Data Manipulation with dplyr

Lucas Mello Schnorr, Jean-Marc Vincent

February 28, 2017

This is a demonstration of how dplyr works.

First, we need some data.

```
library(readr);
df <- read_tsv (file = "dpt2015.txt",
                locale = locale(encoding = "ISO-8859-1"));
```

```
## Parsed with column specification:
## cols(
##   sexe = col_integer(),
##   preusuel = col_character(),
##   annais = col_character(),
##   dpt = col_character(),
##   nombre = col_double()
## )
```

```
head(df);
```

```
## # A tibble: 6 × 5
##    sexe preusuel annais   dpt nombre
##   <int>    <chr>  <chr> <chr>  <dbl>
## 1     1        A   XXXX    XX     27
## 2     1    AADEL   XXXX    XX     53
## 3     1    AADIL   1983    84      3
## 4     1    AADIL   1992    92      3
## 5     1    AADIL   XXXX    XX    162
## 6     1   AAKASH   XXXX    XX     24
```

Load the necessary packages:

```
library(dplyr);
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(magrittr);
```

Now, let's use the *filter()* verb : On filtre les lignes dont l'année de naissance n'est pas renseignée (XXXX):

```
df %>% filter(annais != 'XXXX');
```

```
## # A tibble: 3,372,275 × 5
##     sexe preusuel annais   dpt nombre
##    <int>    <chr>  <chr> <chr>  <dbl>
```

```
## 1      1    AADIL   1983    84      3
## 2      1    AADIL   1992    92      3
## 3      1    AARON   1962    75      3
## 4      1    AARON   1982    75      3
## 5      1    AARON   1984    75      3
## 6      1    AARON   1985    75      4
## 7      1    AARON   1989    75      3
## 8      1    AARON   1990    69      3
## 9      1    AARON   1990    75      4
## 10     1    AARON   1990    93      4
## # ... with 3,372,265 more rows
```

On va afficher le nombre d'occurences de nos noms à chacun :

Marie :

```
df %>% filter(preusuel=='MARIE')%>% summarise(N=sum(nombre));
```

```
## # A tibble: 1 × 1
##         N
##     <dbl>
## 1 2261915
```

Kathleen :

```
df %>% filter(preusuel=='KATHLEEN')%>% summarise(N=sum(nombre));
```

```
## # A tibble: 1 × 1
##       N
##   <dbl>
## 1  4959
```

Lucas :

```
df %>% filter(preusuel=='LUCAS')%>% summarise(N=sum(nombre));
```

```
## # A tibble: 1 × 1
##        N
##    <dbl>
## 1 156149
```

On va voir l'étendue des donnée en terme d'années. Tout d'abord on va transformer l'année de naissance en nombre.

```
df %>% filter(annais!='XXXX') %>% mutate(annaisbis=as.integer(annais));
```

```
## # A tibble: 3,372,275 × 6
##     sexe preusuel annais   dpt nombre annaisbis
##    <int>    <chr>  <chr> <chr>  <dbl>     <int>
## 1      1    AADIL   1983    84      3      1983
## 2      1    AADIL   1992    92      3      1992
## 3      1    AARON   1962    75      3      1962
## 4      1    AARON   1982    75      3      1982
## 5      1    AARON   1984    75      3      1984
## 6      1    AARON   1985    75      4      1985
## 7      1    AARON   1989    75      3      1989
## 8      1    AARON   1990    69      3      1990
## 9      1    AARON   1990    75      4      1990
## 10     1    AARON   1990    93      4      1990
```

```
## # ... with 3,372,265 more rows
```

```
df %>% filter(annais!='XXXX') %>% mutate(annaisbis=as.integer(annais)) %>% summarise(N=min(annais));
```

```
## # A tibble: 1 × 1
##       N
##   <chr>
## 1  1900
```

```
df %>% filter(annais!='XXXX') %>% mutate(annaisbis=as.integer(annais)) %>% summarise(N=max(annais));
```

```
## # A tibble: 1 × 1
##       N
##   <chr>
## 1  2015
```

Les données vont de 1900 à 2015.

On va regarder le prénom le plus représenté, hommes et femmes confondus.

```
df %>% filter(annais!='XXXX') %>% group_by(preusuel)%>%summarise(N=sum(nombre))%>%filter(N==max(N))
```

```
## # A tibble: 1 × 2
##   preusuel       N
##      <chr>   <dbl>
## 1    MARIE 2259067
```

Marie est donc le prénom le plus représenté.

On va regarder le prénom le plus représenté pour les hommes.

```
df %>% filter(annais!='XXXX') %>% filter(sexe==1)%>% group_by(preusuel)%>%summarise(N=sum(nombre))%>%fi
```

```
## # A tibble: 1 × 2
##   preusuel       N
##      <chr>   <dbl>
## 1     JEAN 1918796
```

Jean est le prénom le plus représenté chez les hommes.

