French given names per year per department by SOUAD ELAISSAOUI

The aim of the activity is to develop a methodology to answer a specific question on a given dataset.

The dataset is the set of Firstname given in France on a large period of time. given names data set of INSEE, we choose this dataset because it is sufficiently large, you can't do the analysis by hand, the structure is simple

You need to use the *tidyverse* for this analysis. Unzip the file *dpt2019_txt.zip* (to get the **dpt2019.csv**). Read in R with this code. Note that you might need to install the `readr` package with the appropriate command.

Download Raw Data from the website

```
file = "dpt2019_txt.zip"
if(!file.exists(file)){
download.file("https://www.insee.fr/fr/statistiques/fichier/2540004/
dpt2019_csv.zip",
destfile=file)
}
unzip(file)
```

Build the Dataframe from file

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────
tidyverse 1.3.0 ──
```

```
## ✓ ggplot2 3.3.3     ✓ purrr   0.3.4
## ✓ tibble  3.0.5     ✓ dplyr   1.0.3
## ✓ tidyr   1.1.2     ✓ stringr 1.4.0
## ✓ readr   1.4.0     ✓ forcats 0.5.0
```

```
## ── Conflicts ────────────────────────────────────────
tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
options(dplyr.summarise.inform=F)
# FirstNames <- read_delim("dpt2019.csv",delim=";");
namedata <- read.csv(file = 'dpt2019.csv', sep = ';')
```

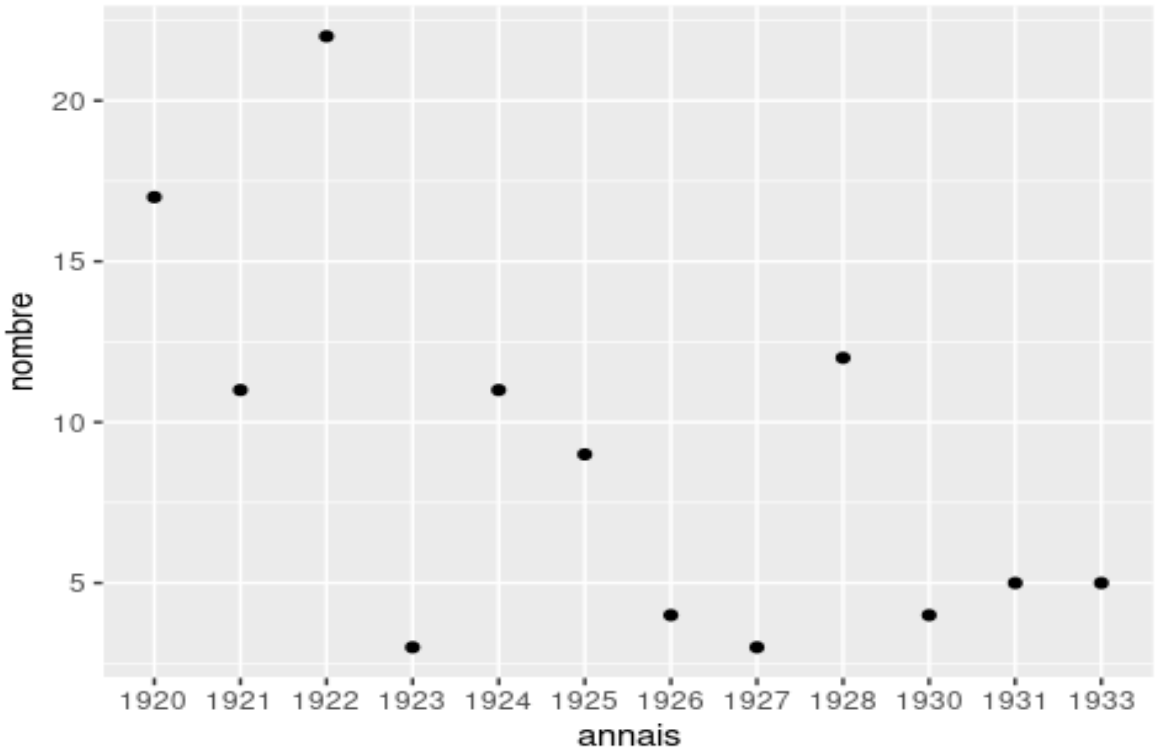
First step is : Filter out incomplete data

