Heat"
## [3093] "Edward"
## [3097] "Kingston"
## [3101] "Rogue Brew"
## [3105] "Alice"
## [3109] "Charlie"
## [3113] "Elle"
## [3117] "Gage"
## [3121] "Lil Dat"
## [3125] "SEMPER FI 11"
## [3129] "BATTLE BORN 27"
## [3133] "Canned Heat"
## [3137] "Edward"
## [3141] "Kingston"
## [3145] "Rogue Brew"
## [3149] "Alice"
## [3153] "Charlie"
## [3157] "Elle"
## [3161] "Gage"
## [3165] "Lil Dat"
## [3169] "SEMPER FI 11"
## [3173] "BATTLE BORN 27"
## [3177] "Canned Heat"
## [3181] "Edward"
## [3185] "Kingston"
## [3189] "Rogue Brew"
## [3193] "Alice"
## [3197] "Charlie"
## [3201] "Elle"
## [3205] "Gage"
## [3209] "Lil Dat"
## [3213] "SEMPER FI 11"
## [3217] "BATTLE BORN 27"
## [3221] "Canned Heat"
## [3225] "Edward"
## [3229] "Kingston"
## [3233] "Rogue Brew"
## [3237] "Alice"
## [3241] "Charlie"
## [3245] "Elle"
## [3249] "Gage"
## [3253] "Lil Dat"
## [3257] "SEMPER FI 11"
## [3261] "BATTLE BORN 27"
## [3265] "Canned Heat"
## [3269] "Edward"
## [3273] "Kingston"
## [3277] "Rogue Brew"
## [3281] "Alice"
## [3285] "Charlie"
## [3289] "Elle"
## [3293] "Gage"
## [3297] "Lil Dat"
## [3301] "SEMPER FI 11"
## [3305] "BATTLE BORN 27"
## [3309] "Canned Heat"
## [3313] "Edward"
## [3317] "Kingston"
## [3321] "Rogue Brew"
## [3325] "Alice"
## [3329] "Charlie"
## [3333] "Elle"
## [3337] "Gage"
## [3341] "Lil Dat"
## [3345] "SEMPER FI 11"
## [3349] "BATTLE BORN 27"
## [3353] "Canned Heat"
## [3357] "Edward"
## [3361] "Kingston"
## [3365] "Rogue Brew"
## [3369] "Alice"
## [3373] "Charlie"
## [3377] "Elle"
## [3381] "Gage"
## [3385] "Lil Dat"
## [3389] "SEMPER FI 11"
## [3393] "BATTLE BORN 27"
## [3397] "Canned Heat"
## [3401] "Edward"
## [3405] "Kingston"
## [3409] "Rogue Brew"
## [3413] "Alice"
## [3417] "Charlie"
## [3421] "Elle"
## [3425] "Gage"
## [3429] "Lil Dat"
## [3433] "SEMPER FI 11"
## [3437] "BATTLE BORN 27"
## [3441] "Canned Heat"
## [3445] "Edward"
## [3449] "Kingston"
## [3453] "Rogue Brew"
## [3457] "Alice"
## [3461] "Charlie"
## [3465] "Elle"
## [3469] "Gage"
## [3473] "Lil Dat"
## [3477] "SEMPER FI 11"
## [3481] "BATTLE BORN 27"
## [3485] "Canned Heat"
## [3489] "Edward"
## [3493] "Kingston"
## [3497] "Rogue Brew"
## [3501] "Alice"
## [3505] "Charlie"
## [3509] "Elle"
## [3513] "Gage"
## [3517] "Lil Dat"
## [3521] "SEMPER FI 11"
## [3525] "BATTLE BORN 27"
## [3529] "Canned Heat"
## [3533] "Edward"
## [3537] "Kingston"
## [3541] "Rogue Brew"
## [3545] "Alice"
## [3549] "Charlie"
## [3553] "Elle"
## [3557] "Gage"
## [3561] "Lil Dat"
## [3565] "SEMPER FI 11"
## [3569] "BATTLE BORN 27"
## [3573] "Canned Heat"
## [3577] "Edward"
## [3581] "Kingston"
## [3585] "Rogue Brew"
## [3589] "Alice"
## [3593] "Charlie"
## [3597] "Elle"
## [3601] "Gage"
## [3605] "Lil Dat"
## [3609] "SEMPER FI 11"
## [3613] "BATTLE BORN 27"
## [3617] "Canned Heat"
## [3621] "Edward"
## [3625] "Kingston"
## [3629]
```

```
pigeon_tb
```

```
## # A tibble: 400 x 12
##   Pos Breeder Pigeon Name Color Sex Ent Arrival Speed To.Win
##   <int> <fct>   <fct>   <fct> <fct> <fct> <fct> <fct>   <dbl> <fct>
## 1     1     1 Texas ~ 19633~ <NA> BCWF H    1   42:14.0  172. 0:00:~
## 2     2     2 Junior~ 0402~~ <NA> SIWF H    1   47:36.0  164. 0:05:~
## 3     3     3 Jerry ~ 0404~~ Perc~ BB   H    1   47:41.0  163. 0:05:~
## 4     4     4 Alias~~ 2013~~ <NA> BBSP H    1   47:43.0  163. 0:05:~
## 5     5     5 Greg G~ 5749~~ <NA> BC   H    1   47:44.0  163. 0:05:~
## 6     6     6 Dal-Te~ 0032~~ <NA> BC   H    1   47:51.0  163. 0:05:~
## 7     7     7 Greg G~ 5768~~ <NA> BBWF C    2   47:53.0  163. 0:05:~
## 8     8     8 N C Sy~ 1067~~ <NA> BC   H    1   47:57.0  163. 0:05:~
## 9     9     9 Baldwi~ 1194~~ <NA> BB   H    1   48:02.0  163. 0:05:~
## 10    10    10 Texas ~ 19632~ <NA> BC   H    2   48:03.0  163. 0:05:~
## # ... with 390 more rows, and 2 more variables: Eligible <fct>, Ini <chr>
```

References

Van der Loo, M. and De Jonge, E. (2013) An introduction to data cleaning with R. https://cran.r-project.org/doc/contrib/de_Jonge+van_der_Loo-Introduction_to_data_cleaning_with_R.pdf

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2019). dplyr: A Grammar of Data Manipulation. R package version 0.8.3. <https://CRAN.R-project.org/package=dplyr>

Hadley Wickham and Lionel Henry (2019). tidyr: Easily Tidy Data with ‘spread()’ and ‘gather()’ Functions. R package version 0.8.3. <https://CRAN.R-project.org/package=tidyr>

Kirill Müller and Hadley Wickham (2019). tibble: Simple Data Frames. R package version 2.1.3. <https://CRAN.R-project.org/package=tibble>

Hadley Wickham and Jennifer Bryan (2019). readxl: Read Excel Files. R package version 1.3.1. <https://CRAN.R-project.org/package=readxl>

Hadley Wickham (2017). tidyverse: Easily Install and Load the ‘Tidyverse’. R package version 1.2.1. <https://CRAN.R-project.org/package=tidyverse>

Usefull resources

- Presentation data cleaning Jonge. https://www.r-project.ro/conference2017/presentations/uRos2017_data-cleaning-workshop.pdf