On chercher les prénoms qui n'ont qui n'ont que le minimum d'occurences.

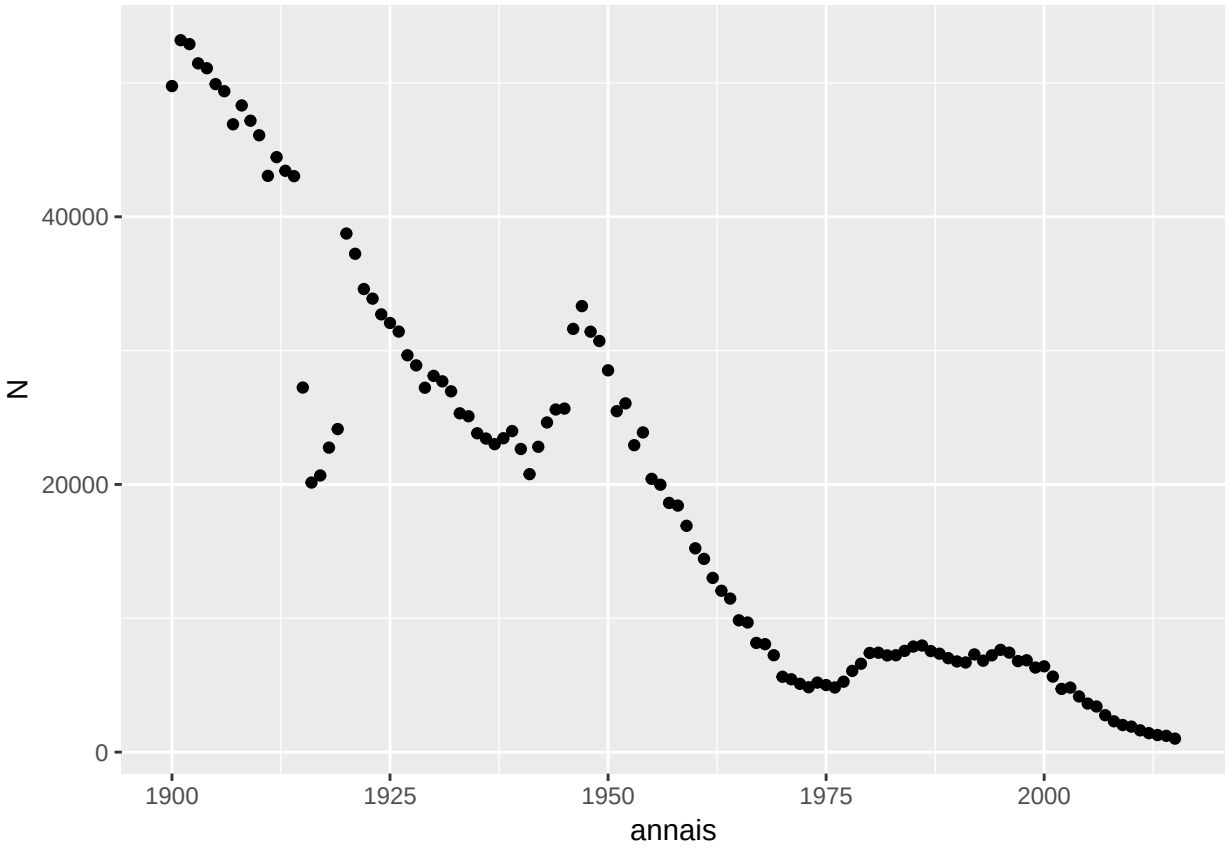
```
df %>% filter(annais!='XXXX') %>% group_by(preusuel)%>%summarise(N=sum(nombre))%>%filter(N==min(N))
```

```
## # A tibble: 2,826 × 2
##       preusuel     N
##          <chr> <dbl>
## 1        AALYA     3
## 2        AAYAN     3
## 3  ABDARRAHMAN     3
## 4       AB-DEL     3
## 5      ABDELAH     3
## 6  ABDELDJALIL     3
## 7  ABDEL-JALIL     3
## 8      ABDELLA     3
## 9   ABDELNACER     3
## 10 ABDEL-RAHIM     3
## # ... with 2,816 more rows
```

Voici la liste des prénoms les moins représentés (beaucoup).

```
library(ggplot2);
df %>% mutate(annais=as.integer(annais)) %>%filter(annais!='XXXX') %>%filter(preusuel=='MARIE') %>% grou
```

```
## Warning in eval(substitute(expr), envir, enclos): NAs introduits lors de la
## conversion automatique
```



On peut voir qu'il y a de moins en moins de Marie.