```
FirstNames = filter(namedata, annais != "XXXX" & dpt != "XX" &
preusuel != "_PRENOMS_RARES")
tail(FirstNames[complete.cases(FirstNames),],30)
```

```
##         sexe preusuel annais dpt nombre
## 3618372    2    ZULMA   1930  62      4
## 3618373    2    ZULMA   1931  59      5
## 3618374    2    ZULMA   1933  62      5
## 3618375    2   ZULMEE   1901  59      7
## 3618376    2   ZULMEE   1903  59      4
## 3618377    2   ZULMEE   1905  59      3
## 3618378    2   ZULMEE   1908  62      3
## 3618379    2   ZULMEE   1912  59      3
## 3618380    2   ZULMEE   1913  59      4
## 3618381    2   ZULMEE   1914  59      3
## 3618382    2    ZUMRA   2009  67      5
## 3618383    2    ZUMRA   2010  67      3
## 3618384    2    ZUMRA   2013  68      3
## 3618385    2    ZUMRA   2019  71      3
## 3618386    2    ZÜMRA   2019  01      3
## 3618387    2    ZÜMRA   2019  91      3
## 3618388    2  ZUZANNA   2009  75      6
## 3618389    2  ZUZANNA   2010  75      3
## 3618390    2  ZUZANNA   2013  75      3
## 3618391    2  ZUZANNA   2015  75      4
## 3618392    2  ZUZANNA   2015  94      3
## 3618393    2  ZUZANNA   2018  75      4
## 3618394    2      ZYA   2011  85      4
## 3618395    2      ZYA   2011  91      3
## 3618396    2      ZYA   2011 974      3
## 3618397    2      ZYA   2013  44      4
## 3618398    2      ZYA   2013  59      3
## 3618399    2      ZYA   2017 974      3
## 3618400    2      ZYA   2018  59      3
## 3618401    2     ZYNA   2013  93      3
```

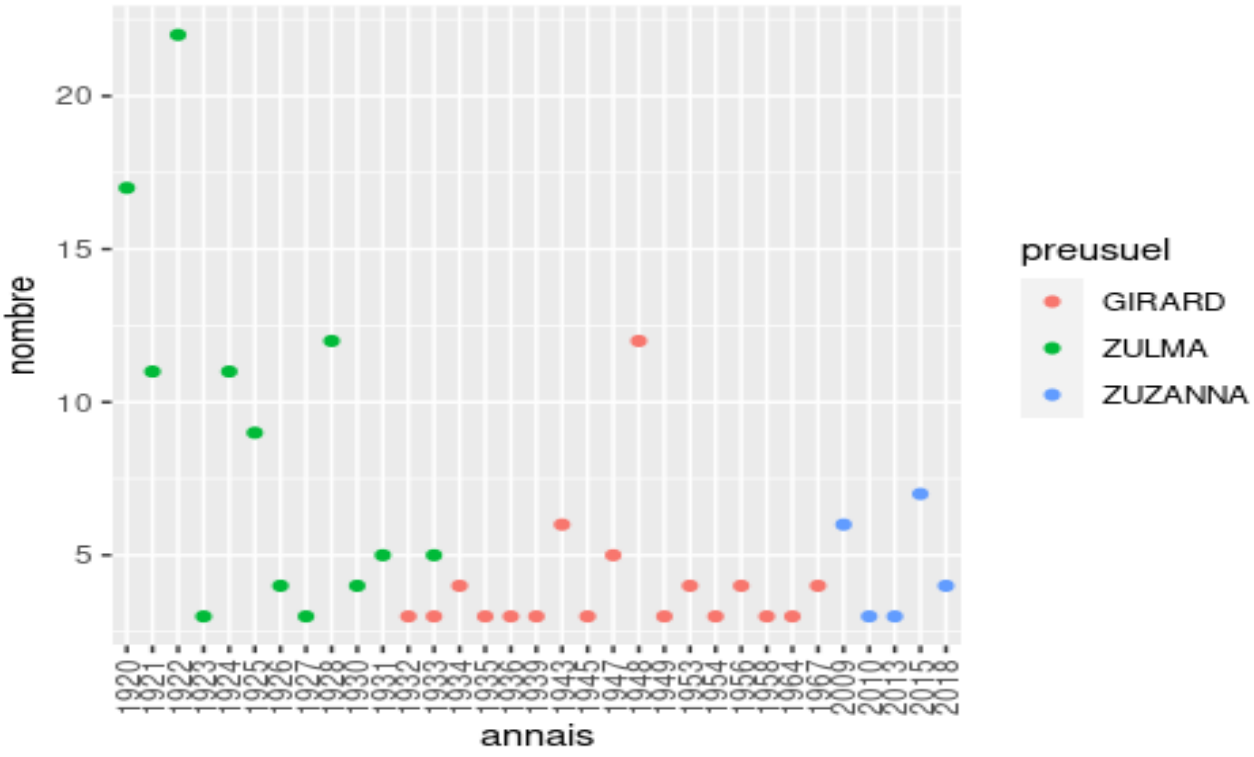
1.1 Choose a firstname and analyse its frequency along time :

```
ChoosenFirstName = filter(FirstNames, as.numeric(as.character(annais))
>= 1920 & (preusuel == "ZULMA"))
ChoosenFirstName = ChoosenFirstName %>%
group_by(annais) %>%
summarise(nombre = sum(nombre))
ggplot(data = ChoosenFirstName, aes(x=annais, y=nombre))+geom_point()
```



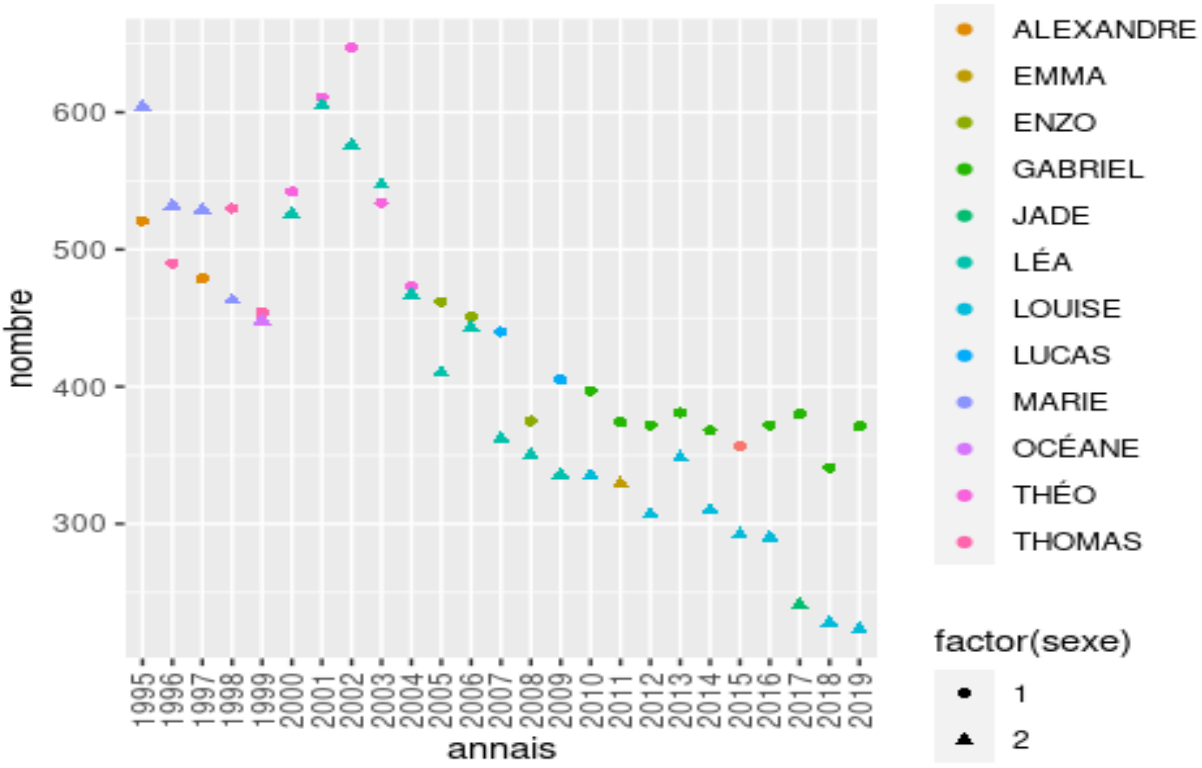
1.2 Compare several firstnames frequency :

```
CompareFirstNames = filter(FirstNames,
as.numeric(as.character(annais)) >= 1920 & (preusuel == "ZUZANNA" |
preusuel == "ZULMA" | preusuel == "GIRARD"))
CompareFirstNames = CompareFirstNames %>%
group_by(annais, preusuel) %>%
summarise(nombre = sum(nombre))
q <- ggplot(data = CompareFirstNames, aes(x=annais, y=nombre, color =
preusuel))+geom_point()
q + theme(axis.text.x = element_text(angle = 90, vjust = 0.5,
hjust=1))
```



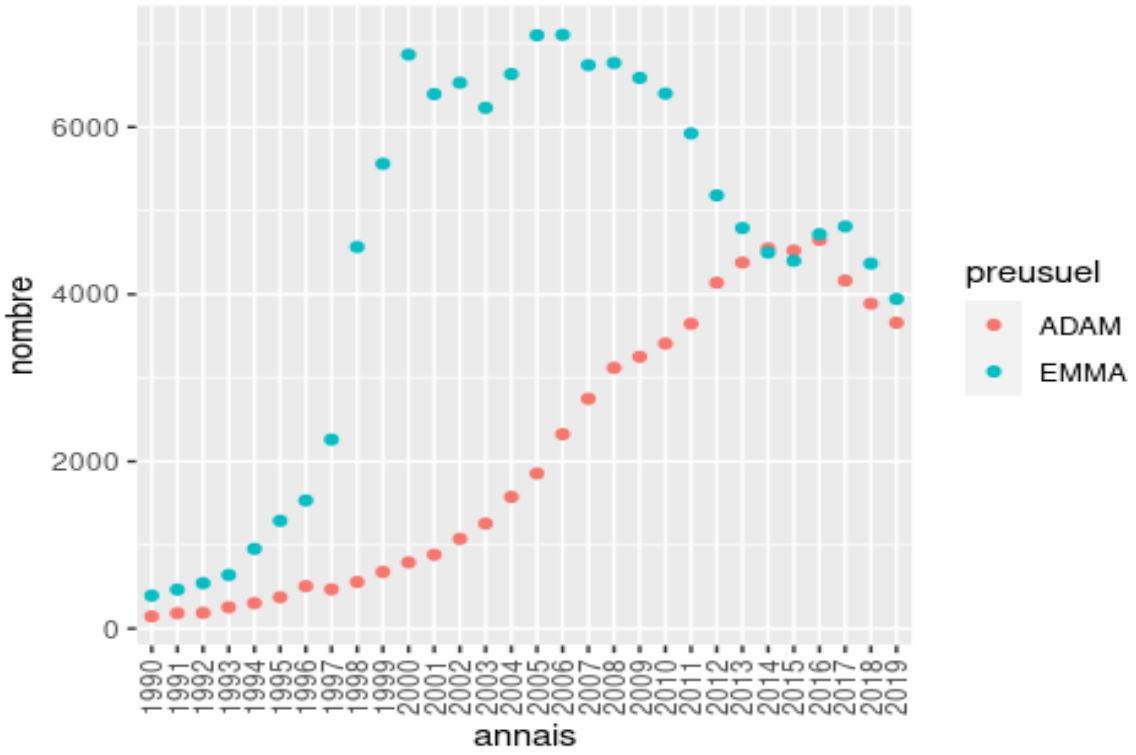
2.1 Establish by gender the most given firstname by year.

```
MostGivenFirstName = FirstNames %>%
group_by(sexe, annais) %>%
filter(nombre == max(nombre) & as.numeric(as.character(annais)) >=
1995)
q <- ggplot(data = MostGivenFirstName, aes(x=annais, y=nombre, shape =
factor(sexe), color = preusuel))+geom_point()
q + theme(axis.text.x = element_text(angle = 90, vjust = 0.5,
hjust=1))
```



2.2 Analyse the evolution of the most frequent firstname for each gender.

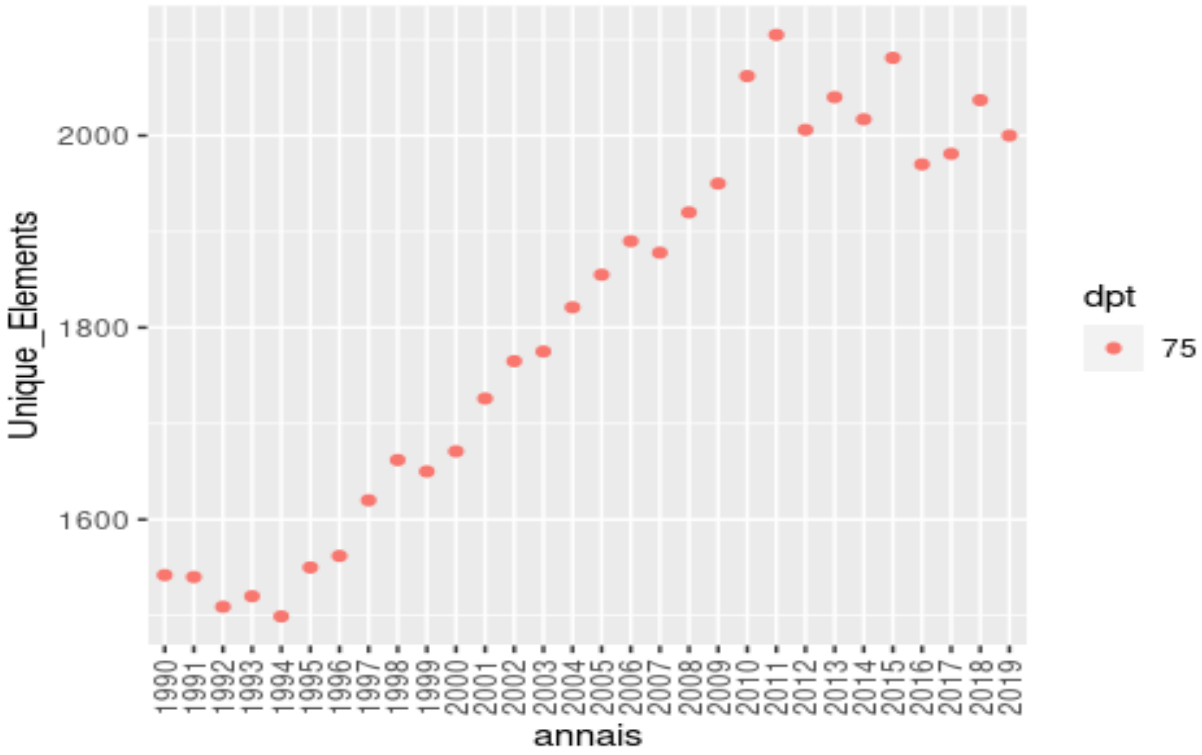
```
MostGivenFirstName = filter(FirstNames,
as.numeric(as.character(annais)) >= 1990 & (preusuel == "ADAM" |
preusuel == "EMMA"))
MostGivenFirstName = MostGivenFirstName %>%
group_by(annais, preusuel) %>%
summarise(nombre = sum(nombre))
q <- ggplot(data = MostGivenFirstName, aes(x=annais, y=nombre, color =
preusuel))+geom_point()
q + theme(axis.text.x = element_text(angle = 90, vjust = 0.5,
hjust=1))
```



We can conclude from the graphs, these two names were not the most popular before 1990 , then they become one of the most popular one, and finally from 2014, they start to fading again.

3. Optional : Which department has a larger variety of names along time ? Is there some sort of geographical correlation with the data?

```
CountUniqueNames = FirstNames %>%
filter(as.numeric(as.character(annais)) >= 1990) %>%
group_by(annais, dpt) %>%
summarise(Unique_Elements = n_distinct(preusuel))
CountUniqueNamesFiltered = CountUniqueNames %>%
filter(Unique_Elements == max(Unique_Elements))
q <- ggplot(data = CountUniqueNamesFiltered, aes(x=annais,
y=Unique_Elements, color = dpt))+geom_point()
q + theme(axis.text.x = element_text(angle = 90, vjust = 0.5,
hjust=1))
```



-Based on the result of data analysis, we can conclude tha the department 75 has the highest variety.

-Yes, because the department 75 is the second populated department in France.